Physics for the 21st Century

Introduction To Online Text *By Christopher Stubbs*

The intellectual heritage of modern physics

Physicists have been trying to figure out the world around them for centuries. Our society has inherited a rich intellectual heritage from this ongoing effort. The accumulated knowledge, the mathematical framework, and the concepts that we have inherited from giants like Galileo, Newton and Einstein are traditionally taught in a roughly historical sequence. It takes most people many years to make the progression from mechanics (forces, masses, and accelerations) to electromagnetism (fields, charges and potentials) to quantum mechanics (propagating waves of probability) to the current research frontier of physics. Most people claw their way to the boundary of modern knowledge only after a few years of graduate study.

Figure 1: Robert Kirshner during his interview.

The approach pursued here is different. We intend to jump directly to "the new stuff." The goal of this course is to present some of the fascinating topics that are currently being actively studied by today's physics community, at a level that is accessible to an interested high school student, or the high-school-student-at-heart.

The course has three components: i) written material, arranged as units, ii) video segments that present case studies related to the units, and iii) interactive Web modules. The different units can be approached in any desired sequence, but taking the time to explore the video and interactive segments associated with the units you find the most interesting is recommended.

The choice of research subjects presented here is representative, not exhaustive, and is meant to convey the excitement, the mystery, and the human aspects of modern physics. Numerous other threads of modern research could have been included. Hopefully, the topics selected will provide incentive for the reader to pursue other topics of interest, and perhaps it will prove possible to cover a number of these other topics in subsequent versions of the course.

Physics for the 21st Century

Introduction To Online Text *By Christopher Stubbs*

The "physics approach" to understanding nature: simplicity, reductionism, and shrewd approximations

So what is physics, anyway? It's an experiment-based way of thinking about the world that attempts to make shrewd simplifying assumptions in order to extract the essential ingredients of a physical system or situation. Physicists try to develop the ability to distinguish the important aspects of a problem from the unimportant and distracting ones. Physicists learn to factor complex problems into tractable subsets that can be addressed individually, and are then brought back together to understand the broader system. An important ingredient in this approach is the attempt to find unifying principles that apply to a wide range of circumstances. The conservation laws of quantities like energy and electric charge are good examples of these broad principles, that help us understand and predict the properties of systems and circumstances.

A core ingredient in the way physicists look at the world is the central role of experiment and observation to determine which concepts or theories best describe the world we are privileged to inhabit. While there are many possible theories one might conjecture about the nature of reality, only those that survive confrontation with experiment endure. This ongoing interplay between theory and experiment distinguishes physics from other intellectual disciplines, even near neighbors like philosophy or mathematics.

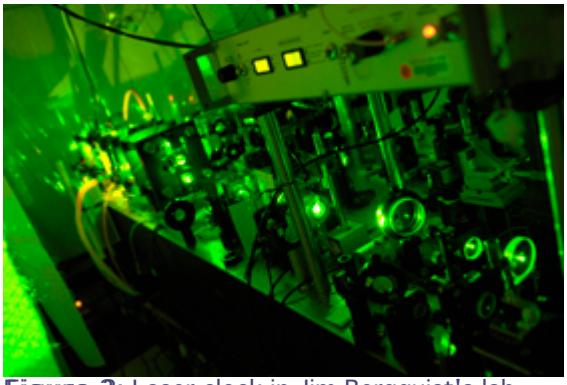


Figure 2: Laser clock in Jim Bergquist's lab.

Physics has had a strong tradition of reductionism, where complex systems are seen as aggregates of simpler subunits. All the substances you see around you are made of compound substances that are combinations of the elements (carbon, oxygen, hydrogen...) that comprise the periodic table. But the atoms in these elements, which are defined by the number of

protons in the respective atomic nuclei, are themselves made of protons, neutrons, and electrons. We now know that protons and neutrons are in turn composite objects, made up of yet more elemental objects called quarks. If the basic ingredients and their mutual interactions are well understood, the properties of remarkably complex situations can be understood and even predicted. For example, the structure of the periodic table can be understood by combining the laws of quantum mechanics with the properties of protons, neutrons, and electrons.

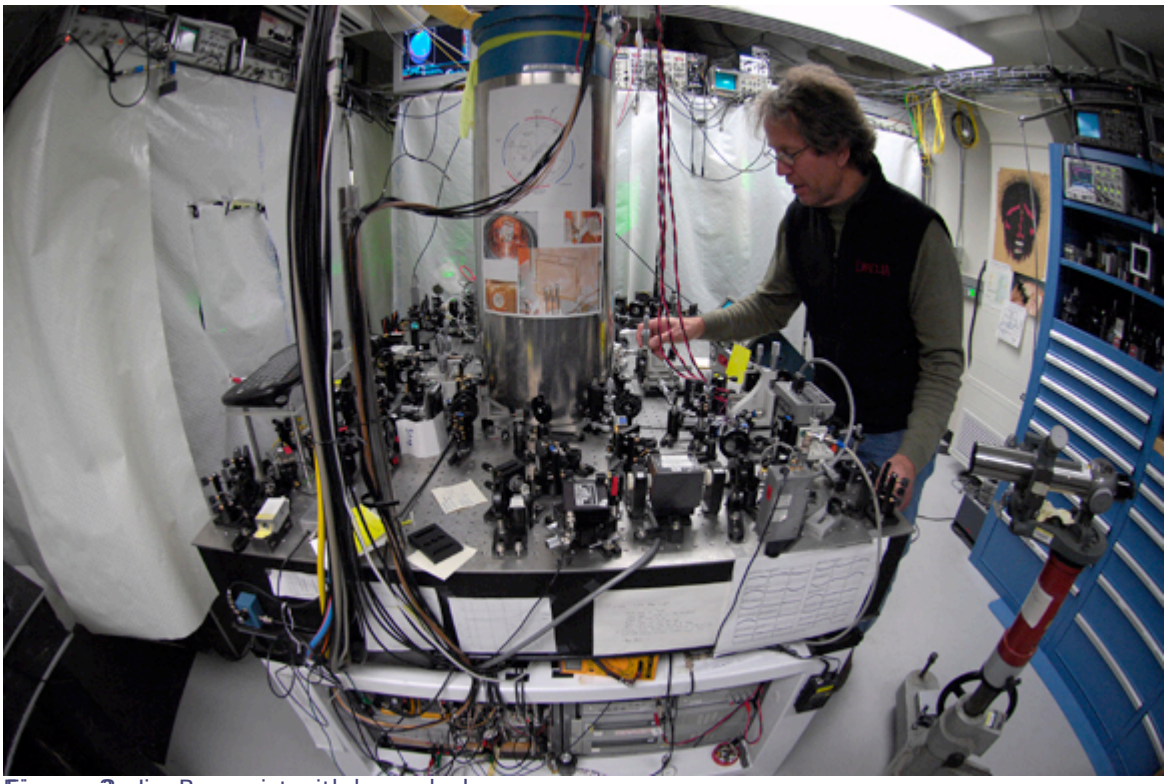


Figure 3: Jim Bergquist with laser clock.

Physical systems that contain billions of atoms or particles acquire bulk properties that appear at first sight to be amenable to the traditional reductionist approach, although concepts like temperature and entropy are really only meaningful (or perhaps more accurately, useful) for these aggregate systems with large numbers of particles. The behavior of these many-body systems can often be described in terms of different levels of abstraction. For example, some aspects of the complicated interactions between light and glass can be summarized in terms of an index of refraction, that is independent of the details of the complex underlying phenomena.

Introduction To Online Text *By Christopher Stubbs*

Emergence

Figure 4: Fermilab researchers.

While physicists have had remarkable successes in understanding the world in terms of its basic constituents and their fundamental interactions, physicists also now recognize that the reductionist approach has very real limitations. For example, even if we knew the inventory of all the objects that made up some physical system, and their initial configuration and interactions, there are both practical and intrinsic limitations to our ability to predict the system's behavior at all future times. Even the most powerful computers have a limitation to the resolution with which numbers can be represented, and eventually computational roundoff errors come into play, degrading our ability to replicate nature in a computer once we are dealing with more than a few dozen objects in a system for which the fundamental interactions are well-known.

As one of our authors, David Pines, writes:

> An interesting and profound change in perspective is the issue of the emergent properties of matter. When we bring together the component parts of any system, be it people in a society or matter in bulk, the behavior of the whole is very different from that of its parts, and we call the resulting behavior emergent. Emergence is a bulk property. Thus matter in bulk acquires properties that are different from those of its fundamental constituents (electrons and nuclei) and we now recognize that a knowledge of their interactions does not make it possible to predict its properties, whether one is trying to determine whether a material becomes, say, an antiferromagnet or a novel superconductor, to say nothing of the behavior of a cell in living matter or the behavior of the neurons in the human brain. Feynman famously said: "life is nothing but the wiggling and jiggling of atoms," but this does not tell us how these gave rise to LUCA, the last universal ancestor that is the progenitor of living matter, to say nothing of its subsequent evolution. It follows that we need to rethink the role of reductionism in understanding emergent behavior in physics or

biology.

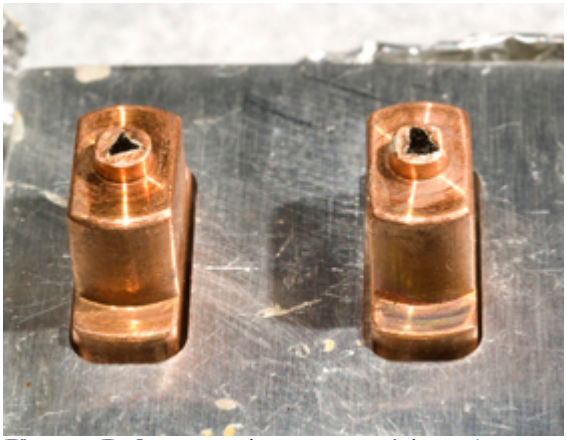


Figure 5: Superconductor materials at Jenny Hoffman's Lab.

Understanding emergent behavior requires a change of focus. Instead of adopting the traditional reductionist approach that begins by identifying the individual constituents (quarks, electrons, atoms, individuals) and uses these as the basic building blocks for constructing a model that describes emergent behavior, we focus instead on identifying the collective organizing concepts and principles that lead to or characterize emergent behavior, and treat these as the basic building blocks of models of emergence.

Both the reductionist and the scientist with an emergent perspective focus on fundamentals. For the reductionist these are the individual constituents of the system, and the forces that couple them. For the scientist with an emergent perspective on matter in bulk, the fundamentals are the collective organizing principles that bring about the emergent behavior of the system as a whole, from the second law of thermodynamics to the mechanisms producing the novel coherent states of matter that emerge as a material seeks to reduce its entropy as its temperature is lowered.

Introduction To Online Text *By Christopher Stubbs*

The plan of this course

The course begins with the classic reductionist perspective, the search for the basic building blocks of nature. Their identification is described by Natalie Roe, in Unit 1, and the fundamental interactions on the subatomic scale are reviewed by David Kaplan in Unit 2. On human length scales and larger, one of the main forces at work is gravity, discussed in Chapter 3 by Blayne Heckel.

Figure 6: Penzias and Wilson horn antenna at Holmdel, NJ.

In Unit 4, Shamut Kachru takes on the issue of developing a theoretical framework that might be capable of helping us understand the very early universe, when gravity and quantum mechanics play equally important roles. He describes the current status of the string theories that seek to develop a unified quantum theory of gravitation and the other forces acting between elementary particles. Note, however, that while these exciting developments are regarded as advances in physics, they are presently in tension with the view that physics is an experiment-based science, in that we have yet to identify accessible measurements that can support or refute the string theory viewpoint.

Conveying the complexities of quantum mechanics in an accessible and meaningful way is the next major challenge for the course. The beginning of the 20th century was the advent of quantum mechanics, with the recognition that the world can't always be approximated as collection of billiard balls. Instead, we must accept the fact that experiments demand a counter-intuitive and inherently probabilistic description. In Unit 5, Daniel Kleppner introduces the basic ideas of quantum mechanics. This is followed in Unit 6, by Bill Reinhart with a description of instances where quantum properties are exhibited on an accessible (macroscopic) scale, while in Unit 7, Lene Hau shows how the subtle interactions between light

and matter can be exploited to produce remarkable effects, such as slowing light to a speed that a child could likely outrun on a bicycle.

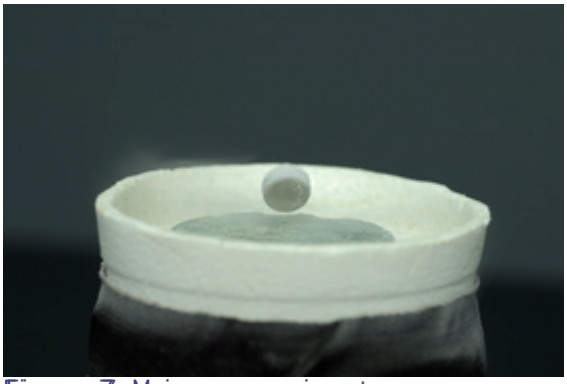


Figure 7: Meissner experiment.

Emergence is introduced in Unit 8, where David Pines presents an emergent perspective on basic topics that are included in many existing courses on condensed matter, and then describes some of the exciting new results on quantum matter that require new organizing principles and the new experimental probes that have helped generate these. The methodology of physics, the instrumentation that is derived from physics labs, and the search for the organizing principles responsible for emergent behavior in living matter, can provide valuable insights in biological systems, and the extent to which these are doing so is discussed by Robert Austin in Unit 9, "Biophysics."

About 90% of the mass in galaxies like the Milky Way comprises "dark matter" whose composition and distribution is unknown. The evidence for dark matter, and the searches under way to find it, are described by Peter Fisher in Unit 10.

Another indication of the work that lies ahead in constructing a complete and consistent intellectual framework by which to understand the universe is found in Unit 11, by Robert Kirshner. In the 1920s astronomers found that the universe was expanding. Professor Kirshner describes the astonishing discovery in the late 1990s that the expansion is happening at an ever-increasing rate. This seems to come about due to a repulsive gravitational interaction between regions of empty space. Understanding the nature of the "dark energy" that is driving the accelerating expansion is a complete mystery, and will likely occupy the astrophysical community for years to come.

Physics for the 21st Century

Introduction To Online Text *By Christopher Stubbs*

An anthropic, accidental universe?

The long term goal of the reductionist agenda in physics is to arrive at a single, elegant, unified "theory of everything" (TOE), some mathematical principle that can be shown to be the single unique description of the physical universe. An aspiration of this approach is that the parameters of this model, such as the charge of the electron, the mass ratio of quarks to neutrinos, and the strengths of the fundamental interactions, would be determined by some grand principle, which has, so far, remained elusive.

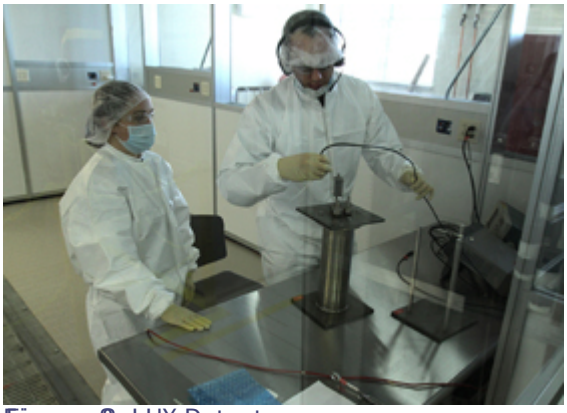


Figure 8: LUX Detector.

In the past decade an alternative point of view has arisen: that the basic parameters of the universe come about not from some grand principle, but are instead constrained by the simple fact that we are here to ask the questions. Universes with ten times as much dark energy would have been blown apart before stars and galaxies had a chance to assemble, and hence would not be subjected to scientific scrutiny, since no scientists would be there to inquire. Conversely, if gravity were a thousand times stronger, stellar evolution would go awry and again no scientists would have appeared on the scene.

The "anthropic" viewpoint, that the basic parameters of the universe we see are constrained by the presence of humans rather than some grand physical unification principle, is seen by many physicists as a retreat from the scientific tradition that has served as the intellectual beacon for generations of physicists. Other scientists accept the notion that the

Figure 9: Andreas Hirstius with some CERN computers.

properties of the universe we see are an almost accidental consequence of the conditions needed for our feeble life forms to evolve. Indeed, whether this issue is a scientific question, which can be resolved on the basis of informed dialogue based upon observational evidence, is itself a topic of debate.

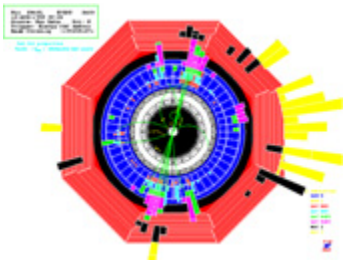
Exploring the anthropic debate is arguably a sensible next step, following the dark energy discussion in Unit 11.

Welcome to the research frontier.

Enjoy.

Physics for the 21st Century

Unit 2: The Fundamental Interactions



© SLAC National Accelerator Laboratory.

Unit Overview

This unit takes the story of the basic constituents of matter beyond the fundamental particles that we encountered in unit 1. It focuses on the interactions that hold those particles together or tear them asunder. Many of the forces responsible for those interactions are basically the same even though they manifest themselves in different ways. Today we recognize four fundamental forces: gravity, electromagnetism, and the strong and weak nuclear forces. Detailed studies of those forces suggest that the last three—and possibly all four—were themselves identical when the universe was young, but have since gone their own way. But while physicists target a grand unification theory that combines all four forces, they also seek evidence of the existence of new forces of nature.

Content for This Unit

Sections:

Section 1: Introduction

The underlying theory of the physical world has two fundamental components: Matter and its interactions. We examined the nature of matter in the previous unit. Now we turn to interactions, or the forces between particles. Just as the forms of matter we encounter on a daily basis can be broken down into their constituent fundamental particles, the forces we experience can be broken down on a microscopic level. We know of four fundamental forces: electromagnetism, gravity, the strong nuclear force, and the weak force.

Electromagnetism causes almost every physical phenomenon we encounter in our everyday life: light, sound, the existence of solid objects, fire, chemistry, all biological phenomena, and color, to name a few. Gravity is, of course, responsible for the attraction of all things on the Earth toward its center, as well as tides—due to the pull of the Moon and the Sun on the oceans—the motions within the solar system, and even the formation of large structures in the universe, such as galaxies. The strong force takes part in all nuclear phenomena, such as fission and fusion, the latter of which occurs at the core of our Sun and all other stars. Finally, the weak force is involved in radioactivity, causing unstable atomic nuclei to decay. The latter two operate only at microscopic distances, while the former two clearly have significant effects on macroscopic scales.

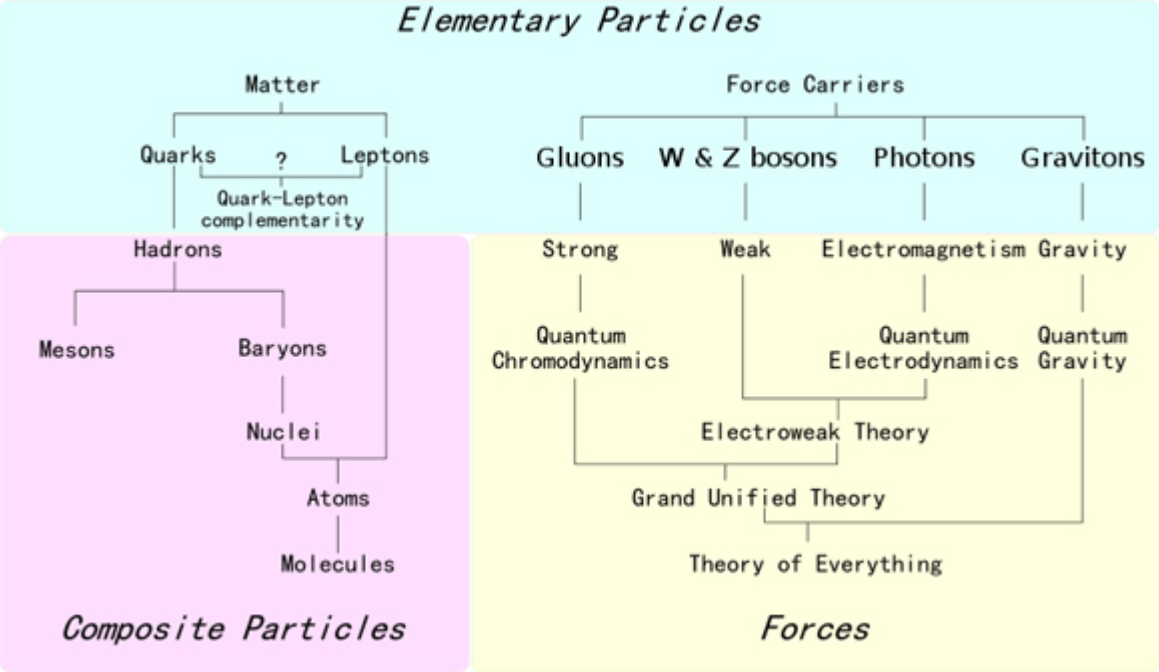


Figure 1: This chart shows the known fundamental particles, those of matter and those of force.
Source: © Wikimedia Commons, Creative Commons Attribution ShareAlike 3.0, Author: Headbomb, 21 January 2009.

The primary goal of physics is to write down theories—sets of rules cast into mathematical equations—that describe and predict the properties of the physical world. The eye is always toward simplicity and unification—simple rules that predict the phenomena we experience (e.g., all objects fall to the Earth with the same acceleration), and unifying principles which describe vastly different phenomena (e.g., the force that keeps objects on Earth is the same as the force that predicts the motion of planets in the solar system). The search is for the "Laws of Nature." Often, the approach is to look at the smallest constituents, because that is where the action of the laws is the simplest. We have already seen that the fundamental particles of the Standard Model naturally fall into a periodic table-like order. We now need a microscopic theory of forces to describe how these particles interact and come together to form larger chunks of matter such as protons, atoms, grains of sand, stars, and galaxies.

In this unit, we will discover a number of the astounding unifying principles of particle physics: First, that forces themselves can be described as particles, too—force particles exchanged between matter particles. Then, that particles are not fundamental at all, which is why they can disappear and reappear at particle colliders. Next, that subatomic physics is best described by a new mathematical framework called quantum field theory (QFT), where the distinction between particle and force is no longer clear. And

finally, that all four fundamental forces seem to operate under the same basic rules, suggesting a deeper unifying principle of all forces of Nature.

Many forms of force

How do we define a force? And what is special about the fundamental forces? We can start to answer those questions by observing the many kinds of forces at work in daily life—gravity, friction, "normal forces" that appear when a surface presses against another surface, and pressure from a gas, the wind, or the tension in a taut rope. While we normally label and describe these forces differently, many of them are a result of the same forces between atoms, just manifesting in different ways.

Figure 2: An example of conservative (right) and non-conservative (left) forces.
Source: © Left: Jon Ovington, Creative Commons Attribution-ShareAlike 2.0 Generic License. Right: Ben Crowell, lightandmatter.com, Creative Commons Attribution-ShareAlike 3.0 License.

At the macroscopic level, physicists sometimes place forces in one of two categories: conservative forces that exchange potential and kinetic energy, such as a sled sliding down a snowy hill; and non-conservative forces that transform kinetic energy into heat or some other dissipative type of energy. The former is characterized by its reversibility, and the latter by its irreversibility: It is no problem to push the sled back up the snowy hill, but putting heat back into a toaster won't generate an electric current.

But what is force? Better yet, what is the most useful description of a force between two objects? It depends significantly on the size and relative velocity of the two objects. If the objects are at rest or moving much more slowly than the speed of light with respect to each other, we have a perfectly fine

description of a static force. For example, the force between the Earth and Sun is given to a very good approximation by Newton's law of universal gravitation, which only depends on their masses and the distance between them. In fact, the formulation of forces by Isaac Newton in the 17th century best describes macroscopic static forces. However, Newton did not characterize the rules of other kinds of forces beyond gravity. In addition, the situation gets more complicated when fundamental particles moving very fast interact with one another.

Forces at the microscopic level

At short distances, forces can often be described as individual particles interacting with one another. These interactions can be characterized by the exchange of energy (and momentum). For example, when a car skids to a stop, the molecules in the tires are crashing into molecules that make up the surface of the road, causing them to vibrate more on the tire (i.e., heat up the tire) or to be ripped from their bonds with the tire (i.e., create skid marks).

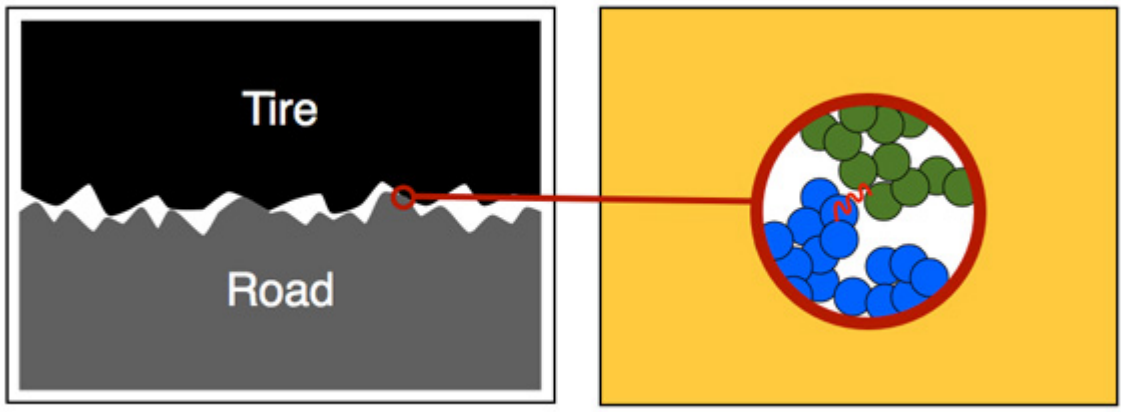


Figure 3: A microscopic view of friction.
Source: © David Kaplan.

When particles interact, they conserve energy and momentum, meaning the total energy and momentum of a set of particles before the interaction occurs is the same as the total energy and momentum afterward. At the particle level, a conservative interaction would be one where two particles come together, interact, and then fly apart after exchanging some amount of energy. After a non-conservative interaction, some of the energy would be carried off by radiation. The radiation, as we shall see, can also be described as particles (such as photons and particles of light).

Light speed and light-meters

The constant c, the speed of light, serves in some sense to convert units of mass to units of energy. When we measure length in meters and time in seconds, then c^2 ~ 90,000,000,000,000,000. However, as it appears in Einstein's famous equation, $E=mc^2$, the c^2 converts mass to energy, and could be measured in ergs per gram. In that formulation, the value of c^2 tells us that the amount of energy stored in the mass of a pencil is roughly equal to the amount of energy used by the entire state of New York in 2007. Unfortunately, we do not have the capacity to make that conversion due to the stability of the proton, and the paucity of available anti-matter.

When conserving energy, however, one must take into account Einstein's relativity—especially if the speeds of the particles are approaching the speed of light. For a particle in the vacuum, one can characterize just two types of energy: The energy of motion and the energy of mass. The latter, summarized in the famous equation $E=mc^2$, suggests that mass itself is a form of energy. However, Einstein's full equation is more complicated. In particular, it involves an object's momentum, which depends on the object's mass and its velocity. This applies to macroscopic objects as well as those at the ultra-small scale. For example, chemical and nuclear energy is energy stored in the mass difference between molecules or nuclei before and after a reaction. When you switch on a flashlight, for example, it loses its mass to the energy of the photons leaving it—and actually becomes lighter! See the math

When describing interactions between fundamental particles at very high energies, it is helpful to use an approximation called the relativistic limit, in which we ignore the mass of the particles. In this situation, the momentum energy is much larger than the mass energy, and the objects are moving at nearly the speed of light. These conditions occur in particle accelerators. But they also existed soon after the Big Bang when the universe was at very high temperatures and the particles that made up the universe had large momenta. As we will explain later in this unit, we expect new fundamental forces of nature to reveal themselves in this regime. So, as in Unit 1, we will focus on high-energy physics as a way to probe the underlying theory of force.

Section 2: Forces and Fundamental Interactions

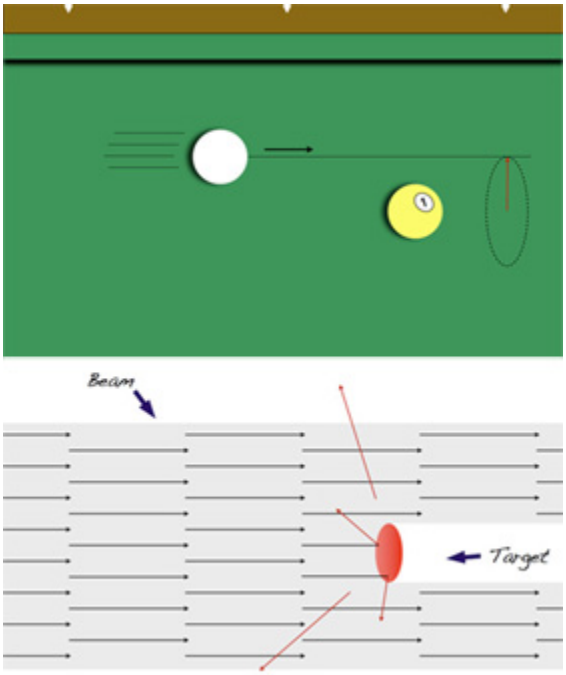


Figure 4: Two examples of a scattering cross section.
Source: © David Kaplan.

A way to measure the fundamental forces between particles is by measuring the probability that the particles will scatter off each other when one is directed toward the other at a given energy. We quantify this probability as an effective cross sectional area, or cross section, of the target particle. The concept of a cross section applies in more familiar examples of scattering as well. For example, the cross section of a billiard ball (See Figure 4) is area at which the on coming ball's center has to be aimed in order for the balls to collide. In the limit that the white ball is infinitesimally small, this is simply the cross sectional area of the target (yellow) ball.

The cross section of a particle in an accelerator is similar conceptually. It is an effective size of the particle—like the size of the billiard ball—that not only depends on the strength and properties of the force between the scattering particles, but also on the energy of the incoming particles. The beam of particles comes in, sees the cross section of the target, and some fraction of them scatter, as illustrated in the bottom of Figure 4. Thus, from a cross section, and the properties of the beam, we can derive a probability of scattering.

Figure 5: This movie shows the simplest way two electrons can scatter.
Source: © David Kaplan

The simplest way two particles can interact is to exchange some momentum. After the interaction, the particles still have the same internal properties but are moving at different speeds in different directions. This is what happens with the billiard balls, and is called "elastic scattering." In calculating the elastic scattering cross section of two particles, we can often make the approximation depicted in Figure 5. Here, the two particles move freely toward each other; they interact once at a single point and exchange some momentum; and then they continue on their way as free particles. The theory of the interaction contains information about the probabilities of momentum exchange, the interactions between the quantum mechanical properties of the two particles known as spins, and a dimensionless parameter, or coupling, whose size effectively determines the strength of the force at a given energy of an incoming particle.

Spin in the Quantum Sense

In the everyday world, we identify items through their physical characteristics: size, weight, and color, for example. Physicists have their own identifiers for elementary particles. Called "quantum numbers," these portray attributes of particles that are conserved, such as energy and momentum. Physicists describe one particular characteristic as "spin."

The name stemmed from the original interpretation of the attribute as the amount and direction in which a particle rotated around its axis. The spin of, say, an electron could take two values, corresponding to clockwise or counterclockwise rotation along a given axis. Physicists now understand that the concept is more complex than that, as we will see later in this unit and in Unit 6. However, it has critical importance in interactions between particles.

The concept of spin has value beyond particle physics. Magnetic resonance imaging, for example, relies on changes in the spin of hydrogen nuclei from one state to another. That enables MRI machines to locate the hydrogen atoms, and hence water molecules, in patients' bodies—a critical factor in diagnosing ailments.

Such an approximation, that the interaction between particles happens at a single point in space at a single moment in time, may seem silly. The force between a magnet and a refrigerator, for example, acts over a distance much larger than the size of an atom. However, when the particles in question are moving fast enough, this approximation turns out to be quite accurate—in some cases extremely so. This is in part due to the probabilistic nature of quantum mechanics, a topic treated in depth in Unit 5. When we are working with small distances and short times, we are clearly in the quantum mechanical regime.

We can approximate the interaction between two particles as the exchange of a new particle between them called a force carrier. One particle emits the force carrier and the other absorbs it. In the intermediate steps of the process—when the force carrier is emitted and absorbed—it would normally be impossible to conserve energy and momentum. However, the rules of quantum mechanics govern particle interactions, and those rules have a loophole.

The loophole that allows force carriers to appear and disappear as particles interact is called the Heisenberg uncertainty principle. German physicist Werner Heisenberg outlined the uncertainty principle named for him in 1927. It places limits on how well we can know the values of certain physical

parameters. The uncertainty principle permits a distribution around the "correct" or "classical" energy and momentum at short distances and over short times. The effect is too small to notice in everyday life, but becomes powerfully evident over the short distances and times experienced in high-energy physics. While the emission and absorption of the force carrier respect the conservation of energy and momentum, the exchanged force carrier particle itself does not. The force carrier particle does not have a definite mass and in fact doesn't even know which particle emitted it and which absorbed it. The exchanged particles are unobservable directly, and thus are called virtual particles.

Feynman, Fine Physicist

Richard Feynman, 1962.
Source: © AIP Emilio Segrè Visual Archives, Segrè Collection.

During a glittering physics career, Richard Feynman did far more than create the diagrams that carry his name. In his 20s, he joined the fraternity of atomic scientists in the Manhattan Project who developed the atom bomb. After World War II, he played a major role in developing quantum electrodynamics, an achievement that won him the Nobel Prize in physics. He made key contributions to understanding the nature of superfluidity and to aspects of particle physics. He has also been credited with pioneering the field of quantum computing and introducing the concept of nanotechnology.

Feynman's contributions went beyond physics. As a member of the panel that investigated the 1986 explosion of the space shuttle Challenger, he unearthed serious misunderstandings of basic concepts by NASA's managers that helped to foment the disaster. He took great interest in biology and did much to popularize science through books and lectures. Eventually, Feynman became one of the world's most recognized scientists, and is considered the best expositor of complex scientific concepts of his generation.

Physicists like to draw pictures of interactions like the ones shown in Figure 6. The left side of Figure 6, for example, represents the interaction between two particles through one-particle exchange. Named a Feynman diagram for American Nobel Laureate and physicist Richard Feynman, it does more than provide a qualitative representation of the interaction. Properly interpreted, it contains the instructions for

calculating the scattering cross section. Linking complicated mathematical expressions to a simple picture made the lives of theorists a lot easier.

Even more important, Feynman diagrams allow physicists to easily organize their calculations. It is in fact unknown how to compute most scattering cross sections exactly (or analytically). Therefore, physicists make a series of approximations, dividing the calculation into pieces of decreasing significance. The Feynman diagram on the left side of Figure 6 corresponds to the first level of approximation—the most significant contribution to the cross section that would be evaluated first. If you want to calculate the cross section more accurately, you will need to evaluate the next most important group of terms in the approximation, given by diagrams with a single loop, like the one on the right side of Figure 6. By drawing every possible diagram with the same number of loops, physicists can be sure they haven't accidentally left out a piece of the calculation.

Feynman diagrams are far more than simple pictures. They are tools that facilitate the calculation of how particles interact in situations that range from high-energy collisions inside particle accelerators to the interaction of the constituent parts of a single, trapped ion. As we will see in Unit 5, one of the most precise experimental tests of quantum field theory compares a calculation based on hundreds of Feynman diagrams to the behavior of an ion in a trap. For now, we will focus on the conceptually simpler interaction of individual particles exchanging a virtual force carrier.

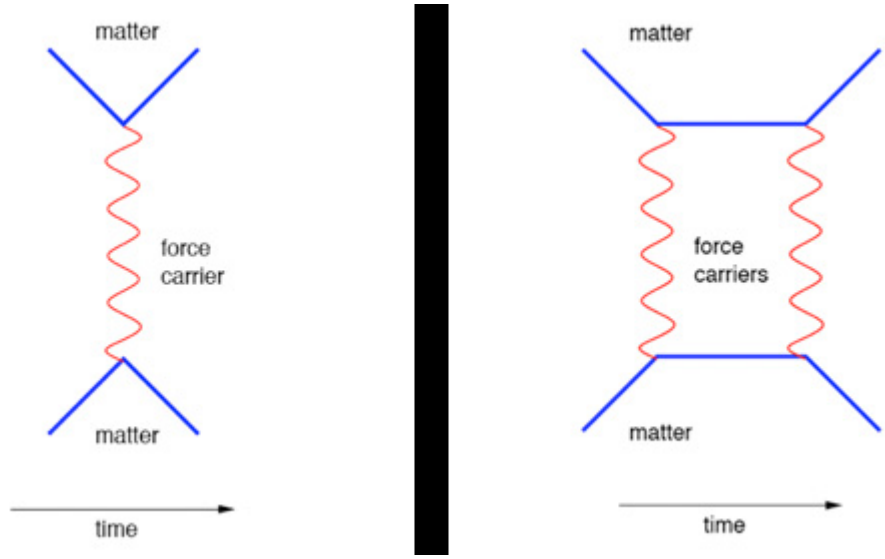


Figure 6: Feynman diagram representing a simple scattering of two particles (left) and a more complicated scattering process involving two particles (right).
Source: © David Kaplan.

Section 3: Fields Are Fundamental

Figure 7: When an electron and its antiparticle collide, they annihilate and new particles are created.
Source: © David Kaplan.

At a particle collider, it is possible for an electron and an antielectron to collide at a very high energy. The particles annihilate each other, and then two new particles, a muon and an antimuon, come out of the collision. There are two remarkable things about such an event, which has occurred literally a million times at the LEP collider that ran throughout the 1990s at CERN. First, the muon is 200 times heavier than the electron. We see in a dramatic way that mass is not conserved—that the kinetic energy of the electrons can be converted into mass for the muon. $E = mc^2$, again. Mass is not a fundamental quantity.

The second remarkable thing is that particles like electrons and muons can appear and disappear, and thus they are, in some sense, not fundamental. In fact, all particles seem to have this property. Then what is fundamental? In response to this question, physicists define something called a field. A field fills all of space, and the field can, in a sense, vibrate in a way that is analogous to ripples on a lake. The places a field vibrates are places that contain energy, and those little pockets of energy are what we call (and have the properties of) particles.

As an analogy, imagine a lake. A pebble is dropped in the lake, and a wave from the splash travels away from the point of impact. That wave contains energy. We can describe that package of localized energy living in the wave as a particle. One can throw a few pebbles in the lake at the same time and create multiple waves (or particles). What is fundamental then is not the particle (wave), it is the lake itself (field).

In addition, the wave (or particle) would have different properties if the lake were made of water or of, say, molasses. Different fields allow for the creation of different kinds of particles.

Figure 8: Ripples in lake from a rock.
Source: © Adam Kleppner.

To describe a familiar particle such as the electron in a quantum field theory, physicists consider the possible ways the electron field can be excited. Physicists say that an electron is the one particle state of the electron field—a state well defined before the electron is ever created. The quantum field description of particles has one important implication: Every electron has exactly the same internal properties—the charge, spin, and mass for each electron exactly matches that for every other one. In addition, the symmetries inherent in relativity require that every particle has an antiparticle with opposite spin, electric charge, and other charges. Some uncharged particles, such as photons, act as their own antiparticles.

A crucial distinction

In general, the fact that all particles—matter or force carriers—are excitations of fields is the great unifying concept of quantum field theory. The excitations all evolve in time like waves, and they interact at points in spacetime like particles. However, the theory contains one crucial distinction between matter and force carriers. This relates to the internal spin of the particles.

By definition, all matter particles, such as electrons, protons, and neutrons, as well as quarks, come with a half-unit of spin. It turns out in quantum mechanics that a particle's spin is related to its angular momentum, which, like energy and linear momentum, is a conserved quantity. While a particle's linear momentum depends on its mass and velocity, its angular momentum depends on its mass and the speed at which it rotates about its axis. Angular momentum is quantized—it can take on values only

in multiples of Planck's constant, $\hbar$ = 1.05 x 10^{-34} Joule-seconds. So the smallest amount by which an object's angular momentum can change is $\hbar$. This value is so small that we don't notice it in normal life. However, it tightly restricts the physical states allowed for the tiny angular moment in atoms. Just to relate these amounts to our everyday experience, a typical spinning top can have an angular momentum of 1,000,000,000,000,000,000,000,000,000,000 times $\hbar$. If you change the angular momentum of the top by multiples of $\hbar$, you may as well be changing it continuously. This is why we don't see the quantum-mechanical nature of spin in everyday life.

Now, force carriers all have integer units of internal spin; no fractions are allowed. When these particles are emitted or absorbed, their spin can be exchanged with the rotational motion of the particles, thus conserving angular momentum. Particles with half-integer spin cannot be absorbed, because the smallest unit of rotational angular momentum is one times $\hbar$. Physicists call particles with half-integer spin fermions. Those with integer (including zero) spins they name bosons.

Not your grandmother's ether theory

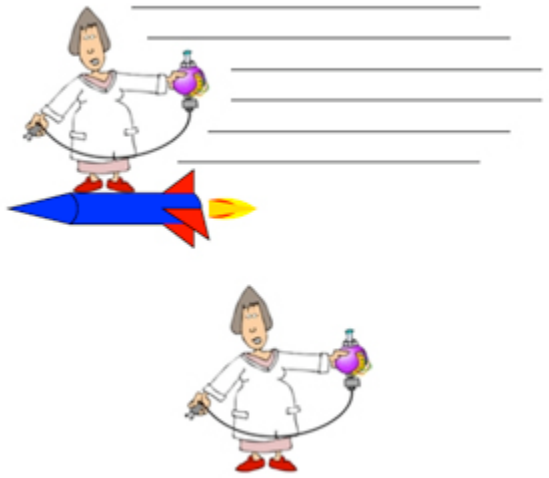


Figure 9: Experimental results remain the same whether they are performed at rest or at a constant velocity.
Source: © David Kaplan.

An important quantum state in the theory is the "zero-particle state," or vacuum. The fact that spacetime is filled with quantum fields makes the vacuum much more than inactive empty space. As we shall see later in this unit, the vacuum state of fields can change the mass and properties of particles. It also

contributes to the energy of spacetime itself. But the vacuum of spacetime appears to be relativistic; in other words, it is best described by the theory of relativity. For example, in Figure 9, a scientist performing an experiment out in empty space will receive the same result as a scientist carrying out the same experiment while moving at a constant velocity relative to the first. So we should not compare the fields that fill space too closely with a material or gas. Moving through air at a constant velocity can affect the experiment because of air resistance. The fields, however, have no preferred "at rest" frame. Thus, moving relative to someone else does not give a scientist or an experiment a distinctive experience. This is what distinguishes quantum field theory from the so-called "ether" theories of light of a century ago.

Section 4: Early Unification for Electromagnetism

The electromagnetic force dominates human experience. Apart from the Earth's gravitational pull, nearly every physical interaction we encounter involves electric and/or magnetic fields. The electric forces we constantly experience have to do with the nature of the atoms we're made of. Particles can carry electric charge, either positive or negative. Particles with the same electric charge repel one another, and particles with opposite electric charges attract each other. An atom consists of negatively charged electrons in the electric field of a nucleus, which is a collection of neutrons and positively charged protons. The negatively charged electrons are bound to the positively charged nucleus.

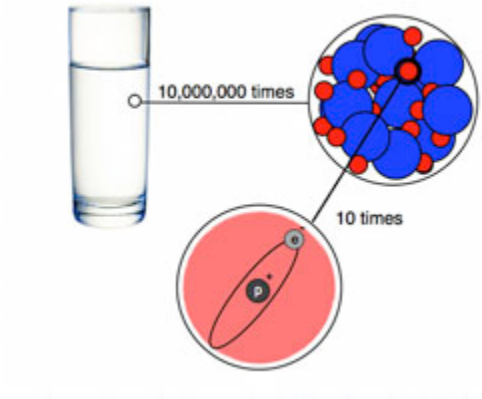


Figure 10: The electromagnetic force and the constituents of matter.
Source: © David Kaplan.

Although atoms are electrically neutral, they can attract each other and bind together, partly because atoms do have oppositely charged component parts and partly due to the quantum nature of the states in which the electrons find themselves (see Unit 6). Thus, molecules exist owing to the electric force. The residual electric force from electrons and protons in molecules allows the molecules to join up in macroscopic numbers and create solid objects. The same force holds molecules together more weakly in liquids. Similarly, electric forces allow waves to travel through gases. Thus, sound is a consequence of electric force, and so are many other common phenomena, including electricity, friction, and car accidents.

We experience magnetic force from materials such as iron and nickel. At the fundamental level, however, magnetic fields are produced by moving electric charges, such as electric currents in wires, and spinning particles, such as electrons in magnetic materials. So, we can understand both electric and magnetic

forces as the effects of classical electric and magnetic fields produced by charged particles acting on other charged particles.

The close connection between electricity and magnetism emerged in the 19th century. In the 1830s, English scientist Michael Faraday discovered that changing magnetic fields produced electric fields. In 1861, Scottish physicist James Clerk Maxwell postulated that the opposite should be true: A changing electric field would produce a magnetic field. Maxwell developed equations that seemed to describe all electric and magnetic phenomena. His solutions to the equations described waves of electric and magnetic fields propagating through space—at speeds that matched the experimental value of the speed of light. Those equations provided a unified theory of electricity, magnetism, and light, as well as all other types of electromagnetic radiation, including infrared and ultraviolet light, radio waves, microwaves, x-rays, and gamma rays.

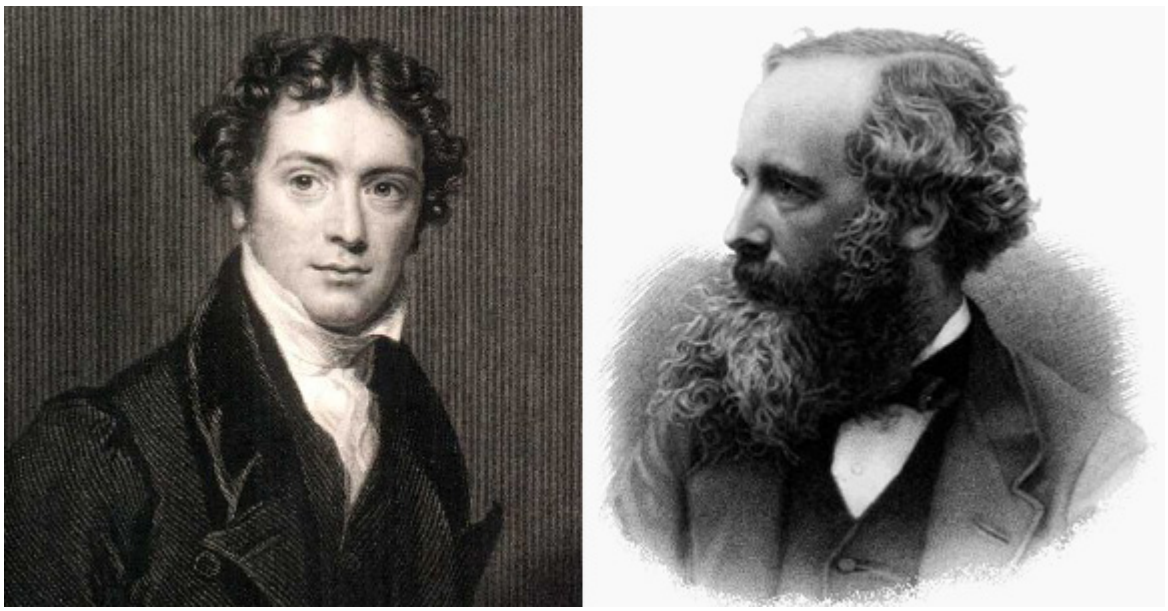


Figure 11: Michael Faraday (left) and James Clerk Maxwell (right) unified electricity and magnetism in classical field theory.
Source: © Wikimedia Commons, Public Domain.

Maxwell's description of electromagnetic interactions is an example of a classical field theory. His theory involves fields that extend everywhere in space, and the fields determine how matter will interact; however, quantum effects are not included.

The photon field

Einstein's Role in the Quantum Revolution

The Nobel Prize for physics that Albert Einstein received in 1921 did not reward his special or general theory of relativity. Rather, it recognized his counterintuitive theoretical insight into the photoelectric effect—the emission of electrons when light shines on the surface of a metal. That insight, developed during Einstein's "miracle year" of 1905, inspired the development of quantum theory.

Experiments by Philipp Lenard in 1902, 15 years after his mentor Heinrich Hertz first observed the photoelectric effect, showed that increasing the intensity of the light had no effect on the average energy carried by each emitted electron. Further, only light above a certain threshold frequency stimulated the emission of electrons. The prevailing concept of light as waves couldn't account for those facts.

Einstein made the astonishing conjecture that light came in tiny packets, or quanta, of the type recently proposed by Max Planck. Only those packets with sufficient frequency would possess enough energy to dislodge electrons. And increasing the light's intensity wouldn't affect individual electrons' energy because each electron is dislodged by a single photon. American experimentalist Robert Millikan took a skeptical view of Einstein's approach. But his precise studies upheld the theory, proving that light existed in wave and particle forms, earning Millikan his own Nobel Prize in 1923 and—as we shall see in Unit 5—laying the foundation of full-blown quantum mechanics.

In the quantum description of the electromagnetic force, there is a particle which plays the role of the force carrier. That particle is called the photon. When the photon is a virtual particle, it mediates the force between charged particles. Real photons, though, are the particle version of the electromagnetic wave, meaning that a photon is a particle of light. It was Albert Einstein who realized particle-wave duality—his study of the photoelectric effect showed the particle nature of the electromagnetic field and won him the Nobel Prize.

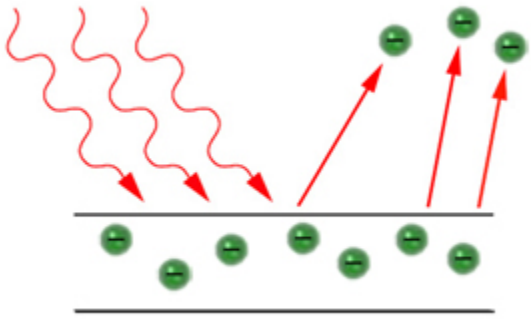


Figure 12: When light shines on a metal, electrons pop out.
Source:

Here, we should make a distinction between what we mean by the electromagnetic field and the fields that fill the vacuum from the last section. The photon field is the one that characterizes the photon particle, and photons are vibrations in the photon field. However, charged particles—for instance, those in the nucleus of an atom—are surrounded by an electromagnetic field, which is in fact the photon field "turned on". An analogy can be made with the string of a violin. An untouched string would be the dormant photon field. If one pulls the middle of the string without letting go, tension (and energy) is added to the string and the shape is distorted—this is what happens to the photon field around a stationary nucleus. And in that circumstance for historical reasons it is called the "electromagnetic field." If the string is plucked, vibrations move up and down the string. If we jiggle the nucleus, an electromagnetic wave leaves the nucleus and travels the speed of light. That wave, a vibration of the photon field, can be called a "photon."

So in general, there are dormant fields that carry all the information about the particles. Then, there are static fields, which are the dormant fields turned on but stationary. Finally, there are the vibrating fields (like the waves in the lake), which (by their quantum nature) can be described as particles.

The power of QED

The full quantum field theory describing charged particles and electromagnetic interactions is called quantum electrodynamics, or QED. In QED, charged particles, such as electrons, are fermions with half-integer spin that interact by exchanging photons, which are bosons with one unit of spin. Photons can be radiated from charged particles when they are accelerated, or excited atoms where the spin of the atom

changes when the photon is emitted. Photons, with integer spin, are easily absorbed by or created from the photon field.

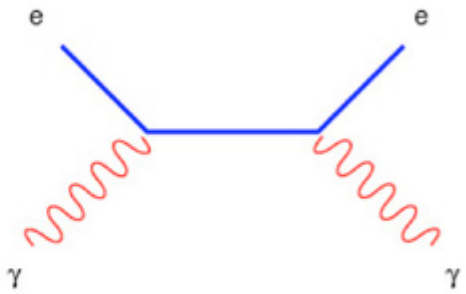


Figure 13: Arthur Holly Compton (left) discovered that the frequency of light can change as it scatters off of matter.
Source: © Left: NASA, Right: David Kaplan.

QED describes the hydrogen atom beautifully. It also describes the high-energy scattering of charged particles. Physicists can accurately compute the familiar Rutherford scattering (see Unit 1) of a beam of electrons off the nuclei of gold atoms by using a single Feynman diagram to calculate the exchange of a virtual photon between the incoming electron and the nucleus. QED also gives, to good precision, the cross section for photons scattered off electrons. This Compton scattering has value in astrophysics as well as particle physics. It is important, for example, in computing the cosmic microwave background of the universe that we will meet in Unit 4. QED also correctly predicts that gamma rays, which are high-energy photons, can annihilate and produce an electron-positron pair when their total energy is greater than the mass energy of the electron and positron, as well as the reverse process in which an electron and positron annihilate into a pair of photons.

Physicists have tested QED to unprecedented accuracy, beyond any other theory of nature. The most impressive result to date is the calculation of the anomalous magnetic moment, a_μ, a parameter related to the magnetic field around a charged particle. Physicists have compared theoretical calculations and

experimental tests that have taken several years to perform. Currently, the experimental and theoretical numbers for the muon are:

$$a_\mu^{exp} = .0011659208 \pm .0000000006$$

$$a_\mu^{th} = .0011659183 \pm .0000000006$$

These numbers reveal two remarkable facts: The sheer number of decimal places, and the remarkably close but not quite perfect match between them. The accuracy (compared to the uncorrected value of the magnetic moment) is akin to knowing the distance from New York to Los Angeles to within the width of a dime. While the mismatch is not significant enough to proclaim evidence that nature deviates from QED and the Standard Model, it gives at least a hint. More important, it reveals an avenue for exploring physics beyond the Standard Model. If a currently undiscovered heavy particle interacts with the muon, it could affect its anomalous magnetic moment and would thus contribute to the experimental value. However, the unknown particle would not be included in the calculated number, possibly explaining the discrepancy. If this discrepancy between the experimental measurement and QED calculation becomes more significant in the future, as more precise experiments are performed and more Feynman diagrams are included in the calculation, undiscovered heavy particles could make up the difference. The discrepancy would thus provide the starting point of speculation for new phenomena that physicists can seek in high-energy colliders.

Changing force in the virtual soup

The strength of the electromagnetic field around an electron depends on the charge of the electron—a bigger charge means a stronger field. The charge is often called the coupling because it represents the strength of the interaction that couples the electron and the photon (or more generally, the matter particle and the force carrier). Due to the quantum nature of the fields, the coupling actually changes with distance. This is because virtual pairs of electrons and positrons are effectively popping in and out of the vacuum at a rapid rate, thus changing the perceived charge of that single electron depending on how close you are when measuring it. This effect can be precisely computed using Feynman diagrams. Doing so reveals that the charge or the electron-photon coupling grows (gets stronger) the closer you get to the electron. This fact, as we will see in the following section, has much more important implications about the theory of the strong force. In addition, it suggests how forces of different strength could have the same strength at very short distances, as we will see in the section on the unification of forces.

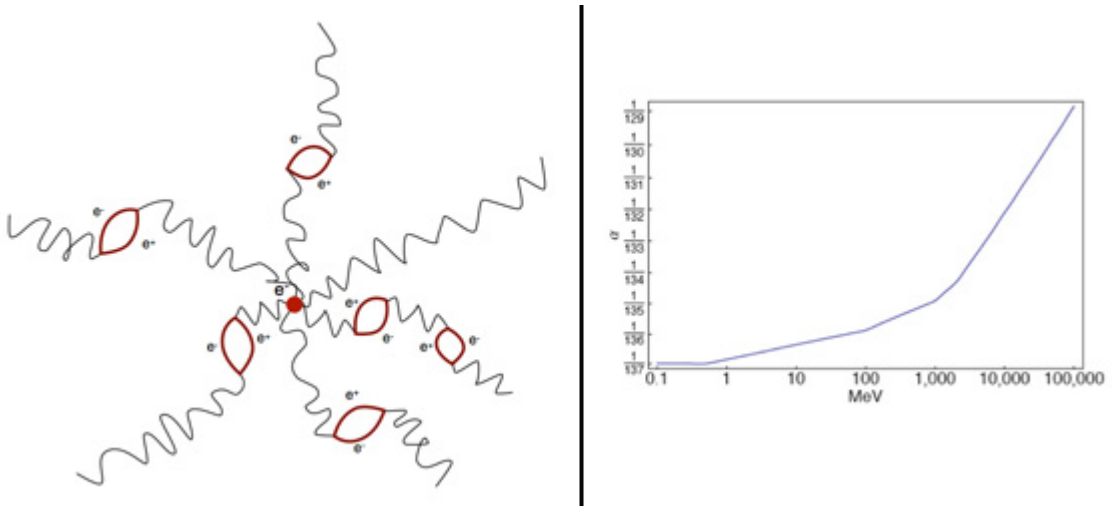


Figure 14: QED at high energies and short distances.
Source: © David Kaplan.

Section 5: The Strong Force: QCD, Hadrons, and the Lightness of Pions

The other force, in addition to the electromagnetic force, that plays a significant role in the structure of the atom is the strong nuclear force. Like the electromagnetic force, the strong force can create bound states that contain several particles. Their bound states, such as nuclei, are around 10^{-15} meters in diameter, much smaller than atoms, which are around 10^{-10} meters across. It is the energy stored in the bound nuclei that is released in nuclear fission, the reaction that takes place in nuclear power plants and nuclear weapons, and nuclear fusion, which occurs in the center of our Sun and of other stars.

Confined quarks

We can define charge as the property particles can have that allow them to interact via a particular force. The electromagnetic force, for example, occurs between particles that carry electric charge. The value of a particle's electric charge determines the details of how it will interact with other electrically charged particles. For example, electrons have one unit of negative electric charge. They feel electromagnetic forces when they are near positively charged protons, but not when they are near electrically neutral neutrinos, which have an electric charge of zero. Opposite charges attract, so the electromagnetic forces tends to create electrically neutral objects: Protons and electrons come together and make atoms, where the positive and negative charges cancel. Neutral atoms can still combine into molecules, and larger objects, as the charged parts of the atoms attract each other.

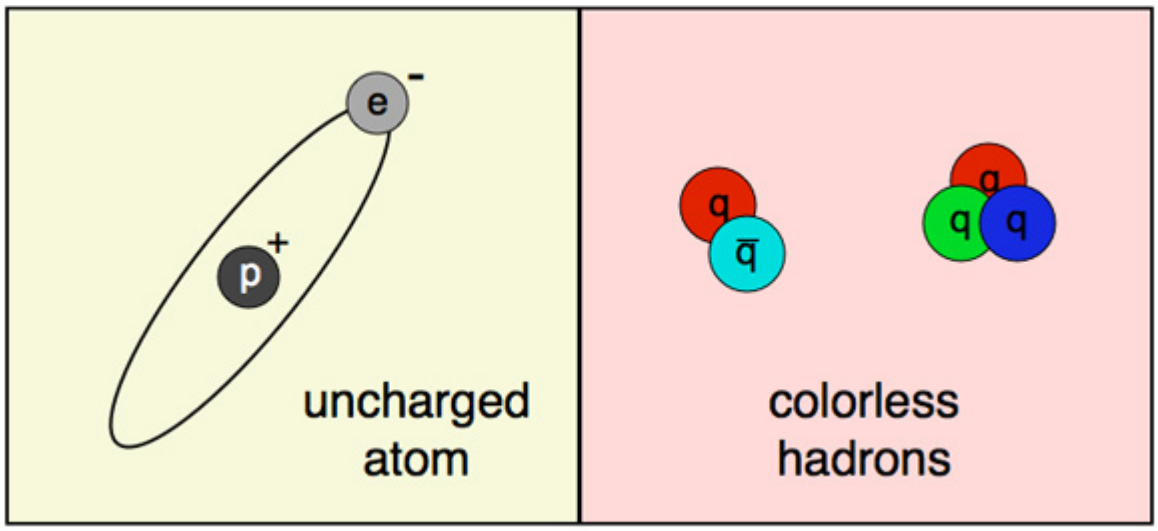


Figure 15: Neutralized charges in QED and QCD.
Source: © David Kaplan.

At the fundamental particle level, it is quarks that feel the strong force. This is because quarks have the kind of charge that allows the strong force to act on them. For the strong force, there are three types of positive charge and three types of negative charge. The three types of charge are labeled as colors—a quark can come in red, green, or blue. Antiquarks have negative charge, labeled as anti-red, etc. Quarks of three different colors will attract each other and form a color-neutral unit, as will a quark of a given color and an antiquark of the same anti-color. As with the atom and the electromagnetic force, baryons such as protons and neutrons are color-neutral (red+green+blue=white), as are mesons made of quarks and antiquarks, such as pions. Protons and neutrons can still bind and form atomic nuclei, again, in analogy to the electromagnetic force binding atoms into molecules. Electrons and other leptons do not carry color charge and therefore do not feel the strong force.

In analogy to quantum electrodynamics, the theory of the strong force is called quantum chromodynamics, or QCD. The force carrier of the strong force is the gluon, analogous to the photon of electromagnetism. A crucial difference, however, is that while the photon itself does not carry electromagnetic charge, the gluon does carry color charge—when a quark emits a gluon, that actually changes its color. Because of this, the strong force binds particles together much more tightly. Unlike the electromagnetic force, whose strength decreases as the inverse square distance between two charged particles (that is, as $1/r^2$, where r is the distance between particles), the strong force between a quark and antiquark remains constant as the distance between them grows.

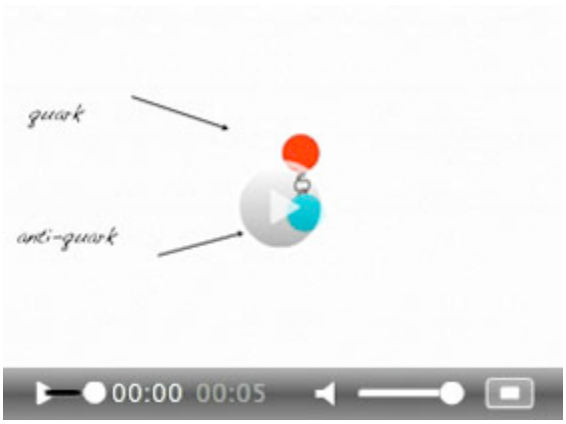


Figure 16: As two bound quarks are pulled apart, new quarks pop out of the vacuum.
Source: © David Kaplan.

The gluon field is confined to a tube that extends from the quark to the antiquark because, in a sense, the exchanged gluons themselves are attracted to each other. These gluon tubes have often been called strings. In fact, the birth of string theory came from an attempt to describe the strong interactions. It has moved on to bigger and better things, becoming the leading candidate for the theory of quantum gravity as we'll see in Unit 4.

As we pull bound quarks apart, the gluon tube cannot grow indefinitely. That is because it contains energy. Once the energy in the tube is greater than the energy required to create a new quark and antiquark, the pair pops out of the vacuum and cuts the tube into two smaller, less energetic, pieces. This fact—that quarks pop out of the vacuum to form new hadrons—has dramatic implications for collider experiments, and explains why we do not find single quarks in nature.

Particle jets

Particle collisions involving QCD can look very different than those involving QED. When a proton and an antiproton collide, one can imagine it as two globs of jelly hurling toward each other. Each glob has a few marbles embedded in them. When they collide, once in a while two marbles find each other, make a hard collision, and go flying out in some random direction with a trail of jelly following. The marbles represent quarks and gluons, and in the collision, they are being torn from the jelly that is the proton.

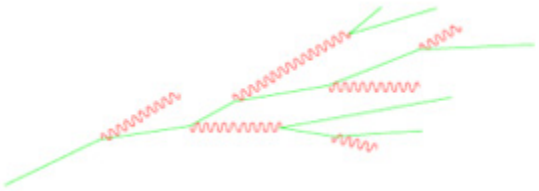


Figure 17: In this Feynman diagram of a jet, a single quark decays into a shower of quarks and gluons.
Source: © David Kaplan.

However, we know quarks cannot be free, and that if quarks are produced or separated in a high-energy collision, the color force starts ripping quark/anti-quark pairs out of the vacuum. The result is a directed spray, or jet of particles headed off in the direction the individual quark would have gone. This can be partially described by a Feynman diagram where, for example, a quark becomes a shower of quarks and gluons.

In the early days of QCD, it became clear that if a gluon is produced with high energy after a collision, it, too, would form a jet. At that point, experimentalists began to look for physical evidence of gluons. In 1979, a team at the newly built PETRA electron-positron storage ring at DESY, Germany's Deutsches Elektronen-Synchrotron, found the evidence, in the form of several of the tell-tale three-jet events. Other groups quickly confirmed the result, and thus established the reality of the gluon.

A confining force

As we have seen, the strength of the strong force changes depending on the energy of the interaction, or the distance between particles. At high energies, or short distances, the strong force actually gets weaker. This was discovered by physicists David Gross, David Politzer, and Frank Wilczek, who received the 2004 Nobel Prize for this work. In fact, the color charge (or coupling) gets so weak at high energies, you can describe the interactions between quarks in colliding protons as the scattering of free quarks; marbles in jelly are a good metaphor.

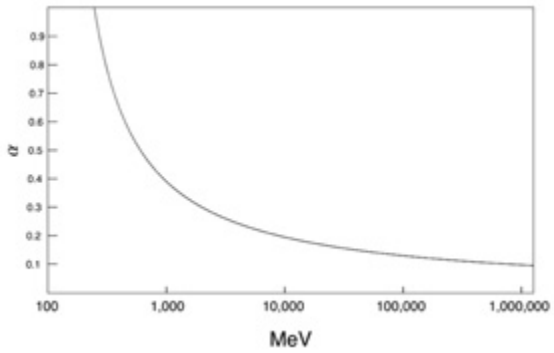


Figure 18: The QCD coupling depends on energy.
Source: © David Kaplan.

At lower energies, or longer distances, the charge strength appears to hit infinity, or blows up as physicists like to say. As a result, protons may as well be a fundamental particle in low-energy proton-proton collisions because the collision energy isn't high enough to probe their internal structure. In this case, we say that the quarks are confined. This qualitative result is clear in experiments, however, "infinity" doesn't make for good quantitative predictions. This difficulty keeps QCD a lively and active area of research.

Physicists have not been able to use QCD theory to make accurate calculations of the masses and interactions of the hadrons made of quarks. Theorists have developed a number of techniques to overcome this issue, the most robust being lattice gauge theory. This takes a theory like QCD, and puts it on a lattice, or grid of points, making space and time discrete rather than continuous. And because the number of points is finite, the situation can be simulated on a computer. Amazingly enough, physicists studying phenomena at length scales much longer than the defined lattice point spacing find that the simulated physics acts as if it is in continuous space. So, in theory, all one needs to do to calculate the mass of a hadron is to space the lattice points close enough together. The problem is that the computing power required for a calculation grows exponentially with the number of points on the lattice. One of the main hurdles to overcome in lattice gauge theory at this point is the computer power needed for accurate calculations.

The pion puzzle

The energy scale where the QCD coupling blows up is in fact the mass of most hadrons—roughly 1 GeV. There are a few exceptions, however. Notably, pions are only about a seventh the mass of the proton.

These particles turn out to be a result of spontaneous symmetry breaking in QCD as predicted by the so-called Nambu-Goldstone theorem that we will learn more about in Section 8.

Japanese physicist Hideki Yukawa predicted the existence of the pion, a light spinless particle, in 1935. Yukawa actually thought of the pion as a force carrier of the strong force, long before QCD and the weak forces were understood, and even before the full development of QED. Yukawa believed that the pion mediated the force that held protons and neutrons together in the nucleus. We now know that pion exchange is an important part of the description of low-energy scattering of protons and neutrons.

Yukawa's prediction came from using the Heisenberg uncertainty principle in a manner similar to what we did in Section 2 when we wanted to understand the exchange of force carriers. Heisenberg's uncertainty principle suggests that a virtual particle of a certain energy (or mass) tends to exist for an amount of time (and therefore tends to travel a certain distance) that is proportional to the inverse of its energy. Yukawa took the estimated distance between protons and neutrons in the nucleus and converted it into an energy, or a mass scale, and predicted the existence of a boson of that mass. This idea of a heavy exchange particle causing the force to only work at short distances becomes the central feature in the next section.

Section 6: The Weak Force and Flavor Changes

Neither the strong force nor the electromagnetic force can explain the fact that a neutron can decay into a proton, electron, and an (invisible) antineutrino. For example carbon–14, an atom of carbon that has six protons and eight neutrons, decays to nitrogen–14 by switching a neutron to a proton and emitting an electron and antineutrino. Such a radioactive decay (called beta decay) led Wolfgang Pauli to postulate the neutrino in 1930 and Enrico Fermi to develop a working predictive theory of the particle three years later, leading eventually to its discovery by Clyde Cowan, Jr. and Frederick Reines in 1956. For our purposes here, what is important is that this decay is mediated by a new force carrier—that of the weak force.

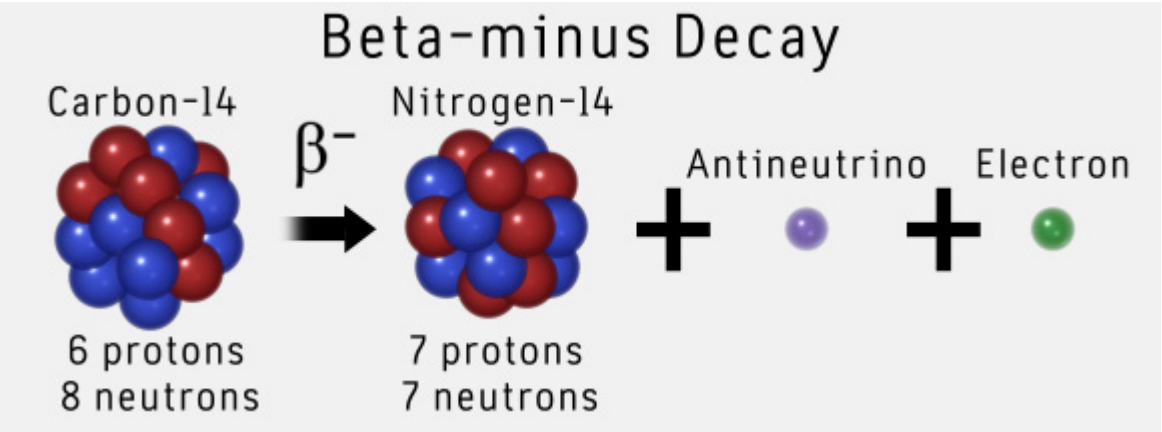


Figure 19: An example of beta decay.
Source:

As with QED and QCD, the weak force carriers are bosons that can be emitted and absorbed by matter particles. They are the electrically charged W^+ and W^-, and the electrically neutral Z^0. There are many properties that distinguish the weak force from the electromagnetic and strong forces, not the least of which is the fact that it is the only force that can mediate the decay of fundamental particles. Like the strong force, the theory of the weak force first appeared in the 1930s as a very different theory.

Fermi theory and heavy force carriers

Setting the Stage for Unification

Yang and Mills: Office mates at Brookhaven National Laboratory who laid a foundation for the unification of forces.
Source: © picture taken by A.C.T. Wu (Univ. of Michigan) at the 1999 Yang Retirement Symposium at Stony Brook, Courtesy of AIP, Emilio Segrè Visual Archives.

When they shared an office at the Brookhaven National Laboratory in 1954, Chen-Ning Yang and Robert Mills created a mathematical construct that lay the groundwork for future efforts to unify the forces of nature. The Yang-Mills theory generalized QED to have more complicated force carriers —ones that interact with each other—in a purely theoretical construct. The physics community originally showed little enthusiasm for the theory. But in the 1960s and beyond the theory proved invaluable to physicists who eventually won Nobel Prizes for their work in uniting electromagnetism and the weak force and understanding the strong force. Yang himself eventually won a Nobel Prize with T.D. Lee for the correct prediction of (what was to become) the weak-force violation of parity invariance. Yang's accomplishments place him as one of the greatest theoretical physicists of the second half of the 20th century.

Fermi constructed his theory of beta decay in 1933, which involved a direct interaction between the proton, neutron, electron,and antineutrino (quarks were decades away from being postulated at that point). Fermi's theory could be extended to other particles as well, and successfully describes the decay of the muon to an electron and neutrinos with high accuracy. However, while the strength of QED (its coupling) was a pure number, the strength of the Fermi interaction depended on a coupling that had the units of one over energy squared. The value of the Fermi coupling (often labeled G_F) thus suggested a

new mass/energy scale in nature associated with its experimental value: roughly 250 times the proton mass (or ~250 GeV).

In 1961, a young Sheldon Glashow fresh out of graduate school, motivated by experimental data at the time, and inspired by his advisor Julian Schwinger's work on Yang-Mills theory, proposed a set of force carriers for the weak interactions. They were the W and Z bosons, and had the masses necessary to reproduce the success of Fermi theory. The massive force carriers are a distinguishing feature of the weak force when compared with the massless photon and essentially massless (yet confined) gluon. Thus, when matter is interacting via the weak force at low energies, the virtual W and Z can only exist for a very short time due to the uncertainty principle, making the weak interactions an extremely short-ranged force.

Another consequence of heavy force carriers is the fact that it requires a large amount of energy to produce them. The energy scale required is associated with their mass ($M_W c^2$) and is often called the weak scale. Thus, it was only in the early 1980s, nearly a century after seeing the carriers' effects in the form of radioactivity, that scientists finally discovered the W and Z particles at the UA1 experiment at CERN.

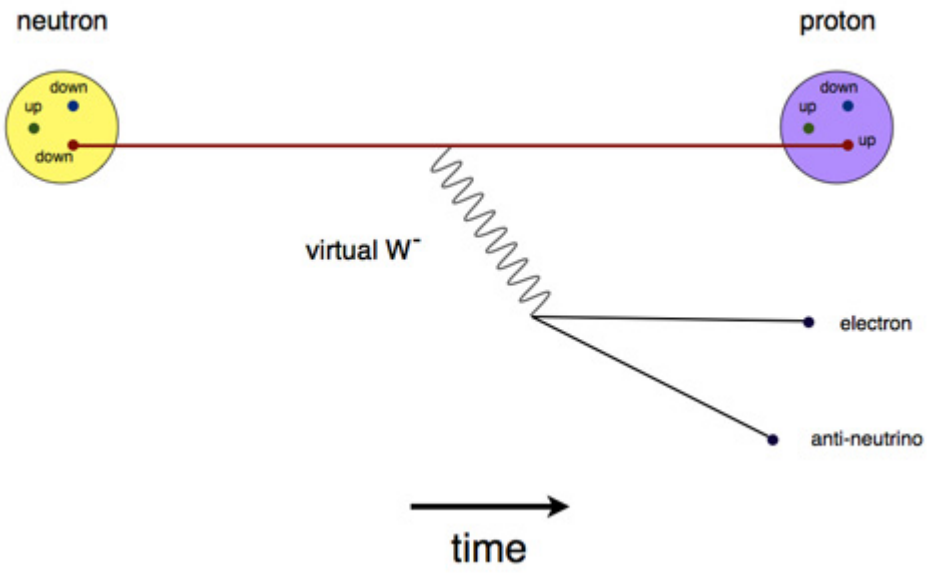


Figure 20: Neutron decay from the inside.
Source: © David Kaplan.

Force of change

An equally important difference between the weak force and the others is that when some of the force carriers are emitted or absorbed (specifically, the $W^{+/-}$), the particle doing the emitting/absorbing changes its flavor. For example, if an up quark emits a W^{+}, it changes into a down quark. By contrast, the electron stays an electron after it emits or absorbs QED's photon. And while the gluon of QCD changes the color of the quark from which it is emitted, the underlying symmetry of QCD makes quarks of different colors indistinguishable. The weak force does not possess such a symmetry because its force carrier, the W, changes one fermion into a distinctly different one. In our example above, the up and down quarks have different masses and electric charges. However, physicists have ample theoretical and indirect experimental evidence that the underlying theory has a true symmetry. But that symmetry is dynamically broken because of the properties of the vacuum, as we shall see later on.

That fact that the W boson changes the flavor of the matter particle has an important physical implication: The weak force is not only responsible for interactions between particles, but it also allows heavy particles to decay. Because the weak force is the only one that changes quarks' flavors, many decays in the Standard Model, such as that of the heavy top quark, could not happen without it. In its absence, all six quark flavors would be stable, as would the muon and the tau particles. In such a universe, stable matter would consist of a much larger array of fundamental particles, rather than the three (up and down quarks and the electron) that make up matter in our universe. In such a universe, it would have taken much less energy to discover the three generations, as we would simply detect them. As it is, we need enough energy to produce them, and even then they decay rapidly and we only get to see their byproducts.

Weak charge?

In the case of QED and QCD, the particles carried the associated charges that could emit or absorb the force carriers. QED has one kind of charge (plus its opposite, or conjugate charge), which is carried by all fundamental particles except neutrinos. In QCD, there are three kinds of color charge (and their conjugates) which are only carried by quarks. Therefore, only quarks exchange gluons. In the case of the weak force, all matter particles interact with and thus can exchange the W and Z particles—but then what exactly is weak charge?

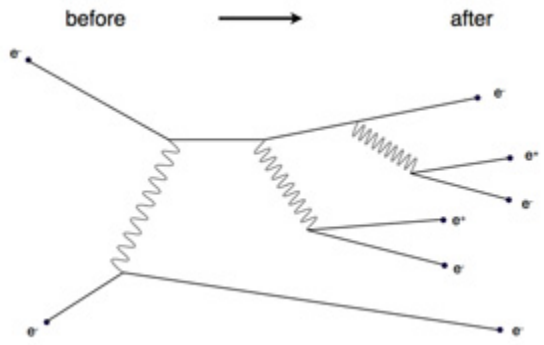


Figure 21: The total amount of electric charge is conserved, even in complicated interactions like this one.
Source: © David Kaplan.

An important characteristic feature of electromagnetic charge is that it is conserved. This means, for any physical process, the total amount of positive charge minus the total amount of negative charge in any system never changes, assuming no charge enters or leaves the system. Thus, positive and negative charge can annihilate each other, or be created in pairs, but a positive charge alone can never be destroyed. Similarly for the strong force, the total amount of color charge minus the total anti-color charge typically stays the same, there is one subtlety. In principle, color charge can also be annihilated in threes, except for the fact that baryon number—the number of baryons like protons and neutrons—is almost exactly conserved as well. This makes color disappearance so rare that it has never been seen.

Weak charge, in this way, does not exist—there is no conserved quantity associated with the weak force like there is for the other two. There is a tight connection between conserved quantities and symmetries. Thus, the fact that there is no conserved charge for the weak force is again suggestive of a broken symmetry.

Look in the mirror—it's not us

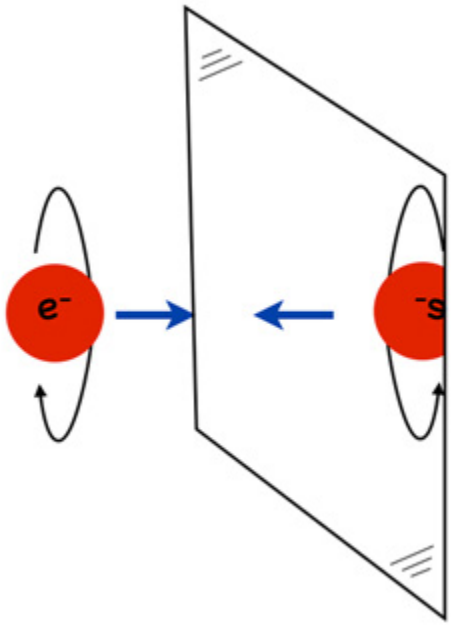


Figure 22: For the weak force, an electron's mirror image is a different type of object.
Source: © David Kaplan.

The weak interactions violate two more symmetries that the strong and electromagnetic forces preserve. As discussed in the previous unit, these are parity (P) and charge conjugation (C). The more striking one is parity. A theory with a parity symmetry is one in which any process or interaction that occurs (say particles scattering off each other, or a particle decaying), its exact mirror image also occurs with the same probability. One might think that such a symmetry must obviously exist in Nature. However, it turns out that the weak interactions *maximally violate* this symmetry.

As a physical example, if the W^- particle is produced at rest, it will—with roughly 10% probability—decay into an electron and an antineutrino. What is remarkable about this decay is that the electron that comes out is almost always left-handed. A left-handed (right-handed) particle is one in which when viewed along the direction it is moving, its spin is in the counterclockwise (clockwise) direction. It is this fact that violates parity symmetry, as the mirror image of a left-handed particle is a right-handed particle.

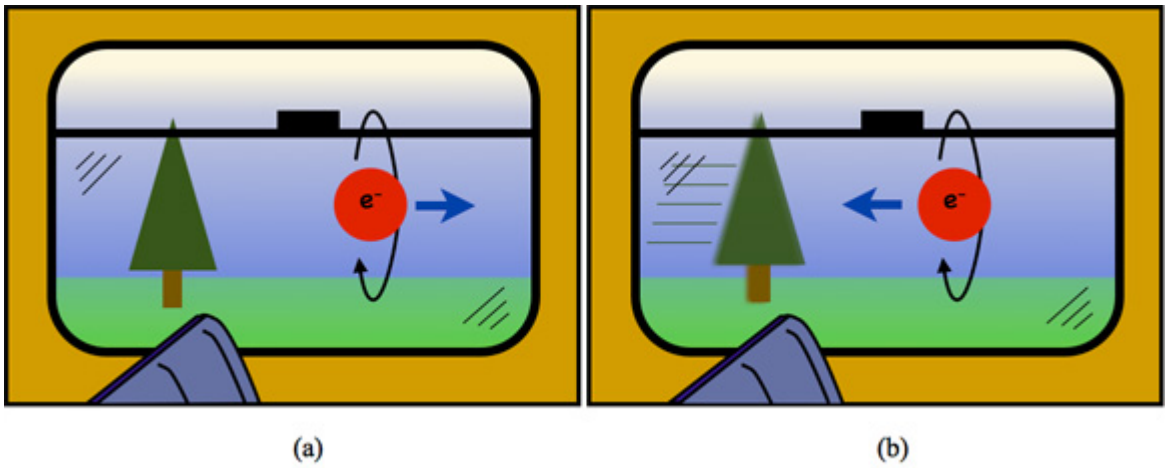


Figure 23: Spin flipping on the train.
Source: © David Kaplan.

The electron mass is very tiny compared to that of the W boson. It turns out that the ability of the W^- to decay into a right-handed electron depends on the electron having a mass. If the mass of the electron were zero in the Standard Model, the W^- would only decay into left-handed electrons. It is the mass, in fact, that connects the left-handed and right-handed electrons as two parts of the same particle. To see why, imagine an electron moving with a left-handed spin. If you were to travel in the same direction as the electron, but faster, then the electron to you would look as if it were moving in the other direction, but its spin would be in the original direction. Thus, you would now see a right-handed electron. However, if the electron had no mass, Einstein's relativity would predict that it moves at the speed of light (like the photon), and you would never be able to catch up to it. Thus, the left-handed massless electron would always look left-handed.

The mixing of the left- and right-handed electrons (and other particles) is again a result of a symmetry breaking. The symmetry is sometimes called chiral symmetry, from the Greek word *chiral*, meaning hand. The masses of the force carriers, the flavor-changing nature of the weak force, and the masses of all matter particles, all have a single origin in the Standard Model of particle physics—the Higgs mechanism —as we will see in the next section.

Section 7: Electroweak Unification and the Higgs

While the W particles are force carriers of the weak force, they themselves carry charges under the electromagnetic force. While it is not so strange that force carriers are themselves charged—gluons carry color charges, for example—the fact that it is electromagnetic charge suggests that QED and the weak force are connected. Glashow's theory of the weak force took this into account by allowing for a mixing between the weak force and the electromagnetic force. The amount of mixing is labeled by a measurable parameter, θ_W.

Unifying forces

The full theory of electroweak forces includes four force carriers: W^+, W^-, and two uncharged particles that mix at low energies—that is, they evolve into each other as they travel. This mixing is analogous to the mixing of neutrinos with one another discussed in the previous unit. One mixture is the massless photon, while the other combination is the Z. So at high energies, when all particles move at nearly the speed of light (and masses can be ignored), QED and the weak interactions unify into a single theory that we call the electroweak theory. A theory with four massless force carriers has a symmetry that is broken in a theory where three of them have masses. In fact, the Ws and Z have different masses. Glashow put these masses into the theory by hand, but did not explain their origin.

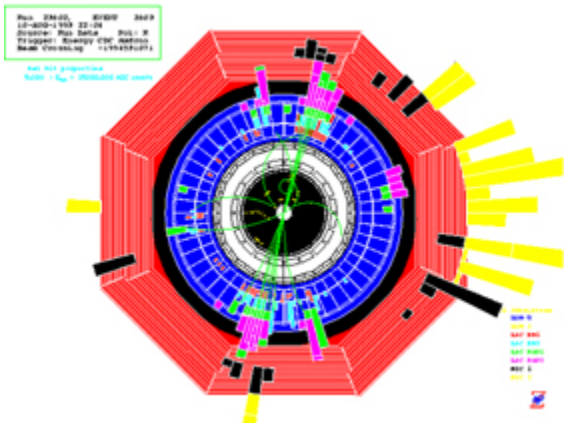


Figure 24: The Z particle at SLAC.
Source: © SLAC National Accelerator Laboratory.

The single mixing parameter predicts many different observable phenomena in the weak interactions. First, it gives the ratio of the W and Z masses (it is the cosine of θ_W). It also gives the ratio of the coupling

strength of the electromagnetic and weak forces (the sine of θ_W). In addition, many other measurable quantities, such as how often electrons or muons or quarks are spinning one way versus another when they come from a decaying Z particle, depend on the single mixing parameter. Thus, the way to test this theory is to measure all of these things and see if you get the same number for the one parameter.

Testing of the electroweak theory has been an integral part of particle physics experimental research from the late 1980s until today. For example, teams at LEP (the Large Electron-Positron collider, which preceded the Large Hadron Collider (LHC) at CERN) produced 17 million Z bosons and watched them decay in different ways, thus measuring their properties very precisely, and putting limits on possible theories beyond the Standard Model. The measurements have been so precise that they needed an intensive program on the theoretical side to calculate the small quantum effects (loop diagrams) so theory and experiment could be compared at similar accuracy.

A sickness and a cure

While the electroweak theory could successfully account for what was observed experimentally at the time of its inception, one could imagine an experiment that could not be explained. If one takes this theory and tries to compute what happens when Standard Model particles scatter at very high energies (above 1 TeV) using Feynman diagrams, one gets nonsense. Nonsense looks like, for example, probabilities greater than 100%, measurable quantities predicted to be infinity, or simply approximations where the next correction to a calculation is always bigger than the last. If a theory produces nonsense when trying to predict a physical result, it is the wrong theory.

A "fix" to a theory can be as simple as a single new field (and therefore, particle). We need a particle to help Glashow's theory, so we'll call it H. If a particle like H exists, and it interacts with the known particles, then it must be included in the Feynman diagrams we use to calculate things like scattering cross sections. Thus, though we may never have seen such a particle, its virtual effects change the results of the calculations. Introducing H in the right way changes the results of the scattering calculation and gives sensible results.

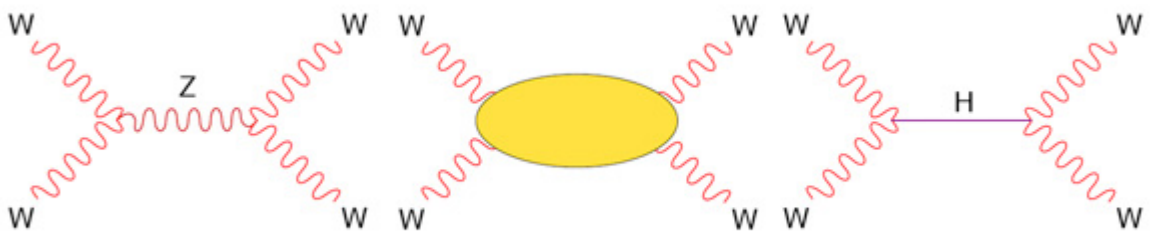


Figure 25: Scattering of W particles in Feynman diagrams.
Source: © David Kaplan.

In the mid-1960s, a number of physicists, including Scottish physicist Peter Higgs, wrote down theories in which a force carrier could get a mass due to the existence of a new field. In 1967, Steven Weinberg (and independently, Abdus Salam), incorporated this effect into Glashow's electroweak theory producing a consistent, unified electroweak theory. It included a new particle, dubbed the Higgs boson, which, when included in the scattering calculations, completed a new theory—the Standard Model—which made sensible predictions even for very high-energy scattering.

A mechanism for mass

The way the Higgs field gives masses to the W and Z particles, and all other fundamental particles of the Standard Model (the Higgs mechanism), is subtle. The Higgs field—which like all fields lives everywhere in space—is in a different phase than other fields in the Standard Model. Because the Higgs field interacts with nearly all other particles, *and* the Higgs field affects the vacuum, the space (vacuum) particles travel through affects them in a dramatic way: It gives them mass. The bigger the coupling between a particle and the Higgs, the bigger the effect, and thus the bigger the particle's mass.

In our earlier description of field theory, we used the analogy of waves traveling across a lake to represent particles moving through the vacuum. A stone thrown into a still lake will send ripples across the surface of the water. We can imagine those ripples as a traveling packet of energy that behaves like a particle when it is detected on the other end. Now, imagine the temperature drops and the lake freezes; waves can still exist on the surface of the ice, but they move at a completely different speed. So, while it is the same lake made of the same material (namely, water), the waves have very different properties. Things

attempting to move through the lake (like fish) will have a very different experience trying to get through the lake. The change in the state of the lake itself is called a phase transition.

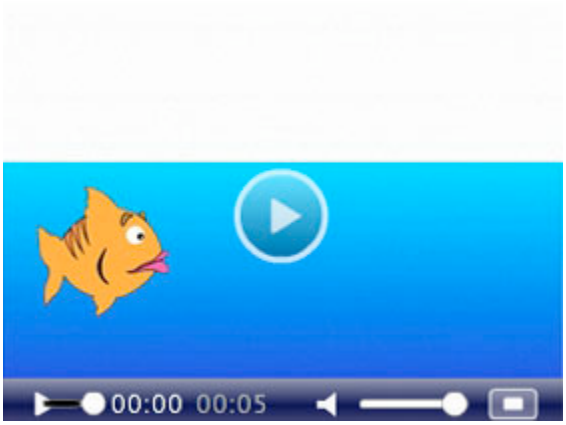


Figure 26: The Higgs mechanism is analogous to a pond freezing over.
Source: © David Kaplan.

This situation with the Higgs has a direct analogy with the freezing lake. At high enough temperatures, the Higgs field does not condense, which means that it takes on a constant value everywhere, and the W and Z are effectively massless. Lower temperatures can cause a transition in which the Higgs doublet condenses, the W and Z gain mass, and it becomes more difficult for them to move through the vacuum, as it is for fish in the lake, or boats on the surface when the lake freezes. In becoming massive, the W and Z absorb parts of the Higgs field. The remaining Higgs field has quantized vibrations that we call the Higgs boson that are analogous to vibrations on the lake itself. This effect bears close analogy with the theory of superconductivity that we will meet in Unit 8. In a sense, the photon in that theory picks up a mass in the superconducting material.

Not only do the weak force carriers pick up a mass in the Higgs phase, so do the fundamental fermions —quarks and leptons—of the Standard Model. Even the tiny neutrino masses require the Higgs effect in order to exist. That explains why physicists sometimes claim that the Higgs boson is the origin of mass. However, the vast majority of mass in our world comes from the mass of the proton and neutron, and thus comes from the confinement of the strong interactions. On the other hand, the Higgs mechanism is responsible for the electron's mass, which keeps it from moving at the speed of light and therefore allows atoms to exist. Thus, we can say that the Higgs is the origin of structure.

Closing in on the Higgs

There is one important parameter in the electroweak theory that has yet to be measured, and that is the mass of the Higgs boson. Throughout the 1990s and onward, a major goal of the experimental particle physics community has been to discover the Higgs boson. The LEP experiments searched for the Higgs to no avail and have put a lower limit on its mass of 114 Giga-electron-volts (GeV), or roughly 120 times the mass of the proton. For the Standard Model not to produce nonsense, the Higgs must appear in the theory at energies (and therefore at a mass) below 1,000 GeV.

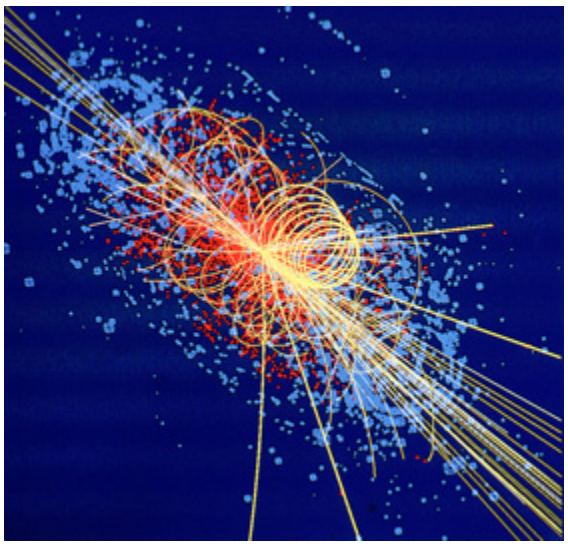


Figure 27: Simulation of a Higgs event at the LHC.
Source: © CERN.

However, there have been stronger, more indirect ways to narrow in on the Higgs. When LEP and other experiments were testing the electroweak theory by making various measurements of the mixing angle, the theory calculations needed to be very precise, and that required the computing of more complicated Feynman diagrams. Some of these diagrams included a virtual Higgs particle, and thus the results of these calculations depend on the existence of the Higgs.

Though the effects of virtual Higgs bosons in Feynman diagrams are subtle, the experimental data is precise enough to be sensitive to the mass to the Higgs. Thus, though never seen, as of 2010, there is a prediction that the Higgs boson mass must be less than roughly 200 times the proton mass. With a successful high-energy run of the Large Hadron Collider, and with the support of a full analysis of data from the Tevatron experiments at Fermilab, we should know a lot about the Higgs boson, whether it exists, and what its mass is by 2015.

Section 8: Symmetries of Nature

Symmetries are a central tool in theoretical physics. They can play the role of an organizing principle in a new theory, or can allow for tremendous simplification of otherwise difficult problems. In particle physics, theorists speculate about new symmetry principles when they seek deeper explanations or theories of fundamental particles or forces. Condensed matter physicists use symmetries to characterize the molecular structure of different materials. Atomic physicists organize atomic states in terms of rotational symmetry. Without symmetries—even approximate symmetries—it is extremely difficult to characterize the properties of physical systems.

Exact and approximate symmetries in particle physics

A system has a symmetry when changing the system in some way leaves it in an identical state. For example, a perfect circle, when rotated around the center, looks the same. We call rotating the circle—or any change of the system that leaves it looking the same—a symmetry transformation. The set of all symmetry transformations—all things that can be done to the system and leave it looking the same—form a group, a word with a precise mathematical definition. The transformations can be continuous, as a rotation by an arbitrary angle, or discrete, as a flip to a mirror image.

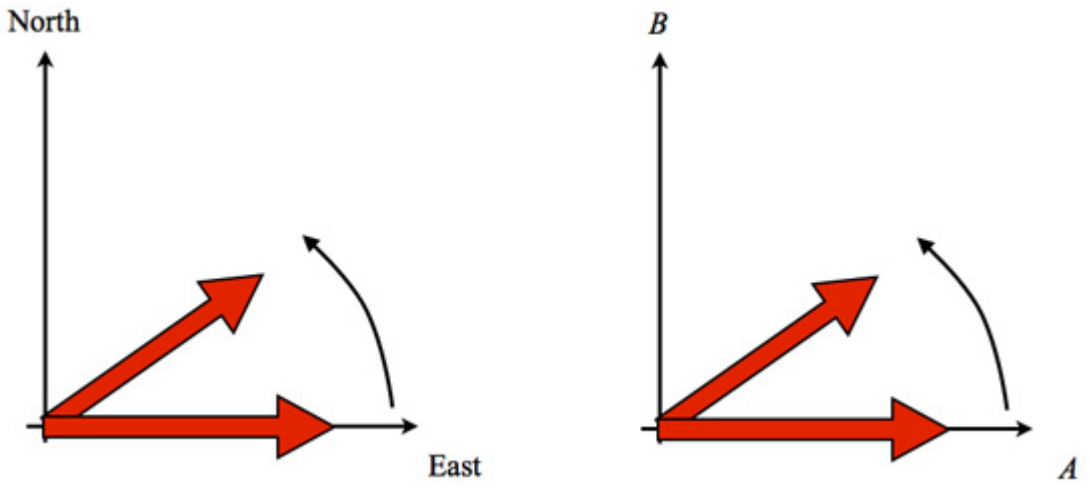


Figure 28: Rotations in physical space and "particle space."
Source: © David Kaplan.

Symmetries can apply not only to external properties, like physical rotations, but also to internal properties, like particle type. For example, a symmetry could exist where all physics experiments done with particle A would yield the same results with the same probabilities if they were done with particle B. This implies an exchange symmetry between A and B: You get the same result if you exchange particle A for particle B, and vice versa. In this case, the two particles have precisely the same properties.

More general and stranger symmetries can exist as well. For example, instead of simply exchanging particles A and B, one could replace particle A with a particle that is partially A and partially B. This occurs in the case of neutrinos, where one flavor is produced—for example, electron neutrinos in the sun—and another is measured far away—muon or tau neutrinos on earth. It is also true in the case of uncharged mesons made of quarks, like neutral Kaons or B-mesons. Thus, this kind of symmetry of A and B could be described as a "rotation" between particle types—it could be one or the other or a mixture of the two. It is very similar to physical direction, in the sense that one could be facing north or east or in some mixture of the two (e.g., east-north-east).

If one wanted to do a measurement to tell whether a particle is A or B, there would have to be something to distinguish the two—some difference. But a difference would mean the symmetry is not exact. One example is the three neutrinos of the Standard Model. They are almost, but not quite, the same. The distinction is that electron neutrinos interact in a special way with electrons, whereas muon and tau neutrinos interact in that same way with muon and tau particles, respectively. So here, the symmetry is approximate because the particles the neutrinos are associated with have very different masses. Exchanging a neutrino of one species with another changes how it interacts with the electron, for example. Also, when a neutrino scatters off of matter, the matter can 'pick out' the flavor of neutrino and change it to one type or another.

Figure 29: A combination particle that's a mixture of type A and type B can be turned into a pure type A or pure type B particle when it interacts with matter.
Source: © David Kaplan.

What if the three neutrinos of the Standard Model were exactly the same—how would we ever know that there are three? It turns out that we can determine the number of light neutrinos from experimental measurements of Z bosons. The decay rate of the Z boson depends on how many light particles are coupled to it and the size of their couplings. Since the couplings can be measured in other ways, and the decays of the Z that don't involve neutrinos are visible, i.e., they leave energy in the detector, one can infer the number of neutrinos. The number measured in this way is ~2.984, in agreement (within errors) with the three neutrinos of the Standard Model.

Spontaneous symmetry breaking

It might seem as though a system either has a symmetry or it doesn't: We can rotate a circle by any angle and it looks the same, but that doesn't work for a square. However, it is possible for a physical theory to have a symmetry that isn't reflected in the current state of the system it describes. This can happen when a symmetry is spontaneously broken.

What does that mean? Consider, for example, a spinning top, whose point remains stationary on a table. The system (the top) is perfectly symmetric around its spinning axis—looking at the top from any side of the table, one sees the same image. Once the top has finished spinning, it lies on the table. The symmetry is gone, and the top no longer looks the same when viewed from any angle around the table. The top's handle now points in a specific direction, and we see different things from different vantage points. However, the top could have fallen in any direction—in fact, one could say that the top has equal probability of pointing in any direction. Thus, the symmetry is inherent in the theory of the top, while

that state of the system breaks the symmetry because the top has fallen in a particular direction. The symmetry was spontaneously broken because the top just fell over naturally as its rotational speed decreased.

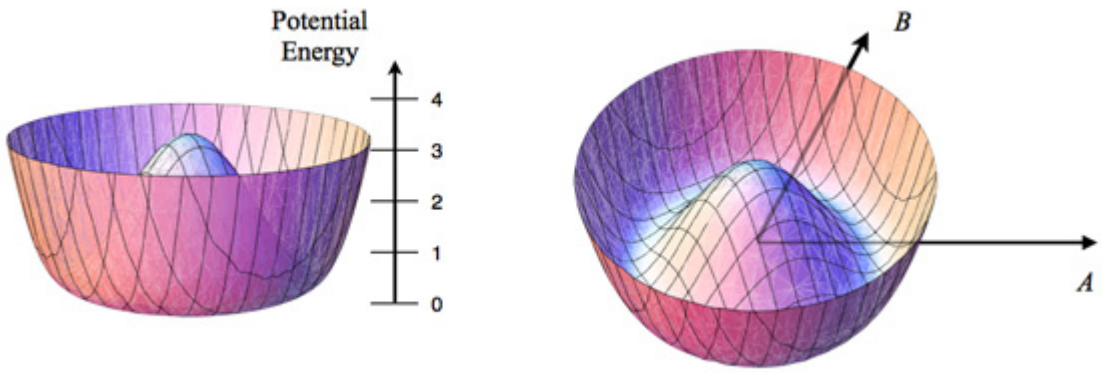


Figure 30: The wine-bottle potential that is characteristic of spontaneous symmetry breaking.
Source: © David Kaplan.

One can have a similar spontaneous breaking of an internal symmetry. Imagine two fields, A and B, whose potential energies depend on each other in the way illustrated in Figure 30. While typically in theories, the lowest energy value of a field is zero, here we see the minimum energy value lies along the circular ring at the bottom. While the potential energy shape is symmetric—it looks the same rotated around the center—the fields take a particular value along the minimum-energy ring at every point in space, thus breaking the symmetry.

In Section 2, we learned that particles are simply fluctuations of a field. Our fields A and B can fluctuate in a very special way because the potential energy minimum forms a ring. If the fields are valued such that they sit in that minimum, and they fluctuate only around the ring, the potential energy does not change as the field fluctuates. Because the field vibrations involve no potential energy, the waves of the field can be as long and as low-energy as one wishes. Thus, they correspond to one or more massless particles. The fact that spontaneously breaking a symmetry results in a massless particle is known as the Nambu-Goldstone theorem, after physicists Yoichiro Nambu and Jeffrey Goldstone.

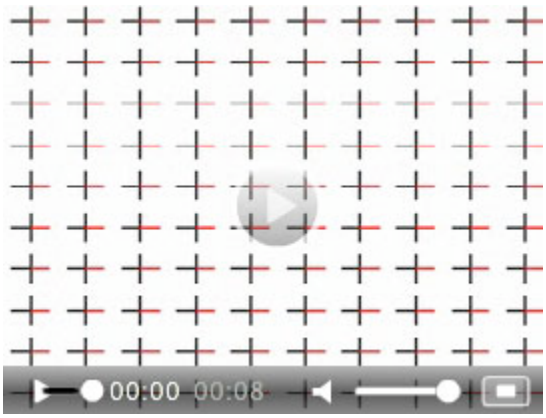


Figure 31: A wave in field space corresponds to a physical particle.
Source: © David Kaplan.

Pions, originally thought to be carriers of the strong force, are a real-life example of Nambu-Goldstone bosons. They are associated with the breaking of a complicated symmetry that involves the change of left-handed quarks into each other, with a simultaneous opposite change to right-handed quarks. This symmetry is spontaneously broken by the dynamics of the strong force in, as of today's knowledge, some inexplicable way. Pions are light, rather than massless, because the symmetry is approximate rather than exact. Knowing that the pions are Nambu-Goldstone bosons allows physicists to determine some of the mysteries of how the strong force actually works.

Recovering symmetry at high temperatures

In our initial example of the spinning top, the theory had an underlying symmetry that was broken when the top fell over. When the top had a lot of energy and was spinning quickly, the symmetry was obvious. It was only when the top lost enough energy that it fell over that the symmetry was broken. The high-energy top displays a symmetry that the low-energy top does not. Something similar happens in more complicated systems such as magnets, superconductors, and even the universe.

Let's take a magnet as an example. Make it a ball of iron to keep things nice and symmetric. The magnetization of iron comes about because the electrons, which are themselves tiny magnets, tend to want to line up their spine, and thus their magnetic fields, such that collectively, the entire material is magnetic. If one dumps energy in the form of heat into the magnet, the electrons effectively vibrate and twist more and more violently until at a critical temperature, 768 degrees Celsius for iron, the directions

of their individual spins are totally randomized. At that point, the magnetic field from all of the electrons averages out to zero, and the iron ball is no longer magnetic.

When the iron is magnetized it is like the fallen top, having selected a particular direction as different from the rest. Once heated to the critical temperature, however, symmetry is restored and any direction within the magnet is equivalent to any other. Many symmetries that are spontaneously broken in the minimum energy state are restored at high temperatures. The Higgs mechanism we encountered in the previous section is a significant, if complicated, example. The restoration of symmetry at high temperatures has significant implications for the early universe, when the temperatures were extremely hot—it implies that at times very soon after the Big Bang, most or all of the spontaneously broken symmetries of the Standard Model (and its underlying theory) were intact.

Section 9: Gravity: So Weak, Yet So Pervasive

The electromagnetic, strong, and weak interactions fit nicely into the Standard Model, and unify into a single theory at high temperatures. Gravity is still an outlier, in more ways than one. It is the first force to have a quantitative model thanks to Isaac Newton, and it happens to be the weakest force at the particle level. The electric force that binds a proton and electron into an atom is 10,000,000,000,000,000,000,000,000,000,000,000,000,000 (10^{40}) times larger than the gravitational force that attracts them.

Despite its weakness, gravity has a significant impact at macroscopic distances because of one crucial unique feature: All matter has the same sign gravitational charge. The charge for the gravitational force is mass, or energy in the full theory of general relativity. The gravitational force between two massive particles is always positive and it always attracts. Thus, unlike say electromagnetism, in which opposite charges attract and can thus screen the long-distant effects, gravitational charge always adds, and large objects can produce large gravitational fields.

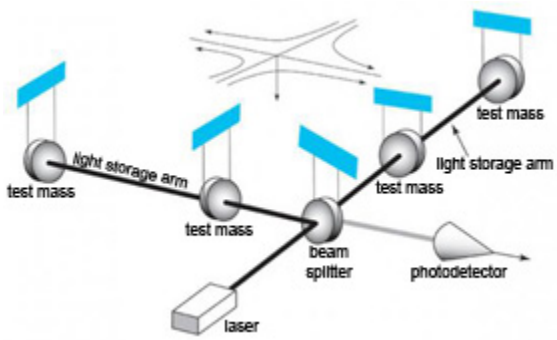


Figure 32: Schematic of a laser interferometer that can detect gravitational waves.
Source: © LIGO Laboratories.

General relativity, Einstein's theory of gravity and its current best description, works in effect as a theory of a gravitational field coupled to matter. In the full quantum theory, one would expect the existence of a particle associated with the field—the graviton—to be the force carrier. Nobody has yet detected an individual graviton. Nor is anyone likely to, owing to the extremely small likelihood that gravitons will interact with matter. However, astrophysicists have mounted several experiments to detect gravitational waves, which represent clusters of many gravitons. The most prominent, as we shall see in Unit 3, involve the use of lasers to measure how gravitational waves stretch space.

The graviton and gravitational coupling

In the absence of experimental evidence, theorists have devised a general picture of the graviton's characteristics. According to that picture, the graviton resembles the photon more closely than other force carriers, as it has no mass and is not confined at low energies, meaning that it can travel long distances freely. It is distinct from the photon in three ways. First, it is a spin-2 rather than spin-1 particle, though it still only comes in two types, analogous to left-handed and right-handed particles. Second, like the gluon, the graviton itself carries (gravitational) charge, in the form of energy (mass). Thus, gravitons attract each other. However, this does not lead to a constant force at arbitrarily long distances. The force still falls off with one over the square of the distance between objects as happens in QED. Third, while the QED coupling is a dimensionless number, the gravitational coupling to matter, Newton's constant, carries the dimensions of meters cubed divided by kilograms times seconds squared. The fact that it carries dimensions is important because it suggests that there is a fundamental mass, energy, length, and duration associated with gravity.

Physicists call the characteristic energy scale for gravity the "Planck scale," whose value is approximately 10^{19} GeV. Using a simple approximation to estimate the cross section, the probability of gravitational scattering of two particles at energy E would be proportional to $E^4/M_{Pl}{}^4c^8$. At the energies the LHC will produce, that amounts to about 10^{-60}—so small as to make it totally irrelevant. That means if a trillion LHCs were packed onto a trillion different planets and they ran for a trillion years, it would still be extremely unlikely for any of them to see the gravitational scattering of two particles.

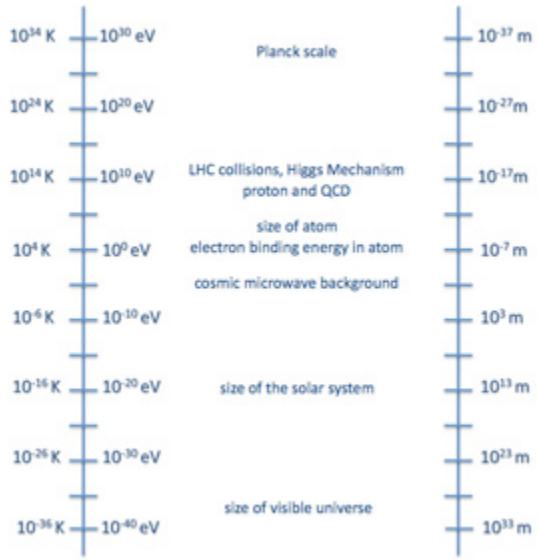


Figure 33: Energies, sizes and temperatures in physics, and in nature.
Source:

Thus, at energies of experimental particle physics, now and anytime in the foreseeable future, one can include gravitons in the Feynman diagram calculations, but their effect is negligible. However, it also suggests that for scattering close to the Planck energy, the gravitons become very important and cannot be neglected. In fact when the coupling (which could be interpreted as $E/M_{Pl}c^2$) is large (much bigger than one), then the simplest Feynman diagrams are no longer the biggest, and one would in principle need to calculate an infinite number of diagrams. It is again the case that the theory becomes nonsense, and a new theory that incorporates quantum theory and Einstein's general relativity must be found. The leading candidate for such a theory is called "string theory," which will be explored in Unit 4.

If one were actually able to build a collider that could scatter particles at the Planck energy, then the simplest assumption, and prediction of general relativity, is that the two particles would form a particle-sized black hole. In a quantum theory, even a black hole will decay, as predicted by British physicist Stephen Hawking. One would in principle study those decays and hope information about the quantum theory of gravity was contained in the spectrum of particles that came out.

However, while a simple estimate points to the energy scale of 10^{19} GeV, we have never probed experimentally beyond energies of about 10^3 GeV. In fact, because gravity is so weak, we have not

probed gravity beyond energy scales on the order of 10^{-2} eV, which corresponds to wavelengths or distances shorter than about 10 microns, which is one of the motivations for the ongoing experimental program exploring gravity at short distances described in Unit 3. Thus, while the Planck scale represents the first guess at and simplest picture of the region in which quantum gravity becomes strongly coupled, it may be wrong-headed to assume that Einstein's theory of gravity continues to work between 10 microns and 0.00000000000000000000000000001 microns, the Planck length. We will discuss alternatives at the end of this unit.

Section 10: The Prospect of Grand Unification

Let's put gravity aside for the moment. Although each of the other three fundamental forces has its own very distinctive characteristics, they all share a common structure: They are all mediated by the exchange of particles, each with one unit of spin. Amazingly enough, as we ask more questions about these forces at higher energies, their differences melt away, while their similarities remain.

Unification in the young universe

The unresolved story of unification

In 1974, Howard Georgi and Sheldon Glashow found a mathematical structure—called a Lie (pronounced "lee") Algebra—into which all of the Standard Model fit. Later that same year, Georgi, with Helen Quinn and Steven Weinberg, computed the values of the couplings at high energies and found (using data at the time) that they unified at an energy between 10^{14} and 10^{15} GeV With the unification scale in mind, the rate of proton decay could be predicted, and experiments such as Kamioka were constructed to look for proton decay. In 1981, while proton decay experiments seemed to see hints of a signal, Georgi and Savas Dimopoulos, wrote down the "minimal supersymmetric standard model", with a different prediction for unification and proton decay. By the mid-1980s, proton decay not being discovered experimentally, the standard unification model was ruled out, leaving theorists thinking this beautiful theory would need to be scrapped. Yet, in 1990, measurements at the LEP experiments recast the unification predictions, favoring the supersymmetric model. Since then, experiments like Super-Kamiokande (Super-K) continue to look for proton decay, while collider experiments like the LHC will search for supersymmetry, potentially giving a hint about a unified field theory at otherwise unreachable energies.

In this unit, we have frequently used the phrase "at high energies" in connection with force carriers. While the term generally arises in connection with scattering, it also refers to higher temperatures. A gas at a particular temperature consists of particles moving with a specific average momentum; the higher the temperature, the higher the energy of the particles. At earlier cosmological times, the (expanding) universe was much smaller and much hotter. Thus, when we say that the forces act in such-and-such way at high energies, we also mean at high temperatures, or in the early universe.

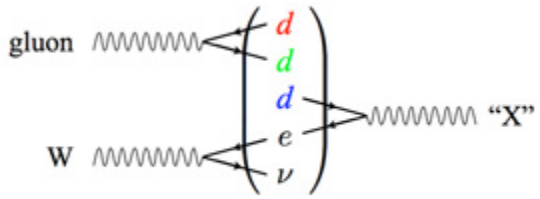


Figure 34: Quarks and leptons, unified.
Source: © David Kaplan.

At high energies, the three forces "tend toward" each other. As we have seen in previous sections, the electromagnetic coupling is bigger (stronger) at higher energies, while the strong coupling is smaller (weaker). Also at higher temperatures, the Higgs mechanism is no longer in effect (the phase change doesn't happen—the lake doesn't freeze), and thus the W and Z particles are massless, just like the gluons and photon. Thus, above the temperature associated with the Higgs energy, all three forces have massless force carriers. If one calculates the strength of each of the force's couplings, one finds that their values are coalescing at high energies.

Other things occur at energies above the Higgs mass (which correspond to temperatures above the phase change). First, the electromagnetic and weak forces "unmix." Above that energy, the Ws are not charged under the unmixed electric charge, which is called "hypercharge." Also, under the new hypercharge, the left-handed up and down quarks have the same charge of 1/6, while the left-handed electrons and associated neutrinos have a common charge of -1/2. In addition, all of these particles' masses vanish at high temperature. Thus, some pairs of particles tend to look more and more identical as the temperature increases, thus restoring a kind of symmetry that is otherwise broken by the Higgs mechanism.

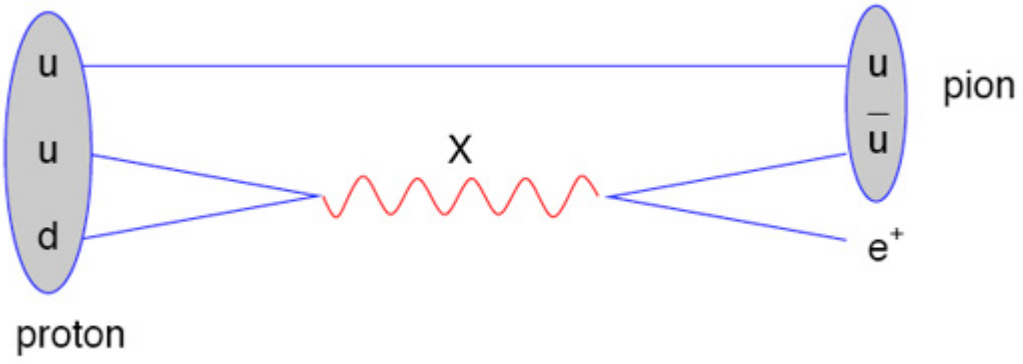


Figure 35: The X boson mediates the decay of the proton.
Source: © David Kaplan.

A true unification of the nongravitational forces would involve an extension of what the forces can do. For example, there are force carriers that, when emitted, change a quark from one color to another (gluons), and there are force carriers that change the electron into a neutrino (W^-). One could imagine force carriers that change quarks into electrons or neutrinos. Let's call them X particles. As depicted in Figure 35, right-handed down quarks, the electron, and the electron neutrino could then all turn into one another by emitting a spin-1 particle. A symmetry would exist among these five particles if they all had precisely the same properties.

At high temperatures, all the force carriers of the Standard Model have no mass. If the X particle got its mass through another, different Higgs mechanism, it too (at temperatures above that other Higgs mass) would become massless. Thus, a true symmetry—and true unification of forces, could occur at some energy when all force carriers are massless, and the strength of each of the forces (their couplings) are the same.

The possibility of proton decay

These proposed new force carriers have one quite dramatic implication. Just as the Ws can cause the decay of particles, so would the Xs. Both quarks and leptons are now connected by the force and by our new symmetry, allowing a quark to transform into a lepton and vice versa. Therefore, the system permits new processes such as the conversion of a proton to a pion and a positron. But this process is proton decay. If the new force carriers had the same mass and couplings as the W, every atom in the entire observable universe would fall apart in a fraction of a second.

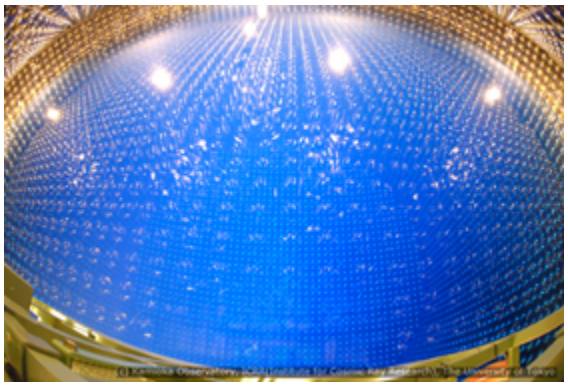


Figure 36: The nearly full water tank of the Super-Kamiokande experiment, which searches for nucleon decay.
Source: © Kamioka Observatory, ICRR (Institute for Cosmic Ray Research), The University of Tokyo.

Perhaps the proton decays very slowly, thereby saving both the theory and the universe. Physics teams aiming to develop unified theories of forces have sought evidence of proton decay since the 1980s. The most prominent search takes place at the Super-Kamiokande (Super-K) nucleon decay experiment in Hida, Japan. This is the same experiment searching for neutrinos from the Sun, as described in the previous unit. Buried in a deep mine, the experiment uses a stainless-steel tank containing 50,000 tons of water. Photomultiplier tubes and other detectors mounted around the tank identify the so-called Cerenkov light generated when neutrinos scatter charged particles. That light provides clues that can indicate whether a proton has decayed into a positron and a pion.

Super-Kamiokande and other experiments have not discredited the possibility that the proton decays, but they have put severe restrictions on the process. The current lower limit on the mass of the Xs is roughly 10^{15} GeV. Remarkably, this is roughly around the energy where the couplings get close to each other.

Unification and physics at the LHC

When one takes the low-energy values of the three couplings and theoretically computes them at high scales to see if they unify, the procedure for doing that depends on what particles exist in the theory. If we assume that only the Standard Model particles exist up to very high energies, we find that the couplings run toward each other, but do not exactly meet. But for various reasons, as we will see in the next section, theorists have been looking at theories beyond the Standard Model which predict new particles at the Large Hadron Collider (LHC) energies. One such example is called supersymmetry. It

predicts the existence of a host of new particles, superpartners, with the same charges as Standard Model particles, but different spins. In 1991, a new accurate measurement was made of the couplings in the electroweak theory, which allowed for precise extrapolation of the forces to high energies. When those couplings are theoretically extrapolated to high energies in a theory with superpartners just above the mass of the Higgs, one finds that the three couplings meet within the current experimental accuracy. Thus, supersymmetry makes the idea of unification more compelling and vice versa.

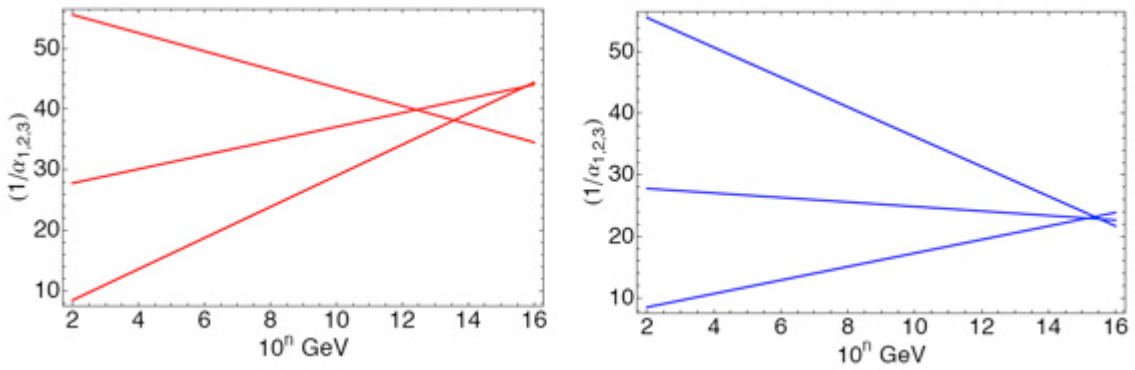


Figure 37: In the Standard Model (left), the couplings for the strong, weak and electromagnetic forces never meet, while in Supersymmetry (right), these forces unify near 1015 GeV.
Source: © David Kaplan.

The couplings meet roughly at the energy scale of 10^{16} GeV. That is comfortably above the lower limit on the new force carriers that would stem from proton decay. In addition, the scale is not too far from the Planck scale below which we expect the appearance of a quantum theory of gravity, such as the string theory that we will encounter in Unit 4. This means that we may be potentially seeing a hint of the unification of all forces, including gravity.

Section 11: Beyond the Standard Model: New Forces of Nature?

The experiments at the LHC will help extend the reach of our knowledge by being sensitive to new particles around and somewhat above the weak scale. If Nature is kind to us, the collider will reveal physics beyond the Standard Model—information about the underlying structure of the theory. Since the Standard Model requires the Higgs to have a mass below 1,000 GeV, physicists expect that the Higgs will appear at the LHC. Since the LHC will represent a significant jump in collider energy, one might naturally expect that new physics will reveal itself, in addition to the Higgs, as often occurs when experimental sensitivity increases. However, beyond typical expectations, there are compelling theoretical motivations to believe that there are new phenomena lurking just around the corner.

One motivation for physics beyond the Standard Model stems from the quantum effects on the Higgs field. While the Higgs mechanism gives masses to Standard Model particles, the actual calculated value of those masses is dramatically affected by quantum corrections, or Feynman diagrams with loops. When one computes these diagrams, they contribute infinity to the physical value of the mass of the Higgs (and W, Z). So one assumes right away that the Standard Model isn't the whole story. The infinity comes from the fact that a rule for computing these diagrams is to sum up all possible momenta in the loop, up to infinity. A solution to this type of issue in quantum field theory is to assume something significant happens at an energy (say, at energy M), in such a way that you only have to sum up to M. If you do this, the quantum correction to the Higgs mass from diagrams with one loop gives a result around M, suggesting that the mass of the Higgs should be around M, and thus new physics should be discovered at the same energy scale at the Higgs.

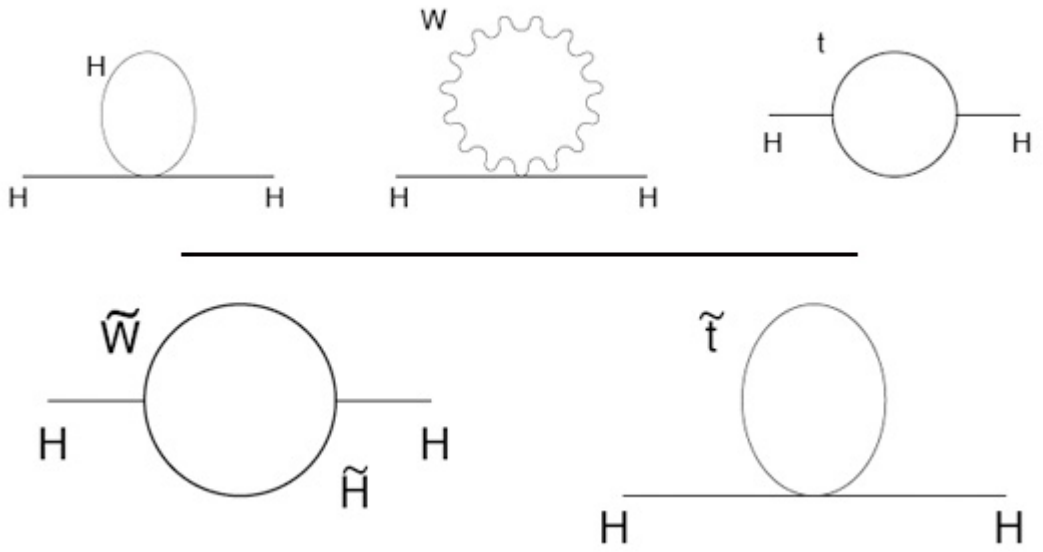


Figure 38: Canceling loops in supersymmetry.
Source: © David Kaplan.

One example of new physics that could get rid of the infinity in the Higgs mass is to have new particles appear at a mass around the mass of the Higgs such that the additional Feynman diagrams required in the complete calculations cancel the infinities. Such a perfect cancellation would imply a symmetry of couplings. A leading possibility for that symmetry is called "supersymmetry." In supersymmetric field theories, there is a symmetry between particles of different spin—specifically between fermions and bosons. Making the Standard Model supersymmetric would give every particle a "superpartner" with the same mass and couplings, but with a spin that differs by half of a unit. For example, the electron would have a partner with the same mass and charge but zero spin. Supersymmetry cannot be a perfect symmetry of Nature; if it were, we would have discovered both particles and superpartners. But what if the symmetry is "softly" broken, so the superpartners have heavier masses while their couplings still match those of the Standard Model? The soft breaking of supersymmetry would be the source of the mass of the Higgs boson and the energy scale of the Higgs mechanism. Such a model would then predict the discovery of superpartners at the LHC.

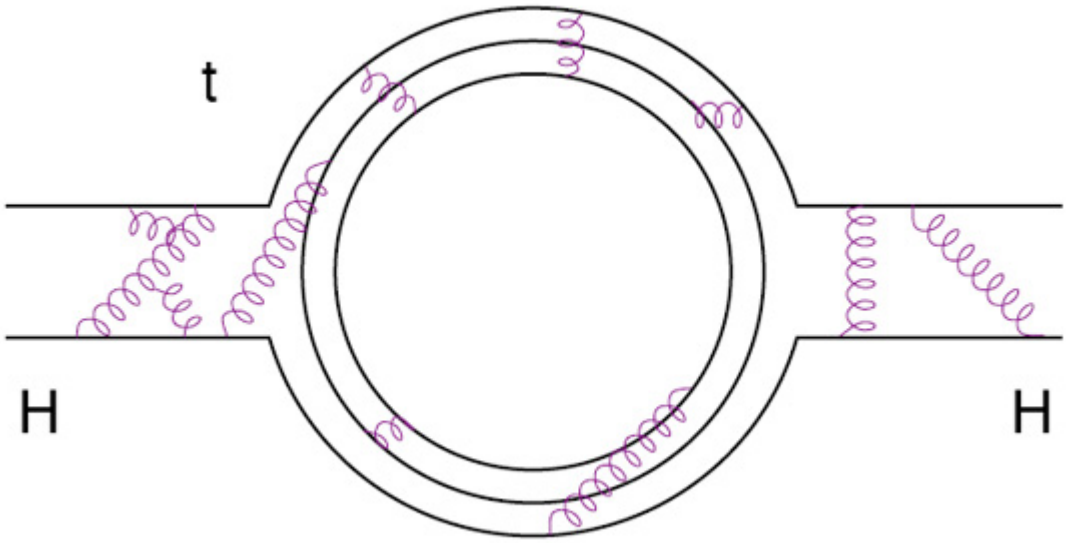


Figure 39: This Feynman diagram representing a composite Higgs and top quark is a part of the Higgs mass calculation in a supersymmetric model.
Source: © David Kaplan.

As discussed in the previous section, the supersymmetric version of the Standard Model predicts the unification of couplings. That is because the superpartners have an effect on the coupling strengths at short distances. An additional motivation for supersymmetry comes from the fact that most versions of the theory predict the existence of a stable, uncharged, weakly interacting particle. Using known and inferred information about the evolution of the universe, one can predict the abundance of these stable particles in our galaxy and beyond. Such estimates seem to predict amounts consistent with the amount of dark matter in the universe, which will be explored in Unit 10.

Another possibility is that the Higgs boson is a composite particle. If a new strong force existed at the 1 TeV scale, the Higgs could naturally have a mass of 100 GeV—and the loop diagrams would no longer be fundamental, and by their rules, would not require summing momenta up to infinity. The electroweak scale would then be populated with hadrons of the new force and its associated quarks. In the extreme limit of this model (and the original version from the late 1970s), the confinement of the new strong color force itself breaks the electroweak symmetry, or causes the condensation that gives mass to the W, Z, and the rest of the fermions, and no Higgs exists. Such a model is disfavored by precision data on the Z boson due to corrections from the new physics. The original name for this type of model is "technicolor."

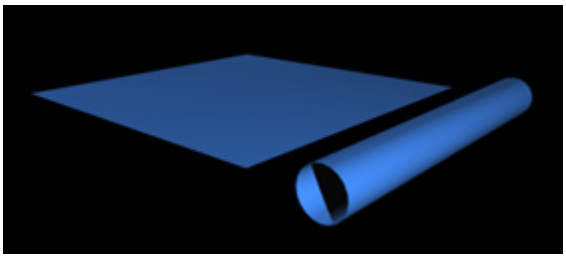


Figure 40: An extra dimension can curl up in a manner that is nearly impossible to discern for an inhabitant of the larger, uncurled dimensions.
Source:

A more exotic-sounding possibility for new physics is extra dimensions. We experience particle physics (and life) entirely in four dimensions—three space and one time—up to energy scales of around 1,000 GeV, which correspond to length scales of about 0.00000000000000002 centimeters, or 2×10^{-17} centimeters. However, because gravity is so weak, physicists have not tested it at distances shorter than 100 microns. Why is this important? Extra dimensions could be finite in size and curled up, like the circular direction on a cylinder. Thus, one or more extra dimensions could exist within which gravity operates, but the Standard Model particles and the other fundamental forces, while remaining four-dimensional, live only on the boundary. Such extra dimensions could thus be as large as 100 microns. Tests of gravity at this scale are discussed in Unit 3. The extra dimensions would dilute gravity in such a way that experiments at the LHC could directly test quantum gravity, as described in Unit 4.

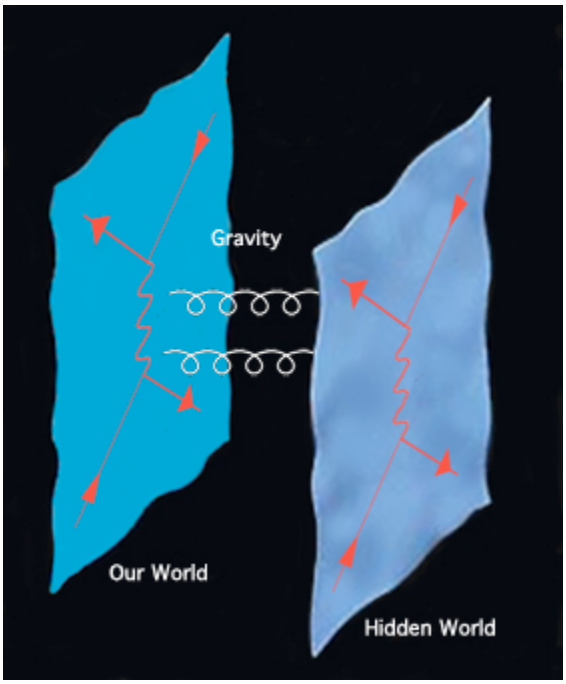


Figure 41: The Standard Model particles could be confined to the surface of a membrane, while gravity is free to leak into other dimensions.
Source:

Theorists have also imagined an extra dimension that is warped. The warping would allow four dimensional gravity to be weak while the total of five dimensions produces a quantum gravity theory at 1 TeV of energy. In this type of case, the LHC will probe a strongly coupled theory that is not four dimensional gravity. This can be understood by a revolutionary speculation, made by Argentine theorist Juan Maldacena in 1997, that certain four dimensional quantum field theories without gravity are equivalent to string theories with gravity in a larger number of dimensions. We will come to this remarkable conjecture in Unit 4. It implies future discoveries at the LHC similar to those of a new strong force.

Where we are now, with the near completion of the Standard Model, is simply a step along the way. Physicists hope that the LHC will shed light on the next step, or on the deeper principles at play not immediately visible with current data. But what we learn at the energy frontier will not simply teach us more information about matter and the vacuum—it will better guide us towards the questions we should be asking.

Section 12: Further Reading

- Richard Phillips Feynman, "Quantum Electrodynamics," *Westview Press*, New Edition, 1998.
- Richard Phillips Feynman: Nobel Prize lecture, available here: http://nobelprize.org/nobel_prizes/physics/laureates/1965/feynman-lecture.html.
- Brian Greene, "The Fabric of the Cosmos: Space, Time, and the Texture of Reality," *Random House*, 2004.
- David Griffiths, "Introduction to Elementary Particles," *John Wiley & Sons Inc.*, 1987.
- Particle Adventure animation available here: http://www.particleadventure.org/.
- Bruce Schumm, "Deep Down Things: The Breathtaking Beauty of Particle Physics," *The Johns Hopkins University Press*, 2004.
- Steven Weinberg, "Dreams of a Final Theory," *Pantheon*, 1993.
- Steven Weinberg, Nobel Prize lecture, available here: http://nobelprize.org/nobel_prizes/physics/laureates/1979/weinberg-lecture.html.

Glossary

angular momentum: In classical physics, the angular momentum of a system is the momentum associated with its rotational motion. It is defined as the system's moment of inertia multiplied by its angular velocity. In quantum mechanics, a system's total angular momentum is the sum of the angular momentum from its rotational motion (called orbital angular momentum) and its spin.

baryon: The term "baryon" refers to any particle in the Standard Model that is made of three quarks. Murray Gell-Mann arranged the baryons into a periodic table-like structure according to their baryon number and strangeness (see Unit 1, Fig. 1). Protons and neutrons are the most familiar baryons.

beta decay: Beta decay is a type of radioactive decay in which a beta particle (electron or positron) is emitted together with a neutrino. Beta decay experiments provided the first evidence that neutrinos exist, which was unexpected theoretically at the time. Beta decay proceeds via the weak interaction.

boson: A boson is a particle with integer, rather than half-integer, spin. In the Standard Model, the force-carrying particles such as photons are bosons. Composite particles can also be bosons. Mesons such as pions are bosons, as are ^{4}He atoms. See: fermion, meson, spin.

charge conjugation: Charge conjugation is an operation that changes a particle into its antiparticle.

chiral symmetry: A physical theory has chiral symmetry if it treats left-handed and right-handed particles on equal footing. Chiral symmetry is spontaneously broken in QCD.

color: In QCD, color is the name given to the charge associated with the strong force. While the electromagnetic force has positive and negative charges that cancel one another out, the strong force has three types of color, red, green, and blue, that are canceled out by anti-red, anti-green, and anti-blue.

Compton scattering: Compton scattering is the scattering of photons from electrons. When Arthur Compton first explored this type of scattering experimentally by directing a beam of electrons onto a target crystal, he found that the wavelength of the scattered photons was longer than the wavelength of the photons incident on the target, and that larger scattering angles were associated with longer

wavelengths. Compton explained this result by applying conservation of energy and momentum to the photon-electron collisions.

cross section: A cross section, or scattering cross section, is a measure of the probability of two particles interacting. It has units of area, and depends on the initial energies and trajectories of the interacting particles as well as the details of the force that causes the particles to interact.

electromagnetic interaction: The electromagnetic interaction, or electromagnetic force, is one of the four fundamental forces of nature. Maxwell first understood at the end of the 19th century that the electric and magnetic forces we experience in daily life are different manifestations of the same fundamental interaction. In modern physics, based on quantum field theory, electromagnetic interactions are described by quantum electrodynamics or QED. The force-carrier particle associated with electromagnetic interactions is the photon.

fermion: A fermion is a particle with half-integer spin. The quarks and leptons of the Standard Model are fermions with a spin of 1/2. Composite particles can also be fermions. Baryons, such as protons and neutrons, and atoms of the alkali metals are all fermions. See: alkali metal, baryon, boson, lepton, spin.

field: In general, a field is a mathematical function that has a value (or set of values) at all points in space. Familiar examples of classical fields are the gravitational field around a massive body and the electric field around a charged particle. These fields can change in time, and display wave-like behavior. In quantum field theory, fields are fundamental objects, and particles correspond to vibrations or ripples in a particular field.

flavor: In particle physics, the flavor of a particle is a set of quantum numbers that uniquely identify the type of particle it is. The quark flavors are up, down, charm, strange, top, and bottom. The lepton flavors are electron, muon, tau, and their corresponding neutrinos. A particle will have a flavor quantum number of +1 in its flavor, and its antiparticle has a quantum number of -1 in the same flavor. For example, an electron has electron flavor +1, and a positron has electron flavor of -1.

force carrier: In quantum field theory, vibrations in the field that correspond to a force give rise to particles called force carriers. Particles that interact via a particular force do so by exchanging these force carrier particles. For example, the photon is a vibration of the electromagnetic field and the carrier of the electromagnetic force. Particles such as electrons, which have negative electric charge, repel one another

by exchanging virtual photons. The carrier of the strong force is the gluon, and the carrier particles of the weak force are the W and Z bosons. Force carriers are always bosons, and may be either massless or massive.

gluons: Gluons are particles in the Standard Model that mediate strong interactions. Because gluons carry color charge, they can participate in the strong interaction in addition to mediating it. The term "gluon" comes directly from the word *glue*, because gluons bind together into mesons.

graviton: The graviton is the postulated force carrier of the gravitational force in quantum theories of gravity that are analogous to the Standard Model. Gravitons have never been detected, nor is there a viable theory of quantum gravity, so gravitons are not on the same experimental or theoretical footing as the other force carrier particles.

gravity: Gravity is the least understood of the four fundamental forces of nature. Unlike the strong force, weak force, and electromagnetic force, there is no viable quantum theory of gravity. Nevertheless, physicists have derived some basic properties that a quantum theory of gravity must have, and have named its force-carrier particle the graviton.

group: Group is a mathematical term commonly used in particle physics. A group is a mathematical set together with at least one operation that explains how to combine any two elements of the group to form a third element. The set and its operations must satisfy the mathematical properties of identity (there is an element that leaves other group elements unchanged when the two are combined), closure (combining any two group elements yields another element in the group), associativity (it doesn't matter in what order you perform a series of operations on a list of elements so long as the order of the list doesn't change), and invertability (every operation can be reversed by combining the result with another element in the group). For example, the set of real numbers is a group with respect to the addition operator. A symmetry group is the set of all transformations that leave a physical system in a state indistinguishable from the starting state.

Heisenberg uncertainty principle: The Heisenberg uncertainty principle states that the values of certain pairs of observable quantities cannot be known with arbitrary precision. The most well-known variant states that the uncertainty in a particle's momentum multiplied by the uncertainty in a particle's position must be greater than or equal to Planck's constant divided by 4π. This means that if you measure a particle's position to better than Planck's constant divided by 4π, you know that there is a larger uncertainty in the particle's momentum. Energy and time are connected by the uncertainty principle in

the same way as position and momentum. The uncertainty principle is responsible for numerous physical phenomena, including the size of atoms, the natural linewidth of transitions in atoms, and the amount of time virtual particles can last.

Higgs mechanism: The Higgs mechanism, named for Peter Higgs but actually proposed independently by several different groups of physicists in the early 1960s, is a theoretical framework that explains how fundamental particles acquire mass. The Higgs field underwent a phase transition as the universe expanded and cooled, not unlike liquid water freezing into ice. The condensed Higgs field interacts with the different massive particles with different couplings, giving them their unique masses. This suggests that particles that we can measure to have various masses were massless in the early universe. Although the Higgs mechanism is an internally consistent theory that makes successful predictions about the masses of Standard Model particles, it has yet to be experimentally verified. The clearest signature of the Higgs mechanism would be the detection of a Higgs boson, the particle associated with vibrations of the Higgs field.

jet: In the terminology of particle physics, a jet is a highly directed spray of particles produced and detected in a collider experiment. A jet appears when a heavy quark is produced and decays into a shower of quarks and gluons flying away from the center of the collision.

kinetic energy: Kinetic energy is the energy associated with the motion of a particle or system. In classical physics, the total energy is the sum of potential and kinetic energy.

LEP: The Large Electron-Positron Collider (LEP) is a particle accelerator that was operated at CERN on the outskirts of Geneva, Switzerland, from 1989 to 2000. LEP accelerated counterpropagating beams of electrons and positrons in a 27 km diameter synchrotron ring. With a total collision energy of 209 GeV, LEP was the most powerful electron-positron collider ever built. Notably, LEP enabled a precision measurement of the mass of W and Z bosons, which provided solid experimental support for the Standard Model. In 2000, LEP was dismantled to make space for the LHC, which was built in its place.

Large Hadron Collider (LHC): The Large Hadron Collider (LHC) is a particle accelerator operated at CERN on the outskirts of Geneva, Switzerland. The LHC accelerates two counter-propagating beams of protons in the 27 km synchrotron beam tube formerly occupied by Large Electron-Positron Collider (LEP). It is the largest and brightest accelerator in the world, capable of producing proton-proton collisions with a total energy of 14 TeV. Commissioned in 2008–09, the LHC is expected to find the Higgs boson, the last undiscovered particle in the Standard Model, as well as probe physics beyond the Standard Model.

handedness: Handedness, also called "chirality," is a directional property that physical systems may exhibit. A system is "right handed" if it twists in the direction in which the fingers of your right hand curl if your thumb is directed along the natural axis defined by the system. Most naturally occurring sugar molecules are right handed. Fundamental particles with spin also exhibit chirality. In this case, the twist is defined by the particle's spin, and the natural axis by the direction in which the particle is moving. Electrons produced in beta-decay are nearly always left handed.

leptons: The leptons are a family of fundamental particles in the Standard Model. The lepton family has three generations, shown in Unit 1, Fig. 1: the electron and electron neutrino, the muon and muon neutrino, and the tau and tau neutrino.

meson: The term meson refers to any particle in the Standard Model that is made of one quark and one anti-quark. Murray Gell-Mann arranged the leptons into a periodic-table-like structure according to their electric charge and strangeness (see Unit 1, Fig. 1). Examples of mesons are pions and kaons.

Nambu-Goldstone theorem: The Nambu-Goldstone theorem states that the spontaneous breaking of a continuous symmetry generates new, massless particles.

Newton's law of universal gravitation: Newton's law of universal gravitation states that the gravitational force between two massive particles is proportional to the product of the two masses divided by the square of the distance between them. The law of universal gravitation is sometimes called the "inverse square law." See: universal gravitational constant.

nuclear fission: Nuclear fission is the process by which the nucleus of an atom decays into a lighter nucleus, emitting some form of radiation. Nuclear fission reactions power nuclear reactors, and provide the explosive energy in nuclear weapons.

nuclear fusion: Nuclear fusion is the process by which the nucleus of an atom absorbs other particles to form a heavier nucleus. This process releases energy when the nucleus produced in the fusion reaction is not heavier than iron. Nuclear fusion is what powers stars, and is the source of virtually all the elements lighter than iron in the universe.

parity: Parity is an operation that turns a particle or system of particles into its mirror image, reversing their direction of travel and physical positions.

phase: In physics, the term phase has two distinct meanings. The first is a property of waves. If we think of a wave as having peaks and valleys with a zero-crossing between them, the phase of the wave is defined as the distance between the first zero-crossing and the point in space defined as the origin. Two waves with the same frequency are "in phase" if they have the same phase and therefore line up everywhere. Waves with the same frequency but different phases are "out of phase." The term phase also refers to states of matter. For example, water can exist in liquid, solid, and gas phases. In each phase, the water molecules interact differently, and the aggregate of many molecules has distinct physical properties. Condensed matter systems can have interesting and exotic phases, such as superfluid, superconducting, and quantum critical phases. Quantum fields such as the Higgs field can also exist in different phases.

Planck's constant: Planck's constant, denoted by the symbol h, has the value 6.626×10^{-34} m^2 kg/s. It sets the characteristic scale of quantum mechanics. For example, energy is quantized in units of h multiplied by a particle's characteristic frequency, and spin is quantized in units of $h/2\pi$. The quantity $h/2\pi$ appears so frequently in quantum mechanics that it has its own symbol: $\hbar$.

potential energy: Potential energy is energy stored within a physical system. A mass held above the surface of the Earth has gravitational potential energy, two atoms bound in a molecule have chemical potential energy, and two electric charges separated by some distance have electric potential energy. Potential energy can be converted into other forms of energy. If you release the mass, its gravitational potential energy will be converted into kinetic energy as the mass accelerates downward. In the process, the gravitational force will do work on the mass. The force is proportional to the rate at which the potential energy changes. It is common practice to write physical theories in terms of potential energy, and derive forces and interactions from the potential.

quantized: Any quantum system in which a physical property can take on only discrete values is said to be quantized. For instance, the energy of a confined particle is quantized. This is in contrast to a situation in which the energy can vary continuously, which is the case for a free particle.

quantum electrodynamics: Quantum electrodynamics, or QED, is the quantum field theory that describes the electromagnetic force. In QED, electromagnetically charged particles interact by exchanging virtual photons, where photons are the force carried of the electromagnetic force. QED is one of the most stringently tested theories in physics, with theory matching experiment to a part in 10^{12}.

relativistic: A relativistic particle is traveling close enough to the speed of light that classical physics does not provide a good description of its motion, and the effects described by Einstein's theories of special and general relativity must be taken into account.

relativistic limit: In general, the energy of an individual particle is related to the sum of its mass energy and its kinetic energy by Einstein's equation $E^2 = p^2c^2 + m^2c^4$, where *p* is the particle's momentum, *m* is its mass, and *c* is the speed of light. When a particle is moving very close to the speed of light, the first term (p^2c^2) is much larger than the second (m^2c^4), and for all practical purposes the second term can be ignored. This approximation—ignoring the mass contribution to the energy of a particle—is called the "relativistic limit."

Rutherford scattering: The term Rutherford scattering comes from Ernest Rutherford's experiments that led to the discovery of the atomic nucleus. Rutherford directed a beam of alpha particles (which are equivalent to helium nuclei) at a gold foil and observed that most of the alpha particles passed through the foil with minimal deflection, but that occasionally one bounced back as if it had struck something solid.

spacetime: In classical physics, space and time are considered separate things. Space is three-dimensional, and can be divided into a three-dimensional grid of cubes that describes the Euclidean geometry familiar from high-school math class. Time is one-dimensional in classical physics. Einstein's theory of special relativity combines the three dimensions of space and one dimension of time into a four-dimensional grid called "spacetime." Spacetime may be flat, in which case Euclidean geometry describes the three space dimensions, or curved. In Einstein's theory of general relativity, the distribution of matter and energy in the universe determines the curvature of spacetime.

spontaneous symmetry breaking: Spontaneous symmetry breaking is said to occur when the theory that describes a system contains a symmetry that is not manifest in the ground state. A simple everyday example is a pencil balanced on its tip. The pencil, which is symmetric about its long axis and equally likely to fall in any direction, is in an unstable equilibrium. If anything (spontaneously) disturbs the pencil, it will fall over in a particular direction and the symmetry will no longer be manifest.

strong interaction: The strong interaction, or strong nuclear force, is one of the four fundamental forces of nature. It acts on quarks, binding them together into mesons. Unlike the other forces, the strong force between two particles remains constant as the distance between them grows, but actually gets

weaker when the particles get close enough together. This unique feature ensures that single quarks are not found in nature. True to its name, the strong force is a few orders of magnitude stronger than the electromagnetic and weak interactions, and many orders of magnitude stronger than gravity.

superpartner: In the theory of supersymmetry, every Standard Model particle has a corresponding "sparticle" partner with a spin that differs by 1/2. Superpartner is the general term for these partner particles. The superpartner of a boson is always a fermion, and the superpartner of a fermion is always a boson. The superpartners have the same mass, charge, and other internal properties as their Standard Model counterparts. See: supersymmetry.

supersymmetry: Supersymmetry, or SUSY, is a proposed extension to the Standard Model that arose in the context of the search for a viable theory of quantum gravity. SUSY requires that every particle have a corresponding superpartner with a spin that differs by 1/2. While no superpartner particles have yet been detected, SUSY is favored by many theorists because it is required by string theory and addresses other outstanding problems in physics. For example, the lightest superpartner particle could comprise a significant portion of the dark matter.

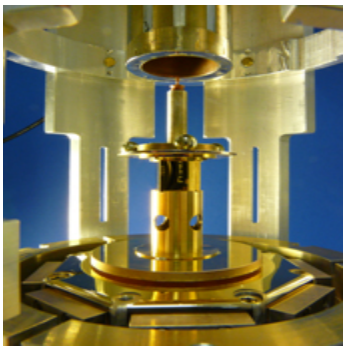
symmetry transformation: A symmetry transformation is a transformation of a physical system that leaves it in an indistinguishable state from its starting state. For example, rotating a square by 90 degrees is a symmetry transformation because the square looks exactly the same afterward.

virtual particle: A virtual particle is a particle that appears spontaneously and exists only for the amount of time allowed by the Heisenberg uncertainty principle. According to the uncertainty principle, the product of the uncertainty of a measured energy and the uncertainty in the measurement time must be greater than Planck's constant divided by 2π. This means that a particle with a certain energy can spontaneously appear out of the vacuum and live for an amount of time inversely proportional to its energy. The force carriers exchanged in an interaction are virtual particles. Virtual particles cannot be observed directly, but their consequences can be calculated using Feynman diagrams and are verified experimentally.

weak interaction: The weak interaction, or weak force, is one of the four fundamental forces of nature. It is called "weak" because it is significantly weaker than both the strong force and the electromagnetic force; however, it is still much stronger than gravity. The weak changes one flavor of quark into another, and is responsible for radioactive decay.

Physics for the 21st Century

Unit 3: Gravity



© Ted Cook.

Unit Overview

Although by far the weakest of the known forces in nature, gravity pervades the universe and played an essential role in the evolution of the universe to its current state. Newton's law of universal gravitation and its elegant successor, Einstein's theory of general relativity, represent milestones in the history of science and provide the best descriptions we have of gravity. General relativity is founded on the principle of equivalence of gravity and acceleration; an inescapable consequence is that gravity governs the very geometry of space and time. This property of gravity distinguishes it from the other forces and makes attempts to unify all of the forces into a "theory of everything" exceedingly difficult. How well do we really understand gravity? Do the same laws of gravity apply to objects on the opposite sides of the universe as to particles in the microscopic quantum world? Current research is attempting to improve the precision to which the laws of gravity have been tested and to expand the realm over which tests of gravity have been made. Gravitational waves, predicted by general relativity, are expected to be observed in the near future. This unit will review what we know about gravity and describe many of the directions that research in gravitation is following.

Content for This Unit

Sections:

Section 1: Introduction

Any two objects, regardless of their composition, size, or distance apart, feel a force that attracts them toward one another. We know this force as gravity. The study of gravity has played a central role in the history of science from the 17th century, during which Galileo Galilei compared objects falling under the influence of gravity and Sir Isaac Newton proposed the law of universal gravitation, to the 20th century and Albert Einstein's theory of general relativity, to the present day, when intense research in gravitational physics focuses on such topics as black holes, gravitational waves, and the composition and evolution of the universe.

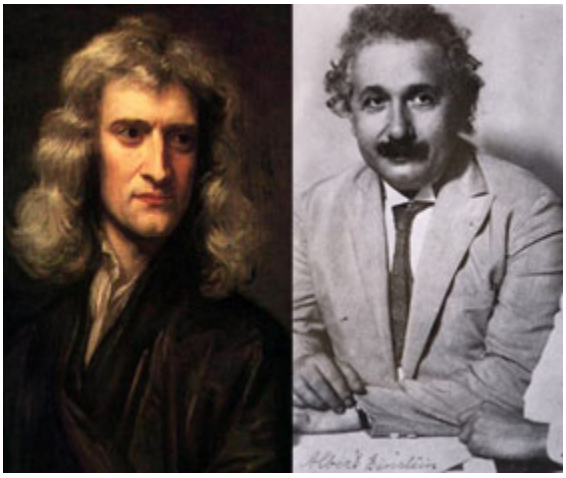


Figure 1: Portraits of Sir Isaac Newton (left) and Albert Einstein (right).
Source: © Image of Newton: Wikimedia Commons, Public Domain; Image of Einstein: Marcelo Gleiser.

Any study of gravity must accommodate two antithetical facts. In many ways, gravity is the dominant force in the universe. Yet, of the four forces known in nature, gravity is by far the weakest. The reason for that weakness remains a major unanswered question in science. Gravity also forms the central focus of efforts to create a "theory of everything" by unifying all four forces of nature. Ironically, gravity was responsible for the first unification of forces, when Newton identified the force that caused an apple to fall to Earth to be the same as the force that held the Moon in orbit.

Current research on gravity takes several forms. Experiments with ever-greater precision seek to test the foundations of gravitational theory such as the universality of free fall and the inverse square law. Other experimentalists are developing ways to detect the gravitational waves predicted by Einstein's general relativity theory and to understand the fundamental nature of gravity at the largest and smallest

units of length. At the same time, theorists are exploring new approaches to gravity that extend Einstein's monumental work in the effort to reconcile quantum mechanics and general relativity.

Section 2: Nature's Strongest and Weakest Force

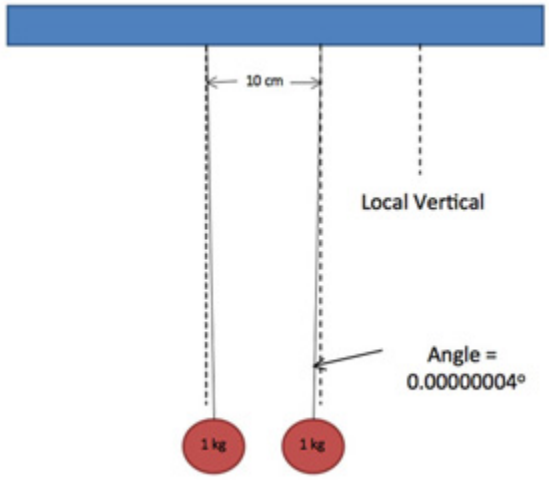


Figure 2: Gravitational attraction between two spheres causes a tiny change in their positions.
Source: © Blayne Heckel.

How weak is gravity? We can find out by comparing the gravitational force with the electromagnetic force, the other long-range force in nature, in the case of a hydrogen atom. By using Coulomb's law of electrical attraction and repulsion we can compute the magnitude of the attractive electrical force, F_E, between the electron and proton and Newton's Law of universal gravitation, which we will discuss in the next section, to calculate the magnitude of the gravitational force, F_G, between the two particles. We find that $F_G/F_E \approx 4 \times 10^{-40}$. Because both forces decrease as the square of the distance between the objects, the gravitational force between the electron and proton remains almost 39 orders of magnitude weaker than the electric force at all distances. That is a number so large that we can hardly fathom it: roughly the ratio of the size of the observable universe to the size of an atomic nucleus. Relatively speaking, at short distances the strong, weak, and electromagnetic forces all have comparable strengths, 39 orders of magnitude stronger than gravity.

The contrast has practical consequences. We can easily feel the magnetic force between two refrigerator magnets, yet we don't feel the gravitational force of attraction between our hands when they are near to one another. The force is there, but too weak to notice. Physicists use sensitive instruments such as the torsion balances that we discuss below to detect the gravitational force between small objects. But the measurements require great care to ensure that residual electric and magnetic forces do not overwhelm the feeble gravitational effects.

Nevertheless, gravity is the force we experience most often. Whether lifting our arms, climbing a staircase, or throwing a ball, we routinely feel and compensate for the effects of our gravitational attraction to the Earth in our daily lives. We call the direction opposite to Earth's gravity "up." Removing the effects of Earth's gravity in a free fall off a diving board or the weightlessness of space leaves us disoriented. Gravity holds the Moon in orbit about the Earth, the Earth in orbit about the Sun, and the Sun in orbit about the center of our Milky Way galaxy. Gravity holds groups of galaxies together in clusters and, we believe, governs the largest structures in the universe.

Gravity's role in forming stars and galaxies

Gravity also caused stars and galaxies to form in the first place. The standard model of cosmology has the universe beginning in a Big Bang roughly 14 billion years ago, followed by an expansion that continues today. At an early age, before stars existed, the universe could be described as a nearly homogeneous gas of matter and radiation. The matter consisted mostly of hydrogen atoms, helium atoms, neutrinos, and dark matter (an unknown form of matter that interacts via gravity but whose exact nature is currently a field of intense research, as we shall see in Unit 10). In regions of space where the density of matter slightly exceeded the average, the gravitational attraction between the constituents of the matter caused the gas to coalesce into large clouds. Friction inside the clouds due to collisions between the atoms and further gravitational attraction caused regions of the clouds to coalesce to densities so high as to ignite nuclear fusion, the energy source of stars.

Figure 3: Hubble Space Telescope image of a star-forming region in the Small Magellanic Cloud.
Source: © NASA/ESA and A.Nota (STScI/ESA).

A massive star that has burnt all of its nuclear fuel can collapse under the influence of gravity into a black hole, a region of space where gravity is so strong that not even light can escape the gravitational pull. Near to a black hole, therefore, nature's weakest interaction exerts the strongest force in the universe.

How do physicists reconcile the incredible weakness of gravity relative to the electromagnetic force with the observation that gravity dominates the interactions between the largest objects in the universe? How can it take the gravitational attraction billions of years, as calculations show, to cause two hydrogen atoms starting just 10 cm apart to collide when we know that the hydrogen gas of the early universe condensed into enormous clouds and stars on a much quicker time scale? Why does Earth's gravity feel so strong while the gravitational forces between objects on the Earth are so small as to be difficult to detect? The answer, common to all of these questions, arises from the relative masses of the objects in question. Gravity is weak between objects that have small masses, but it grows in strength as the objects grow in mass. This seemingly simple answer reflects a profound difference between gravity and the other forces in nature.

Attraction without repulsion

Gravity is an attractive force that acts between any objects at any distance regardless of their composition. The property of matter that gives rise to this attraction is essentially the mass of the object.

The gravitational force between each atom in the Earth and each atom in our bodies is incredibly small. However, every one of the roughly 10^{50} atoms in the Earth attracts each of the approximately 10^{27} atoms in our bodies, leading to the appreciable force that we experience. In contrast, the other forces in nature can be both attractive and repulsive. The electric force is attractive between unlike charges and equally repulsive between like charges. Because ordinary matter, such as the Earth or our bodies, consists of equal numbers of positive and negative charges bound closely together in atoms, the net electric force between electrically neutral objects essentially vanishes.

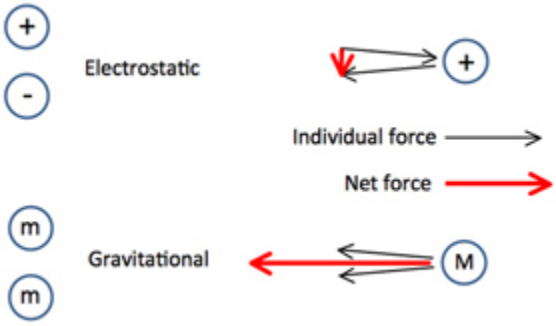


Figure 4: Electrostatic and gravitational shielding.
Source: © Blayne Heckel.

If we place an electric charge inside an otherwise empty grounded metal box, then a charge outside of the box is unaffected by the charge inside. This "electrical shielding" arises from the movement of charges within the metal that rearrange themselves to cancel the electric force of the charge inside. If instead, we place a mass inside a grounded metal box or any other kind of box, another mass placed outside of the box will always feel its gravitational pull, as well as the pull from the mass of the box itself. Because the gravitational force has only one sign—attractive—it cannot be shielded. Every particle, whether normal or dark matter, in regions of the early universe that had slightly higher than average density gravitationally attracted nearby particles more strongly than did regions with less than average density. Gravity caused the matter to coalesce into the structures in the universe that we see today.

As we stand at rest, the few square inches of our feet in contact with the ground oppose the downward gravitational pull of all 10^{50} atoms in the Earth. What counteracts gravity is the electrical repulsion between the outermost electrons of the soles of our shoes and the electrons at the ground's surface. The mere act of standing embodies the contrast between the weak but cumulative gravitational attraction and the much stronger but self-canceling electric force.

Section 3: Newton's Law of Universal Gravitation

An underlying theme in science is the idea of unification—the attempt to explain seemingly disparate phenomena under the umbrella of a common theoretical framework. The first major unification in physics was Sir Isaac Newton's realization that the same force that caused an apple to fall at the Earth's surface—gravity—was also responsible for holding the Moon in orbit about the Earth. This universal force would also act between the planets and the Sun, providing a common explanation for both terrestrial and astronomical phenomena.

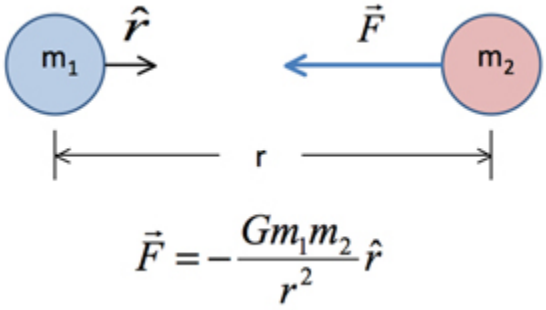


Figure 5: Newton's law of universal gravitation.
Source: © Blayne Heckel.

Newton's law of universal gravitation states that *every two particles attract one another with a force that is proportional to the product of their masses and inversely proportional to the square of the distance between them*. The proportionality constant, denoted by G, is called the universal gravitational constant. We can use it to calculate the minute size of the gravitational force inside a hydrogen atom. If we assign m_1 the mass of a proton, 1.67×10^{-27} kilograms and m_2 the mass of an electron, 9.11×10^{-31} kilograms, and use 5.3×10^{-11} meters as the average separation of the proton and electron in a hydrogen atom, we find the gravitational force to be 3.6×10^{-47} Newtons. This is approximately 39 orders of magnitude smaller than the electromagnetic force that binds the electron to the proton in the hydrogen nucleus. See the math

Local gravitational acceleration

The law of universal gravitation describes the force between point particles. Yet, it also accurately describes the gravitational force between the Earth and Moon if we consider both bodies to be points with all of their masses concentrated at their centers. The fact that the gravitational force from a spherically

symmetric object acts as if all of its mass is concentrated at its center is a property of the inverse square dependence of the law of universal gravitation. If the force depended on distance in any other way, the resulting behavior would be much more complicated. A related property of an inverse square law force is that the net force on a particle inside of a spherically symmetric shell vanishes.

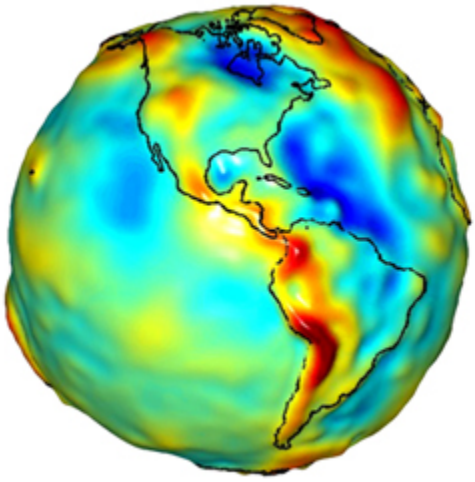


Figure 6: GRACE mission gravity map of the Earth.
Source: © Courtesy of The University of Texas Center for Space Research (NASA/DLR Gravity Recovery and Climate Experiment).

Just as we define an electric field as the electric force per unit charge, we define a gravitational field as the gravitational force per unit mass. The units of a gravitational field are the same units as acceleration, meters per second squared (m/s^2). For a point near the surface of the Earth, we can use Newton's law of universal gravitation to find the local gravitational acceleration, g. If we plug in the mass of the Earth for one of the two masses and the radius of the Earth for the separation between the two masses, we find that g is 9.81 m/s^2. This is the rate at which an object dropped near the Earth's surface will accelerate under the influence of gravity. Its velocity will increase by 9.8 meters per second, each second. Unlike big G, the universal gravitational constant, little g is not a constant. As we move up further from the Earth's surface, g decreases (by 3 parts in 10^5 for each 100 meters of elevation). But it also decreases as we descend down a borehole, because the mass that influences the local gravitational field is no longer that of the entire Earth but rather the total mass within the radius to which we have descended.

Even at constant elevation above sea level, g is not a constant. The Earth's rotation flattens the globe into an oblate spheroid; the radius at the equator is nearly 20 kilometers larger than at the poles, leading

to a 0.5 percent larger value for g at the poles than at the equator. Irregular density distributions within the Earth also contribute to variations in g. Scientists can use maps of the gravitational field across the Earth's surface to infer what structures lay below the surface.

Gravitational fields and tides

Every object in the universe creates a gravitational field that pervades the universe. For example, the gravitational acceleration at the surface of the Moon is about one-sixth of that on Earth's surface. The gravitational field of the Sun at the position of the Earth is 5.9×10^{-3} m/s^2, while that of the Moon at the position of the Earth is 3.3×10^{-5} m/s^2, 180 times weaker than that of the Sun.

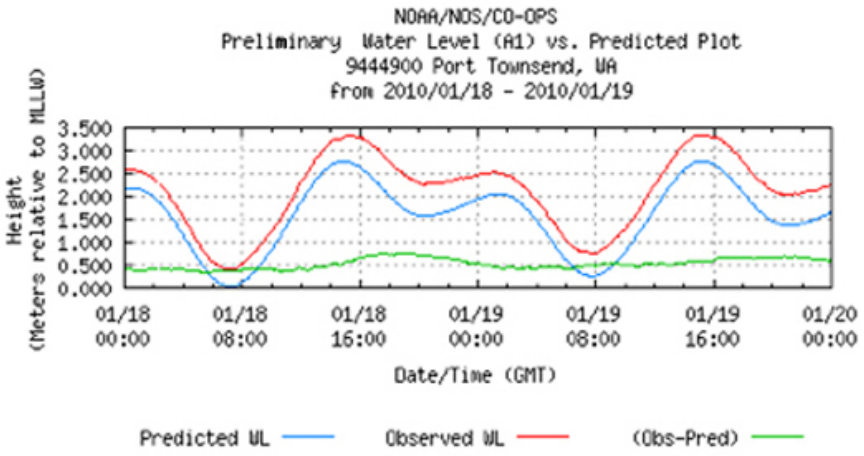


Figure 7: Plot of the tidal Water Level (WL) at Port Townsend, Washington.
Source: © NOAA.

The tides on Earth result from the gravitational pull of the Moon and Sun. Despite the Sun's far greater gravitational field, the lunar tide exceeds the solar tide. That's because it is not the gravitational field itself that produces the tides but its gradient—the amount the field changes from Earth's near side to its far side. If the Sun's gravitational field were uniform across the Earth, all points on and within the Earth would feel the same force, and there would be no relative motion (or tidal bulge) between them. However, because the gravitational field decreases as the inverse of the distance squared, the side of the Earth facing the Sun or Moon feels a larger field and the side opposite feels a smaller field than the field acting at Earth's center. The result is that water (and the Earth itself to a lesser extent) bulges toward the Moon or Sun on the near side and away on the far side, leading to tides twice a day. Because the Moon is much

closer to Earth than the Sun, its gravitational gradient between the near and far sides of the Earth is more than twice as large as that of the Sun.

Section 4: Gravitational and Inertial Mass

A subtlety arises when we compare the law of universal gravitation with Newton's second law of motion. The mass that appears in the law of universal gravitation is the property of the particle that creates the gravitational force acting on the other particle; for if we double m_2, we double the force on m_1. Similarly, the mass in the law of universal gravitation is the property of the particle that responds to the gravitational force created by the other particle. The law of universal gravitation provides a definition of gravitational mass as the property of matter that creates and responds to gravitational forces. Newton's second law of motion, F=ma, describes how any force, gravitational or not, changes the motion of an object. For a given force, a large mass responds with a small acceleration and vice versa. The second law provides a definition of inertial mass as the property of matter that resists changes in motion or, equivalently, as an object's inertia.

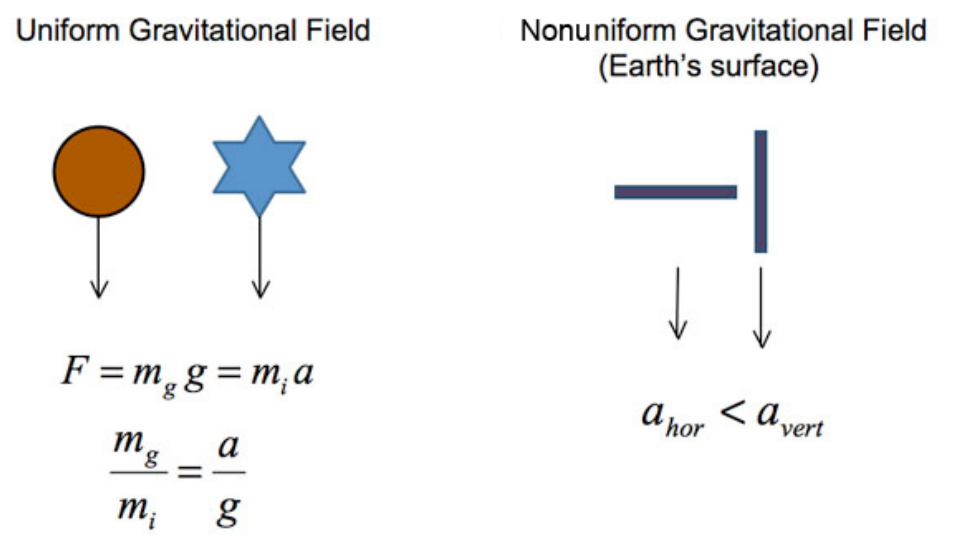


Figure 8: Equality of gravitational and inertial mass.
Source: © Blayne Heckel.

Is the inertial mass of an object necessarily the same as its gravitational mass? This question troubled Newton and many others since his time. Experiments are consistent with the premise that inertial and gravitational mass are the same. We can measure the weight of an object by suspending it from a spring balance. Earth's gravity pulls the object down with a force (weight) of $m_g g$, where g is the local gravitational acceleration and m_g the gravitational mass of the object. Gravity's pull on the object is

balanced by the upward force provided by the stretched spring. We say that two masses that stretch identical springs by identical amounts have the same gravitational mass, even if they possess different sizes, shapes, or compositions. But will they have the same inertial mass? We can answer this question by cutting the springs, letting the masses fall, and measuring the accelerations. The second law says the net force acting on the mass is the product of the inertial mass, m_i, and acceleration, a, giving us: $m_g g = m_i a$ or $g/a = m_i/m_g$. But g is a property of the Earth alone and does not depend upon which object is placed at its surface, while experiments find the acceleration, a, to be the same for all objects falling from the same point in the absence of air friction. Therefore, g/a is the same for all objects and thus for m_i/m_g. We define the universal gravitational constant, G, to make $m_i = m_g$.

The principle of the universality of free fall is the statement that all materials fall at the same rate in a uniform gravitational field. This principle is equivalent to the statement that $m_i = m_g$. Physicists have found the principle to be valid within the limits of their experiments' precision, allowing them to use the same mass in both the law of universal gravitation and Newton's second law.

Measuring G

Measurements of planets' orbits about the Sun provide a value for the product GM_s, where M_S is the mass of the Sun. Similarly, earthbound satellites and the Moon's orbit provide a value for GM_E, where M_E is the mass of the Earth. To determine a value for G alone requires an *a priori* knowledge of both masses involved in the gravitational attraction. Physicists have made the most precise laboratory measurements of G using an instrument called a "torsion balance," or torsion pendulum. This consists of a mass distribution suspended by a long thin fiber. Unbalanced forces that act on the suspended mass distribution can rotate the mass distribution; the reflection of a light beam from a mirror attached to the pendulum measures the twist angle. Because a very weak force can twist a long thin fiber, even the tiny torques created by gravitational forces lead to measurable twist angles.

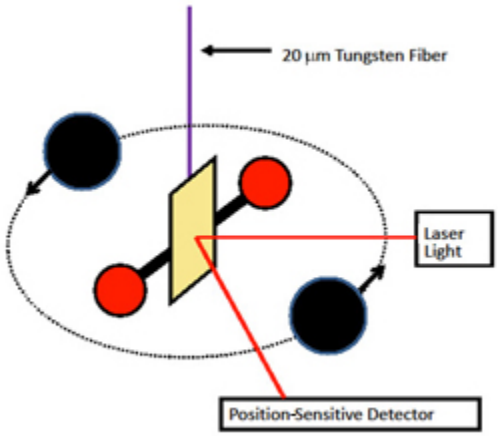


Figure 9: Schematic of a torsion balance to measure the gravitational constant, G.
Source: © Blayne Heckel.

To measure *G*, physicists use a dumbbell-shaped mass distribution (or more recently a rectangular plate) suspended by the fiber, all enclosed within a vacuum vessel. Precisely weighed and positioned massive spheres are placed on a turntable that surrounds the vacuum vessel. Rotating the turntable with the outer spheres about the fiber axis modulates the gravitational torque that the spheres exert on the pendulum and changes the fiber's twist angle.

This type of experiment accounts in large part for the currently accepted value of $(6.67428 \pm 0.00067) \times 10^{-11}$ N-m^2/kg^2 for the universal gravitational constant. It is the least precisely known of the fundamental constants because the weakness of gravity requires the use of relatively large masses, whose homogeneities and positioning are challenging to determine with high precision. Dividing GM_E found from satellite and lunar orbits by the laboratory value for G allows us to deduce the mass of the Earth: 5.98 X 10^{24} kilograms.

Section 5: Testing the Law of Universal Gravitation

Early successes of the law of universal gravitation included an explanation for Kepler's laws of planetary orbits and the discovery of the planet Neptune. Like any physical law, however, its validity rests on its agreement with experimental observations. Although the theory of general relativity has replaced the law of universal gravitation as our best theory of gravity, the three elements of the universal law—the universal constant of gravitation, the equality of gravitational and inertial mass, and the inverse square law—are also key elements of general relativity. To test our understanding of gravity, physicists continue to examine these elements of the universal law of gravitation with ever-increasing experimental sensitivity.

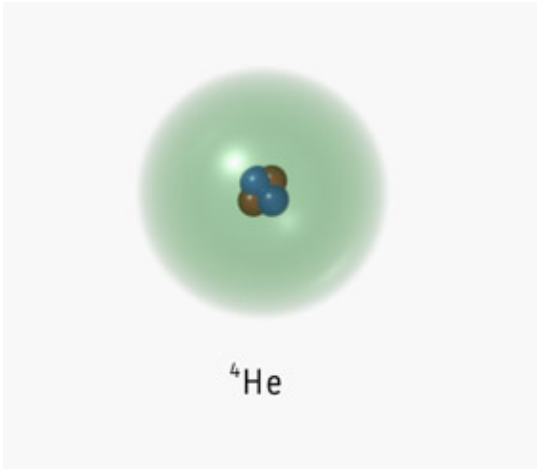


Figure 10: Mass of a helium atom is not equal to the sum of the individual masses of its constituent parts.
Source:

The mass of an object does not equal the sum of the masses of its constituents. For example, the mass of a helium atom is about one part per thousand less than the sum of the masses of the two neutrons, two protons, and two electrons that comprise it. The mass of the Earth is about five parts in 10^{10} smaller than the sum of the masses of the atoms that make up our planet. This difference arises from the nuclear and electrostatic binding—or potential—energy that holds the helium atom together and the gravitational binding (potential) energy that holds the earth together.

The inertial mass of an object therefore has contributions from the masses of the constituents and from all forms of binding energy that act within the object. If $m_i = m_g$, gravity must act equally on the constituent masses and the nuclear, electrostatic, and gravitational binding energies. Is this indeed the case? Does

the Sun's gravity act on both the atoms in the Earth and the gravitational binding energy that holds the Earth together? These are questions that have to be answered by experimental measurements. Modern tests of the universality of free fall tell us that the answer to these questions is yes, at least to within the precision that the measurements have achieved to date.

Tests of the universality of free fall

To test the universality of free fall (UFF), experimentalists compare the accelerations of different materials under the influence of the gravitational force of a third body, called the "source." Many of the most sensitive tests have come from torsion balance measurements. A recent experiment used eight barrel-shaped test bodies attached to a central frame, with four made of beryllium (Be) on one side and four of titanium (Ti) on the other. The denser titanium bodies were hollowed out to make their masses equal to those of the beryllium bodies while preserving the same outer dimensions. All surfaces on the pendulum were coated by a thin layer of gold. The vacuum vessel that surrounded and supported the torsion fiber and pendulum rotated at a slow uniform rate about the tungsten fiber axis. Any differential acceleration of the two types of test bodies toward an external source would have led to a twist about the fiber that changed in sign as the apparatus rotated through 180°. Essential to the experiment was the removal of all extraneous (nongravitational) forces acting on the test bodies.

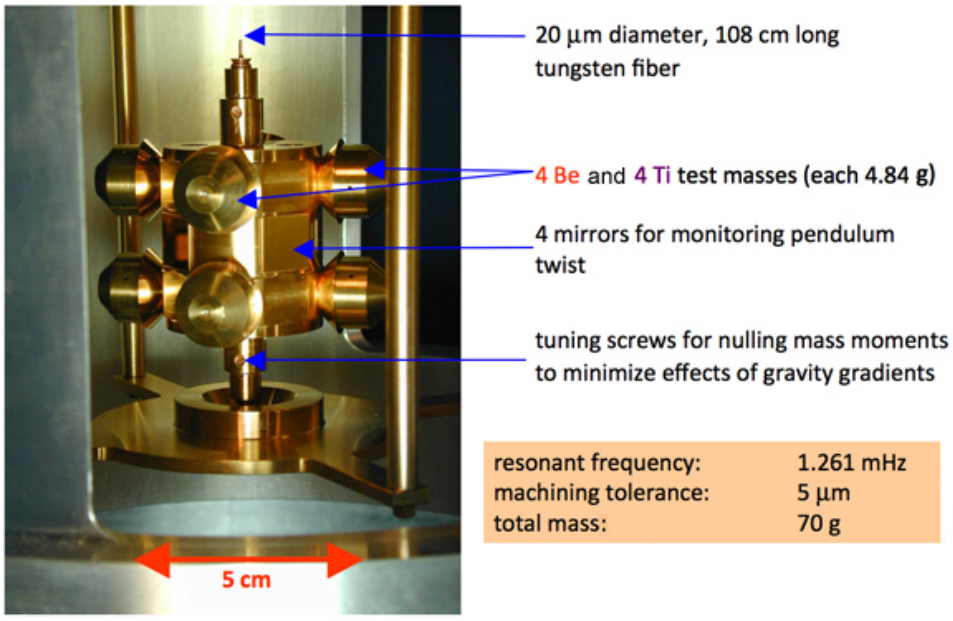


Figure 11: Torsion pendulum to test the universality of free fall.
Source: © Blayne Heckel.

For source masses, experiments have used locally constructed masses within the laboratory, local topographic features such as a hillside, the Earth itself, the Sun, and the entire Milky Way galaxy. Comparing the differential acceleration of test bodies toward the galactic center is of particular interest. Theorists think that dark matter causes roughly 30 percent of our solar system's acceleration about the center of the galaxy. The same dark matter force that helps to hold the solar system in orbit about the galactic center acts on the test bodies of a torsion pendulum. A dark matter force that acts differently on different materials would then lead to an apparent breakdown of the UFF. Because physicists have observed no differential acceleration in the direction of the galactic center, they conclude that dark matter interacts with ordinary matter primarily through gravity.

No violation of the UFF has yet been observed. Physicists use tests of the UFF to search for very weak new forces that may act between objects. Such forces would lead to an apparent violation of the UFF and would be associated with length scales over which the new forces act. Different experimental techniques have been used to test the UFF (and search for new forces) at different length scales. For example, there is a region between 10^3 meters and 10^5 meters over which torsion balances fail to produce reliable constraints on new weak forces. This is because over this length scale, we do not have sufficient knowledge of the density homogeneity of the Earth to calculate reliably the direction of the new force—it might point directly parallel to the fiber axis and not produce a torque on the pendulum. In this length range, the best limits on new forces come from modern "drop tower" experiments that directly compare the accelerations of different materials in free fall at the Earth's surface.

UFF tests in space

The future for tests of the UFF may lie in space-based measurements. In a drag-free satellite, concentric cylinders of different composition can be placed in free fall in the Earth's gravitational field. Experimentalists can monitor the relative displacement (and acceleration) of the two cylinders with exquisite accuracy for long periods of time using optical or superconducting sensors. Satellite-based measurements might achieve a factor of 1,000 times greater sensitivity to UFF violation than ground-based tests.

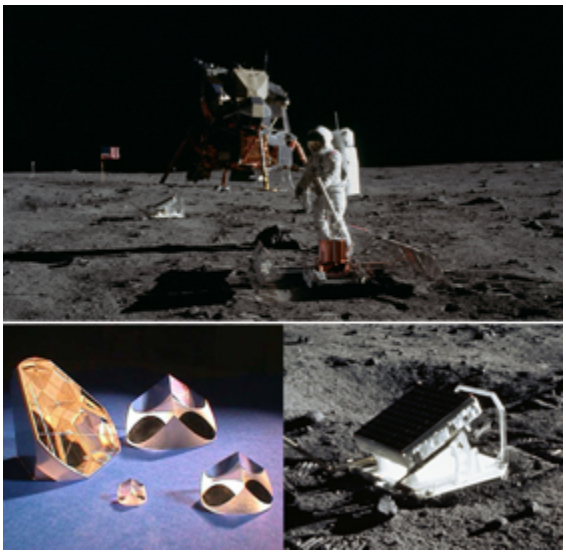


Figure 12: Apollo mission astronauts deploy corner cube reflectors.
Source: © NASA.

One source of space-based tests of the UFF already exists. The Apollo space missions left optical corner mirror reflectors on the Moon that can reflect Earth-based laser light. Accurate measurements of the time of flight of a laser pulse to the Moon and back provide a record of the Earth-Moon separation to a precision that now approaches 1 millimeter. Because both the Earth and the Moon are falling in the gravitational field of the Sun, this lunar laser ranging (LLR) experiment provides a test of the relative accelerations of the Earth and Moon toward the Sun with precision of 2×10^{-13} of their average accelerations. Gravitational binding energy provides a larger fraction of the Earth's mass than it does for the Moon. Were the UFF to be violated because gravity acts differently on gravitational binding energy than other types of mass or binding energy, then one would expect a result about 2,000 times larger than the experimental limit from LLR.

Validating the inverse square law

Physicists have good reason to question the validity of the inverse square law at both large and short distances. Short length scales are the domain of the quantum world, where particles become waves and we can no longer consider point particles at rest. Finding a theory that incorporates gravity within quantum mechanics has given theoretical physicists a daunting challenge for almost a century; it remains an open question. At astronomical length scales, discrepancies between observations and the

expectations of ordinary gravity require dark matter and dark energy to be the dominant constituents of the universe. How sure are we that the inverse square law holds at such vast distances?

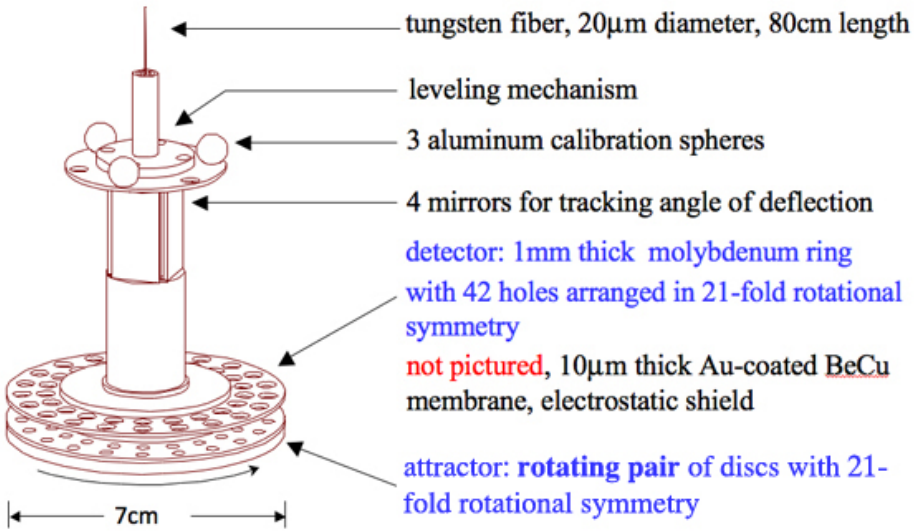


Figure 13: Torsion pendulum to test the inverse square law of gravity at submillimeter distances.
Source: © Blayne Heckel.

The inverse square law has been tested over length scales ranging from 5×10^{-5} to 10^{15} meters. For the large lengths, scientists monitor the orbits of the planets, Moon, and spacecraft with high accuracy and compare them with the orbits calculated for a gravitational force that obeys the inverse square law (including small effects introduced by the theory of general relativity). Adding an additional force can lead to measurable modifications of the orbits. For example, general relativity predicts that the line connecting the perihelia and aphelia of an elliptical gravitational orbit (the points of closest and furthest approach to the Sun for planetary orbits, respectively) should precess slowly. Any violation of the inverse square law would change the precession rate of the ellipse's semi-major axis. So far, no discrepancy has been found between the observed and calculated orbits, allowing scientists to place tight limits on deviations of the inverse square law over solar system length scales.

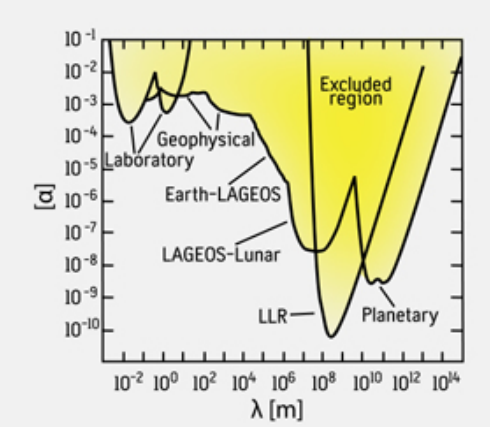


Figure 14: Experimental limits on the universality of free fall.
Source: © Blayne Heckel.

At the shortest distances, researchers measure the gravitational force between plates separated by about 5×10^{-5} meters, a distance smaller than the diameter of a human hair. A thin conducting foil stretched between the plates eliminates any stray electrical forces. Recent studies using a torsion pendulum have confirmed the inverse square law at submillimeter distances. To probe even shorter distances, scientists have etched miniature (micro) cantilevers and torsion oscillators from silicon wafers. These devices have measured forces between macroscopic objects as close as 10^{-8} meters, but not yet with enough sensitivity to isolate the gravitational force.

Does the inverse square law hold at the tiny distances of the quantum world and at the large distances where dark matter and dark energy dominate? We don't know the answer to that question. Definitive tests of gravity at very small and large length scales are difficult to perform. Scientists have made progress in recent years, but they still have much to learn.

Section 6: The Theory of General Relativity

The law of universal gravitation describes a force that acts instantaneously between objects separated by arbitrarily large distances. This behavior is in conflict with the theory of special relativity, which forbids the transfer of information faster than the speed of light. So how does one construct a theory of gravity that is consistent with special relativity?

Einstein found the key: the apparent equivalence between gravity and acceleration. Imagine that you are in a windowless rocket ship far from any stars or planets. With the rocket engines turned off, you and everything else not secured to the rocket float freely in weightlessness within the rocket's cabin. When turned on, the rocket engines provide a constant acceleration—9.8 m/s^2, say—and you stand firmly on the floor directly above the engines. In fact, the floor pushes against your feet with the same force that the ground pushes against your feet when you stand on Earth. Einstein posed the question: Is there any experiment you could perform within your sealed rocket ship that could distinguish between being in a rocket with a constant acceleration of 9.8 m/s^2, or a rocket at rest on the launch pad on Earth?

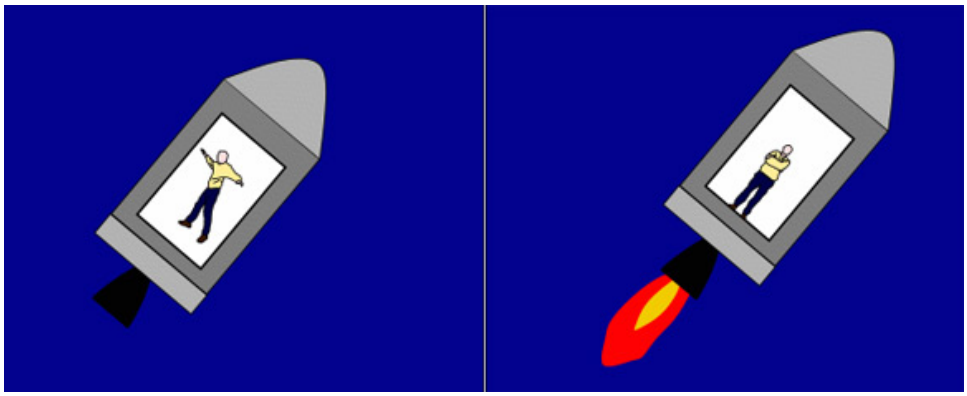


Figure 15: A person in an accelerating rocket feels the same downward pull as a person on Earth feels from gravity.
Source: © M. Poessel/Einstein Online/Max Planck Institute for Gravitational Physics.

Einstein concluded that the answer was no: There is no way to tell the difference between the presence of a uniform gravitational field and a frame of reference that has a constant acceleration. This observation embodies Einstein's principle of equivalence, the equivalence of gravity and acceleration, on which he built the theory of general relativity.

Gravitational lensing and the bending of light

We can use the equivalence between an accelerated reference frame and a frame with a uniform gravitational field to infer the behavior of light in a gravitational field. Imagine a beam of light traveling horizontally from one side of the rocket cabin to the other. With the rocket engines off, the light follows a straight path across the cabin in accordance with the laws of special relativity. With the engines on, causing constant acceleration, the cabin moves slightly upward in the time it takes the light to travel across the cabin. Hence, the light beam strikes a point lower on the cabin wall than when the engines were off. In the frame of the accelerating rocket, the light beam follows a curved (parabolic) path. Because an accelerating rocket is equivalent to a rocket at rest in a uniform gravitational field, a light beam will follow a curved path in a gravitational field; in other words, light is bent by gravity. A famous observation during a solar eclipse in 1919 confirmed that prediction: Measurements showed that starlight passing near the edge of the eclipsed Sun was deflected by an amount consistent with the principle of equivalence.

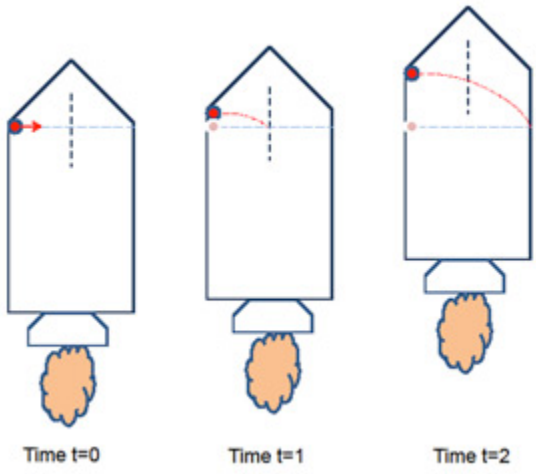


Figure 16: Bending of light in an accelerating rocket.
Source: © Blayne Heckel.

In the absence of gravity, a distant galaxy will appear to an observer on Earth to be a tiny source of light. However, if there are mass distributions such as other galaxies or clouds of dark matter near to the line sight between the Earth and the distant light source, the gravity from these mass distributions will bend the light from the distant galaxy. The image of the distant galaxy on Earth can then become a ring, one or multiple arcs, or even appear as several galaxies depending upon the location and distribution of the intervening mass. This distortion of light from distant sources is called gravitational lensing and is well established in observations from modern telescopes. The observed gravitational lensing is used to infer what sources of gravity lie between the Earth and distant light sources. A related phenomenon is an

increase in intensity of the light observed from a distant source due to the passage of a massive object near to the line of sight. The gravitational field of the moving object acts as a lens, focusing more light into the telescope during the time that the massive object is near to the line of sight.

Gravitational time dilation

Returning to our rocket ship thought-experiment, imagine that a light beam travels from the ceiling to the floor of the accelerating rocket. In the time the beam takes to traverse the cabin, the cabin floor has acquired a larger velocity than it had when the light left the ceiling. A device on the floor measuring the frequency of the light would find a higher frequency than that of the emitted beam because of the Doppler shift, a phenomenon noticed most commonly in an ambulance siren that has a higher pitch as the ambulance approaches and a lower pitch as it recedes. The principle of equivalence then asserts that, in a gravitational field, a light beam traveling opposite to the field acquires a higher frequency, shifted toward the blue end of the spectrum; while a light beam shining upward from the Earth's surface decreases in frequency as it rises, the effect that we know as the gravitational redshift. Again, experiments have confirmed this phenomenon.

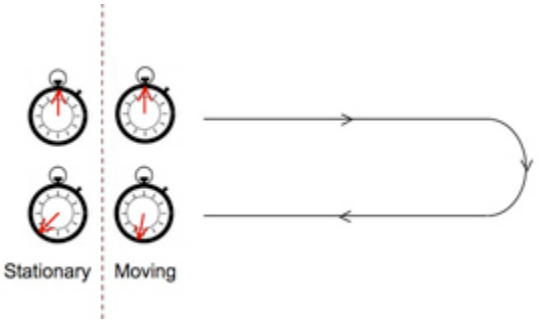


Figure 17: Time dilation and the twin paradox.
Source: © Blayne Heckel.

An inertial (nonaccelerating) observer sees no change in the light's frequency—the frequency associated with the atomic transition generating the light—as the light moves across the cabin, because it is traveling freely through empty space. Yet, an observer on the rocket floor, accelerating with the rocket, can use the same, now accelerating, atoms and atomic transitions as a clock (see Unit 5 for details); the observer defines a second as the time required for the fixed number of oscillations of a specific atomic transition. We concluded in the last paragraph that this accelerating observer will see the frequency of the light beam to be higher than the frequency of the same atomic transition in the measuring device. The

inescapable conclusion is that the atomic clock (like all clocks) ticks more slowly in the accelerating frame of reference.

By the principle of equivalence, clocks in a gravitational field tick more slowly than in the absence of the field; the stronger the field, the more slowly the clock ticks. An atomic clock at sea level loses five microseconds per year relative to an identical clock at an altitude of 5,000 feet. We age more slowly at sea level than on a mountaintop. The global positioning system (GPS) relies heavily on the accuracy of clocks and corrects for the gravitational time dilation to achieve its fantastic precision.

Curved spacetime

The second key ingredient of general relativity is the notion of curved spacetime. Special relativity combines space and time into a four-dimensional spacetime, often referred to as "flat spacetime" or "Minkowski space." In flat spacetime, Euclidean geometry describes the spatial dimensions: Parallel lines never intersect, and the sum of the interior angles of a triangle is always 180 degrees. The two-dimensional analogue of flat space is the Cartesian plane, familiar from high school geometry class.

The surface of a sphere is also a two dimensional surface, but one that must be described by non-Euclidean geometry. Lines of constant longitude are parallel at the equator yet intersect at the poles. If you start at the equator and walk due north to the pole, turn right by 90 degrees, walk south to the equator, and then turn right again and walk along the equator, you will return to your original position having taken a triangular path on the Earth's surface. The sum of the interior angles of your triangular path is 270 degrees. A spherical surface is said to have positive curvature; a saddle-shaped surface has a negative curvature; and the sum of the interior angles of a triangle drawn on a saddle is less than 180 degrees.

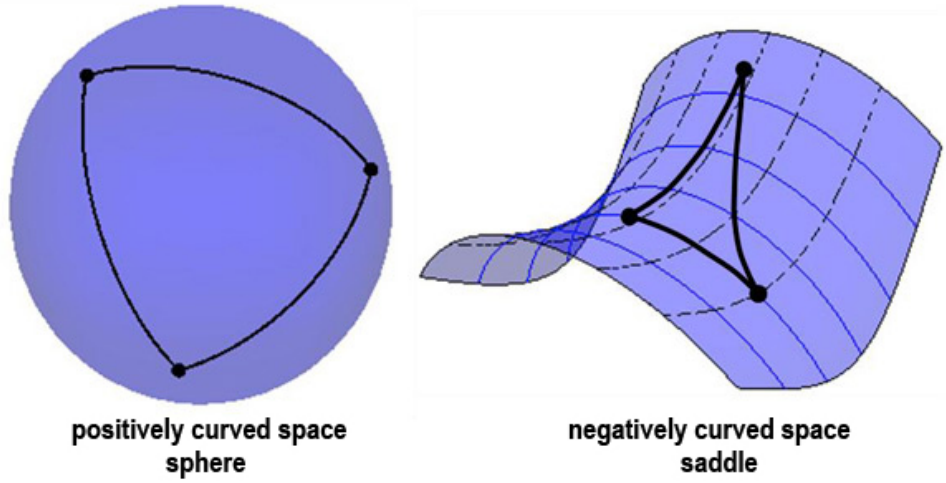


Figure 18: Triangles on curved surfaces.
Source: © Blayne Heckel.

Viewed from three dimensions, the shortest path between two points on a spherical or saddle-shaped surface is a curved line. This "geodesic" is the path that light would follow in that space. Three-dimensional space and four-dimensional spacetime, described by Riemannian geometry, can also be curved. The geodesics in spacetime are the paths that light beams follow—or, equivalently, the paths that observers in free fall follow. The Earth is in free fall about the Sun. We can construct a curved spacetime in which a circular orbit about the Sun is a geodesic. In such a spacetime, the Earth's orbit would stem from the curvature of spacetime rather than from a force acting between the Earth and the Sun.

The theory of general relativity takes the equivalence between motion in a gravitational field and motion in curved spacetime one step further. It asserts that what we call gravity is the bending of spacetime by matter. Rather than viewing gravity as a force acting between objects in flat spacetime, we should understand gravity as the interaction between matter and spacetime. The field equations of general relativity specify how matter and energy determine the curvature of spacetime. In turn, the spacetime curvature determines how the matter and energy will evolve.

Black holes

If enough matter and energy are concentrated in a small enough volume, general relativity predicts that spacetime can become so highly curved that a black hole is formed. A black hole is characterized by an event horizon, a surface surrounding the enclosed matter, through which nothing can escape; the event

horizon represents a surface of no return. To an outside observer, the black hole is completely described by just three numbers: its mass, its electric charge, and its angular momentum.

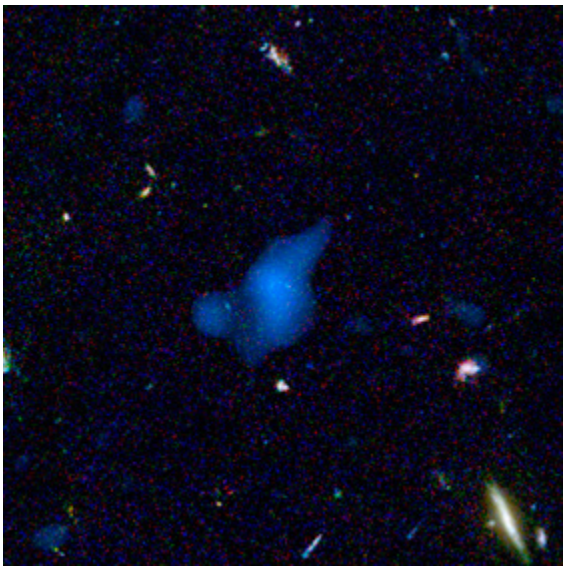


Figure 19: As gas falls into this supermassive black hole, it emits x-rays.
Source: © NASA, ESA, A.M. Koekemoer (STScI), M. Dickinson (NOAO), and the GOODS team.

Black holes can be created when a star of sufficient mass, after having burnt its nuclear fuel, collapses under its own weight. The black hole grows by capturing nearby matter and radiation that is pulled through the event horizon and by merging with astronomical objects such as stars, neutron stars, and other black holes. Massive black holes, millions to billions times more massive than our Sun, have been found near the center of many galaxies, including our own Milky Way. The black holes become visible when they accrete gas from the surrounding regions; the gas is accelerated and heated, producing observable radiation, before falling through the event horizon. The presence of a black hole can also be inferred from its gravitational influence on the orbits of nearby stars.

Section 7: Gravitational Waves

Gravitational waves are the gravitational analogues to electromagnetic waves—electric and magnetic fields that oscillate in the plane perpendicular to the direction that the wave travels. Similarly, gravitational waves are gravitational fields that oscillate perpendicular to the direction of travel. Unlike electromagnetic waves, which can be produced by a single oscillating electric charge, conservation of linear momentum requires at least two masses moving in opposition to produce gravitational waves. In the theory of special relativity, the constant *c*, called the speed of light, connects space with time and is the speed at which all massless particles travel. Like electromagnetic waves, gravitational waves are believed to propagate at the speed *c*.

General relativity predicts the existence of gravitational waves. In matter-free regions of spacetime where gravity is weak, the field equations of general relativity simplify to wave equations for spacetime itself. The solutions to these equations are transverse ripples in spacetime, propagating at the speed of light, which we identify as gravitational waves. The distortion of spacetime caused by a gravitational wave is distinctive: In the plane perpendicular to the direction the wave is travelling, space is stretched along one axis and compressed along the orthogonal axis, and vice versa one half-wave cycle later.

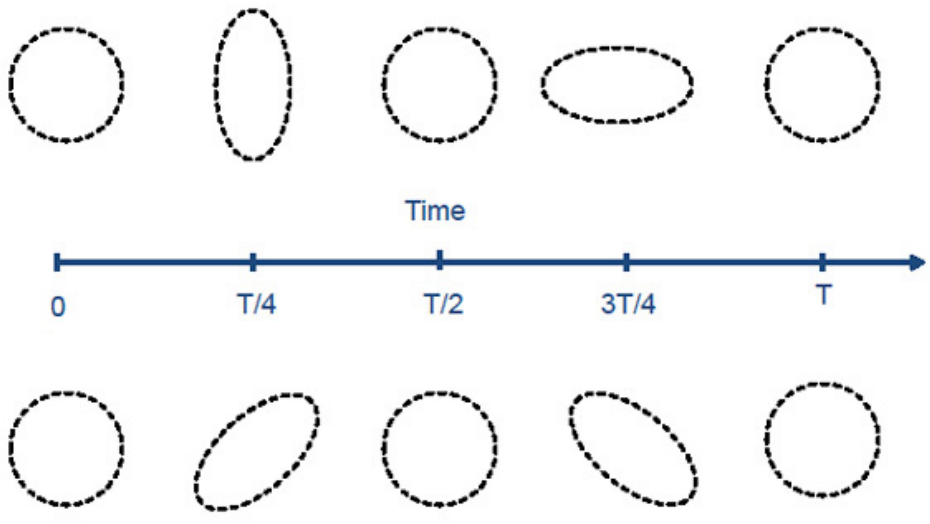


Figure 20: Distortion of space from a gravitational wave.
Source: © Blayne Heckel.

What are the similarities and differences between electromagnetic and gravitational waves? Both waves travel at speed *c* and carry with them energy and momentum. For electromagnetic waves, spacetime

is the background medium in which the waves travel, while for gravitational waves, spacetime itself constitutes the waves. Electromagnetic waves are produced by accelerating or oscillating electric charges, while gravitational waves are produced by accelerating or oscillating mass distributions.

The frequencies of both waves reflect the oscillation frequencies of the sources that produce them. Electronic, vibrational, and rotational transitions (that is, oscillations) in atoms and molecules provide the most common source of electromagnetic waves, producing wave frequencies between roughly 10^7 and 10^{17} Hertz (Hz, or cycles per second). The most efficient sources for gravitational waves are massive objects undergoing rapid acceleration, such as pairs of neutron stars and/or black holes orbiting closely about one another. Considerations of orbital speeds and masses lead us to expect that the strongest gravitational radiation will have frequencies less than 10,000 Hz. Electromagnetic waves interact strongly with matter through absorption or scattering. Gravitational waves, by contrast, interact extremely weakly with matter; they travel essentially unimpeded through spacetime.

Indirect detection of gravitational waves

The most obvious difference between gravitational and electromagnetic waves is the fact that no one has yet directly detected gravitational waves—although this situation should change soon, given the significant progress in the technologies necessary for detection. In the meantime, we have strong indirect evidence that gravitational radiation exists. Astronomers have monitored the orbital frequency of the binary neutron star system PSR1913+16 since 1974, the year that Russell Hulse and Joseph Taylor discovered the system. One of the neutron stars is a pulsar that beams radio waves to the Earth as the neutron star rotates about its axis. Astrophysicists use the arrival times of the radio pulses to reconstruct the orbit of the binary system. The oscillating mass distribution of this binary system should generate gravitational waves and lose orbital energy as the waves radiate outward. A loss in orbital energy moves the neutron stars closer together and decreases the orbital period. The observed decrease of the orbital period over the past 35 years agrees with the energy loss through gravitational radiation predicted by general relativity to better than 1 percent accuracy.

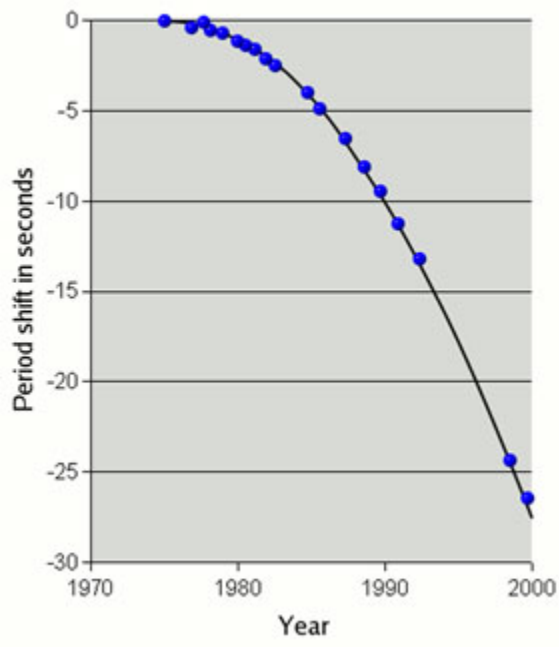


Figure 21: Orbital period of the binary neutron star system PSR 1913+15 measured from 1975 to 2000.
Source: © Wikimedia Commons, Public Domain. Author: M. Haynes et Lorimer, 2001.

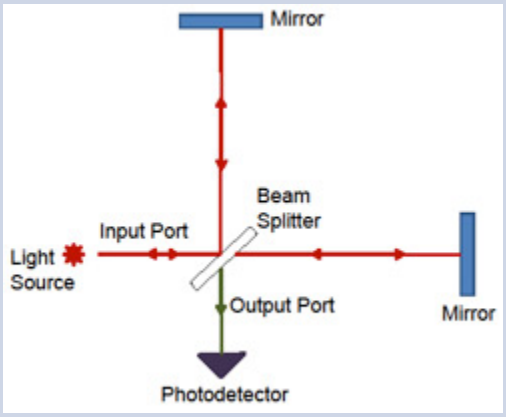
Radio pulses from pulsars arrive at such a regular rate as to provide hope that pulsars may provide a means to detect very low frequency gravitational waves. Waves with frequencies around 10^{-9} Hz (equivalent to wavelengths of around 10 light-years) may persist from mass motions early in the history of the universe. When such a wave passes a pulsar, it slightly alters the arrival time of the radio beam from the pulsar. By comparing the arrival times of signals from perhaps 100 pulsars spread across the sky for many years, astronomers might possibly detect the tell-tale distortion of spacetime that is the signature of a passing gravitational wave.

Researchers believe that even lower frequency (that is, longer wavelength) gravitational waves were created in the early moments of the universe. We have evidence for events around 380,000 years after the Big Bang in the form of extraordinarily precise measurements of the cosmic microwave background (CMB), which is electromagnetic radiation left over from the early universe. Primordial gravitational waves would leave their imprint on the CMB as a distinctive polarization pattern as one compares the polarization of CMB radiation from different regions across the sky. Intense efforts are under way to mount instruments with enough polarization sensitivity to search for the primordial gravitational waves. Both ground-based observations (CLOVER, EBEX, Polarbear, QUIET, SPIDER, and SPUD instruments,

to name a few) and space-based measurements from the Planck satellite launched in 2009 promise rapid progress toward the detection of primordial gravitational waves.

Direct detection of gravitational waves

The Classic Michelson Interferometer



Source: © Blayne Heckel.

Originally devised as part of the fruitless 19th century effort to identify the "ether" that supposedly suffused space, the Michelson interferometer now finds application in a 21st century experiment: the search for gravitational waves. The diagram shows the original version of the instrument.

A beam splitter divides laser light entering the input port into a transmitted beam and a reflected beam, perpendicular to each other. At the end of each beam's path, a mirror reflects the light back toward the beam splitter. If the two beams' paths have exactly the same length, the beams' electric fields oscillate in phase when the light returns to the beam splitter. The beams recombine to produce a beam that exits the beam splitter along the output port.

If the two paths differ in length by half a wavelength, they are out of phase. In that case, they interfere destructively at the beam splitter and no light exits from the output port. The intensity of light leaving the output port changes from a maximum to zero as the relative distance to the end mirrors changes by a quarter of a wavelength—about 2.5 x 10^{-7} meters for typical laser light. Precisely measuring this light intensity allows experimenters to detect even smaller relative displacements of the mirrors.

Any passing gravitational wave should compress spacetime in one direction and stretch it out in the perpendicular direction. Physicists believe that a modern version of the Michelson interferometer has the precise measuring ability that can detect the difference between the two.

The earliest attempts to detect gravitational waves directly used resonant mass detectors, also called "bar detectors," first developed by Joseph Weber. A typical bar detector might be a 5000 kg cylinder, two meters long, suspended in vacuum, and made from a low mechanical loss material such as certain alloys of aluminum. A burst of gravitational radiation could stretch and compress the bar, exciting the roughly one kilohertz lowest frequency vibrational mode of the cylinder. Sensors at the ends of the cylinder would detect the vibrations. A low-loss material would ring for many vibrational cycles, enhancing the ability to identify the excess vibration from a gravitational wave in the presence of background noise. Modern versions of the bar detectors (for example, the NAUTILUS and AURIGA detectors in Italy, miniGRAIL in the Netherlands, and the EXPLORER bar in Switzerland) are cooled to liquid helium temperatures or even lower to reduce the mechanical losses and thermal vibrations, and to reduce the noise inherent in the motion sensors.

Figure 22: Nautilus cryogenic antenna at the Laboratori Nazionali di Frascati, Italy.
Source: © Italian National Institute of Nuclear Physics (INFN)—National Laboratory of Frascati.

The most developed technology for the detection of gravitational waves involves long baseline laser interferometers. These instruments use laser light as a "meter stick" to compare the distances between a central object and distant objects along perpendicular axes. A passing gravitational wave will compress spacetime along one axis while stretching it along a perpendicular axis. An interferometer provides a precise measurement of the relative distance that light travels along different paths.

The long baseline gravitational wave interferometers are refined versions of the Michelson interferometer that, when it failed to detect the ether late in the 19th century, helped to set the scene for the theory of special relativity. But instead of being mounted rigidly on a table, the end mirrors of the gravitational wave instruments are suspended, like pendulums, from thin wires. In addition, the entire laser path occurs

within a vacuum chamber. In the horizontal plane, the end mirrors are essentially objects in freefall, able to follow the stretching and compressing of spacetime from a gravitational wave. (In the classical picture of gravitational waves, the waves produce horizontal forces on the end mirrors; suspended mirrors can move in response to the wave forces.) However, even the strongest gravitational waves that one might hope to detect on Earth stretch space by an extremely small amount: The strain (change in distance divided by the distance) between two objects is expected to be less than 10^{-18}. To make the change in distance large enough for an interferometer to detect, designers must make the baseline as long as possible.

Gravitational wave discovery on Earth and in space

Figure 23: Aerial view of the LIGO Observatory at Hanford, Washington.
Source: © LIGO Laboratory.

The LIGO (Laser Interferometer Gravitational Wave Observatory) interferometers in the states of Louisiana and Washington each have end mirrors 4 kilometers from the beam splitter. VIRGO in Italy, GEO in Germany, and TAMA in Japan have separation distances of 3 kilometers, 600 meters, and 300 meters, respectively. With the 4-kilometer separation, physicists expect a strong gravitational wave to produce a relative change in distance between the mirrors and beam splitter of only about 4 x 10^{-15} meters, roughly the size of an atomic nucleus. Having several gravitational wave interferometers operating simultaneously greatly improves the chances of distinguishing a gravitational wave from the inevitable background sources of noise.

Ground-based gravitational wave interferometers are designed to detect waves with frequencies between roughly 10 Hz and 1,000 Hz. Sources for gravitational waves in this frequency band include the final moments of the in-spiral of orbiting pairs of neutron stars or black holes that lead to their collision and

merger into a single object, violent astronomical events such as supernovae, and constant frequency signals such as those from a rapidly rotating neutron star that has a residual mass quadrupole moment.

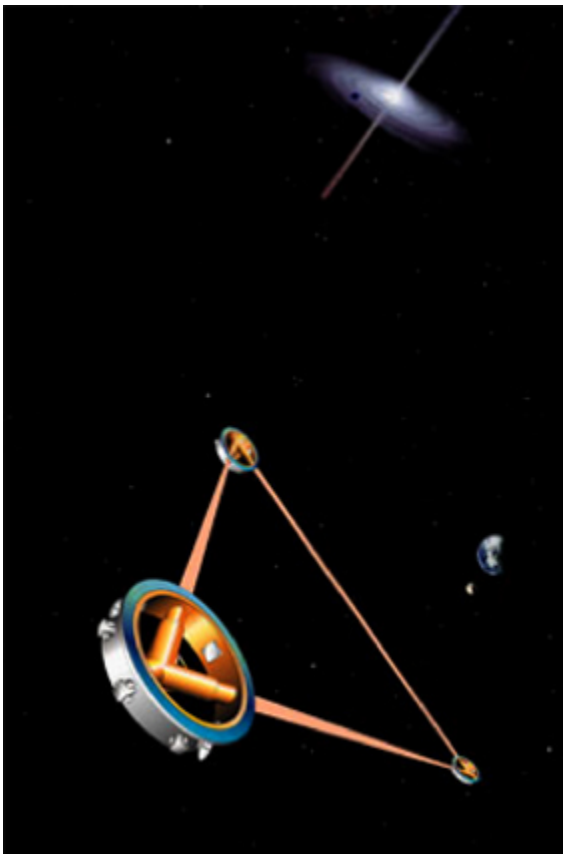


Figure 24: Artist's conception of the LISA satellites in space.
Source: © JPL/NASA.

Ground motion and seismic noise increase rapidly below a frequency of about 10 Hz and prevent Earth-based interferometers from detecting gravitational waves below this frequency limit. However, placing the interferometer on satellites in space allows us to avoid seismic noise and to envision much larger separations between the components of the interferometer. LISA (Laser Interferometer Space Antenna) is a joint NASA/European Space Agency proposal to launch three satellites into orbits to form an equilateral triangle with a distance of 5×10^6 kilometers between each spacecraft. Laser light exchanged between the spacecraft will measure the relative distances between them and may detect gravitational waves within a frequency range of 10^{-4} Hz to 0.1 Hz. Sources for gravitational waves in this frequency band include massive black hole binaries that form after galactic mergers, the orbits of stars as they spiral into black holes, and the gravitational radiation from the orbits of millions of compact binary systems within our

Milky Way galaxy. Once the detection of gravitational waves becomes routine, a new field of gravitational wave astronomy will be born.

Section 8: Gravity and Quantum Mechanics

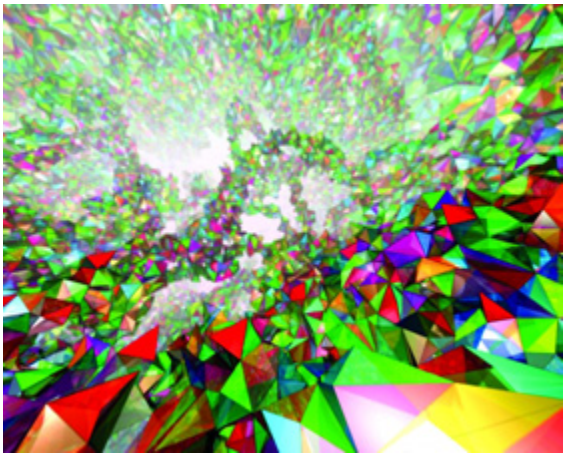


Figure 25: Visualization of a stage in the quantum evolution of geometry, according to Loop Quantum Gravity.
Source: © T. Thiemann (Max Planck Institute for Gravitational Physics (Albert Einstein Institute)) & Mildemarketing Science Communication.

Despite the fact that there is no experimental evidence that conflicts with the predictions of general relativity, physicists have found compelling reasons to suspect that general relativity may be only a good approximation to a more fundamental theory of gravity. The central issue is reconciling general relativity with the demands of quantum mechanics. Well tested by experiment, quantum mechanics is the theory that describes the microscopic behavior of particles. Unit 5 of this course will delve into the details of quantum mechanics. In the quantum world, particles are also waves, the results of measurements are probabilistic in nature, and an uncertainty principle forbids knowing certain pairs of measurable quantities, such as position and momentum, to arbitrary precision. The Standard Model described in the previous two units provides a unified picture of the strong, weak, and electromagnetic forces within the framework of quantum mechanics. Nonetheless, theoretical physicists have found it to be extremely difficult to construct a theory of quantum gravity that incorporates both general relativity and quantum mechanics.

At the atomic scale, gravity is some 40 orders of magnitude weaker than the other forces in nature. In both general relativity and Newtonian gravity, the strength of gravity grows at shorter and shorter distances, while quantum effects prevent the other forces from similarly increasing in strength. At a distance of approximately 10^{-35} m, called the Planck length, gravity becomes as strong as the other forces. At the Planck length, gravity is so strong and spacetime is so highly distorted that our common notions of space and time lose meaning. Quantum fluctuations at this length scale produce energies so

large that microscopic black holes would pop into and out of existence. A theory of quantum gravity is needed to provide a description of nature at the Planck length. Yet, attempts by researchers to construct such a theory, analogous to the Standard Model of particle physics, have lead to serious inconsistencies.

Theories of quantum gravity

A significant difference between a quantum theory of gravity and the Standard Model of particle physics is the role of spacetime in the theory. In the Standard Model, spacetime is a background in which the quantum particles interact. In quantum gravity, spacetime itself participates in the interactions and acquires quantum fluctuations. Theorists have proposed radically new ideas about spacetime at microscopic distances to serve as foundations for theories of quantum gravity. Loop Quantum Gravity is an approach in which spacetime itself arises from the theory as a grid of discrete (quantized) loops of gravitational field lines called "spin networks." In Causal Dynamical Triangulation, spacetime is two-dimensional at the Planck length scale and evolves into our four-dimensional spacetime at larger length scales.

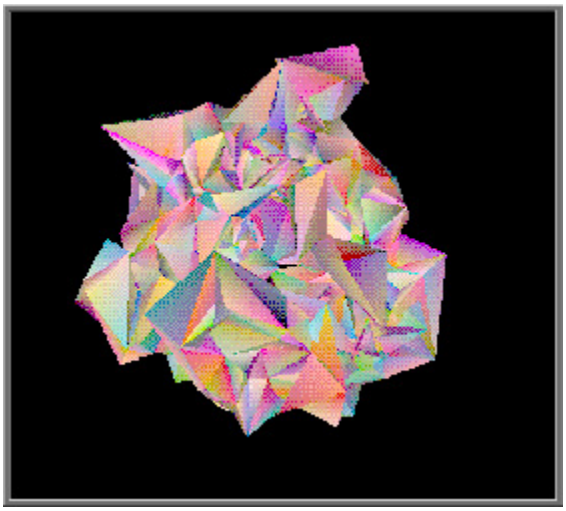


Figure 26: Causal Dynamical Triangulation builds the spacetime in which we live from tiny triangles.
Source: © Paul Coddington, University of Adelaide.

The most studied candidate for a theory of quantum gravity, string theory, posits that elementary particles are not points in spacetime but rather one-dimensional objects like open lengths or closed loops of string. Different modes of vibrations of the elementary strings give rise to the spectrum of particles in nature including the graviton, the particle that carries the gravitational force (analogous to the photon in electromagnetism). To provide a realistic theory of quantum gravity, string theories require extra spatial

dimensions, each normally viewed as being finite in extent, such as a one-dimensional circle with a radius of the Planck length or larger. The presence of extra dimensions and new particles associated with gravity in string theories alters the gravitational inverse square law and the equivalence principle at very short distances. We will learn more about string theory and extra dimensions in Unit 4.

The small length scales and equivalently high energy scales at which quantum effects should modify gravity are far beyond the reach of current experimental techniques. A major challenge to finding the correct theory of quantum gravity is that it will be difficult to find experimental evidence to point us in the right direction.

Gravity at large distances

We can also wonder how well we know the behavior of gravity at very large lengths scales. As we have seen, the inverse square law of gravity has been verified over solar system distances, but the observable universe is 100 billion times larger than that. It requires a leap of faith to believe that our local laws of gravity hold everywhere. Some of the evidence for dark matter relies upon comparing the observed acceleration of objects far apart to that expected from the inverse square law. If the law of universal gravity is invalid for very small accelerations, as proposed in the MOND (Modified Newtonian Dynamics) theory, then our expectations for the interactions of distant objects would change.

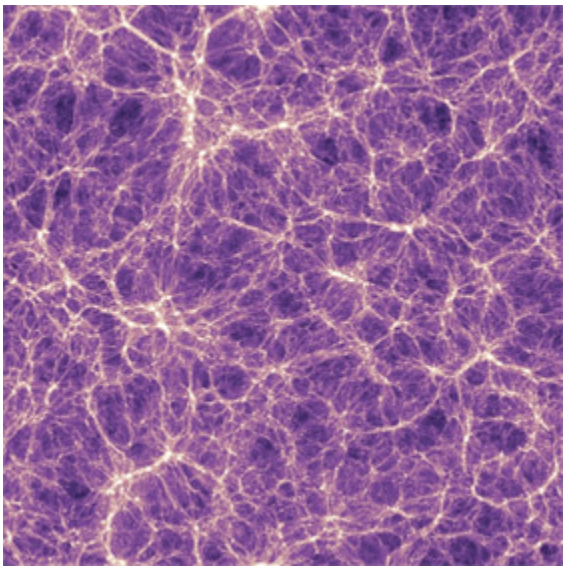


Figure 27: Simulations of structure formation in the universe show the influence of gravity and dark energy.
Source: © Raul Angulo, Max Planck Institute for Astrophysics.

Dark energy, described in detail in Unit 11, has been proposed to explain why the expansion rate of the universe appears to be accelerating. The evidence for dark energy rests upon the comparison of observations with the predictions of general relativity applied to very large length scales. Theorists continue to explore a variety of ways to modify general relativity to circumvent the need for dark energy. As there is no direct experimental evidence one way or another, the behavior of gravity and very large length scales is still an open question.

The first unification in physics was Newton's law of universal gravitation that provided a common explanation for the motion of terrestrial and heavenly objects. It is ironic that for modern attempts to unify all of the forces in nature, gravity is the last and most difficult force to include. The theory of general relativity was a triumph of 20th century physics that revolutionized our concepts of space and time. Yet, even general relativity is not likely to be the ultimate theory of gravity. There is still much to be learned about gravity.

Section 9: Further Reading

- Avery Broderick and Abraham Loeb, "Portrait of a Black Hole," *Scientific American*, December 2009, p. 42.
- George Gamov, "Gravity," Dover Publications, Inc., 2002.
- GRACE Mission website: http://www.csr.utexas.edu/grace/.
- Eduardo Gueron, "Adventures in Curved Spacetime," *Scientific American*, August 2009, p. 38.
- Pankaj S. Joshi, "Naked Singularities," *Scientific American*, February 2009, p. 36.
- Jerzy Jurkiewicz, Renate Loll, and Jan Ambjorn, "The Self-Organizing Quantum Universe," *Scientific American*, July 2008, p. 42.
- Laser Interferometer Gravitational Wave Observatory (LIGO) website: http://www.ligo.caltech.edu/.

Glossary

black hole: A black hole is a region of space where gravity is so strong that nothing can escape its pull. Black holes have been detected through their gravitational influence on nearby stars and through observations of hot gas from surrounding regions accelerating toward them. These black holes are thought to have formed when massive stars reached the end of their cycle of evolution and collapsed under the influence of gravity. If a small volume of space contains enough mass, general relativity predicts that spacetime will become so highly curved that a black hole will form.

cosmic microwave background: The cosmic microwave background (CMB) radiation is electromagnetic radiation left over from when atoms first formed in the early universe, according to our standard model of cosmology. Prior to that time, photons and the fundamental building blocks of matter formed a hot, dense soup, constantly interacting with one another. As the universe expanded and cooled, protons and neutrons formed atomic nuclei, which then combined with electrons to form neutral atoms. At this point, the photons effectively stopped interacting with them. These photons, which have stretched as the universe expanded, form the CMB. First observed by Penzias and Wilson in 1965, the CMB remains the focus of increasingly precise observations intended to provide insight into the composition and evolution of the universe.

Coulomb's Law: Coulomb's Law states that the electric force between two charged particles is proportional to the product of the two charges divided by the square of the distance between the particles.

Doppler shift (Doppler effect): The Doppler shift is a shift in the wavelength of light or sound that depends on the relative motion of the source and the observer. A familiar example of a Doppler shift is the apparent change in pitch of an ambulance siren as it passes a stationary observer. When the ambulance is moving toward the observer, the observer hears a higher pitch because the wavelength of the sound waves is shortened. As the ambulance moves away from the observer, the wavelength is lengthened and the observer hears a lower pitch. Likewise, the wavelength of light emitted by an object moving toward an observer is shortened, and the observer will see a shift to blue. If the light-emitting object is moving away from the observer, the light will have a longer wavelength and the observer will see a shift to red. By observing this shift to red or blue, astronomers can determine the velocity of distant stars and galaxies relative to the Earth. Atoms moving relative to a laser also experience a Doppler shift, which must be taken into account in atomic physics experiments that make use of laser cooling and trapping.

ether: In the late nineteenth century, physicists were putting what they thought were the finishing touches on their theoretical description of electricity and magnetism. In the theory, electromagnetic waves traveled through a medium called "luminiferous ether" just as sound waves travel through the air, or the seismic waves that we experience as earthquakes travel through the Earth. The last remaining detail was to detect the ether and understand its properties. In 1887, Albert Michelson and Edward Morley performed an experiment, verified by many others, that demonstrated that light does not travel through ether. The lack of ether was one of many factors leading Einstein to develop special relativity.

event horizon: A black hole's event horizon is the point of no return for matter falling toward the black hole. Once matter enters the event horizon, it is gravitationally bound to the black hole and cannot escape. However, an external observer will not see the matter enter the black hole. Instead, the gravitational redshift due to the black hole's strong gravitational field causes the object to appear to approach the horizon increasingly slowly without ever going beyond it. Within the event horizon, the black hole's gravitational field warps spacetime so much that even light cannot escape.

general relativity: General relativity is the theory Einstein developed to reconcile gravity with special relativity. While special relativity accurately describes the laws of physics in inertial reference frames, it does not describe what happens in an accelerated reference frame or gravitational field. Since acceleration and gravity are important parts of our physical world, Einstein recognized that special relativity was an incomplete description and spent the years between 1905 and 1915 developing general relativity. In general relativity, we inhabit a four-dimensional spacetime with a curvature determined by the distribution of matter and energy in space. General relativity makes unique, testable predictions that have been upheld by experimental measurements, including the precession of Mercury's orbit, gravitational lensing, and gravitational time dilation. Other predictions of general relativity, including gravitational waves, have not yet been verified. While there is no direct experimental evidence that conflicts with general relativity, the accepted view is that general relativity is an approximation to a more fundamental theory of gravity that will unify it with the Standard Model. See: gravitational lensing. gravitational time dilation, gravitational wave, precession, spacetime, special relativity, Standard Model.

gravitational lensing: Gravitational lensing occurs when light travels past a very massive object. According to Einstein's theory of general relativity, mass shapes spacetime and space is curved by massive objects. Light traveling past a massive object follows a "straight" path in the curved space, and is deflected as if it had passed through a lens. Strong gravitational lensing can cause stars to appear as rings as their light travels in a curved path past a massive object along the line of sight. We observe microlensing when an

object such as a MACHO moves between the Earth and a star. The gravitational lens associated with the MACHO focuses the star' light, so we observe the star grow brighter then dimmer as the MACHO moves across our line of sight to the star.

gravitational mass: The gravitational mass of a particle is the gravitational equivalent of electric charge: the physical property of an object that causes it to interact with other objects through the gravitational force. According to the equivalence principle, gravitational mass is equivalent to inertial mass. See: equivalence principle, inertial mass.

gravitational time dilation: Clocks in a strong gravitational field run slower than clocks in a weaker gravitational field. This effect, predicted by Einstein's theory of general relativity and confirmed by precision experiments both on Earth and in space, is called "gravitational time dilation."

Hertz: Hertz (Hz) is a unit of frequency, defined as the number of complete cycles of a periodic signal that take place in one second. For example, the frequency of sound waves is usually reported in units of Hertz. The normal range of human hearing is roughly 20–20,000 Hz. Radio waves have frequencies of thousands of Hz, and light waves in the visible part of the spectrum have frequencies of over 10^{14} Hz.

inertial mass: Inertia is the measure of an object's reluctance to accelerate under an applied force. The inertial mass of an object is the mass that appears in Newton's second law: the acceleration of an object is equal to the applied force divided by its inertial mass. The more inertial mass an object has, the less it accelerates under a fixed applied force. See: equivalence principle, gravitational mass.

MOND: MOND, or Modified Newtonian Dynamics, is a theory that attempts to explain the evidence for dark matter as a modification to Newtonian gravity. There are many versions of the theory, all based on the premise that Newton's laws are slightly different at very small accelerations. A ball dropped above the surface of the Earth would not deviate noticeably from the path predicted by Newtonian physics, but the stars at the very edges of our galaxy would clearly demonstrate modified dynamics if MOND were correct.

Newton's law of universal gravitation: Newton's law of universal gravitation states that the gravitational force between two massive particles is proportional to the product of the two masses divided by the square of the distance between them. The law of universal gravitation is sometimes called the "inverse square law." See: universal gravitational constant.

polarization: The polarization of a wave is the direction in which it is oscillating. The simplest type of polarization is linear, transverse polarization. Linear means that the wave oscillation is confined

along a single axis, and transverse means that the wave is oscillating in a direction perpendicular to its direction of travel. Laser light is most commonly a wave with linear, transverse polarization. If the laser beam travels along the x-axis, its electric field will oscillate either in the y-direction or in the z-direction. Gravitational waves also have transverse polarization, but have a more complicated oscillation pattern than laser light.

precession: Precession is a systematic change in the orientation of a rotation axis. For example, the orbits of planets in our solar system precess. Each planet follows an elliptical path around the Sun, with the Sun at one of the focal points of the ellipse. The long axis of the ellipse slowly rotates in the plane of the orbit with the Sun as a pivot point, so the planet never follows exactly the same path through space as it continues to orbit in its elliptical path. The precession measured in Mercury's orbit was found to be different from the prediction of Newtonian gravity but matched the prediction of general relativity, providing some of the first concrete evidence that Einstein's version of gravity is correct.

pulsar: A pulsar is a spinning neutron star with a strong magnetic field that emits electromagnetic radiation along its magnetic axis. Because the star's rotation axis is not aligned with its magnetic axis, we observe pulses of radiation as the star's magnetic axis passes through our line of sight. The time between pulses ranges from a few milliseconds to a few seconds, and tends to slow down over time.

spacetime: In classical physics, space and time are considered separate things. Space is three-dimensional, and can be divided into a three-dimensional grid of cubes that describes the Euclidean geometry familiar from high-school math class. Time is one-dimensional in classical physics. Einstein's theory of special relativity combines the three dimensions of space and one dimension of time into a four-dimensional grid called "spacetime." Spacetime may be flat, in which case Euclidean geometry describes the three space dimensions, or curved. In Einstein's theory of general relativity, the distribution of matter and energy in the universe determines the curvature of spacetime.

special relativity: Einstein developed his theory of special relativity in 1905, 10 years before general relativity. Special relativity is predicated on two postulates. First, the speed of light is assumed to be constant in all inertial frames. Second, the laws of physics are assumed to be the same in all inertial frames. An inertial frame, in this context, is defined as a reference frame that is not accelerating or in a gravitational field. Starting from these two postulates, Einstein derived a number of counterintuitive consequences that were later verified by experiment. Among them are time dilation (a moving clock will run slower than a stationary clock), length contraction (a moving ruler will be shorter than a stationary

ruler), the equivalence of mass and energy, and that nothing can move faster than the speed of light. See: general relativity, spacetime.

standard model of cosmology: Our best model for how the universe began and evolved into what we observe now is called the "standard model of cosmology." It contends that the universe began in a Big Bang around 14 billion years ago, which was followed by a short period of exponential inflation. At the end of inflation, quarks, photons, and other fundamental particles formed a hot, dense soup that cooled as the universe continued to expand. Roughly 390,000 years after the end of inflation, the first atoms formed and the cosmic microwave background photons decoupled. Over the course of billions of years, the large structures and astronomical objects we observe throughout the cosmos formed as the universe continued to expand. Eventually the expansion rate of the universe started to increase under the influence of dark energy.

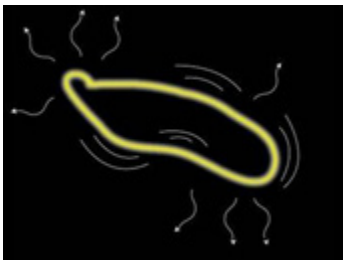
torsion pendulum: A conventional pendulum is a mass suspended on a string that swings periodically. A torsion pendulum is a mass suspended on a string (or torsion fiber) that rotates periodically. When the mass of a torsion pendulum is rotated from its equilibrium position, the fiber resists the rotation and provides a restoring force that causes the mass to rotate back to its original equilibrium position. When the mass reaches its equilibrium position, it is moving quickly and overshoots. The fiber's restoring force, which is proportional to the rotation angle of the mass, eventually causes the mass to slow down and rotate back the other way. Because the restoring force of the torsion fiber is very small, a torsion pendulum can be used to measure extremely small forces affecting the test mass.

universal gravitational constant: The universal gravitational constant, denoted by G, is the proportionality constant in Newton's law of universal gravitation. The currently accepted value for G is $6.67428 \pm 0.00067 \times 10^{-11}$ N-m^2/kg^2.

universality of free fall: The universality of free fall, sometimes abbreviated UFF, is the idea that all materials fall at the same rate in a uniform gravitational field. This is equivalent to stating that inertial and gravitational mass are the same. See: equivalence principle, gravitational mass, inertial mass.

Physics for the 21st Century

Unit 4: String Theory and Extra Dimensions



© Matt DePies, University of Washington.

Unit Overview

This unit continues our movement from experimentally proven understanding of nature's four fundamental forces to the theoretical effort to develop a "theory of everything" that brings all four forces under the same conceptual umbrella. The most prominent aspect of that effort is the family of string theories that envision the basic units of matter as minuscule stretches of threadlike strings rather than point particles. Introduced in the mid-1970s, the string concept has stimulated a great deal of theoretical excitement even though it has no connection to experiment so far. The unit introduces string theory in the context of quantum gravity and outlines its inherent multidimensional nature; the most promising approach involves a total of ten dimensions. The unit then covers the relationship of string theory to particle physics and introduces the idea of "branes," related to strings. Next, the unit focuses on cosmological issues arising from our understanding of the Big Bang, outlines the way in which the concept of rapid inflation very early in the universe can solve some major issues, and details links between string theory and cosmic inflation. Finally, the unit summarizes the understanding that string theory brings to fundamental understanding of gravity.

Content for This Unit

Sections:

Section 1: Introduction

The first two units of this course have introduced us to the four basic forces in nature and the efforts to unify them in a single theory. This quest has already been successful in the case of the electromagnetic and weak interactions, and we have promising hints of a further unification between the strong interactions and the electroweak theory (though this is far from experimentally tested). However, bringing gravity into this unified picture has proven far more challenging, and fundamental new theoretical issues come to the forefront. To reach the ultimate goal of a "theory of everything" that combines all four forces in a single theoretical framework, we first need a workable theory of quantum gravity.

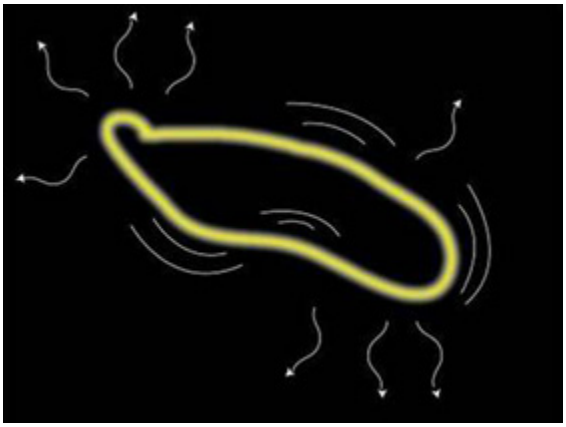


Figure 1: The fundamental units of matter may be minuscule bits of string.
Source: © Matt DePies, University of Washington.

As we saw in Units 2 and 3, theorists attempting to construct a quantum theory of gravity must somehow reconcile two fundamentals of physics that seem irreconcilable—Einstein's general theory of relativity and quantum mechanics. Since the 1920s, that effort has produced a growing number of approaches to understanding quantum gravity. The most prominent at present is string theory—or, to be accurate, an increasing accumulation of *string theories*. Deriving originally from studies of the strong nuclear force, the string concept asserts that the fundamental units of matter are not the traditional point-like particles but miniscule stretches of threadlike entities called "strings."

One of the most striking qualitative features of the string theories is that they predict the existence of extra spatial dimensions, with a total of 10 spacetime dimensions in the best-studied variants of the theory. This multitude of extra dimensions, beyond the familiar three of space plus one of time, suggests new approaches to account for some of the unsolved puzzles in the Standard Model of particle physics.

String theory also has the potential to provide insights into the ultimate puzzles of cosmology, such as the nature of the Big Bang and the origin of dark matter and dark energy that we will learn more about in Units 10 and 11.

However, candidates for an acceptable framework that combines gravity with the other three fundamental forces have one characteristic that separates them, and this unit, from all of the other topics covered in this course: Quantum gravity theories in general, and string theories in particular, have virtually no connection (as yet) to experimental evidence. There is, so far, no evidence that string theory is the correct modification of Einstein's theory, which would render it compatible with quantum mechanics in our world. String theories, or at least most models that follow from string theory, are only predictive at energy scales far from what can be probed with current particle physics and cosmological observations. This is not surprising; it follows from basic dimensional analysis, which we will describe in this unit, and which suggests that we will need a great deal of luck (or a very big accelerator) to directly test any approach to quantum gravity. Enthusiasm for string theory has been based, instead, on the theoretical richness and consistency of the structure it gives rise to, as well as the fruitful connections it has enjoyed with many other areas of physics. But one should keep in mind that its proponents will eventually need to make experimental predictions that can be tested to confirm or deny the validity of the approach as a literal description of quantum gravity in our universe.

In the following sections, we will see some current frontiers of physics where the quantum properties of gravity may be visible in near-term experiments studying the interactions of elementary particles at very high energy, or the physics of the very early universe. We hope, then, to gain one or more experimental windows (albeit very indirect ones) into the world of quantum gravity.

Section 2: The Origins of Quantum Gravity

In the early 20th century, physicists succeeded in explaining a wide range of phenomena on length scales ranging from the size of an atom (roughly 10^{-8} centimeters) to the size of the currently visible universe (roughly 10^{28} centimeters). They accomplished this by using two different frameworks for physical law: quantum mechanics and the general theory of relativity.

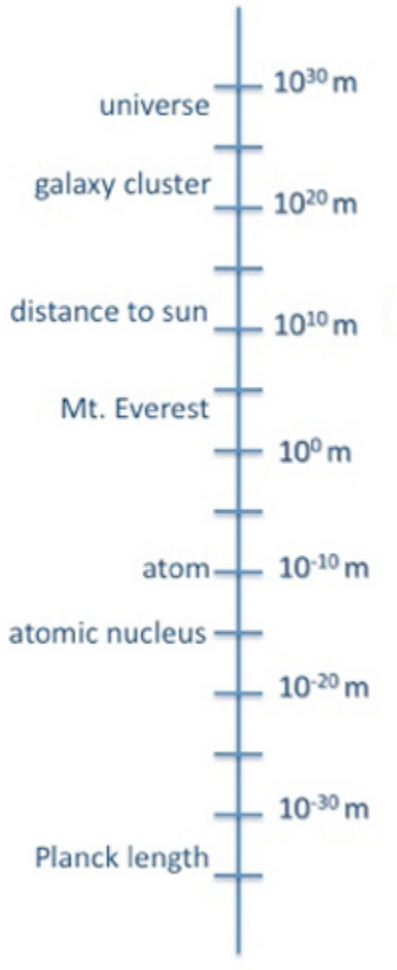


Figure 3: Quantum effects meet general relativity somewhere near the center of this length scale.
Source:

Built on Albert Einstein's use of Max Planck's postulate that light comes in discrete packets called "photons" to explain the photoelectric effect and Niels Bohr's application of similar quantum ideas to explain why atoms remain stable, quantum mechanics quickly gained a firm mathematical footing. ("Quickly" in this context, means over a period of 25 years). Its early successes dealt with systems in which a few elementary particles interacted with each other over short distance scales, of an order the size of an atom. The quantum rules were first developed to explain the mysterious behavior of matter at those distances. The end result of the quantum revolution was the realization that in the quantum world—as opposed to a classical world in which individual particles follow definite classical trajectories—positions, momenta, and other attributes of particles are controlled by a wave function that gives probabilities for different classical behaviors to occur. In daily life, the probabilities strongly reflect the classical behavior we intuitively expect; but at the tiny distances of atomic physics, the quantum rules can behave in surprising and counterintuitive ways. These are described in detail in Units 5 and 6.

In roughly the same time period, but for completely different reasons, an equally profound shift in our understanding of classical gravity occurred. One of the protagonists was again Einstein, who realized that Newton's theory of gravity was incompatible with his special theory of relativity. In Newton's theory, the attractive gravitational force between two bodies involves action at a distance. The two bodies attract each other instantaneously, without any time delay that depends on their distance from one another. The special theory of relativity, by contrast, would require a time lapse of at least the travel time of light between the two bodies. This and similar considerations led Einstein to unify Newtonian gravity with special relativity in his general relativity theory.

Einstein proposed his theory in 1915. Shortly afterward, in the late 1920s and early 1930s, theorists found that one of the simplest solutions of Einstein's theory, called the Friedmann-Lemaître-Robertson-Walker cosmology after the people who worked out the solution, can accommodate the basic cosmological data that characterize our visible universe. As we will see in Unit 11, this is an approximately flat or Euclidean geometry, with distant galaxies receding at a rate that gives us the expansion rate of the universe. This indicates that Einstein's theory seems to hold sway at distance scales of up to 10^{28} centimeters.

In the years following the discoveries of quantum mechanics and special relativity, theorists worked hard to put the laws of electrodynamics and other known forces (eventually including the strong and weak nuclear forces) into a fully quantum mechanical framework. The quantum field theory they developed describes, in quantum language, the interactions of fields that we learned about in Unit 2.

The theoretical problem of quantizing gravity

In a complete and coherent theory of physics, one would like to place gravity into a quantum framework. This is not motivated by practical necessity. After all, gravity is vastly weaker than the other forces when it acts between two elementary objects. It plays a significant role in our view of the macroscopic world only because all objects have positive mass, while most objects consist of both positive and negative electric charges, and so become electromagnetically neutral. Thus, the aggregate mass of a large body like the Earth becomes quite noticeable, while its electromagnetic field plays only a small role in everyday life.

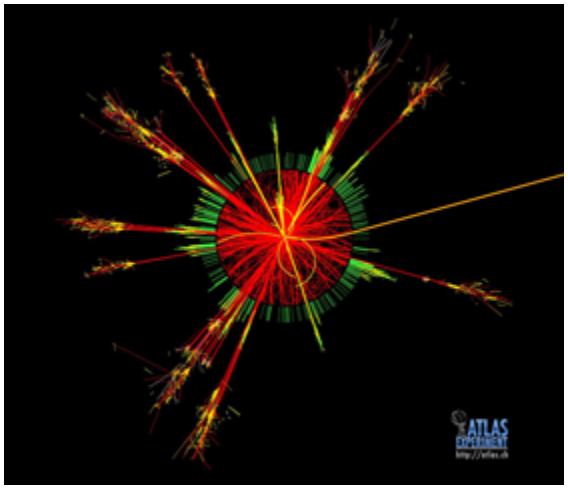


Figure 4: The rules of quantum gravity must predict the probability of different collision fragments forming at the LHC, such as the miniature black hole simulated here.
Source: © ATLAS Experiment, CERN.

But while the problem of quantizing gravity has no obvious practical application, it is inescapable at the theoretical level. When we smash two particles together at some particular energy in an accelerator like the Large Hadron Collider (LHC), we should at the very least expect our theory to give us quantum mechanical probabilities for the nature of the resulting collision fragments.

Gravity has a fundamental length scale—the unique quantity with dimensions of length that one can make out of Planck's constant, Newton's universal gravitational constant, G, and the speed of light, c. The Planck length is 1.61×10^{-35} meters, 10^{20} times smaller than the nucleus of an atom. A related constant is the Planck mass (which, of course, also determines an energy scale); is around 10^{-5} grams, which is equivalent to ~ 10^{19} giga-electron volts (GeV). These scales give an indication of when quantum gravity

is important, and how big the quanta of quantum gravity might be. They also illustrate how in particle physics, energy, mass, and 1/length are often considered interchangeable, since we can convert between these units by simply multiplying by the right combination of fundamental constants. See the math

Since gravity has a built-in energy scale, M_{Planck}, we can ask what happens as we approach the Planckian energy for scattering. Simple approaches to quantum gravity predict that the probability of any given outcome when two energetic particles collide with each other grows with the energy, E, of the collision at a rate controlled by the dimensionless ratio $(E/M_{Planck})^2$. This presents a serious problem: At some energy close to the Planck scale, one finds that the sum of the probabilities for final states of the collision is greater than 100%. This contradiction means that brute-force approaches to quantizing gravity are failing at sufficiently high energy.

We should emphasize that this is not yet an experimentally measured problem. The highest energy accelerator in the world today, the LHC, is designed to achieve center-of-mass collision energies of roughly 10 tera-electron volts (TeV)—15 orders of magnitude below the energies at which we strongly suspect that quantum gravity presents a problem.

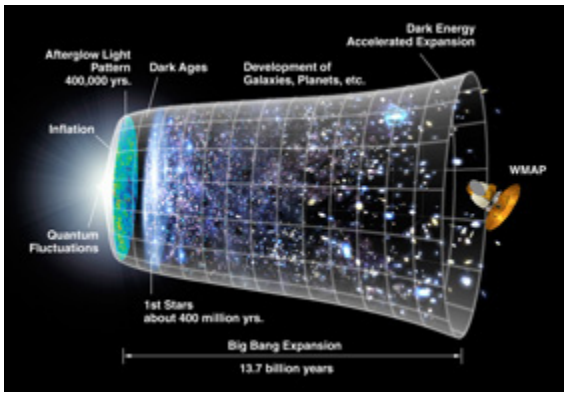


Figure 5: General relativity is consistent with all the cosmological data that characterizes our visible universe.
Source: © NASA, WMAP Science Team.

On the other hand, this does tell us that somewhere before we achieve collisions at this energy scale (or at distance scales comparable to 10^{-32} centimeters), the rules of gravitational physics will fundamentally change. And because gravity in Einstein's general relativity is a theory of spacetime geometry, this also implies that our notions of classical geometry will undergo some fundamental shift. Such a shift in our understanding of spacetime geometry could help us resolve puzzles associated with early universe

cosmology, such as the initial cosmological singularity that precedes the Big Bang in all cosmological solutions of general relativity.

These exciting prospects of novel gravitational phenomena have generated a great deal of activity among theoretical physicists, who have searched long and hard for consistent modifications of Einstein's theory that avoid the catastrophic problems in high-energy scattering and that yield new geometrical principles at sufficiently short distances. As we will see, the best ideas about gravity at short distances also offer tantalizing hints about structures that may underlie the modern theory of elementary particles and Big Bang cosmology.

Section 3: String Theory

Almost by accident in the mid 1970s, theorists realized that they could obtain a quantum gravity theory by postulating that the fundamental building blocks of nature are not point particles, a traditional notion that goes back at least as far as the ancient Greeks, but instead are tiny strands of string. These strings are not simply a smaller version of, say, our shoelaces. Rather, they are geometrical objects that represent a fundamentally different way of thinking about matter. This family of theories grew out of the physics of the strong interactions. In these theories, two quarks interacting strongly are connected by a stream of carriers of the strong force, which forms a "flux tube." The potential energy between the two quarks, therefore, grows linearly with the distance between the quarks. See the math

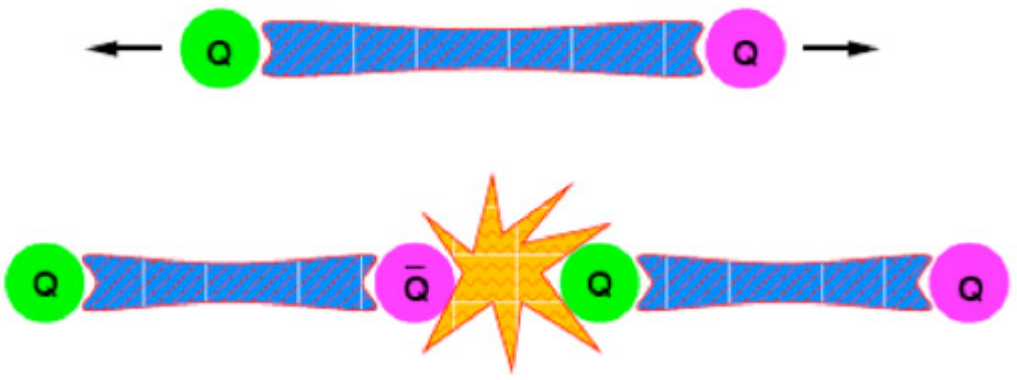


Figure 6: As quarks are pulled apart, eventually new quarks appear.
Source: © David Kaplan.

We choose to call the proportionality constant that turns the distance between the strongly interacting particles into a quantity with the units of energy "T_{string}" because it has the dimensions of mass per unit length as one would expect for a string's tension. In fact, one can think of the object formed by the flux tube of strong force carriers being exchanged between the two quarks as being an effective string, with tension T_{string}.

One of the mysteries of strong interactions is that the basic charged objects—the quarks—are never seen in isolation. The string picture explains this confinement: If one tries to pull the quarks farther and farther apart, the growing energy of the flux tube eventually favors the creation of another quark/anti-quark pair in the middle of the existing quark pair; the string breaks, and is replaced by two new flux tubes connecting the two new pairs of quarks. For this reason and others, string descriptions of the strong

interactions became popular in the late 1960s. Eventually, as we saw in Unit 2, quantum chromodynamics (QCD) emerged as a more complete description of the strong force. However, along the way, physicists discovered some fascinating aspects of the theories obtained by treating the strings not as effective tubes of flux, but as fundamental quantum objects in their own right.

Perhaps the most striking observation was the fact that any theory in which the basic objects are strings will inevitably contain a particle with all the right properties to serve as a graviton, the basic force carrier of the gravitational force. While this is an unwanted nuisance in an attempt to describe strong interaction physics, it is a compelling hint that quantum string theories may be related to quantum gravity.

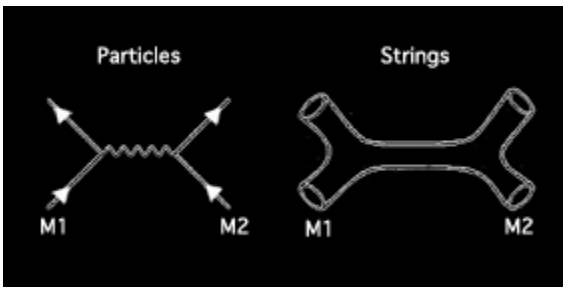


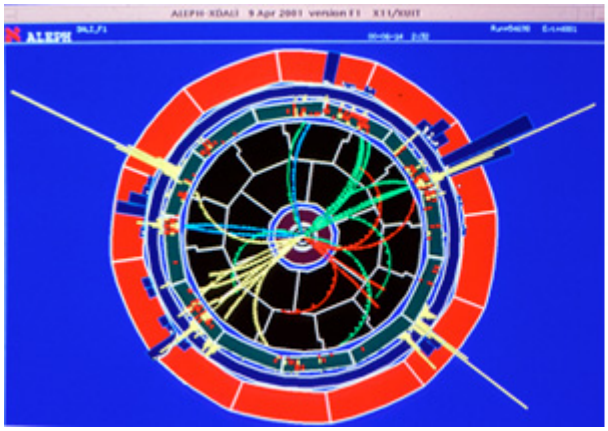
Figure 7: Gravitons arise naturally in string theory, leading to Feynman diagrams like the one on the right.
Source:

In 1984, Michael Green of Queen Mary, University of London and John Schwarz of the California Institute of Technology discovered the first fully consistent quantum string theories that were both free of catastrophic instabilities of the vacuum and capable in principle of incorporating the known fundamental forces. These theories automatically produce quantum gravity and force carriers for interactions that are qualitatively (and in some special cases even quantitatively) similar to the forces like electromagnetism and the strong and weak nuclear forces. However, this line of research had one unexpected consequence: These theories are most naturally formulated in a 10-dimensional spacetime.

We will come back to the challenges and opportunities offered by a theory of extra spacetime dimensions in later sections. For now, however, let us examine how and why a theory based on strings instead of point particles can help with the problems of quantum gravity. We will start by explaining how strings resolve the problems of Einstein's theory with high-energy scattering. In the next section, we discuss how strings modify our notions of geometry at short distances.

Strings at high energy

In the previous section, we learned that for particles colliding at very high energies, the sum of the probabilities for all the possible outcomes of the collision calculated using the techniques of Unit 2 is greater than 100 percent, which is clearly impossible. Remarkably, introducing an extended object whose fundamental length scale is not so different from the Planck length, $\ell_{string} \sim 10^{-32}$ centimeters, seems to solve this basic problem in quantum gravity. The essential point is that in high-energy scattering processes, the size of a string grows with its energy.



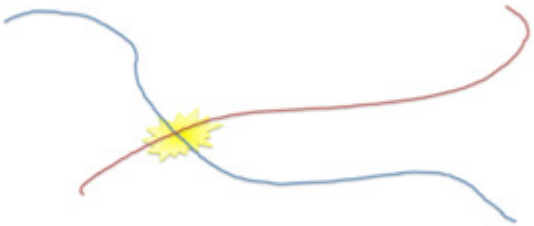


Figure 8: String collisions are softer than particle collisions.
Source: © Left: CERN, Right: Claire Cramer.

This growth-with-energy of an excited string state has an obvious consequence: When two highly energetic strings interact, they are both in the form of highly extended objects. Any typical collision involves some small segment of one of the strings exchanging a tiny fraction of its total energy with a small segment of the other string. This considerably softens the interaction compared with what would happen if two bullets carrying the same energy undergo a direct collision. In fact, it is enough to make the scattering probabilities consistent with the conservation of probability. In principle, therefore, string theories can give rise to quantum mechanically consistent scattering, even at very high energies.

Section 4: Strings and Extra Dimensions

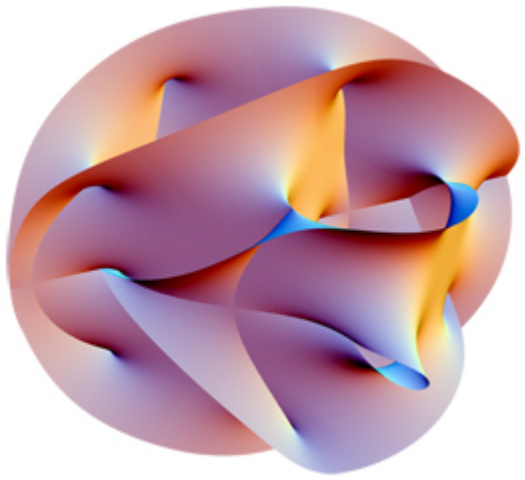


Figure 9: A depiction of one of the six-dimensional spaces that seem promising for string compactification.
Source: © Wikimedia Commons, CC Attribution-Share Alike 2.5 Generic license. Author: Lunch, 23 September 2007.

We have already mentioned that string theories that correspond to quantum gravity together with the three other known fundamental forces seem to require 10 spacetime dimensions. While this may come as a bit of a shock—after all, we certainly seem to live in four spacetime dimensions—it does not immediately contradict the ability of string theory to describe our universe. The reason is that what we call a "physical theory" is a set of equations that is dictated by the fundamental fields and their interactions. Most physical theories have a unique basic set of fields and interactions, but the equations may have many different solutions. For instance, Einstein's theory of general relativity has many nonphysical solutions in addition to the cosmological solutions that look like our own universe. We know that there are solutions of string theory in which the 10 dimensions take the form of four macroscopic spacetime dimensions and six dimensions curled up in such a way as to be almost invisible. The hope is that one of these is relevant to physics in our world.

To begin to understand the physical consequences of tiny, curled-up extra dimensions, let us consider the simplest relevant example. The simplest possibility is to consider strings propagating in nine-dimensional flat spacetime, with the 10th dimension curled up on a circle of size R. This is clearly not a realistic theory

of quantum gravity, but it offers us a tantalizing glimpse into one of the great theoretical questions about gravity: How will a consistent theory of quantum gravity alter our notions of spacetime geometry at short distances? In string theory, the concept of curling up, or compactification, on a circle, is already startlingly different from what it would be in point particle theory.

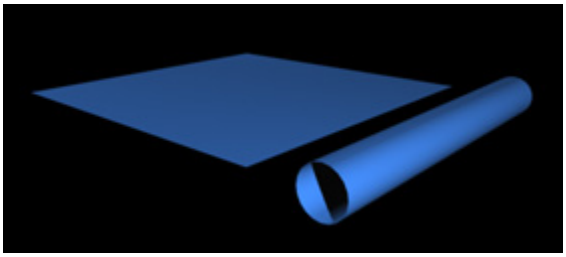


Figure 10: String theorists generally believe that extra dimensions are compactified, or curled up.
Source:

To compare string theory with normal particle theories, we will compute the simplest physical observable in each kind of theory, when it is compactified on a circle from ten to nine dimensions. This simplest observable is just the masses of elementary particles in the lower-dimensional space. It will turn out that a single type of particle (or string) in 10 dimensions gives rise to a whole infinite tower of particles in nine dimensions. But the infinite towers in the string and particle cases have an important difference that highlights the way that strings "see" a different geometry than point particles.

Particles in a curled-up dimension

Let us start by explaining how an infinite tower of nine-dimensional (9D) particles arises in the 10-dimensional (10D) particle theory. To a 9D observer, the velocity and momentum of a given particle in the hidden tenth dimension, which is too small to observe, are invisible. But the motion is real, and a particle moving in the tenth dimension has a nonzero energy. Since the particle is not moving around in the visible dimensions, one cannot attribute its energy to energy of motion, so the 9D observer attributes this energy to the particle's mass. Therefore, for a given particle species in the fundamental 10D theory, each type of motion it is allowed to perform along the extra circle gives rise to a new elementary particle from the 9D perspective.

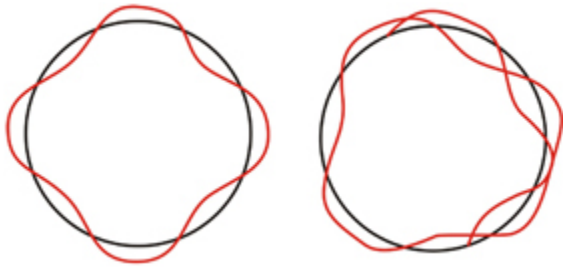


Figure 11: If a particle is constrained to move on a circle, its wave must resemble the left rather than the right drawing.
Source: © CK-12 Foundation.

To understand precisely what elementary particles the 9D observer sees, we need to understand how the 10D particle is allowed to move on the circle. It turns out that this is quite simple. In quantum mechanics, as we will see in Units 5 and 6, the mathematical description of a particle is a "probability wave" that gives the likelihood of the particle being found at any position in space. The particle's energy is related to the frequency of the wave: a higher frequency wave corresponds to a particle with higher energy. When the particle motion is confined to a circle, as it is for our particle moving in the compactified tenth dimension, the particle's probability wave needs to oscillate some definite number of times (0, 1, 2 ...) as one goes around the circle and comes back to the same point. Each possible number of oscillations on the circle corresponds to a distinct value of energy that the 10D particle can have, and each distinct value of energy will look like a new particle with a different mass to the 9D observer. The masses of these particles are related to the size of the circle, and the number of wave oscillations around the circle:

$$m_0 = 0, m_1 = 1/R, m_2 = 2/R \ldots$$

So, as promised, the hidden velocity in the tenth dimension gives rise to a whole tower of particles in nine dimensions.

Strings in a curled-up dimension

Now, let us consider a string theory compactified on the same circle as above. For all intents and purposes, if the string itself is not oscillating, it is just like the 10D particle we discussed above. The 9D experimentalist will see the single string give rise to an infinite tower of 9D particles with distinct masses. But that's not the end of the story. We can also wind the string around the circular tenth dimension. To visualize this, imagine winding a rubber band around the thin part of a doorknob, which is also a circle. If

the string has a tension $T_{string} = 1/\alpha'$, (the conventional notation for the string tension), then winding the string once, twice, three times ... around a circle of size R, costs an energy:

$$m_1 = R/\alpha', m_2 = 2R/\alpha', m_3 = 3R/\alpha' \ldots$$

This is because the tension is defined as the mass per unit length of the string; and if we wind the string n times around the circle, it has a length which is n times the circumference of the circle. Just as a 9D experimentalist cannot see momentum in the 10th dimension, she also cannot see this string's winding number. Instead, she sees each of the winding states above as new elementary particles in the 9D world, with discrete masses that depend on the size of the compactified dimension and the string tension.

Geometry at short distances

One of the problems of quantum gravity raised in Section 2 is that we expect geometry at short distances to be different somehow. After working out what particles our 9D observer would expect to see, we are finally in a position to understand how geometry at short distances is different in a string theory.

The string tension, $1/\alpha'$, is related to the length of the string, ℓ_{string}, via $\alpha' = \ell_{string}^2$. Strings are expected to be tiny, with $\ell_{string} \sim 10^{-32}$ centimeter, so the string tension is very high. If the circle is of moderate to macroscopic size, the winding mode particles are incredibly massive since their mass is proportional to the size of the circle multiplied by the string tension. In this case, the 9D elementary particle masses in the string theory look much like that in the point particle theory on a circle of the same size, because such incredibly massive particles are difficult to see in experiments.

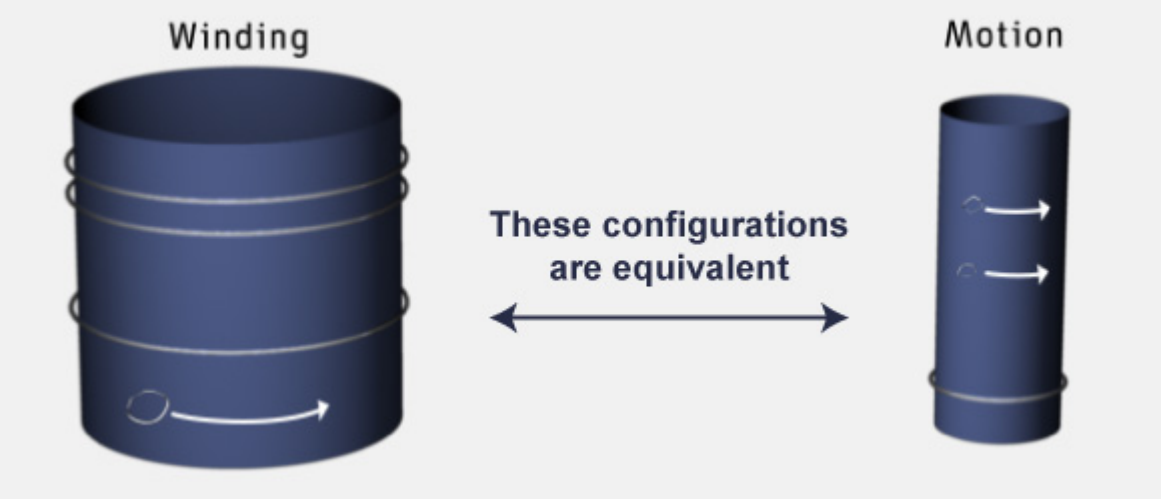


Figure 12: The consequences of strings winding around a larger extra dimension are the same as strings moving around a smaller extra dimension.
Source:

However, let us now imagine shrinking R until it approaches the scale of string theory or quantum gravity, and becomes less than ℓ_{string}. Then, the pictures one sees in point particle theory, and in string theory, are completely different. When R is smaller than ℓ_{string}, the modes m_1, m_2 ... are becoming lighter and lighter. And at very small radii, they are low-energy excitations that one would see in experiments as light 9D particles.

In the string theory with a small, compactified dimension, then, there are two ways that a string can give rise to a tower of 9D particles: motion around the circle, as in the particle theory, and winding around the circle, which is unique to the string theory. We learn something very interesting about geometry in string theory when we compare the masses of particles in these two towers.

For example, in the "motion" tower, $m_1 = 1/R$; and in the "winding" tower, $m_1 = R/\alpha'$. If we had a circle of size α'/R instead of size R, we'd get exactly the same particles, with the roles of the momentum-carrying strings and the wound strings interchanged. Up to this interchange, strings on a very large space are identical (in terms of these light particles, at least) to strings on a very small space. This large/small equivalence extends beyond the simple considerations we have described here. Indeed, the full string theory on a circle of radius R is completely equivalent to the full string theory on a circle of radius α'/R. This is a very simple illustration of what is sometimes called "quantum geometry" in string theory; string theories see geometric spaces of small size in a very different way than particle theories do. This

is clearly an exciting realization, because many of the mysteries of quantum gravity involve spacetime at short distances and high energies.

Section 5: Extra Dimensions and Particle Physics

The Standard Model of particle physics described in Units 1 and 2 is very successful, but leaves a set of lingering questions. The list of forces, for instance, seems somewhat arbitrary: Why do we have gravity, electromagnetism, and the two nuclear forces instead of some other cocktail of forces? Could they all be different aspects of a single unified force that emerges at higher energy or shorter distance scales? And why do three copies of each of the types of matter particles exist—not just an electron but also a muon and a tau? Not just an up quark, but also a charm quark and a top quark? And how do we derive the charges and masses of this whole zoo of particles? We don't know the answers yet, but one promising and wide class of theories posits that some or all of these mysteries are tied to the geometry or topology of extra spatial dimensions.

Perhaps the first attempt to explain properties of the fundamental interactions through extra dimensions was that of Theodor Kaluza and Oskar Klein. In 1926, soon after Einstein proposed his theory of general relativity, they realized that a unified theory of gravity and electromagnetism could exist in a world with 4+1 spacetime dimensions. The fifth dimension could be curled up on a circle of radius R so small that nobody had observed it.

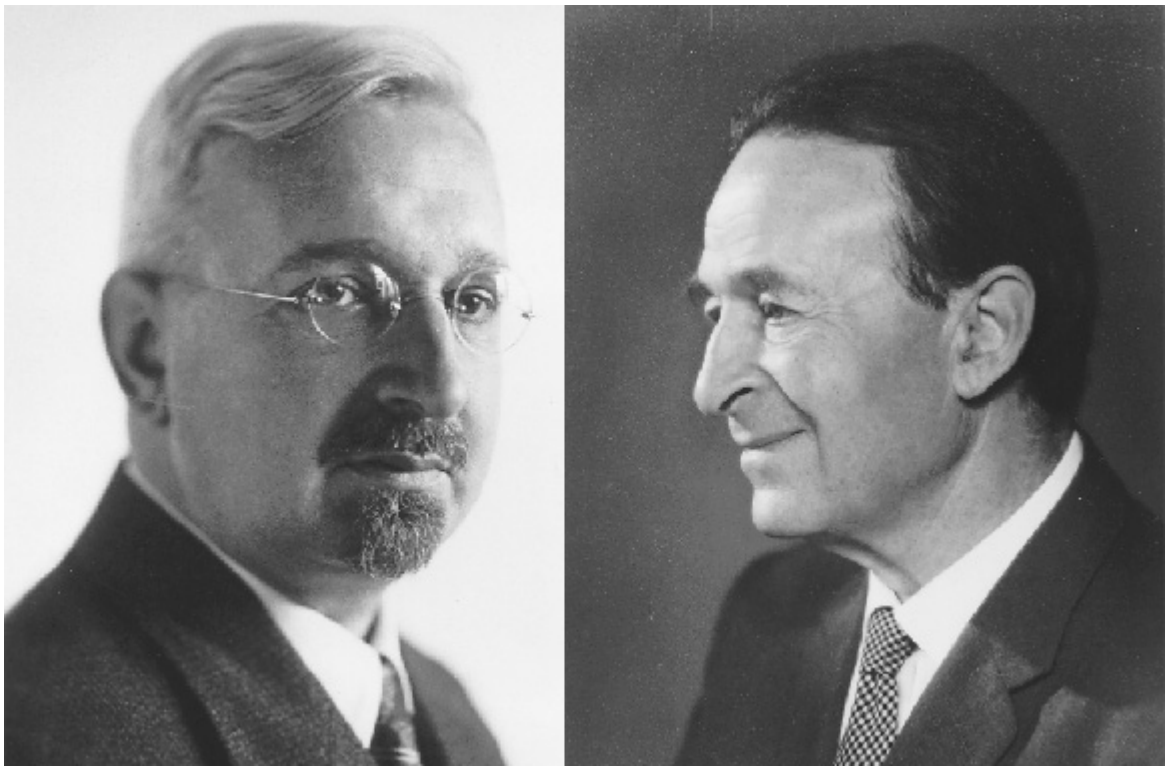


Figure 13: Theodor Kaluza (left) and Oskar Klein (right) made a remarkable theoretical description of gravity in a fifth dimension.
Source: © Left: University of Göttingen, Right: Stanley Deser.

In the 5D world, there are only gravitons, the force carriers of the gravitational field. But, as we saw in the previous section, a single kind of particle in higher dimensions can give rise to many in the lower dimension. It turns out that the 5D graviton would give rise, after reduction to 4D on a circle, to a particle with very similar properties to the photon, in addition to a candidate 4D graviton. There would also be a whole tower of other particles, as in the previous section, but they would be quite massive if the circle is small, and can be ignored as particles that would not yet have been discovered by experiment.

This is a wonderful idea. However, as a unified theory, it is a failure. In addition to the photon, it predicts additional particles that have no counterpart in the known fundamental interactions. It also fails to account for the strong and weak nuclear forces, discovered well after Kaluza and Klein published their papers. Nevertheless, modern generalizations of this basic paradigm, with a few twists, can both account for the full set of fundamental interactions and give enormous masses to the unwanted additional particles, explaining their absence in low-energy experiments.

Particle generations and topology

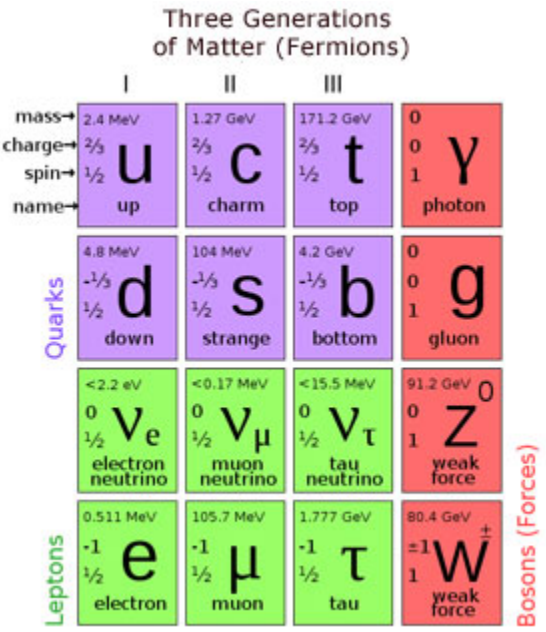


Figure 14: The Standard Model of particle physics. **Source:** © Wikimedia Commons, Creative Commons 3.0 Unported License. Author: MissMJ, 27 June 2006.

One of the most obvious hints of substructure in the Standard Model is the presence of three generations of particles with the same quantum numbers under all the basic interactions. This is what gives the Standard Model the periodic table-like structure we saw in Unit 1. This kind of structure sometimes has a satisfying and elegant derivation in models based on extra dimensions coming from the geometry or topology of space itself. For instance, in string theories, the basic elementary particles arise as the lowest energy states, or ground states, of the fundamental string. The different possible string ground states, when six of the 10 dimensions are compactified, can be classified by their topology.

Because it is difficult for us to imagine six dimensions, we'll think about a simpler example: two extra dimensions compactified on a two-dimensional surface. Mathematicians classified the possible topologies of such compact, smooth two-dimensional surfaces in the 19th century. The only possibilities are so-called "Riemann surfaces of genus g," labeled by a single integer that counts the number of "holes" in the surface. Thus, a beach ball has a surface of genus 0; a donut's surface has genus 1, as does a coffee mug's; and one can obtain genus g surfaces by smoothly gluing together the surfaces of g donuts.

Figure 15: These objects are Riemann surfaces with genus 0, 1, and 2.
Source: © Left: Wikimedia Commons, Public Domain, Author: Norvy, 27 July 2006; Center: Wikimedia Commons, Public Domain, Author: Tijuana Brass, 14 December 2007; Right: Wikimedia Commons, Public Domain. Author, NickGorton, 22 August 2005.

To understand how topology is related to the classification of particles, let's consider a toy model as we did in the previous section. Let's think about a 6D string theory, in which two of the dimensions are compactified. To understand what particles a 4D observer will see, we can think about how to wind strings around the compactified extra dimensions. The answer depends on the topology of the two-dimensional surface. For instance, if it is a torus, we can wrap a string around the circular cross-section of the donut. We could also wind the string through the donut hole. In fact, arbitrary combinations of wrapping the hole N1 times and the cross-section N2 times live in distinct topological classes. Thus, in string theory on the torus, one obtains two basic stable "winding modes" that derive from wrapping the string in those two ways. These will give us two distinct classes of particles.

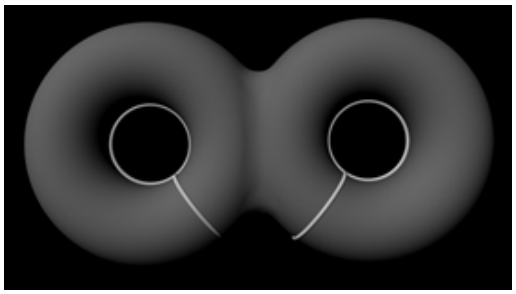


Figure 16: Strings can wind around a double torus in many distinct ways.
Source:

Similarly, a Riemann surface of genus g would permit 2g different basic stable string states. In this way, one could explain the replication of states of one type—by, say, having all strings that wrap a circular cross-section in any of the g different handles share the same physical properties. Then, the replication

of generations could be tied in a fundamental way to the topology of spacetime; there would, for example, be three such states in a genus 3 surface, mirroring the reality of the Standard Model.

Semi-realistic models of particle physics actually exist that derive the number of generations from specfic string compactifications on six-dimensional manifolds in a way that is very similar to our toy discussion in spirit. The mathematical details of real constructions are often considerably more involved. However, the basic theme—that one may explain some of the parameters of particle theory through topology—is certainly shared.

Section 6: Extra Dimensions and the Hierarchy Problem

Figure 17: The weakness of gravity is difficult to maintain in a quantum mechanical theory, much as it is difficult to balance a pencil on its tip.
Source:

At least on macroscopic scales, we are already familiar with the fact that gravity, is 10^{40} times weaker than electromagnetism. We can trace the weakness of gravity to the large value of the Planck mass, or the smallness of Newton's universal gravitational constant relative to the characteristic strength of weak interactions, which set the energy scale of modern-day particle physics. However, this is a description of the situation, rather than an explanation of why gravity is so weak.

This disparity of the scales of particle physics and gravity is known as the hierarchy problem. One of the main challenges in theoretical physics is to explain why the hierarchy problem is there, and how it is quantum mechanically stable. Experiments at the LHC should provide some important clues in this regard. On the theory side, extra dimensions may prove useful.

A speculative example

We'll start by describing a speculative way in which we could obtain the vast ratio of scales encompassed by the hierarchy problem in the context of extra dimensional theories. We describe this here not so much because it is thought of as a likely way in which the world works, but more because it is an extreme

illustration of what is possible in theories with extra spatial dimensions. Let us imagine, as in string theory, that there are several extra dimensions. How large should these dimensions be?

Figure 18: The gravitational field lines spread out radially from a massive particle.
Source:

First, let us think a bit about a simple explanation for Newton's law of gravitational attraction. A point mass in three spatial dimensions gives rise to a spherically symmetrical gravitational field: Lines of gravitational force emanate from the mass and spread out radially in all directions. At a given distance r from the mass, the area that these lines cross is the surface of a sphere of radius r, which grows like r^2. Therefore, the density of field lines of the gravitational field, and hence the strength of the gravitational attraction, falls like $1/r^2$. This is the inverse square law from Unit 3.

Now, imagine there were k extra dimensions, each of size L. At a distance from the point mass that is small compared to L, the field lines of gravitation would still spread out as if they are in 3+k dimensional flat space. At a distance r, the field lines would cross the surface of a hypersphere of radius r, which grows like r^{2+k}. Therefore the density of field lines and the strength of the field fall like $1/r^{2+k}$ —more quickly than in three-dimensional space. However, at a distance large compared to L, the compact dimensions don't matter—one can't get a large distance by moving in a very small dimension—and the field lines again fall off in density like $1/r^2$. The extra-fast fall-off of the density of field lines between

distance of order, the Planck length, and L has an important implication. The strength of gravity is diluted by this extra space that the field lines must thread.

An only slightly more sophisticated version of the argument above shows that with k extra dimensions of size L, one has a 3+1 dimensional Newton's constant that scales like L^{-k}. This means that gravity could be as strong as other forces with which we are familiar in the underlying higher-dimensional theory of the world, if the extra dimensions that we haven't seen yet are large (in Planck units, of course; not in units of meters). Then, the relative weakness of gravity in the everyday world would be explained simply by the fact that gravity's strength is diluted by the large volume of the extra dimensions, where it is also forced to spread.

String theory and brane power

The astute reader may have noticed a problem with the above explanation for the weakness of the gravitational force. Suppose *all* the known forces really live in a 4+k dimensional spacetime rather than the four observed dimensions. Then the field lines of other interactions, like electromagnetism, will be diluted just like gravity, and the observed disparity between the strength of gravity and electromagnetism in 4D will simply translate into such a disparity in 4+k dimensions. Thus, we need to explain why gravity is different.

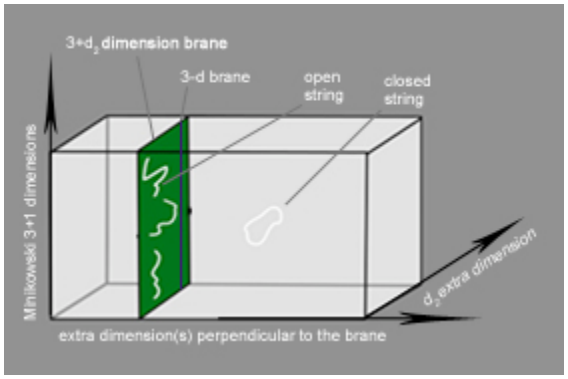


Figure 19: Strings can break open and end on a brane.
Source:

In string theories, a very elegant mechanism can confine all the interactions except gravity, which is universal and is tied directly to the geometry of spacetime, to just our four dimensions. This is because string theories have not only strings, but also branes. Derived from the term "membranes," these act like

dimensions on steroids. A p-brane is a p-dimensional surface that exists for all times. Thus, a string is a kind of 1-brane; for a 2-brane, you can imagine a sheet of paper extending in the x and y directions of space, and so on. In string theory, p-branes exist for various values of p as solutions of the 10D equations of motion.

So far, we have pictured strings as closed loops. However, strings can break open and end on a p-brane. The open strings that end in this manner give rise to a set of particles which live just on that p-brane. These particles are called "open string modes," and correspond to the lowest energy excitations of the open string. In common models, this set of open string modes includes analogs of the photon. So, it is easy to get toy models of the electromagnetic force, and even the weak and strong forces, confined to a 3-brane or a higher dimensional p-brane in 10D spacetime.

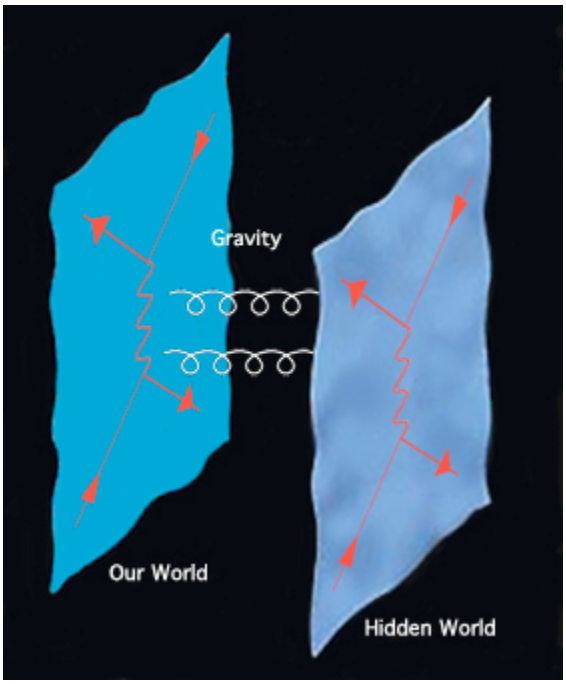


Figure 20: Gravity leaking into extra dimensions could explain the hierarchy problem.
Source:

In a scenario that contains a large number of extra dimensions but confines the fundamental forces other than gravity on a 3-brane, only the strength of gravity is diluted by the other dimensions. In this case, the weakness of gravity could literally be due to the large unobserved volume in extra spacetime dimensions. Then the problem we envisioned at the end of the previous section would not occur: Gravitational

field lines would dilute in the extra dimensions (thereby weakening our perception of gravity), while electromagnetic field lines would not.

While most semi-realistic models of particle physics derived from string theory work in an opposite limit, with the size of the extra dimensions close to the Planck scale and the natural string length scale around 10^{-32} centermeters, it is worth keeping these more extreme possibilities in mind. In any case, they serve as an illustration of how one can derive hierarchies in the strengths of interactions from the geometry of extra dimensions. Indeed, examples with milder consequences abound as explanations of some of the other mysterious ratios in Standard Model couplings.

Section 7: The Cosmic Serpent

Throughout this unit, we have moved back and forth between two distinct aspects of physics: the very small (particle physics at the shortest distance scales) and the very large (cosmology at the largest observed distances in the universe). One of the peculiar facts about any modification of our theories of particle physics and gravity is that, although we have motivated them by thinking of short-distance physics or high-energy localized scattering, any change in short-distance physics also tends to produce profound changes in our picture of cosmology at the largest distance scales. We call this relationship between the very small and the very large the "cosmic serpent."

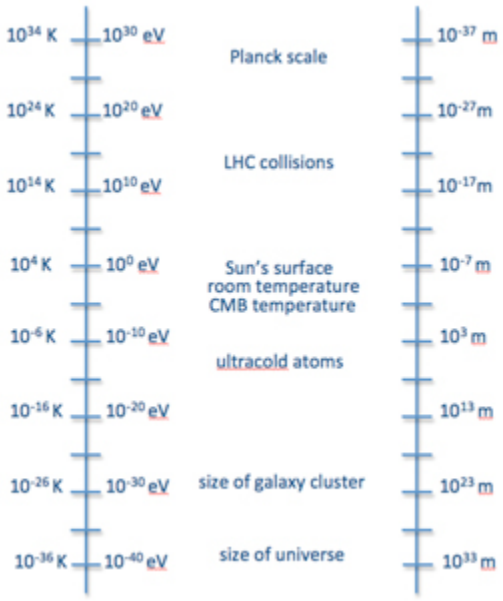


Figure 21: Changes in short distance physics—at the Planck scale—can produce profound changes in cosmology, at the largest imaginable distances.
Source:

This connection has several different aspects. The most straightforward stems from the Big Bang about 13.7 billion years ago, which created a hot, dense gas of elementary particles, brought into equilibrium with each other by the fundamental interactions, at a temperature that was very likely in excess of the TeV scale (and in most theories, at far higher temperatures). In other words, in the earliest phases of cosmology, nature provided us with the most powerful accelerator yet known that attained energies and densities unheard of in terrestrial experiments. Thus, ascertaining the nature of, and decoding the detailed physics of, the Big Bang, is an exciting task for both particle physicists and cosmologists.

The cosmic microwave background

One very direct test of the Big Bang picture that yields a great deal of information about the early universe is the detection of the relic gas of radiation called the cosmic microwave background, or CMB. As the universe cooled after the Big Bang, protons and neutrons bound together to form atomic nuclei in a process called Big Bang nucleosynthesis, then electrons attached to the nuclei to form atoms in a process called recombination. At this time, roughly 390,000 years after the Big Bang, the universe by and large became transparent to photons. Since the charged protons and electrons were suddenly bound in electrically neutral atoms, photons no longer had charged particles to scatter them from their path of motion. Therefore, any photons around at that time freely streamed along their paths from then until today, when we see them as the "surface of last scattering" in our cosmological experiments.

Bell Labs scientists Arno Penzias and Robert Wilson first detected the cosmic microwave background in 1964. Subsequent studies have shown that the detailed thermal properties of the gas of photons are largely consistent with those of a blackbody at a temperature of 2.7 degrees Kelvin, as we will see in Unit 5. The temperature of the CMB has been measured to be quite uniform across the entire sky—wherever you look, the CMB temperature will not vary more than 0.0004 K.

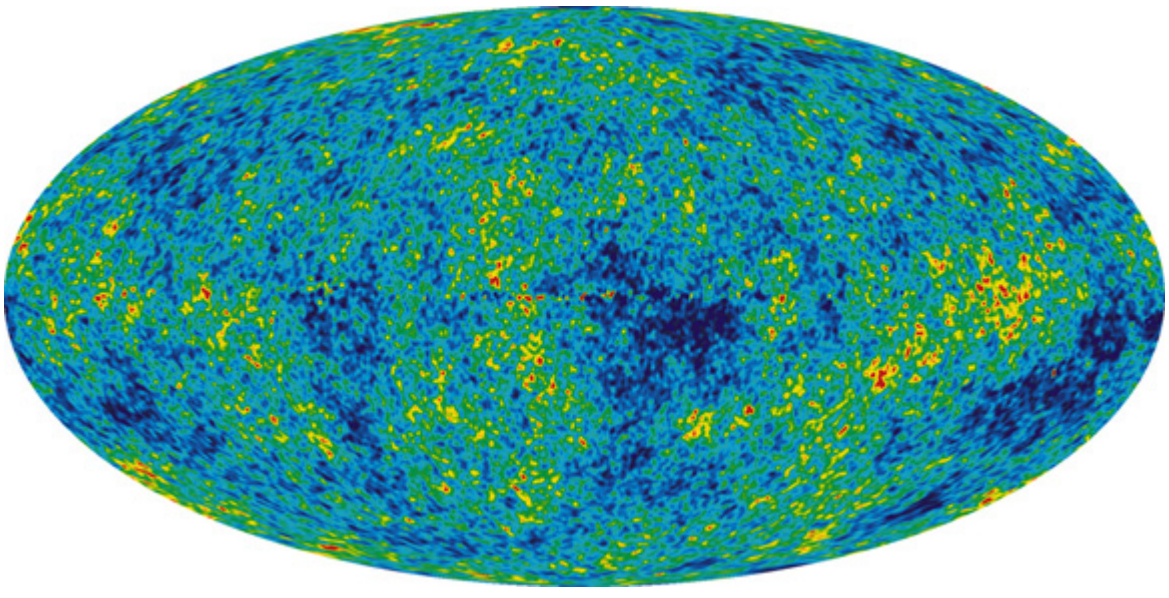


Figure 22: Map of the temperature variations in the cosmic microwave background measured by the WMAP satellite.
Source: © NASA, WMAP Science Team.

So, in the cosmic connection between particle physics and cosmology, assumptions about the temperature and interactions of the components of nuclei or atoms translate directly into epochs in cosmic history like nucleosynthesis or recombination, which experimentalists can then test indirectly or probe directly. This promising approach to testing fundamental theory via cosmological observations continues today, with dark matter, dark energy, and the nature of cosmic inflation as its main targets. We will learn more about dark matter in Unit 10 and dark energy in Unit 11. Here, we will attempt to understand inflation.

Cosmic inflation

Let us return to a puzzle that may have occurred to you in the previous section, when we discussed the gas of photons that started to stream through the universe 390,000 years after the Big Bang. Look up in the sky where you are sitting. Now, imagine your counterpart on the opposite side of the Earth doing the same. The microwave photons impinging on your eye and hers have only just reached Earth after their long journey from the surface of last scattering. And yet, the energy distribution (and hence the temperature) of the photons that you see precisely matches what she discovers.

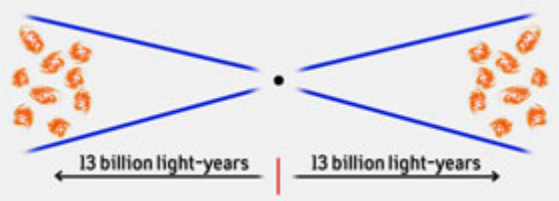


Figure 23: Both sides of the universe look the same, although light could not have traveled from one side to the other.
Source:

How is this possible? Normally, for a gas to have similar properties (such as a common temperature) over a given distance, it must have had time for the constituent atoms to have scattered off of each other and to have transferred energy throughout its full volume. However, the microwave photons reaching you and your doppelgänger on the other side of the Earth have come from regions that do not seem to be in causal contact. No photon could have traveled from one to the other according to the naive cosmology that we are imagining. How then could those regions have been in thermal equilibrium? We call this cosmological mystery the "horizon problem."

To grasp the scope of the problem, imagine that you travel billions of light-years into the past, find a distribution of different ovens with different manufacturers, power sources, and other features in the sky; and yet discover that all the ovens are at precisely the same temperature making Baked Alaska. Some causal process must have set up all the ovens and synchronized their temperatures and the ingredients they are cooking. In the case of ovens, we would of course implicate a chef. Cosmologists, who have no obvious room for a cosmic chef, have a more natural explanation: The causal structure of the universe differs from what we assume in our naive cosmology. We must believe that, although we see a universe expanding in a certain way today and can extrapolate that behavior into the past, *something drastically different happened in the far enough past*.

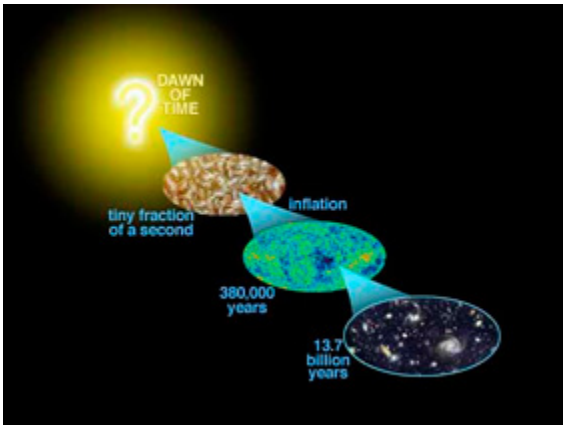


Figure 24: During the period of inflation, the universe grew by a factor of at least 10^{25}.
Source: © NASA, WMAP Science Team.

Cosmic inflation is our leading candidate for that something. Theories of cosmic inflation assert that the universe underwent a tremendously explosive expansion well before the last scattering of photons occurred. The expansion blew up a region of space a few orders of magnitude larger than the Planck scale into the size of a small grapefruit in just a few million Planck times (where a Planck time is 10^{-44} seconds). During that brief period, the universe expanded by a factor of at least 10^{25}. The inflation would thus spread particles originally in thermal contact in the tiny region a few orders of magnitude larger than the Planck length into a region large enough to be our surface of last scattering. In contrast, extrapolation of the post-Big Bang expansion of the universe into the past would never produce a region small enough for causal contact to be established at the surface of last scattering without violating some other cherished cosmological belief.

Inflation and slow-roll inflation

How does this inflation occur? In general relativity, inflation requires a source of energy density that does not move away as the universe expands. As we will see in Unit 11, simply adding a constant term (a cosmological constant) to Einstein's equations will do the trick. But, the measured value of the present-day expansion rate means the cosmological constant could only have been a tiny, tiny fraction of the energy budget of the universe at the time of the Big Bang. Thus, it had nothing to do with this explosive expansion.

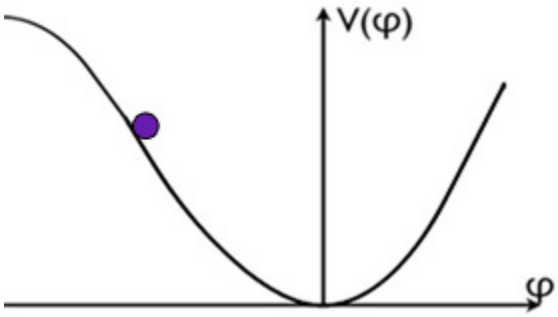


Figure 25: A potential like the one shown above for the inflaton field could have caused inflation.
Source:

However, there could have been another source of constant energy density: not exactly a cosmological constant, but something that mimics one well for a brief period of a few million Planck times. This is possible if there is a new elementary particle, the inflaton, and an associated scalar field φ. The field φ evolves in time toward its lowest energy state. The energy of φ at any spacetime point is given by a function called its "potential." If φ happens to be created in a region where the potential varies extremely slowly, then inflation will proceed. This is quite intuitive; the scalar field living on a very flat region in its potential just adds an approximate constant to the energy density of the universe, mimicking a cosmological constant but with a much larger value of the energy density than today's dark energy. We know that a cosmological constant causes accelerated (in fact, exponential) expansion of the universe.

As inflation happens, φ will slowly roll in its potential as the universe exponentially expands. Eventually, φ reaches a region of the potential where this peculiar flatness no longer holds configuration. As it reaches the ground state, its oscillations result in the production of the Standard Model quarks and leptons

through weak interactions that couple them to the inflaton. This end of inflation, when the energy stored in the inflation field is dumped into quarks and leptons, is what we know as the Big Bang.

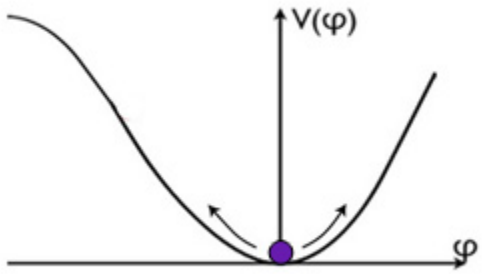


Figure 26: All the Standard Model particles could have been produced by the inflaton oscillating around its ground state like a ball rolling around in a valley.
Source:

We can imagine how this works by thinking about a ball rolling slowly on a very flat, broad hilltop. The period of inflation occurs while the ball is meandering along the top. It ends when the ball reaches the edge of the hilltop and starts down the steep portion of the hill. When the ball reaches the valley at the bottom of the hill, it oscillates there for a while, dissipating its remaining energy. However, the classical dynamics of the ball and the voyage of the inflaton differ in at least three important ways. The inflaton's energy density is a constant; it suffuses all of space, as if the universe were filled with balls on hills (and the number of the balls would grow as the universe expands). Because of this, the inflaton sources an exponentially fast expansion of the universe as a whole. Finally, the inflaton lives in a quantum world, and quantum fluctuations during inflation have very important consequences that we will explore in the next section.

Section 8: Inflation and the Origin of Structure

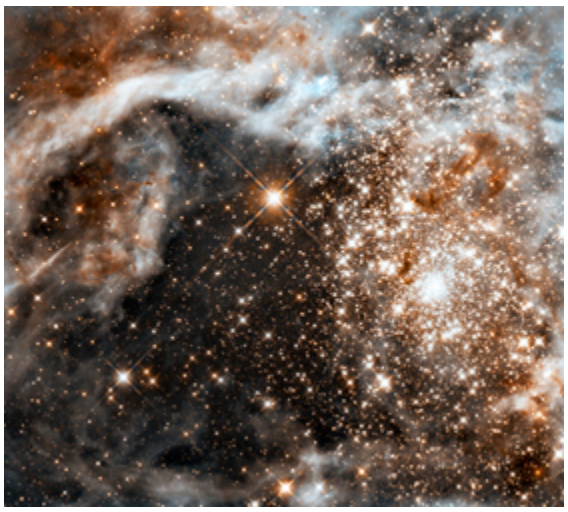


Figure 27: While the universe appears approximately uniform, we see varied and beautiful structure on smaller scales.
Source: © NASA, ESA, and F. Paresce (INAPF-AIASF, Bologna, Italy).

On the largest cosmological scales, the universe appears to be approximately homogeneous and isotropic. That is, it looks approximately the same in all directions. On smaller scales, however, we see planets, stars, galaxies, clusters of galaxies, superclusters, and so forth. Where did all of this structure come from, if the universe was once a smooth distribution of hot gas with a fixed temperature?

The temperature of the fireball that emerged from the Big Bang must have fluctuated very slightly at different points in space (although far from enough to solve the horizon problem). These tiny fluctuations in the temperature and density of the hot gas from the Big Bang eventually turned into regions of a slight overdensity of mass and energy. Since gravity is attractive, the overdense regions collapsed after an unimaginably long time to form the galaxies, stars, and planets we know today. The dynamics of the baryons, dark matter, and photons all played important and distinct roles in this beautiful, involved process of forming structure. Yet, the important point is that, over eons, gravity amplified initially tiny density fluctuations to produce the clumpy astrophysics of the modern era. From where did these tiny density fluctuations originate? In inflationary theory, the hot gas of the Big Bang arises from the oscillations and decay of the inflaton field itself. Therefore, one must find a source of slight fluctuations or differences in the inflaton's trajectory to its minimum, at different points in space. In our analogy with the ball on the hill, remember that inflation works like a different ball rolling down an identically shaped hill at

each point in space. Now, we are saying that the ball must have chosen very slightly different trajectories at different points in space—that is, rolled down the hill in very slightly different ways.

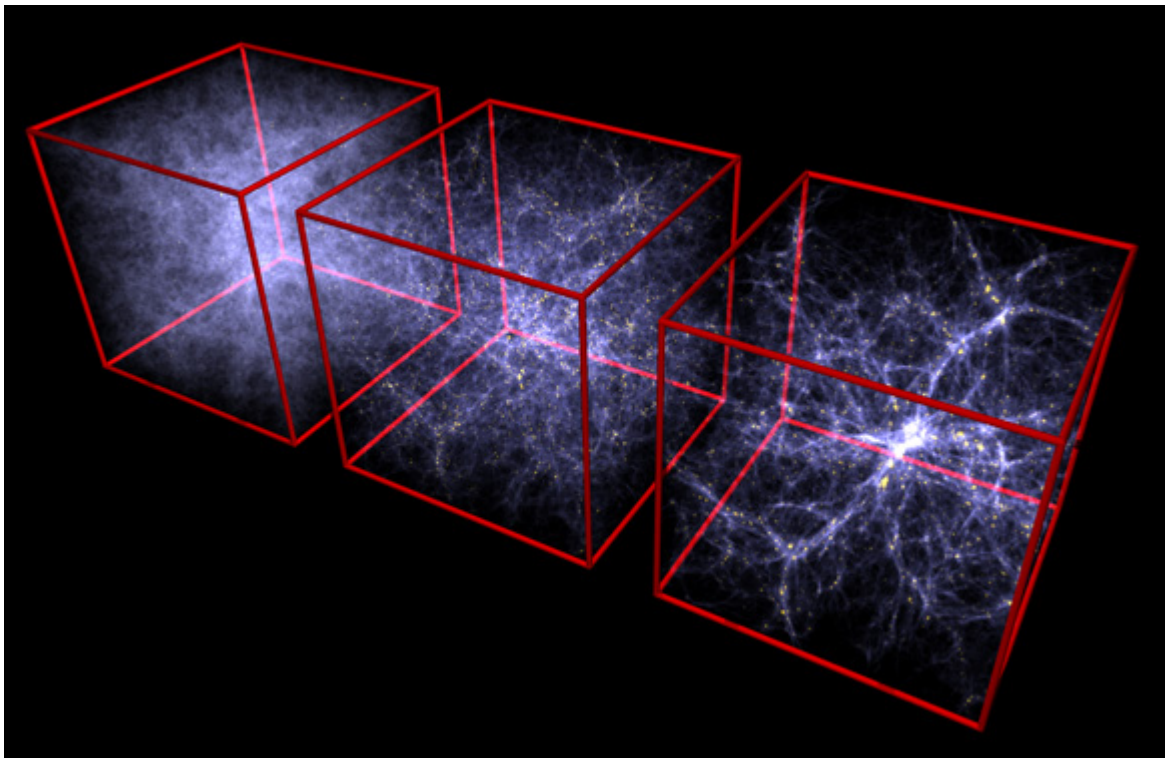


Figure 28: Small fluctuations in density in the far left box collapse into large structures on the right in this computer simulation of the universe.
Source: © Courtesy of V. Springel, Max-Planck-Institute for Astrophysics, Germany.

One source of fluctuations is quantum mechanics itself. The roll of the inflaton down its potential hill cannot be the same at all points in space, because small quantum fluctuations will cause tiny differences in the inflaton trajectories at distinct points. But because the inflaton's potential energy dominates the energy density of the universe during inflation, these tiny differences in trajectory will translate to small differences in local energy densities. When the inflaton decays, the different regions will then reheat the Standard Model particles to slightly different temperatures.

Who caused the inflation?

This leaves our present-day understanding of inflation with the feel of a murder mystery. We've found the body—decisive evidence for what has happened through the nearly uniform CMB radiation and numerous other sources. We have an overall framework for what must have caused the events, but we don't know

precisely which suspect is guilty; at our present level of knowledge, many candidates had opportunity and were equally well motivated.

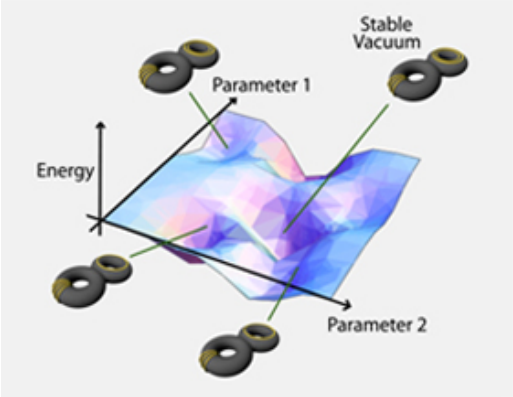


Figure 29: Typical string theories or supersymmetric field theories have many candidate scalar field inflatons.
Source:

In inflationary theory, we try to develop a watertight case to convict the single inflaton that was relevant for our cosmological patch. However, the suspect list is a long one, and grows every day. Theories of inflation simply require a scalar field with a suitable potential and some good motivation for the existence of that scalar and some rigorous derivation of that potential. At a more refined level, perhaps they should also explain why the initial conditions for the field were just right to engender the explosive inflationary expansion. Modern supersymmetric theories of particle physics, and their more complete embeddings into unified frameworks like string theory, typically provide abundant scalar fields.

While inflationary expansion is simple to explain, it is not simple to derive the theories that produce it. In particular, inflation involves the interplay of a scalar field's energy density with the gravitational field. When one says one wishes for the potential to be flat, or for the region over which it is flat to be large, the mathematical version of those statements involves M_{Planck} in a crucial way: Both criteria depend on M_{Planck}^2 multiplied by a function of the potential. Normally, we don't need to worry so much about terms in the potential divided by powers of M_{Planck} because the Planck mass is so large that these terms will be small enough to neglect. However, this is no longer true if we multiply by M_{Planck}^2. Without going into mathematical details, one can see then that even terms in the potential energy suppressed by a few powers of M_{Planck} can qualitatively change inflation, or even destroy it. In the analogy with the rolling ball on the hill, it is as if we need to make sure that the hilltop is perfectly flat with well-mown grass, and with

no gophers or field mice to perturb its flat perfection with minute tunnels or hills, if the ball is to follow the inflation-causing trajectory on the hill that we need it to follow.

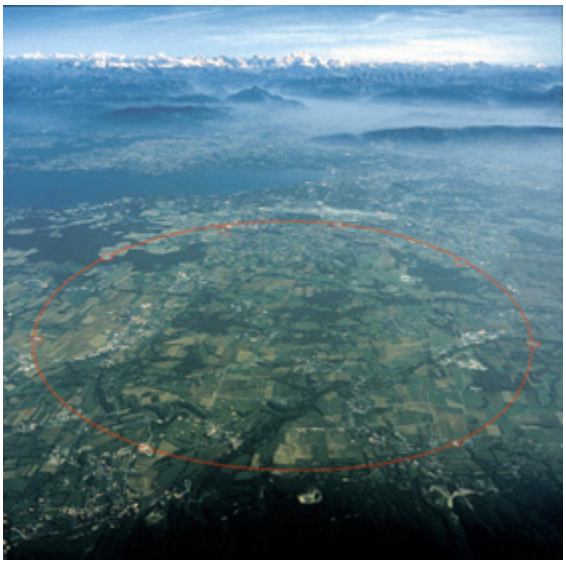


Figure 30: The LHC creates the highest-energy collisions on Earth, but these are irrelevant to Planck-scale physics.
Source: © CERN.

In particle physics we will probe the physics of the TeV scale at the LHC. There, we will be just barely sensitive to a few terms in the potential suppressed by a few TeV. Terms in the potential that are suppressed by M_{Planck} are, for the most part, completely irrelevant in particle physics at the TeV scale. If we use the cosmic accelerator provided by the Big Bang, in contrast, we get from inflation a predictive class of theories that are crucially sensitive to quantum gravity or string theory corrections to the dynamics. This is, of course, because cosmic inflation involves the delicate interplay of the inflaton with the gravitational field.

Section 9: Inflation in String Theory

Because inflation is sensitive to M_{Planck}-suppressed corrections, physicists must either make strong assumptions about Planck-scale physics or propose and compute with models of inflation in theories where they can calculate such gravity effects. String theory provides one class of theories of quantum gravity well developed enough to offer concrete and testable models of inflation—and sometimes additional correlated observational consequences. A string compactification from 10D to 4D often introduces interesting scalar fields. Some of those fields provide intriguing inflationary candidates.

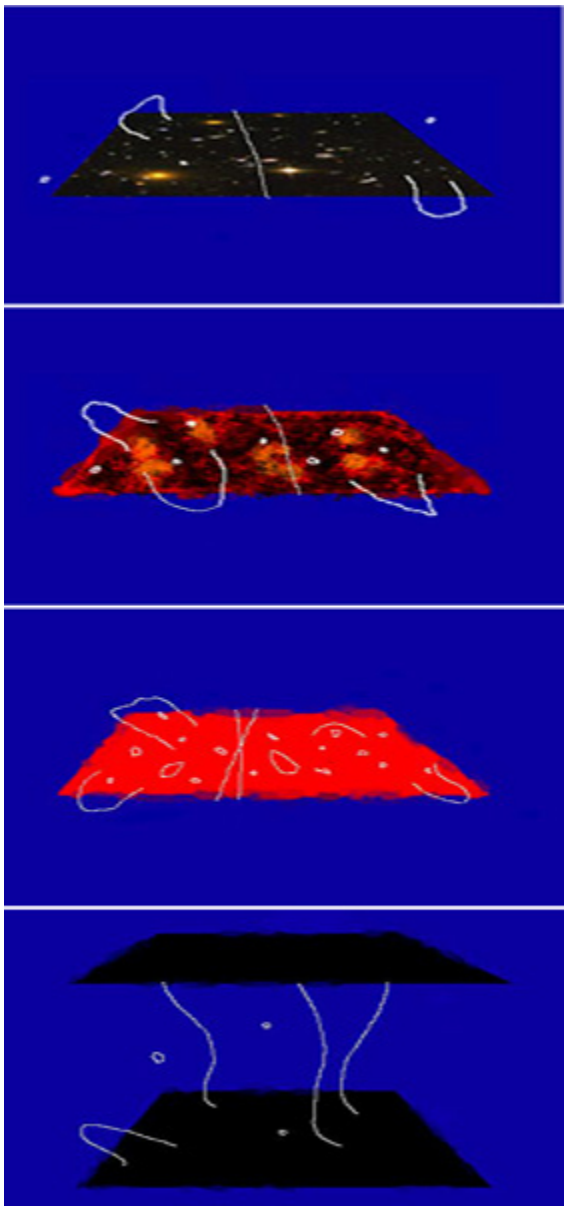


Figure 31: A brane and an anti-brane moving toward one another and colliding could have caused inflation and the Big Bang.
Source: © S.-H. Henry Tye Laboratory at Cornell University.

Perhaps the best-studied classes of models involve p-branes of the sort we described earlier in this unit. So just for concreteness, we will briefly describe this class of models. Imagine the cosmology of a universe that involves string theory compactification, curling up six of the extra dimensions to yield a 4D world. Just as we believe that the hot gas of the Big Bang in the earliest times contained particles and

anti-particles, we also believe that both branes and anti-branes may have existed in a string setting; both are p-dimensional hyperplanes on which strings can end, but they carry opposite charges under some higher-dimensional analog of electromagnetism.

The most easily visualized case involves a 3-brane and an anti-3-brane filling our 4D spacetime but located at different points in the other six dimensions. Just as an electron and a positron attract one another, the brane and anti-brane attract one another via gravitational forces as well as the other force under which they are charged. However, the force law is not exactly a familiar one. In the simplest case, the force falls off as $1/r^4$, where r is the distance separating the brane and anti-brane in the 6D compactified space.

Models with sufficiently interesting geometries for the compactified dimensions can produce slow-roll inflation when the brane and anti-brane slowly fall together, under the influence of the attractive force. The inflaton field is the mode that controls the separation between the brane and the anti-brane. Each of the branes, as a material object filling our 4D space, has a tension that provides an energy density filling all of space. So, a more accurate expression for the inter-brane potential would be $V(r) \sim 2T_3 - 1/r^4$, where T_3 is the brane tension. For sufficiently large r and slowly rolling branes, the term $2T_3$ dominates the energy density of the universe and serves as the effective cosmological constant that drives inflation.

As the branes approach each other and $r \sim \ell_{string}$, this picture breaks down. This is because certain open strings, now with one end on each of the branes as opposed to both ends on a single brane, can become light. In contrast, when $r >> \ell_{string}$, such open strings must stretch a long distance and are quite heavy. Remember that the energy or mass of a string scales with its length. In the regime where r is very small, and the open strings become light, the picture in terms of moving branes breaks down. Instead, some of the light open strings mediate an instability of the brane configuration. In the crudest approximation, the brane and anti-brane simply annihilate (just as an electron and anti-electron would), releasing all of the energy density stored in the brane tensions in the form of closed-string radiation. In this type of model, the Big Bang is related to the annihilation of a brane with an anti-brane in the early universe.

Other consequences of inflation

Any well-specified model of cosmic inflation has a full list of consequences that can include observables beyond just the density fluctuations in the microwave background that result from inflation. Here, we mention some of the most spectacular possible consequences.

Quantum jiggles: We do not know the energy scale of inflation directly from data. In many of the simplest theories, however, this energy scale is very high, close to the Grand Unified Theory scale of 10^{16} GeV. It is therefore quite possible that inflation is probing energies 13 orders of magnitude higher than we'll see at the LHC.

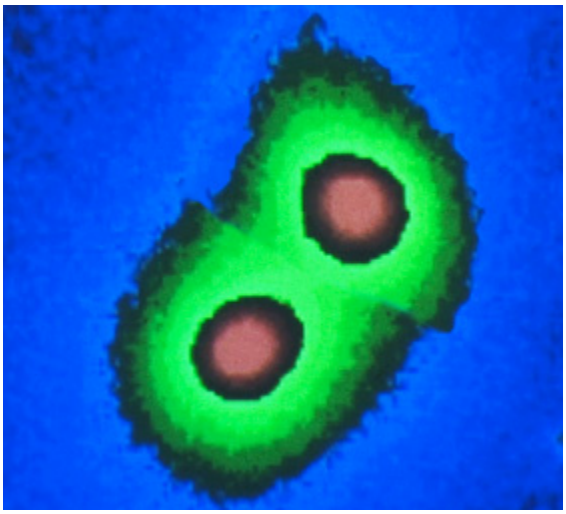


Figure 32: A cosmic string could produce a double image of a galaxy behind it.
Source: © NASA/ESA Hubble Space Telescope.

If the scale of inflation is high enough, we may see further corroborating evidence beyond the solution of the horizon problem and the explanation of density fluctuations. The density fluctuations we discussed in the previous section came from the quantum jiggles of the inflaton field itself. But during inflation, quantum jiggles also originate in the other fields present, including the gravitational field. Future cosmological experiments exploring those phenomena could pin down the scale of inflation to just a few orders of magnitude shy of the Planck scale.

Cosmic strings: Very particular models often come with their own smoking-gun signatures. Take, for example, the class of speculative models we discussed earlier based on the slow attraction and eventual annihilation of a 3-brane and an anti-3-brane. The annihilation process involves the dynamics of open strings that stretch between the 3-brane and its partner, and that eventually "condense." This

condensation process creates cosmic strings as the branes annihilate, which can be thought of as 1-branes or even fundamental strings that thread our 4D spacetime, and have grown to macroscopic size. If these tension-bearing cosmic strings really were created at the end of inflation, they should be present in the universe today, with tell-tale signatures in experiments that study the distribution of matter through gravitational lensing. Future experiments should rule out the presence of such strings or detect them for a wide range of values of the possible scale of inflation.

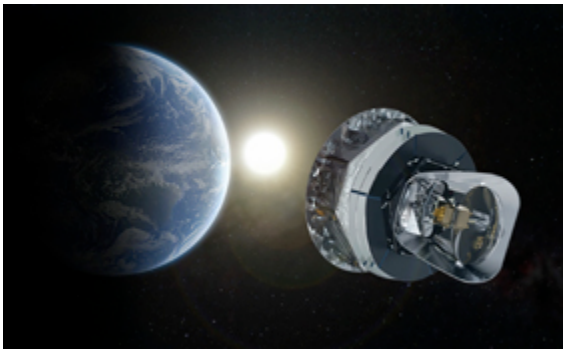


Figure 33: The Planck satellite will make precise measurements of fluctuations in the CMB.
Source: © ESA.

Density fluctuations: The slow-roll approach to inflation, with an inflaton field moving on a flat potential, is the class of models we focused on in the last section, but is not the only model of inflation. In some more modern theories, again partially inspired by branes in superstring theory, the inflaton undergoes rapid motion. Instead of the flatness of the potential, a delicate interplay between the potential and the complicated structure of the inflation kinetic terms produces the inflation. If any such model captures a grain of truth, then the pattern of density fluctuations would bear tell-tale structural signatures. Measurements by the Planck satellite that the European Space Agency launched in May 2009 should put constraints on the validity of those models.

Section 10: Fundamental Questions of Gravity

Figure 34: Some of Einstein's great insights began in thought experiments.
Source: © Marcelo Gleiser.

Quantum gravity is not a mystery that is, as yet, open to experiment. With a few exceptions such as inflationary cosmology and, quite possibly, the correct interpretation of dark energy (see Unit 11), thoughts about quantum gravity remain in the theoretical realm. This does not mean that theories of quantum gravity cannot be important and testable in some circumstances. In a given model, for instance, string theory models of particle physics make very definite statements about the origins of the mass hierarchy of fermions or the number of generations of Standard Model particles. But these problems may

also have other solutions, insensitive to the structure of gravity at short distances; only in very few cases do we suspect that quantum gravity must be a part of the solution to a problem.

Here, we discuss some of these issues that are intrinsically gravitational. They have not yet confronted experiment directly. But we should remember that Einstein formulated special relativity by reconciling different thought experiments, and that therefore even thought experiments about quantum gravity may eventually be useful.

Black holes and entropy

Black holes are objects so dense that the escape velocity from their surface exceeds the speed of light, c. Because of that, one would think that in a relativistic theory, outside observers performing classical experiments can never see their surfaces. As a rough-and-ready definition, we will call the surface defining the region where light itself can no longer escape from the gravitational attraction of a black hole, the event horizon. Nothing, in a classical theory, can be emitted from this horizon, though many things can fall through.

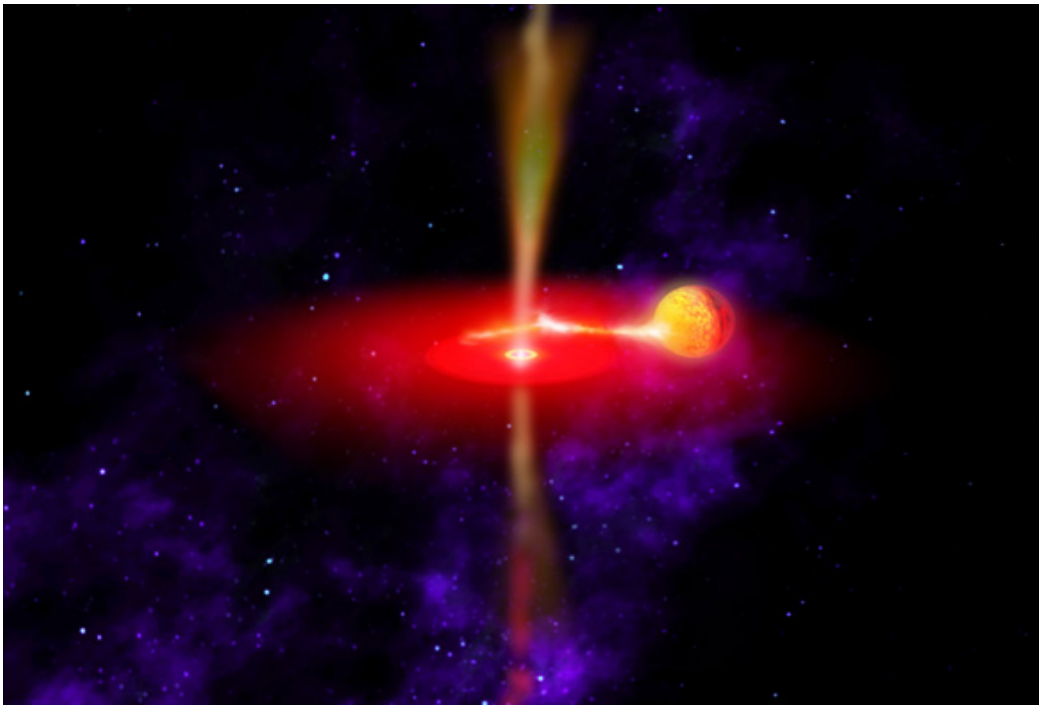


Figure 35: Artist's conception of a black hole accreting matter from a companion star.
Source: © Dana Berry (CfA, NASA).

Careful consideration of the theory of black holes in classical general relativity in the early 1970s led Jacob Bekenstein, Stephen Hawking, and others to a striking set of conclusions. They found that as a chargeless, non-rotating black hole accretes matter, its mass grows by an amount proportional to the strength of gravity at the black hole's surface and the change in its surface area. Also, the black hole's surface area (defined by its event horizon) cannot decrease under any circumstances, and usually increases in time.

At a heuristic level, Bekenstein and Hawking's laws for black holes seem reminiscent of the laws of thermodynamics and statistical mechanics: The change in energy is proportional to the change in entropy and the entropy (a measure of disorder) of a system can only increase. This is no coincidence. The results of general relativity imply what they seem to: A black hole does carry an entropy proportional to its surface area, and, of course, it has an energy that grows with its mass.

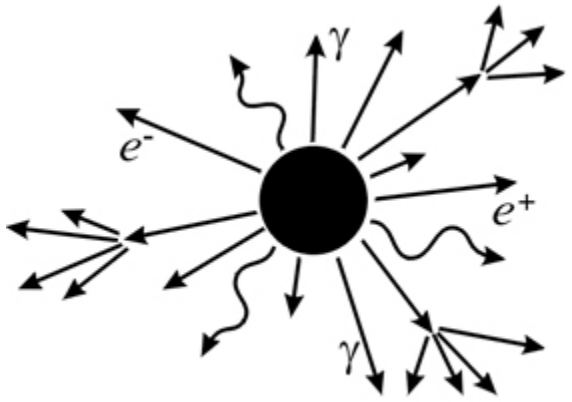


Figure 36: Black holes radiate by a quantum mechanical process.
Source: © Reprinted with permission from Nova Science Publishers, Inc. from: Sabine Hossenfelder, "What Black Holes Can Teach Us," in Focus on Black Hole Research, ed. Paul V. Kreitler (New York: Nova Publishers Inc., 2006), 121-58.

One mystery remains, however. In thermodynamics, the change in energy is proportional to the temperature times the change in entropy; and hot bodies radiate. Even though there is an analogous quantity—the surface gravity—in the black hole mechanics, no classical process can bring radiation through the horizon of a black hole. In a brilliant calculation in 1974, Stephen Hawking showed that,

nevertheless, black holes radiate by a *quantum* process. This quantum effect occurs at just the right level to make the analogy between black hole thermodynamics and normal thermodynamics work perfectly.

Hawking's calculation reinforces our belief that a black hole's entropy should be proportional to its surface area. This is a bit confusing because most theories that govern the interactions of matter and force-carrying particles in the absence of gravity posit that entropy grows in proportion to the *volume* of the system. But in a gravity theory also containing these other degrees of freedom, if one tries to fill a region with enough particles so that their entropy exceeds the area bounding the region, one instead finds gravitational collapse into a black hole, whose entropy is proportional to its surface area. This means that at least in gravity theories, our naive idea that the entropy that can be contained in a space should scale with its volume must be incorrect.

Holography, multiple dimensions, and beyond

This concept that in every theory of quantum gravity, the full entropy is proportional only to the area of some suitably chosen boundary or "holographic screen" in the system, and not the full volume, carries a further implication: that we may be able to formulate a theory of gravity in D +1 spacetime dimensions in just D dimensions—but in terms of a purely non-gravitational quantum field theory. Dutch theorist Gerard 't Hooft and his American colleague Leonard Susskind put forward this loose idea, called holography, in the early 1990s. The idea, as stated, is a bit vague. It begs questions such as: *On which "bounding surface" do we try to formulate the physics? Which quantum field theory is used to capture which quantum gravity theory in the "bulk" of the volume?*

In the late 1990s, through the work of Argentine theorist Juan Maldacena and many others, this idea received its first very concrete realization. We will focus on the case of gravity in four dimensions; for different reasons, much of the actual research has been focused on theories of gravity in five dimensions. This work gives, for the first time, a concrete non-perturbative formulation of quantum gravity in the 4D Anti de Sitter spacetime—the simplest solution of Einstein's theory with a negative cosmological constant.

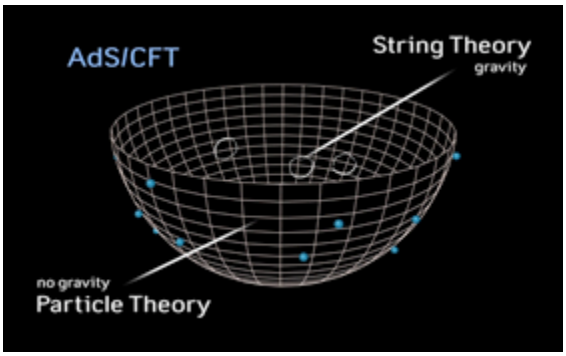


Figure 37: The AdS/CFT duality relates a theory on the boundary of a region to a theory with gravity in the interior.
Source:

It turns out that the symmetries of the 4D space—that is, in 3+1 dimensions—match those of a quantum field theory of a very particular sort, called a "conformal field theory," in 2+1 spacetime dimensions. However, we also expect that quantum gravity in 3+1 dimensions should have the same behavior of its thermodynamic quantities (and in particular, its entropy) as a theory in 2+1 dimensions, without gravity. In fact, these two observations coincide in a beautiful story called the Anti de Sitter/Conformal Field Theory (AdS/CFT) correspondence. Certain classes of quantum field theories in 2+1 dimensions are exactly equivalent to quantum gravity theories in 3+1 dimensional Anti de Sitter space. Physicists say that these two theories are dual to one another.

The precise examples of this duality that we recognize come from string theory. Compactifications of the theory to four dimensions on different compact spaces give rise to different examples of AdS_4 gravity and to different dual field theories. While we do not yet know the gravity dual of every 2+1 dimensional conformal field theory or the field theory dual of every gravity theory in AdS_4, we do have an infinite set of explicit examples derived from string theory.

The value of duality

This duality has a particularly interesting, useful, and, on reflection, necessary aspect. The 2+1 dimensional field theories analogous to electromagnetism have coupling constants g analogous to the electron charge e. The gravity theory also has a natural coupling constant, given by the ratio of the curvature radius of space to the Planck length, which we will call L. In the known examples of the duality between AdS space and quantum field theories, large values of L, for which the gravity theory is weakly curved, and hence involves only weak gravity, correspond to very large values of g for the quantum field

theory. Conversely, when the quantum field theory is weakly coupled (at small values of g), the gravity theory has very small L; it is strongly coupled in the sense that quantum gravity corrections (which are very hard to compute, even in string theory) are important.

This kind of duality, between a strongly coupled theory on the one hand and a weakly coupled theory on the other, is actually a common (though remarkable and beautiful) feature in physical theories. The extra shock here is that one of the theories involves quantum gravity in a different dimension.

This duality has had two kinds of uses to date. One obvious use is that it provides a definition of quantum gravity in terms of a normal field theory for certain kinds of gravitational backgrounds. Another use, however, that has so far proved more fruitful, is that it gives a practical way to compute in classes of strongly coupled quantum field theories: You can use their weakly curved gravitational dual and compute the dual quantities in the gravity theory there.

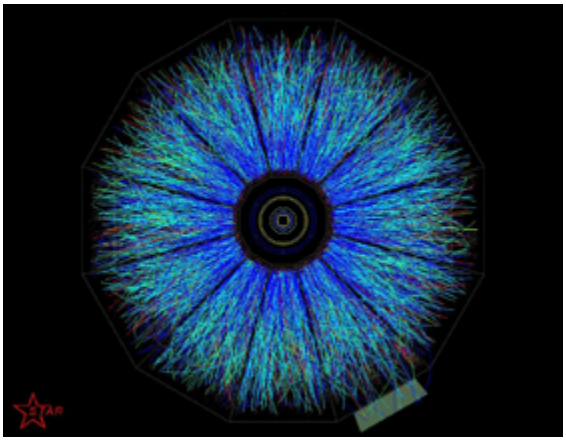


Figure 38: The aftermath of a heavy ion collision at RHIC, the Relativistic Heavy Ion Collider.
Source: © Courtesy of Brookhaven National Laboratory.

Physicists have high hopes that such explorations of very strongly coupled quantum field theories based on gravity duals may provide insight into many of the central problems in strongly coupled quantum field theory; these include a proper understanding of quark confinement in QCD, the ability to compute transport in the strongly coupled QCD plasma created at present-day accelerators like Brookhaven National Laboratory's Relativistic Heavy Ion Collider (RHIC) that smash together heavy ions, the ability to solve for quantities such as conductivity in strongly correlated electron systems in condensed matter physics, and an understanding of numerous zero-temperature quantum phase transitions in such systems. However, we must note that, while this new approach is orthogonal to old ones and promises

to shed new light in various toy models of those systems, it has not yet helped to solve any of the central problems in those subjects.

The initial singularity

Even more intriguing is the question of the initial cosmological singularity. In general relativity, one can prove powerful theorems showing that any expanding cosmology (of the sort we inhabit) must have arisen in the distant past from a point of singularity in which the energy density and curvature are very large and the classical theory of relativity is expected to break down. Intuitively, one should just run the cosmological expansion backwards; then the matter we currently see flying apart would grow into an ever-denser and more highly curved state.

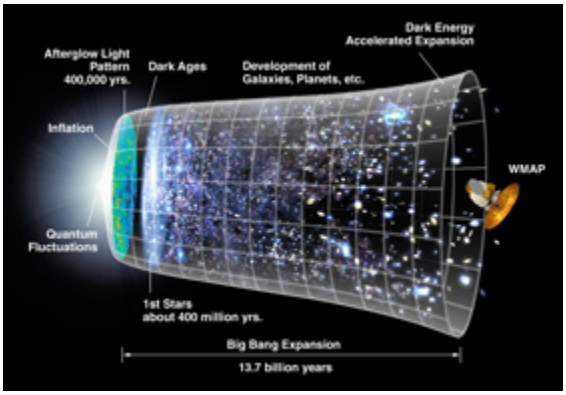


Figure 39: If we run time backwards, the entire universe collapses into a single, infinitely dense point. **Source:** © NASA, WMAP Science Team.

Our current picture of the Big Bang, including the successes of Big Bang nucleosynthesis and the CMB, gives us confidence that we can safely extrapolate our current cosmology back to temperatures of order MeV. In most inflationary theories, the Big Bang engendered temperatures far above the TeV scale. But at some still higher energy scale, probing the start of inflation and beyond, we do not know what happened; we do not even have any ideas that provide hints of testable predictions.

What is the origin of our observable universe? Is it one of many, coming from quantum tunneling events out of regions of space with different macroscopic laws of physics? Or is it a unique state, arising from some unknown initial condition of quantum gravity that we have yet to unravel? And, how does this physics eventually deal with the singularity theorems of general relativity, which assure us that

extrapolation backwards into the past will lead to a state of high density and curvature, where little can be reliably calculated? These mysteries remain at the forefront of modern research in quantum gravity.

Section 11: Further Reading

- Brian Greene, "The Elegant Universe," Vintage Books, 2000.
- Juan Maldacena, "The Illusion of Gravity," *Scientific American*, November, 2005, p. 56–63.
- Lisa Randall, "Warped Passages: Unraveling Mysteries of the Universe's Hidden Dimensions," *Harper Perennial*, 2006.
- Barton Zwiebach, "A First Course in String Theory," *Cambridge University Press*, 2009.

Glossary

Anti-de Sitter/Conformal Field Theory (AdS/CFT): AdS/CFT is a mathematical relationship between two separate descriptions of the same physics. According to AdS/CFT, a string theory in a region of Anti-de Sitter (AdS) space is equivalent to a conformal field theory (CFT) on the boundary of that region. Anti-de Sitter space has negative curvature (a two-dimensional plane is curved in a saddle shape rather than flat), and is one of the simplest geometries in which the equations of general relativity can be solved. A conformal field theory is the type of field theory used in the Standard Model. Although AdS/CFT describes an artificially simple situation—we appear to live in flat space, not Anti-de Sitter space—the mathematical correspondence between the two descriptions of physics has allowed relatively straightforward field theory calculations to shed light on problems associated with the quantum mechanics of black holes. AdS/CFT has also been used the other way, with black hole calculations providing insight into complicated particle collisions and condensed matter systems that are difficult to understand with the conventional field theory approach.

blackbody: A blackbody is an object that absorbs all incident electromagnetic radiation and re-radiates it after reaching thermal equilibrium. The spectrum of light emitted by a blackbody is smooth and continuous, and depends on the blackbody's temperature. The peak of the spectrum is higher and at a shorter wavelength as the temperature increases.

black hole: A black hole is a region of space where gravity is so strong that nothing can escape its pull. Black holes have been detected through their gravitational influence on nearby stars and through observations of hot gas from surrounding regions accelerating toward them. These black holes are thought to have formed when massive stars reached the end of their cycle of evolution and collapsed under the influence of gravity. If a small volume of space contains enough mass, general relativity predicts that spacetime will become so highly curved that a black hole will form.

brane, p-brane: In string theory, branes are fundamental objects that exist in a specific number of spatial dimensions. The "p" in p-brane stands for the number of dimensions that brane has. For example, a string is a 0-brane, a membrane is a 2-brane, and we could live on a 3-brane.

closed string: In string theory, a closed string forms a loop. Unlike open strings, closed strings are not attached to other objects; however, a closed string can be broken apart to form an open string.

Open strings and closed strings have different properties, and give rise to different sets of fundamental particles.

compactification: In string theory, the term compactification refers to how an extra dimension is made small enough that we cannot perceive it. The three spatial dimensions we are familiar with from daily life are essentially infinite, while compactified dimensions are curled up, and have a finite size that ranges from a few microns (10^{-6} m) down to the Planck length.

cosmic microwave background: The cosmic microwave background (CMB) radiation is electromagnetic radiation left over from when atoms first formed in the early universe, according to our standard model of cosmology. Prior to that time, photons and the fundamental building blocks of matter formed a hot, dense soup, constantly interacting with one another. As the universe expanded and cooled, protons and neutrons formed atomic nuclei, which then combined with electrons to form neutral atoms. At this point, the photons effectively stopped interacting with them. These photons, which have stretched as the universe expanded, form the CMB. First observed by Penzias and Wilson in 1965, the CMB remains the focus of increasingly precise observations intended to provide insight into the composition and evolution of the universe.

cosmic string: A cosmic string is a one-dimensional topological defect stretching across the universe, essentially an extremely thin, extremely dense line in space that would deform spacetime around it according to general relativity. Cosmic strings have been predicted by various theories, but never detected. It is possible that if the period of inflation in the early universe ended in a collision between a brane and an anti-brane, cosmic strings were produced in the process.

cosmological constant: The cosmological constant is a constant term that Einstein originally included in his formulation of general relativity. It has the physical effect of pushing the universe apart. Einstein's intent was to make his equations describe a static universe. After astronomical evidence clearly indicated that the size of the universe is changing, Einstein abandoned the cosmological constant though other astrophysicists, such as Georges Lemaître and Sir Arthur Stanley Eddington, thought it might be the source of cosmic expansion. The cosmological constant is a simple explanation of dark energy consistent with the observations; however, it is not the only possible explanation, and the value of the cosmological constant consistent with observation is over 60 orders of magnitude different from what theory predicts.

event horizon: A black hole's event horizon is the point of no return for matter falling toward the black hole. Once matter enters the event horizon, it is gravitationally bound to the black hole and cannot

escape. However, an external observer will not see the matter enter the black hole. Instead, the gravitational redshift due to the black hole's strong gravitational field causes the object to appear to approach the horizon increasingly slowly without ever going beyond it. Within the event horizon, the black hole's gravitational field warps spacetime so much that even light cannot escape.

general relativity: General relativity is the theory Einstein developed to reconcile gravity with special relativity. While special relativity accurately describes the laws of physics in inertial reference frames, it does not describe what happens in an accelerated reference frame or gravitational field. Since acceleration and gravity are important parts of our physical world, Einstein recognized that special relativity was an incomplete description and spent the years between 1905 and 1915 developing general relativity. In general relativity, we inhabit a four-dimensional spacetime with a curvature determined by the distribution of matter and energy in space. General relativity makes unique, testable predictions that have been upheld by experimental measurements, including the precession of Mercury's orbit, gravitational lensing, and gravitational time dilation. Other predictions of general relativity, including gravitational waves, have not yet been verified. While there is no direct experimental evidence that conflicts with general relativity, the accepted view is that general relativity is an approximation to a more fundamental theory of gravity that will unify it with the Standard Model. See: gravitational lensing. gravitational time dilation, gravitational wave, precession, spacetime, special relativity, Standard Model.

gravitational lensing: Gravitational lensing occurs when light travels past a very massive object. According to Einstein's theory of general relativity, mass shapes spacetime and space is curved by massive objects. Light traveling past a massive object follows a "straight" path in the curved space, and is deflected as if it had passed through a lens. Strong gravitational lensing can cause stars to appear as rings as their light travels in a curved path past a massive object along the line of sight. We observe microlensing when an object such as a MACHO moves between the Earth and a star. The gravitational lens associated with the MACHO focuses the star' light, so we observe the star grow brighter then dimmer as the MACHO moves across our line of sight to the star.

graviton: The graviton is the postulated force carrier of the gravitational force in quantum theories of gravity that are analogous to the Standard Model. Gravitons have never been detected, nor is there a viable theory of quantum gravity, so gravitons are not on the same experimental or theoretical footing as the other force carrier particles.

ground state: The ground state of a physical system is the lowest energy state it can occupy. For example, a hydrogen atom is in its ground state when its electron occupies the lowest available energy level.

hierarchy problem: The hierarchy problem in theoretical physics is the fact that there appear to be two distinctly different energy scales in the universe for reasons that are not understood. The first energy scale, called the "electroweak scale," is associated with everything except gravity. The electroweak scale is set by the mass of the W and Z bosons at around 100 GeV, and determines the strength of the strong, electromagnetic, and weak interactions. The second is the Planck scale, at 10^{19} GeV, which is associated with gravitational interactions. Another way of stating the hierarchy problem is to ask why gravity is 39 orders of magnitude weaker than the other fundamental forces of nature.

inflation: Inflation is a period of exponential expansion thought to have occurred around 10^{-36} seconds after the universe began. During this period, which lasted for a few million Planck times, the universe expanded by a factor of at least 10^{25}, smoothing out temperature and density fluctuations to produce the nearly uniform universe we observe today. Although the mechanism driving inflation is still not understood, evidence from the cosmic microwave background supports its existence.

inflaton: The inflaton is a hypothetical scalar field that could drive the period of inflation that took place in the early universe.

nucleosynthesis: The term "nucleosynthesis" refers either to the process of forming atomic nuclei from pre-existing protons and neutrons or to the process of adding nucleons to an existing atomic nucleus to form a heavier element. Nucleosynthesis occurs naturally inside stars and when stars explode as supernovae. In our standard model of cosmology, the first atomic nuclei formed minutes after the Big Bang, in the process termed "Big Bang nucleosynthesis."

open string: In string theory, an open string has two distinct ends. Open strings can have one end attached to another object like a brane, and the two ends of an open string can connect to form a closed string. Open strings and closed strings have different properties, and give rise to different sets of fundamental particles.

brane, p-brane: In string theory, branes are fundamental objects that exist in a specific number of spatial dimensions. The "p" in p-brane stands for the number of dimensions that brane has. For example, a string is a 0-brane, a membrane is a 2-brane, and we could live on a 3-brane.

Planck length: The Planck length is the fundamental unit of length used in high energy physics, and is a combination of Planck's constant, Newton's constant of universal gravitation, and the speed of light. The Planck length is approximately 1.6×10^{-35} m.

Planck mass: The Planck mass is the fundamental unit of mass used in high energy physics, and is a combination of Planck's constant, Newton's constant of universal gravitation, and the speed of light. The Planck mass is approximately 2.2×10^{-8} kg.

Planck time: The Planck time is the time it takes light to travel one Planck length, and is considered the fundamental unit of time in high energy physics. The Planck time is approximately 5.4×10^{-44} seconds.

potential energy: Potential energy is energy stored within a physical system. A mass held above the surface of the Earth has gravitational potential energy, two atoms bound in a molecule have chemical potential energy, and two electric charges separated by some distance have electric potential energy. Potential energy can be converted into other forms of energy. If you release the mass, its gravitational potential energy will be converted into kinetic energy as the mass accelerates downward. In the process, the gravitational force will do work on the mass. The force is proportional to the rate at which the potential energy changes. It is common practice to write physical theories in terms of potential energy, and derive forces and interactions from the potential.

quantum chromodynamics: Quantum chromodynamics, or QCD, is the theory that describes the strong nuclear force. It is a quantum field theory in which quarks interact with one another by exchanging force-carrying particles called "gluons." It has two striking features that distinguish it from the weak and electromagnetic forces. First, the force between two quarks remains constant as the quarks are pulled apart. This explains why single quarks have never been found in nature. Second, quarks and gluons interact very weakly at high energies. QCD is an essential part of the Standard Model and is well tested experimentally; however, calculations in QCD can be very difficult and are often performed using approximations and computer simulations rather than solved directly.

recombination: In the context of cosmology, the term recombination refers to electrons combining with atomic nuclei to form atoms. In our standard model of cosmology, this took place around 390,000 years

after the Big Bang. Prior to the time of recombination, the universe was filled with a plasma of electrically charged particles. Afterward, it was full of neutral atoms.

scalar field: A scalar field is a smoothly varying mathematical function that assigns a value to every point in space. An example of a scalar field in classical physics is the gravitational field that describes the gravitational potential of a massive object. In meteorology, the temperature and pressure distributions are scalar fields. In quantum field theory, scalar fields are associated with spin-zero particles. All of the force-carrying particles as well as the Higgs boson are generated by scalar fields.

singularity: Singularity is a mathematical term that refers to a point at which a mathematical object is undefined, either because it is infinite or degenerate. A simple example is the function 1/x. This function has a singularity at x = 0 because the fraction 1/0 is undefined. Another example is the center of a black hole, which has infinite density. In our standard model of cosmology, the universe we live in began as a spacetime singularity with infinite temperature and density.

special relativity: Einstein developed his theory of special relativity in 1905, 10 years before general relativity. Special relativity is predicated on two postulates. First, the speed of light is assumed to be constant in all inertial frames. Second, the laws of physics are assumed to be the same in all inertial frames. An inertial frame, in this context, is defined as a reference frame that is not accelerating or in a gravitational field. Starting from these two postulates, Einstein derived a number of counterintuitive consequences that were later verified by experiment. Among them are time dilation (a moving clock will run slower than a stationary clock), length contraction (a moving ruler will be shorter than a stationary ruler), the equivalence of mass and energy, and that nothing can move faster than the speed of light. See: general relativity, spacetime.

strong interaction: The strong interaction, or strong nuclear force, is one of the four fundamental forces of nature. It acts on quarks, binding them together into mesons. Unlike the other forces, the strong force between two particles remains constant as the distance between them grows, but actually gets weaker when the particles get close enough together. This unique feature ensures that single quarks are not found in nature. True to its name, the strong force is a few orders of magnitude stronger than the electromagnetic and weak interactions, and many orders of magnitude stronger than gravity.

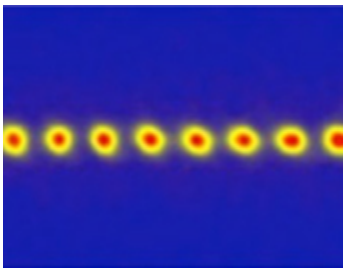
topology: Topology is the mathematical study of what happens to objects when they are stretched, twisted, or deformed. Objects that have the same topology can be morphed into one another smoothly,

without any tearing. For example, a donut and a coffee cup have the same topology, while a beach ball is in a different topological category.

winding mode: In string theory, a winding mode is a distinct way in which a string can wrap around a compactified extra dimension. If we imagine a single extra dimension compactified into a circle, the simplest winding mode is for the string to wind once around the circle.

Physics for the 21st Century

Unit 5: The Quantum World



Unit Overview

This unit covers a field of physics that is simultaneously one of the most powerful, transformational, and precise tools for exploring nature and yet for non-physicists one of the most mysterious and misunderstood aspects of all science. Developed early in the 20th century to solve a crisis in understanding the nature of the atom, quantum mechanics has laid the foundation for theoretical and practical advances in 21st century physics. The unit details the reasoning that led to ever deeper awareness of the nature of particles, waves, and their interrelationships, provides a primer on present-day understanding of the field, and outlines ways in which that understanding has led to significant applications today.

Content for This Unit

Sections:

Section 1: Introduction

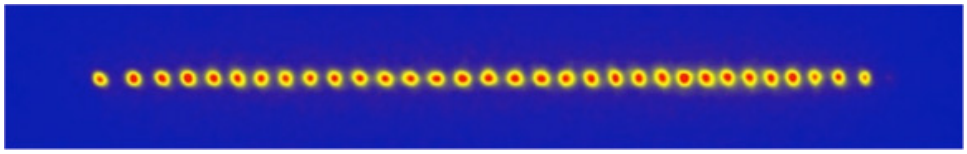


Figure 1: Thirty-two ions, fluorescing under illumination by laser light in an electrodynamic trap.
Source: © R. Blatt at Institut f. Experimentalphysik, Universtaet Innsbruck, Austria.

The creation of quantum mechanics in the 1920s broke open the gates to understanding atoms, molecules, and the structure of materials. This new knowledge transformed our world. Within two decades, quantum mechanics led to the invention of the transistor to be followed by the invention of the laser, revolutionary advances in semiconductor electronics, integrated circuits, medical diagnostics, and optical communications. Quantum mechanics also transformed physics because it profoundly changed our understanding of how to ask questions of nature and how to interpret the answers. An intellectual change of this magnitude does not come easily. The founders of quantum mechanics struggled with its concepts and passionately debated them. We are the beneficiaries of that struggle and quantum mechanics has now been developed into an elegant and coherent discipline. Nevertheless, quantum mechanics always seems strange on first acquaintance and certain aspects of it continue to generate debate today. We hope that this unit provides insight into how quantum mechanics works and why people find it so strange at first. We will also sketch some of the recent developments that have enormously enhanced our powers for working in the quantum world. These advances make it possible to manipulate and study quantum systems with a clarity previously achieved only in hypothetical thought experiments. They are so dramatic that some physicists have described them as a second quantum revolution.

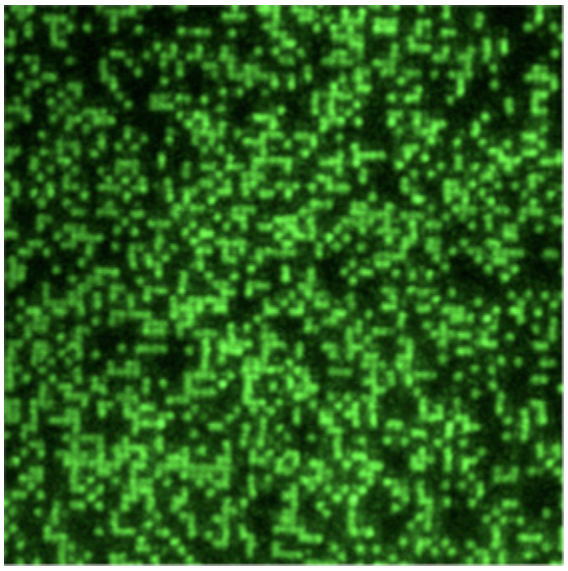


Figure 2: Neutral rubidium atoms in an optical lattice trap.
Source: © M. Greiner.

An early step in the second quantum revolution was the discovery of how to capture and manipulate a single ion in an electromagnetic trap, reduce its energy to the quantum limit, and even watch the ion by eye as it fluoresces. Figure 1 shows an array of fluorescing ions in a trap. Then methods were discovered for cooling atoms to microkelvin temperatures (a microkelvin is a millionth of a degree) and trapping them in magnetic fields or with light waves (Figure 2). These advances opened the way to stunning advances such as the observation of Bose-Einstein condensation of atoms, to be discussed in Unit 6, and the creation of a new discipline that straddles atomic and condensed matter physics.

The goal of this unit is to convey the spirit of life in the quantum world—that is, to give an idea of what quantum mechanics is and how it works—and to describe two events in the second quantum revolution: atom cooling and atomic clocks.

Section 2: Mysteries of Light

Figure 3: This furnace for melting glass is nearly an ideal blackbody radiation source.
Source: © OHM Equipment, LLC.

The nature of light was a profound mystery from the earliest stirrings of science until the 1860s and 1870s, when James Clerk Maxwell developed and published his electromagnetic theory. By joining the two seemingly disparate phenomena, electricity and magnetism, into the single concept of an electromagnetic field, Maxwell's theory showed that waves in the field travel at the speed of light and are, in fact, light itself. Today, most physicists regard Maxwell's theory as among the most important and beautiful theories in all of physics.

Maxwell's theory is elegant because it can be expressed by a short set of equations. It is powerful because it leads to powerful predictions—for instance, the existence of radio waves and, for that matter, the entire electromagnetic spectrum from radio waves to x-rays. Furthermore, the theory explained how light can be created and absorbed, and provided a key to essentially every question in optics.

Given the beauty, elegance, and success of Maxwell's theory of light, it is ironic that the quantum age, in which many of the most cherished concepts of physics had to be recast, was actually triggered by a problem involving light.

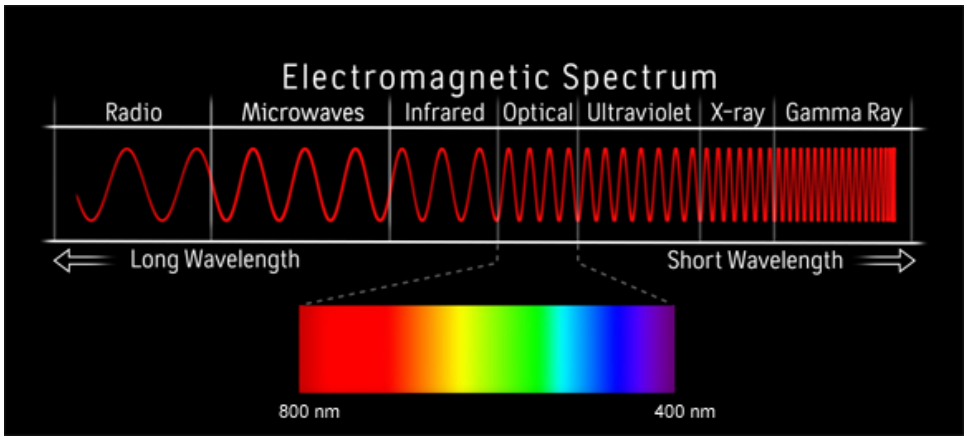


Figure 4: The electromagnetic spectrum from radio waves to gamma rays.
Source:

The spectrum of light from a blackbody—for instance the oven in Figure 3 or the filament of an electric light bulb—contains a broad spread of wavelengths. The spectrum varies rapidly with the temperature of the body. As the filament is heated, the faint red glow of a warm metal becomes brighter, and the peak of the spectrum broadens and shifts to a shorter wavelength, from orange to yellow and then to blue. The spectra of radiation from blackbodies at different temperatures have identical shapes and differ only in the scales of the axes.

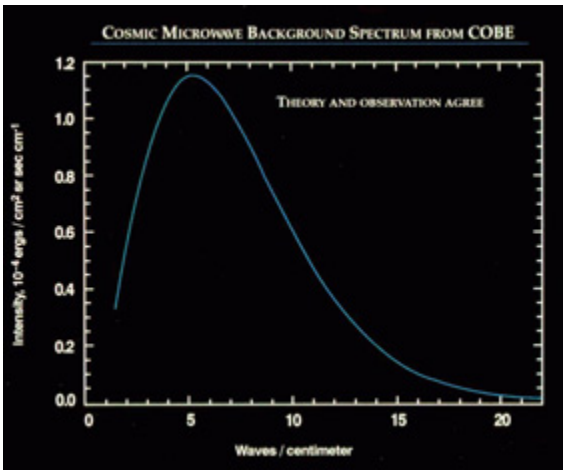


Figure 5: Spectrum of the cosmic microwave background radiation.
Source: © NASA, COBE.

Figure 5 shows the blackbody spectrum from a particularly interesting source: the universe. This is the spectrum of thermal radiation from space—the cosmic microwave background—taken by the Cosmic Background Explorer (COBE) satellite experiment. The radiation from space turns out to be the spectrum

of a blackbody at a temperature of 2.725 Kelvin. The peak of the spectrum occurs at a wavelength of about one millimeter, in the microwave regime. This radiation can be thought of as an echo of the primordial Big Bang.

Enter the quantum

In the final years of the 19th century, physicists attempted to understand the spectrum of blackbody radiation but theory kept giving absurd results. German physicist Max Planck finally succeeded in calculating the spectrum in December 1900. However, he had to make what he could regard only as a preposterous hypothesis. According to Maxwell's theory, radiation from a blackbody is emitted and absorbed by charged particles moving in the walls of the body, for instance by electrons in a metal. Planck modeled the electrons as charged particles held by fictitious springs. A particle moving under a spring force behaves like a harmonic oscillator. Planck found he could calculate the observed spectrum if he hypothesized that the energy of each harmonic oscillator could change only by discrete steps. If the frequency of the oscillator is ν (ν is the Greek letter "nu" and is often used to stand for frequency), then the energy had to be 0, 1 $h\nu$, 2 $h\nu$, 3 $h\nu$, ... n$h\nu$, ... where n could be any integer and h is a constant that soon became known as Planck's constant. Planck named the step $h\nu$ a quantum of energy. The blackbody spectrum Planck obtained by invoking his quantum hypothesis agreed beautifully with the experiment. But the quantum hypothesis seemed so absurd to Planck that he hesitated to talk about it.

Figure 6: Max Planck solved the blackbody problem by introducing quanta of energy.
Source: © The Clendening History of Medicine Library, University of Kansas Medical Center.

The physical dimension—the unit—of Planck's constant h is interesting. It is either [energy] / [frequency] or [angular momentum]. Both of these dimensions have important physical interpretations. The constant's value in S.I. units, 6.6×10^{-34} joule-seconds, suggests the enormous distance between the quantum world and everyday events.

Planck's constant is ubiquitous in quantum physics. The combination $h/2\pi$ appears so often that it has been given a special symbol called "hbar." This symbol appears in the upper-right-hand corner of these pages.

For five years, the quantum hypothesis had little impact. But in 1905, in what came to be called his *miracle year*, Swiss physicist Albert Einstein published a theory that proposed a quantum hypothesis from a totally different point of view. Einstein pointed out that, although Maxwell's theory was wonderfully successful in explaining the known phenomena of light, these phenomena involved light waves interacting

with large bodies. Nobody knew how light behaved on the microscopic scale—with individual electrons or atoms, for instance. Then, by a subtle analysis based on the analogy of certain properties of blackbody radiation with the behavior of a gas of particles, he concluded that electromagnetic energy itself must be quantized in units of $h\nu$. Thus, the light energy in a radiation field obeyed the same quantum law that Planck proposed for his fictitious mechanical oscillators; but Einstein's quantum hypothesis did not involve hypothetical oscillators.

An experimental test of the quantum hypothesis

Whereas Planck's theory led to no experimental predictions, Einstein's theory did. When light hits a metal, electrons can be ejected, a phenomenon called the photoelectric effect. According to Einstein's hypothesis, the energy absorbed by each electron had to come in bundles of light quanta. The minimum energy an electron could extract from the light beam is one quantum, $h\nu$. A certain amount of energy, W, is needed to remove electrons from a metal; otherwise they would simply flow out. So, Einstein predicted that the maximum kinetic energy of a photoelectron, E, had to be given by the equation $E = h\nu - W$.

The prediction is certainly counterintuitive, for Einstein predicted that E would depend only on the frequency of light, not on the light's intensity. The American physicist Robert A. Millikan set out to prove experimentally that Einstein must be wrong. By a series of painstaking experiments, however, Millikan convinced himself that Einstein must be right.

The quantum of light energy is called a photon. A photon possesses energy $h\nu$, and it carries momentum $h\nu$/c, where c is the speed of light. Photons are particle-like because they carry discrete energy and momentum. They are relativistic because they always travel at the speed of light and consequently can possess momentum even though they are massless.

Although the quantum hypothesis solved the problem of blackbody radiation, Einstein's concept of a light quantum—a particle-like bundle of energy—ran counter to common sense because it raised a profoundly troubling question: Does light consist of waves or particles? As we will show, answering this question required a revolution in physics. The issue was so profound that we should devote the next section to reviewing just what we mean by a wave and what we mean by a particle.

Section 3: Waves, Particles, and a Paradox

A particle is an object so small that its size is negligible; a wave is a periodic disturbance in a medium. These two concepts are so different that one can scarcely believe that they could be confused. In quantum physics, however, they turn out to be deeply intertwined and fundamentally inseparable.

Figure 7: A circular wave created by tossing a pebble in a pond.
Source: © Adam Kleppner.

The electron provides an ideal example of a particle because no attempt to measure its size has yielded a value different from zero. Clearly, an electron is small compared to an atom, while an atom is small compared to, for instance, a marble. In the night sky, the tiny points of starlight appear to come from luminous particles, and for many purposes we can treat stars as particles that interact gravitationally. It is evident that "small" is a relative term. Nevertheless, the concept of a particle is generally clear.

The essential properties of a particle are its mass, m; and, if it is moving with velocity v, its momentum, mv; and its kinetic energy, $1/2mv^2$. The energy of a particle remains localized, like the energy of a bullet, until it hits something. One could say, without exaggeration, that nothing could be simpler than a particle.

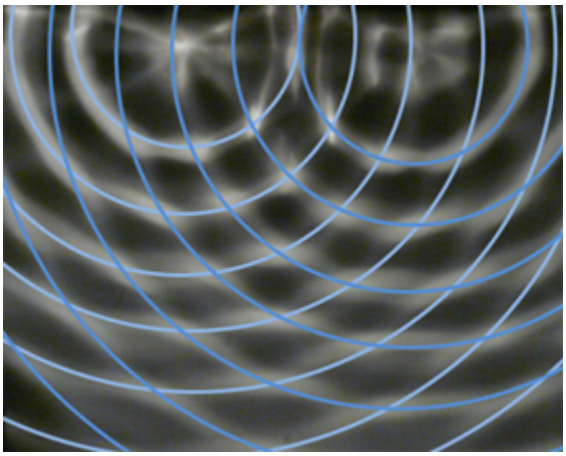


Figure 8: Two waves interfere as they cross paths.
Source: © Eli Sidman, Technical Services Group, MIT.

A wave is a periodic disturbance in a medium. Water waves are the most familiar example (we talk here about gentle waves, like ripples on a pond, not the breakers loved by surfers); but there are numerous other kinds, including sound waves (periodic oscillations of pressure in the air), light waves (periodic oscillations in the electromagnetic field), and the yet-to-be-detected gravitational waves (periodic oscillations in the gravitational field). The nature of the amplitude, or height of the wave, depends on the medium, for instance the pressure of air in a sound wave, the actual height in a water wave, or the electric field in a light wave. However, every wave is characterized by its wavelength λ (the Greek letter "lambda"), the distance from one crest to the next; its frequency ν (the Greek letter "nu"), the number of cycles or oscillations per second; and its velocity v, the distance a given crest moves in a second. This distance is the product of the number of oscillations the wave undergoes in a second and the wavelength.

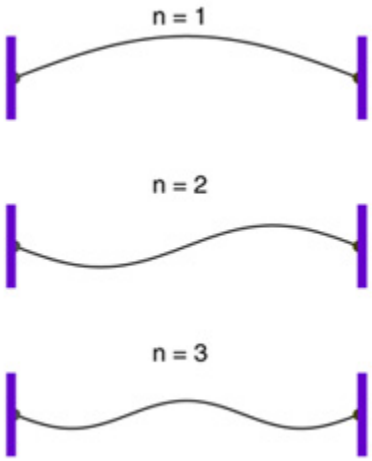


Figure 9: Standing waves on a string between two fixed endpoints.
Source: © Daniel Kleppner.

The energy in a wave spreads like the ripples traveling outward in Figure 7. A surprising property of waves is that they pass freely through each other: as they cross, their displacements simply add. The wave fronts retain their circular shape as if the other wave were not there. However, at the intersections of the circles marking the wave crests, the amplitudes add, producing a bright image. In between, the positive displacement of one wave is canceled by the negative displacement of the other. This phenomenon, called interference, is a fundamental property of waves. Interference constitutes a characteristic signature of wave phenomena.

If a system is constrained, for instance if the medium is a guitar string that is fixed at either end, the energy cannot simply propagate away. As a result, the pattern is fixed in space and it oscillates in time. Such a wave is called a standing wave.

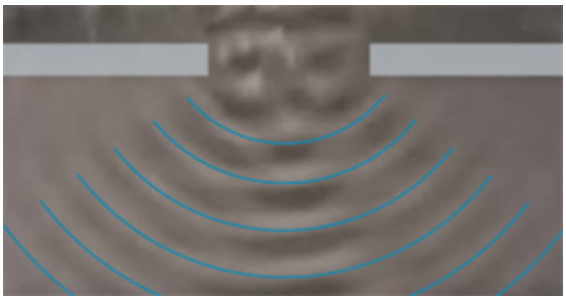


Figure 10: Ripple tank picture of plane waves incident on a slit that is about two wavelengths wide.
Source: © Eli Sidman, Technical Services Group, MIT.

Far from their source, in three dimensions, the wave fronts of a disturbance behave like equally spaced planes, and the waves are called plane waves. If plane waves pass through a slit, the emerging wave does not form a perfect beam but spreads, or diffracts, as in Figure 10. This may seem contrary to experience because light is composed of waves, but light waves do not seem to spread. Rather, light appears to travel in straight lines. This is because in everyday experience, light beams are formed by apertures that are many wavelengths wide. A 1 millimeter aperture, for instance, is about 2,000 wavelengths wide. In such a situation, diffraction is weak and spreading is negligible. However, if the slit is about a wavelength across, the emerging disturbance is not a sharp beam but a rapidly spreading wave, as in Figure 10. To see light diffract, one must use very narrow slits.

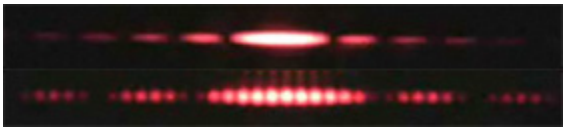


Figure 11: Diffraction of laser light through one (top) and two (bottom) small slits.
Source: © Eli Sidman, Technical Services Group, MIT.

If a plane wave passes through two nearby slits, the emerging beams can overlap and interfere. The points of interference depend only on the geometry and are fixed in space. The constructive interference creates a region of brightness, while destructive interference produces darkness. As a result, the photograph of light from two slits reveals bright and dark fringes, called "interference fringes." An example of two-slit interference is shown in Figure 11.

The paradox emerges

Diffraction, interference, and in fact all of the phenomena of light can be explained by the wave theory of light, Maxwell's theory. Consequently, there can be no doubt that light consists of waves. However, in Section 2 we described Einstein's conjecture that light consists of particle-like bundles of energy, and explained that the photoelectric effect provides experimental evidence that this is true. A single phenomenon that displays contrary properties creates a paradox.

Is it possible to reconcile these two descriptions? One might argue that the bundles of light energy are so small that their discreteness is unimportant. For instance, a one-watt light source, which is quite dim, emits over 10^{18} photons per second. The number of photons captured in visual images or the immages in digital cameras are almost astronomically large. One photon more or less would never make a difference.

However, we will see show examples where wave-like behavior is displayed by *single* particles. We will return to the wave-particle paradox later.

Section 4: Mysteries of Matter

Early in the 20th century, it was known that everyday matter consists of atoms and that atoms contain positive and negative charges. Furthermore, each type of atom, that is, each element, has a unique spectrum—a pattern of wavelengths the atom radiates or absorbs if sufficiently heated. A particularly important spectrum, the spectrum of atomic hydrogen, is shown in Figure 12. The art of measuring the wavelengths, spectroscopy, had been highly developed, and scientists had generated enormous quantities of precise data on the wavelengths of light emitted or absorbed by atoms and molecules.

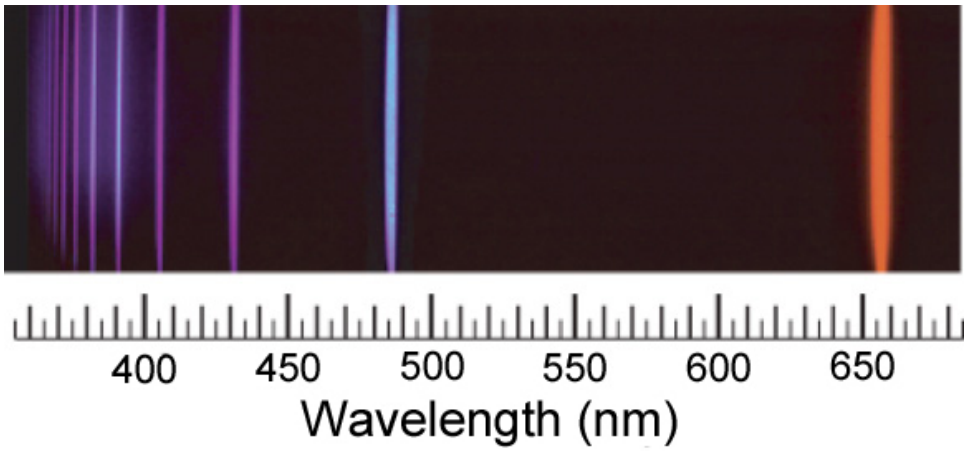


Figure 12: The spectrum of atomic hydrogen.
Source: © T.W. Hänsch.

In spite of the elegance of spectroscopic measurement, it must have been uncomfortable for scientists to realize that they knew essentially nothing about the structure of atoms, much less why they radiate and absorb certain colors of light. Solving this puzzle ultimately led to the creation of quantum mechanics, but the task took about 20 years.

The nuclear atom

In 1910, there was a major step in unraveling the mystery of matter: Ernest Rutherford realized that most of the mass of an atom is located in a tiny volume—a nucleus—at the center of the atom. The positively charged nucleus is surrounded by the negatively charged electrons. Rutherford was forced reluctantly to accept a planetary model of the atom in which electrons, electrically attracted to the nucleus, fly around the nucleus like planets gravitationally attracted to a star. However, the planetary model gave rise to a dilemma. According to Maxwell's theory of light, circling electrons radiate energy. The electrons would

generate light at ever-higher frequencies as they spiraled inward to the nucleus. The spectrum would be broad, not sharp. More importantly, the atom would collapse as the electrons crashed into the nucleus. Rutherford's discovery threatened to become a crisis for physics.

The Bohr model of hydrogen

Niels Bohr, a young scientist from Denmark, happened to be visiting Rutherford's laboratory and became intrigued by the planetary atom dilemma. Shortly after returning home Bohr proposed a solution so radical that even he could barely believe it. However, the model gave such astonishingly accurate results that it could not be ignored. His 1913 paper on what became known as the "Bohr model" of the hydrogen atom opened the path to the creation of quantum mechanics.

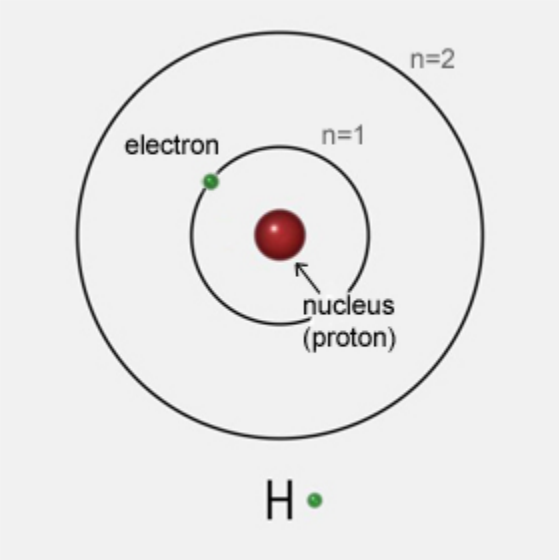


Figure 13: Bohr's model of an atom.
Source: © Daniel Kleppner.

Bohr proposed that—contrary to all the rules of classical physics—hydrogen atoms exist only in certain fixed energy states, called stationary states. Occasionally, an atom somehow jumps from one state to another by radiating the energy difference. If an atom jumps from state b with energy E_b to state a with lower energy, E_a, it radiates light with frequency ν given by $h\nu = E_b - E_a$. Today, we would say that the atom emits a photon when it makes a quantum jump. The reverse is possible: An atom in a lower energy state can absorb a photon with the correct energy and make a transition to the higher state. Each energy state would be characterized by an integer, now called a quantum number, with the lowest energy state described by n = 1.

Bohr's ideas were so revolutionary that they threatened to upset all of physics. However, the theories of physics, which we now call "classical physics," were well tested and could not simply be dismissed. So, to connect his wild proposition with reality, Bohr introduced an idea that he later named the Correspondence Principle. This principle holds that there should be a smooth transition between the quantum and classical worlds. More precisely, in the limit of large energy state quantum numbers, atomic systems should display classical-like behavior. For example, the jump from a state with quantum number n = 100 to the state n = 99 should give rise to radiation at the frequency of an electron circling a proton with approximately the energy of those states. With these ideas, and using only the measured values of a few fundamental constants, Bohr calculated the spectrum of hydrogen and obtained astonishing agreement with observations.

Bohr understood very well that his theory contained too many radical assumptions to be intellectually satisfying. Furthermore, it left numerous questions unanswered, such as why atoms make quantum jumps. The fundamental success of Bohr's model of hydrogen was to signal the need to replace classical physics with a totally new theory. The theory should be able to describe behavior at the microscopic scale—atomic behavior—but it should also be in harmony with classical physics, which works well in the world around us.

Matter waves

By the end of the 1920s, Bohr's vision of a new theory was fulfilled by the creation of quantum mechanics, which turned out to be strange and even disturbing.

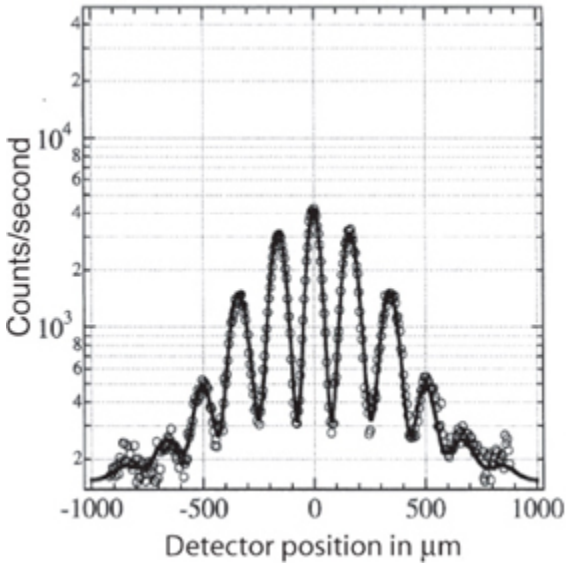


Figure 14: This diffraction pattern appeared when a beam of sodium molecules encountered a series of small slits, showing their wave-like nature.
Source: © D.E. Pritchard.

A key idea in the development of quantum mechanics came from the French physicist Louis de Broglie. In his doctoral thesis in 1924, de Broglie suggested that if waves can behave like particles, as Einstein had shown, then one might expect that particles can behave like waves. He proposed that a particle with momentum p should be associated with a wave of wavelength $\lambda = h/p$, where, as usual, h stands for Planck's constant. The question "Waves of what?" was left unanswered.

de Broglie's hypothesis was not limited to simple particles such as electrons. Any system with momentum p, for instance an atom, should behave like a wave with its particular de Broglie wavelength. The proposal must have seemed absurd because in the entire history of science, nobody had ever seen anything like a de Broglie wave. The reason that nobody had ever seen a de Broglie wave, however, is simple: Planck's constant is so small that the de Broglie wavelength for observable everyday objects is much too small to be noticeable. But for an electron in hydrogen, for instance, the deBroglie wavelength is about the size of the atom.

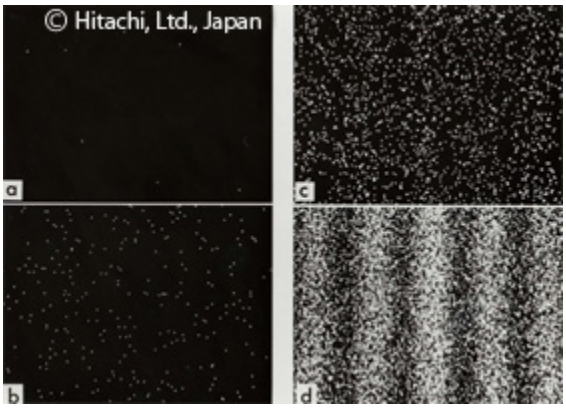


Figure 15: An interference pattern builds up as individual electrons pass through two slits.
Source: © Reprinted courtesy of Dr. Akira Tonomura, Hitachi, Ltd., Japan.

Today, de Broglie waves are familiar in physics. For example, the diffraction of particles through a series of slits (see Figure 14) looks exactly like the interference pattern expected for a light wave through a series of slits. The signal, however, is that of a matter wave—the wave of a stream of sodium molecules. The calculated curve (solid line) is the interference pattern for a wave with the de Broglie wavelength of sodium molecules, which are diffracted by slits with the measured dimensions. The experimental points are the counts from an atom (or molecule) detector. The stream of particles behaves exactly like a wave.

The concept of a de Broglie wave raises troubling issues. For instance, for de Broglie waves one must ask: Waves of what? Part of the answer is provided in the two-slit interference data in Figure 15. The particles in this experiment are electrons. Because the detector is so sensitive, the position of every single electron can be recorded with high efficiency. Panel (a) displays only eight electrons, and they appear to be randomly scattered. The points in panels (b) and (c) also appear to be randomly scattered. Panel (d) displays 60,000 points, and these are far from randomly distributed. In fact, the image is a traditional two-slit interference pattern. This suggests that the probability that an electron arrives at a given position is proportional to the intensity of the interference pattern there. It turns out that this suggestion provides a useful interpretation of a quantum wavefunction: The probability of finding a particle at a given position is proportional to the intensity of its wavefunction there, that is to the square of the wavefunction.

Section 5: Introducing Quantum Mechanics

As we saw in the previous section, there is strong evidence that atoms can behave like waves. So, we shall take the wave nature of atoms as a fact and turn to the questions of how matter waves behave and what they mean.

Figure 16: Werner Heisenberg (left) and Erwin Schrödinger (right).
Source: © Left: German Federal Archive, Creative Commons Attribution ShareAlike 3.0 Germany License. Right: Francis Simon, courtesy AIP Emilio Segrè Visual Archives.

Mathematically, waves are described by solutions to a differential equation called the "wave equation." In 1925, the Austrian physicist Erwin Schrödinger reasoned that since particles can behave like waves, there must be a wave equation for particles. He traveled to a quiet mountain lodge to discover the equation; and after a few weeks of thinking and skiing, he succeeded. Schrödinger's equation opened the door to the quantum world, not only answering the many paradoxes that had arisen, but also providing a method for calculating the structure of atoms, molecules, and solids, and for understanding the structure of all matter. Schrödinger's creation, called wave mechanics, precipitated a genuine revolution in science. Almost simultaneously, a totally different formulation of quantum theory was created by Werner Heisenberg: matrix mechanics. The two theories looked different but turned out to be fundamentally equivalent. Often, they are simply referred to as "quantum mechanics." Schrödinger and Heisenberg were awarded the Nobel Prize in 1932 for their theories.

In wave mechanics, our knowledge about a system is embodied in its wavefunction. A wavefunction is the solution to Schrödinger's equation that fits the particular circumstances. For instance, one can speak of the wavefunction for a particle moving freely in space, or an electron bound to a proton in a hydrogen atom, or a mass moving under the spring force of a harmonic oscillator.

To get some insight into the quantum description of nature, let's consider a mass M, moving in one dimension, bouncing back and forth between two rigid walls separated by distance L. We will refer to this idealized one-dimensional system as a particle in a box. The wavefunction must vanish outside the box because the particle can never be found there. Physical waves cannot jump abruptly, so the wavefunction must smoothly approach zero at either end of the box. Consequently, the box must contain an integral number of half-wavelengths of the particle's de Broglie wave. Thus, the de Broglie wavelength λ must obey $n\lambda/2 = L$, where L is the length of the box and n = 1, 2, 3... . The integer n is called the quantum number of the state. Once we know the de Broglie wavelength, we also know the particle's momentum and energy. See the math

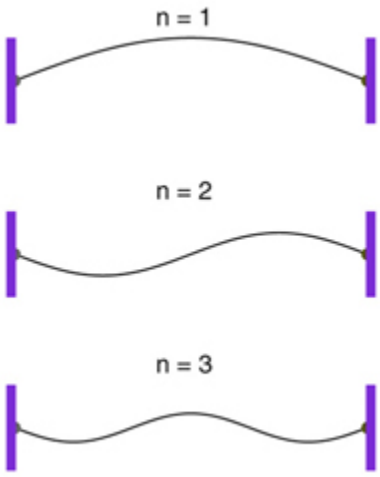


Figure 17: The first three allowed de Broglie wave modes for a particle in a box.
Source: © Daniel Kleppner.

The mere existence of matter waves suggests that in any confined system, the energy can have only certain discrete values, that is the energy is quantized. The minimum energy is called the ground state energy. For the particle in the box, the ground state energy is $(hL)^2/8M$. The energy of the higher-lying states increases as n^2. For a harmonic oscillator, it turns out that the energy levels are equally spaced, and the allowed energies increase linearly with n. For a hydrogen atom, the energy levels are found to get closer and closer as n increases, varying as $1/n^2$.

If this is your first encounter with quantum phenomena, you may be confused as to what the wavefunction means and what connection it could have with the behavior of a particle. Before discussing the

interpretation, it will be helpful to look at the wavefunction for a system slightly more interesting than a particle in a box.

The harmonic oscillator

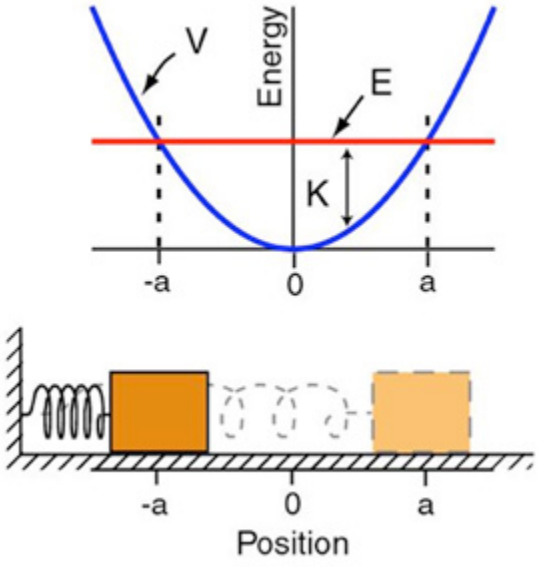


Figure 18: A simple harmonic oscillator (bottom) and its energy diagram (top).
Source: © Daniel Kleppner.

In free space, where there are no forces, the momentum and kinetic energy of a particle are constant. In most physically interesting situations, however, a particle experiences a force. A harmonic oscillator is a particle moving under the influence of a spring force as shown in Figure 18. The spring force is proportional to how far the spring is stretched or compressed away from its equilibrium position, and the particle's potential energy is proportional to that distance squared. Because energy is conserved, the total energy, E = K + V, is constant. These relations are shown in the energy diagram in Figure 18.

The energy diagram in Figure 18 is helpful in understanding both classical and quantum behavior. Classically, the particle moves between the two extremes (-a, a) shown in the drawing. The extremes are called "turning points" because the direction of motion changes there. The particle comes to momentary rest at a turning point, the kinetic energy vanishes, and the potential energy is equal to the total energy. When the particle passes the origin, the potential energy vanishes, and the kinetic energy is equal to the total energy. Consequently, as the particle moves back and forth, its momentum oscillates between zero and its maximum value.

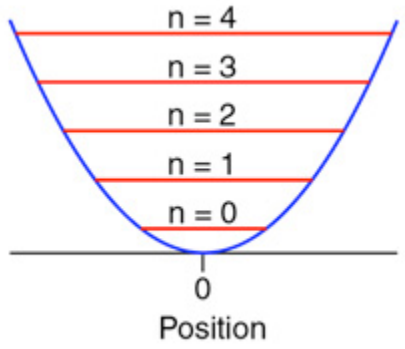


Figure 19: Low-lying energy levels of a harmonic oscillator.
Source: © Daniel Kleppner.

Solutions to Schrödinger's equation for the harmonic oscillator show that the energy is quantized, as we expect for a confined system, and that the allowed states are given by $E_n = (n + 1/2)h\nu$, where ν is the frequency of the oscillator and n = 0, 1, 2... . The energy levels are separated by $h\nu$, as Planck had conjectured, but the system has a ground state energy 1/2 $h\nu$, which Planck could not have known about. The harmonic oscillator energy levels are evenly spaced, as shown in Figure 19.

What does the wavefunction mean?

If we measure the position of the mass, for instance by taking a flash photograph of the oscillator with a meter stick in the background, we do not always get the same result. Even under ideal conditions, which includes eliminating thermal fluctuations by working at zero temperature, the mass would still jitter due to its zero point energy. However, if we plot the results of successive measurements, we find that they start to look reasonably orderly. In particular, the fraction of the measurements for which the mass is in some interval, s, is proportional to the area of the strip of width s lying under the curve in Figure 20, shown in blue. This curve is called a probability distribution curve. Since the probability of finding the mass somewhere is unity, the height of the curve must be chosen so that the area under the curve is 1. With this convention, the probability of finding the mass in the interval s is equal to the area of the shaded strip. It turns out that the probability distribution is simply the wavefunction squared.

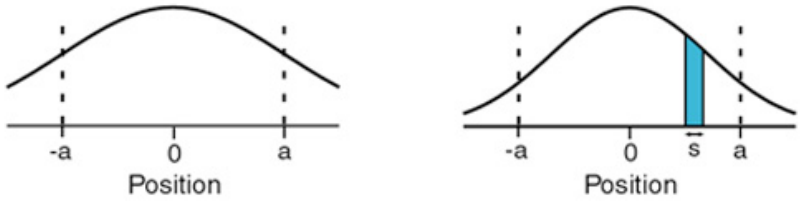


Figure 20: The ground state wavefunction of a harmonic oscillator (left) and the corresponding probability distribution (right).
Source: © Daniel Kleppner.

Here, we have a curious state of affairs. In classical physics, if one knows the state of a system, for instance the position and speed of a marble at rest, one can predict the result of future measurements as precisely as one wishes. In quantum mechanics, however, the harmonic oscillator cannot be truly at rest: The closest it can come is the ground state energy 1/2 $h\nu$. Furthermore, we cannot predict the precise result of measurements, only the probability that a measurement will give a result in a given range. Such a probabilistic theory was not easy to accept at first. In fact, Einstein never accepted it.

Aside from its probabilistic interpretation, Figure 20 portrays a situation that could hardly be less like what we expect from classical physics. A classical harmonic oscillator moves fastest near the origin and spends most of its time as it slows down near the turning points. Figure 20 suggests the contrary: The most likely place to find the mass is at the origin where it is moving fastest. However, there is an even more bizarre aspect to the quantum solution: The wavefunction extends *beyond* the turning points. This means that in a certain fraction of measurements, the mass will be found in a place where it could never go if it obeyed the classical laws. The penetration of the wavefunction into the classically forbidden region gives rise to a purely quantum phenomenon called tunneling. If the energy barrier is not too high, for instance if the energy barrier is a thin layer of insulator in a semiconductor device, then a particle can pass from one classically allowed region to another, tunneling through a region that is classically forbidden.

The quantum description of a harmonic oscillator starts to look a little more reasonable for higher-lying states. For instance, the wavefunction and probability distribution for the state n = 10 are shown in Figure 21.

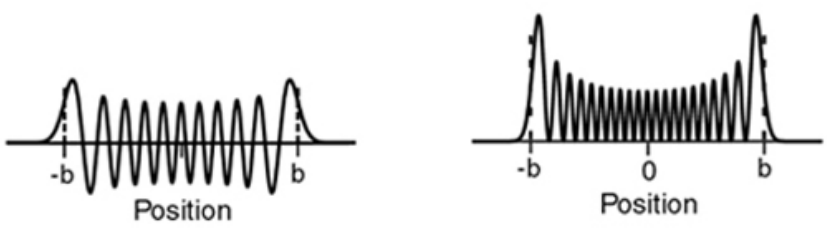


Figure 21: The wavefunction (left) and probability distribution (right) of a harmonic oscillator in the state n = 10.
Source: © Daniel Kleppner.

Although the n = 10 state shown in Figure 21 may look weird, it shows some similarities to classical behavior. The mass is most likely to be observed near a turning point and least likely to be seen near the origin, as we expect. Furthermore, the fraction of time it spends outside of the turning points is much less than in the ground state. Aside from these clues, however, the quantum description appears to have no connection to the classical description of a mass oscillating in a real harmonic oscillator. We turn next to showing that such a connection actually exists.

Section 6: The Uncertainty Principle

The idea of the position of an object seems so obvious that the concept of position is generally taken for granted in classical physics. Knowing the position of a particle means knowing the values of its coordinates in some coordinate system. The precision of those values, in classical physics, is limited only by our skill in measuring. In quantum mechanics, the concept of position differs fundamentally from this classical meaning. A particle's position is summarized by its wavefunction. To describe a particle at a given position in the language of quantum mechanics, we would need to find a wavefunction that is extremely high near that position and zero elsewhere. The wavefunction would resemble a very tall and very thin tower. None of the wavefunctions we have seen so far look remotely like that. Nevertheless, we can construct a wavefunction that approximates the classical description as precisely as we please.

Let's take the particle in a box described in Section 4 as an example. The possible wavefunctions, each labeled by an integer quantum number, n, obey the superposition principle, and so we are free to add solutions with different values of n, adjusting the amplitudes as needed. The sum of the individual wavefunctions yields another legitimate wavefunction that could describe a particle in a box. If we're clever, we can come up with a combination that resembles the classical solution. If, for example, we add a series of waves with n = 1, 3, 5, and 7 and the carefully chosen amplitudes shown in Figure 22, the result appears to be somewhat localized near the center of the box.

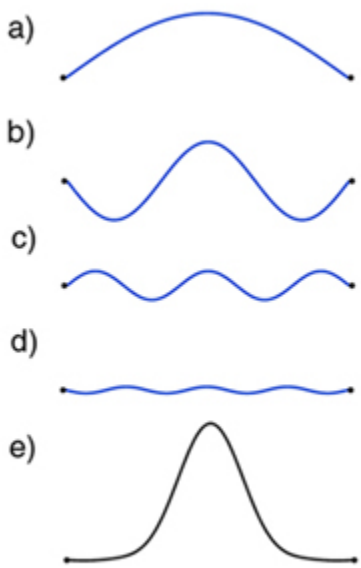


Figure 22: Some wavefunctions for a particle in a box. Curve (e) is the sum of curves (a-d).
Source: © Daniel Kleppner.

Localizing the particle has come at a cost, however, because each wave we add to the wavefunction corresponds to a different momentum. If the lowest possible momentum is p_0, then the wavefunction we created has components of momentum at p_0, $3p_0$, $5p_0$, and $7p_0$. If we measure the momentum, for instance, by suddenly opening the ends of the box and measuring the time for the particle to reach a detector, we would observe one of the four possible values. If we repeat the measurement many times and plot the results, we would find that the probability for a particular value is proportional to the square of the amplitude of its component in the wavefunction.

If we continue to add waves of ever-shortening wavelengths to our solution, the probability curve becomes narrower while the spread of momentum increases. Thus, as the wavefunction sharpens and our uncertainty about the particle's position decreases, the spread of values observed in successive measurements, that is, the uncertainty in the particle's momentum, increases.

This state of affairs may seem unnatural because energy is not conserved: Often, the particle is observed to move slowly but sometimes it is moving very fast. However, there is no reason energy should be

conserved because the system must be freshly prepared before each measurement. The preparation process requires that the particle has the given wavefunction before each measurement. All the information that we have about the state of a particle is in its wavefunction, and this information does not include a precise value for the energy.

The reciprocal relation between the spread in repeated measurements of position and momentum was first recognized by Werner Heisenberg. If we denote the scatter in results for repeated measurements of a position of a particle by Δx (Δ, Greek letter "delta"), and the scatter in results in repeated measurements of the momentum by Δp, then Heisenberg showed that $\Delta x \Delta p \geq h/4\pi$, a result famously known as the Heisenberg uncertainty principle. The uncertainty principle means that in quantum mechanics, we cannot simultaneously know both the position and the momentum of an object arbitrarily well.

Measurements of certain other quantities in quantum mechanics are also governed by uncertainty relations. An important relation for quantum measurements relates the uncertainty in measurements of the energy of a system, ΔE, to the time τ (τ, Greek letter "tau") during which the measurement is made: $\tau \Delta E \geq h/4\pi$.

Some illustrations of the uncertainty principle

Harmonic oscillator. The ground state energy of the harmonic oscillator, $1/2h\nu$, makes immediate sense from the uncertainty principle. If the ground state of the oscillator were more highly localized, that is sharper than in Figure 20, the oscillator's average potential energy would be lower. However, sharpening the wavefunction requires introducing shorter wavelength components. These have higher momentum, and thus higher kinetic energy. The result would be an increase in the total energy. The ground state represents the optimum trade-off between decreasing the potential energy and increasing the kinetic energy.

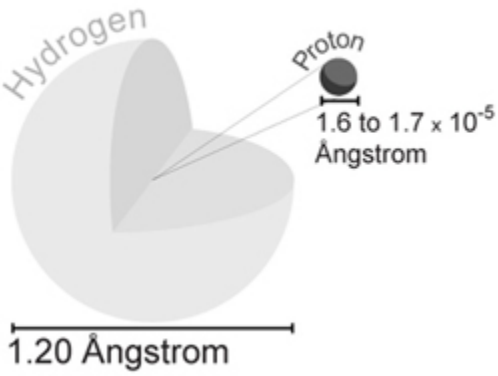


Figure 23: The size of a hydrogen atom is determined by the uncertainty principle.
Source: © Wikimedia Commons, Public Domain. Author: Bensaccount, 10 July 2006.

Hydrogen atom. The size of a hydrogen atom also represents a trade-off between potential and kinetic energy, dictated by the uncertainty principle. If we think of the electron as smeared over a spherical volume, then the smaller the radius, the lower the potential energy due to the electron's interaction with the positive nucleus. However, the smaller the radius, the higher the kinetic energy arising from the electron's confinement. Balancing these trade-offs yields a good estimate of the actual size of the atom. The mean radius is about 0.05 nm.

Natural linewidth. The most precise measurements in physics are frequency measurements, for instance the frequencies of radiation absorbed or radiated in transitions between atomic stationary states. Atomic clocks are based on such measurements. If we designate the energy difference between two states by E, then the frequency of the transition is given by Bohr's relation: $E = h\nu$. An uncertainty in energy ΔE leads to an uncertainty in the transition frequency given by $\Delta E = h\Delta\nu$. The time-energy uncertainty principle can be written $\Delta E \geq h/(4\pi\tau)$, where τ is the time during which the measurement is made. Combining these, we find that the uncertainty in frequency is $\Delta\nu \geq 1/(4\pi\tau)$.

It is evident that the longer the time for a frequency measurement, the smaller the possible uncertainty. The time τ may be limited by experimental conditions, but even under ideal conditions τ would still be limited. The reason is that an atom in an excited state eventually radiates to a lower state by a process called spontaneous emission. This is the process that causes quantum jumps in the Bohr model. Spontaneous emission causes an intrinsic energy uncertainty, or width, to an energy level. This width is

called the natural linewidth of the transition. As a result, the energies of all the states of a system, except for the ground states, are intrinsically uncertain. One might think that this uncertainty fundamentally precludes accurate frequency measurement in physics. However, as we shall see, this is not the case.

Figure 24: Gerald Gabrielse (left) is shown with the apparatus he used to make some of the most precise measurements of a single electron.
Source: © Gerald Gabrielse.

Myths about the uncertainty principle. Heisenberg's uncertainty principle is among the most widely misunderstood principles of quantum physics. Non-physicists sometimes argue that it reveals a fundamental shortcoming in science and poses a limitation to scientific knowledge. On the contrary, the uncertainty principle is seminal to quantum measurement theory, and quantum measurements have achieved the highest accuracy in all of science. It is important to appreciate that the uncertainty principle does not limit the precision with which a physical property, for instance a transition frequency, can be measured. What it does is to predict the scatter of results of a single measurement. By repeating the measurements, the ultimate precision is limited only by the skill and patience of the experimenter. Should there be any doubt about whether the uncertainty principle limits the power of precision in physics, measurements made with the apparatus shown in Figure 24 should put them to rest. The experiment confirmed the accuracy of a basic quantum mechanical prediction to an accuracy of one part in 10^{12}, one of the most accurate tests of theory in all of science.

The uncertainty principle and the world about us

Because the quantum world is so far from our normal experience, the uncertainty principle may seem remote from our everyday lives. In one sense, the uncertainty principle really is remote. Consider, for instance, the implications of the uncertainty principle for a baseball. Conceivably, the baseball could fly off unpredictably due to its intrinsically uncertain momentum. The more precisely we can locate the baseball in space, the larger is its intrinsic momentum. So, let's consider a pitcher who is so sensitive that he can tell if the baseball is out of position by, for instance, the thickness of a human hair, typically 0.1 mm or 10^{-4} m. According to the uncertainty principle, the baseball's intrinsic speed due to quantum effects is about 10^{-29} m/s. This is unbelievably slow. For instance, the time for the baseball to move quantum mechanically merely by the diameter of an atom would be roughly 20 times the age of the universe. Obviously, whatever might give a pitcher a bad day, it will not be the uncertainty principle.

Figure 25: The effect of quantum mechanical jitter on a pitcher, luckily, is too small to be observable.
Source: © Clive Grainger, 2010.

Nevertheless, effects of the uncertainty principle are never far off. Our world is composed of atoms and molecules; and in the atomic world, quantum effects rule everything. For instance, the uncertainty principle prevents electrons from crashing into the nucleus of an atom. As an electron approaches a nucleus under the attractive Coulomb force, its potential energy falls. However, localizing the electron

near the nucleus requires the sharpening of its wavefunction. This sharpening causes the electron's momentum spread to get larger and its kinetic energy to increase. At some point, the electron's total energy would start to increase. The quantum mechanical balance between the falling potential energy and rising kinetic energy fixes the size of the atom. If we magically turned off the uncertainty principle, atoms would vanish in a flash. From this point of view, you can see the effects of the uncertainty principle everywhere.

Section 7: Atom Cooling and Trapping

The discovery that laser light can cool atoms to less than a millionth of a degree above absolute zero opened a new world of quantum physics. Previously, the speeds of atoms due to their thermal energy were always so high that their de Broglie wavelengths were much smaller than the atoms themselves. This is the reason why gases often behave like classical particles rather than systems of quantum objects. At ultra-low temperatures, however, the de Broglie wavelength can actually exceed the distance between the atoms. In such a situation, the gas can abruptly undergo a quantum transformation to a state of matter called a Bose-Einstein condensate. The properties of this new state are described in Unit 6. In this section, we describe some of the techniques for cooling and trapping atoms that have opened up a new world of ultracold physics. The atom-cooling techniques enabled so much new science that the 1997 Nobel Prize was awarded to three of the pioneers: Steven Chu, Claude Cohen-Tannoudji, and William D. Phillips.

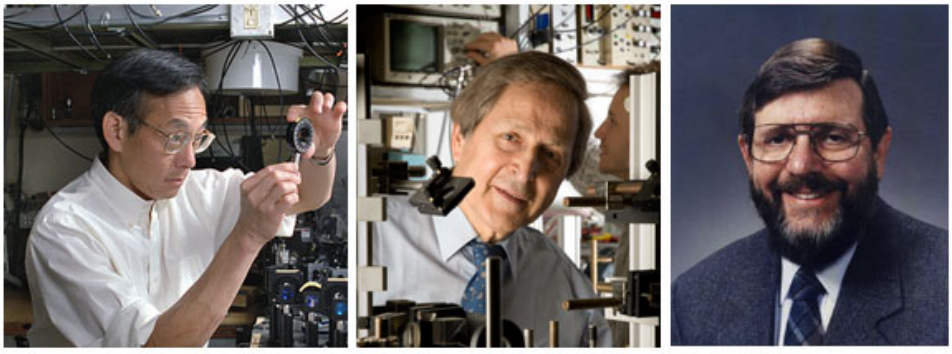


Figure 26: Recipients of the 1997 Nobel Prize, for laser cooling and trapping of atoms.
Source: © Left: Steven Chu, Stanford University; Middle: Claude Cohen-Tannoudji, Jean-Francois DARS, Laboratoire Kastler Brossel; Right: William D. Phillips, NIST.

Doppler cooling

As we learned earlier, a photon carries energy and momentum. An atom that absorbs a photon recoils from the momentum kick, just as you experience recoil when you catch a ball. Laser cooling manages the momentum transfer so that it constantly slows the atom's motion, slowing it down. In absorbing a photon, the atom makes a transition from its ground state to a higher energy state. This requires that the photon has just the right energy. Fortunately, lasers can be tuned to precisely match the difference between energy levels in an atom. After absorbing a photon, an atom does not remain in the excited state but returns to the ground state by a process called spontaneous emission, emitting a photon in the

process. At optical wavelengths, the process is quick, typically taking a few tens of nanoseconds. The atom recoils as it emits the photon, but this recoil, which is opposite to the direction of photon emission, can be in any direction. As the atom undergoes many cycles of absorbing photons from one direction followed by spontaneously emitting photons in random directions, the momentum absorbed from the laser beam accumulates while the momentum from spontaneous emission averages to zero.

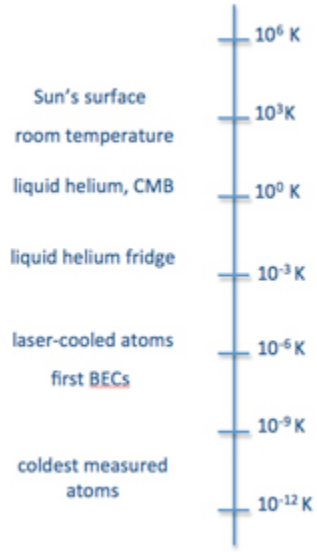


Figure 27: Temperature scale in physics
Source:

This diagram of temperatures of interest in physics uses a scale of factors of 10 (a logarithmic scale). On this scale, the difference between the Sun's surface temperature and room temperature is a small fraction of the range of temperatures opened by the invention of laser cooling. Temperature is in the Kelvin scale at which absolute zero would describe particles in thermal equilibrium that are totally at rest. The lowest temperature measured so far by measuring the speeds of atoms is about 450 picokelvin (one picokelvin is 10^{-12} K). This was obtained by evaporating atoms in a Bose-Einstein condensate.

The process of photon absorption followed by spontaneous emission can heat the atoms just as easily as cool them. Cooling is made possible by a simple trick: Tune the laser so that its wavelength is slightly too long for the atoms to absorb. In this case, atoms at rest cannot absorb the light. However, for an atom moving toward the laser, against the direction of the laser beam, the wavelength appears to be slightly shortened due to the Doppler effect. The wavelength shift can be enough to permit the atom to absorb the light. The recoil slows the atom's motion. To slow motion in the opposite direction, away from the light

source, one merely needs to employ a second laser beam, opposite to the first. These two beams slow atoms moving along a single axis. To slow atoms in three dimensions, six beams are needed (Figure 28). This is not as complicated as it may sound: All that is required is a single laser and mirrors.

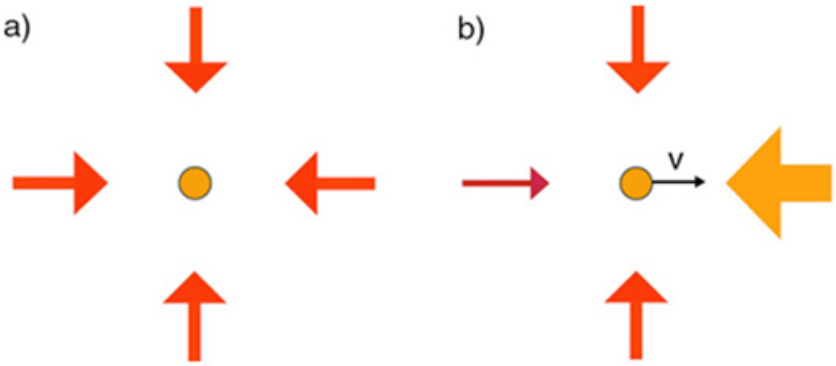


Figure 28: Red-detuned lasers don't affect an atom at rest (left) but will slow an atom moving toward the light source (right).
Source: © Daniel Kleppner.

Laser light is so intense that an atom can be excited just as soon as it gets to the ground state. The resulting acceleration is enormous, about 10,000 times the acceleration of gravity. An atom moving with a typical speed in a room temperature gas, thousands of meters per second, can be brought to rest in a few milliseconds. With six laser beams shining on them, the atoms experience a strong resistive force no matter which way they move, as if they were moving in a sticky fluid. Such a situation is known as optical molasses.

Numbers: time for atom cooling

A popular atom for laser cooling is rubidium-87. Its mass is $m = 1.45 \times 10^{-25}$ kg. The wavelength for excitation is $\lambda = 780$ nm. The momentum carried by the photon is $p = h\nu/c = h/\lambda$, and the change in the atom's velocity from absorbing a photon is $\Delta v = p/m = 5.9 \times 10^{-3}$ m/s. The lifetime for spontaneous emission is 26×10^{-9} s, and the average time between absorbing photons is about $t_{abs} = 52 \times 10^{-9}$ s. Consequently, the average acceleration is $a = \Delta v / t_{abs} = 1.1 \times 10^{5}$ m/s^2, which is about 10,000 times the acceleration of gravity. At room temperature, the rubidium atom has a mean thermal speed of $v_{th} = 290$ m/s. The time for the atom to come close to rest is $v_{th}/a = 2.6 \times 10^{-3}$ s.

As one might expect, laser cooling cannot bring atoms to absolute zero. The limit of Doppler cooling is actually set by the uncertainty principle, which tells us that the finite lifetime of the excited state due to spontaneous emission causes an uncertainty in its energy. This blurring of the energy level causes a spread in the frequency of the optical transition called the natural linewidth. When an atom moves so slowly that its Doppler shift is less than the natural linewidth, cooling comes to a halt. The temperature at which this occurs is known as the Doppler cooling limit. The theoretical predictions for this temperature are in the low millikelvin regime. However, by great good luck, it turned out that the actual temperature limit was lower than the theoretical prediction for the Doppler cooling limit. Sub-Doppler cooling, which depends on the polarization of the laser light and the spin of the atoms, lowers the temperature of atoms down into the microkelvin regime.

Atom traps

Like all matter, ultracold atoms fall in a gravitational field. Even optical molasses falls, though slowly. To make atoms useful for experiments, a strategy is needed to support and confine them. Devices for confining and supporting isolated atoms are called "atom traps." Ultracold atoms cannot be confined by material walls because the lowest temperature walls might just as well be red hot compared to the temperature of the atoms. Instead, the atoms are trapped by force fields. Magnetic fields are commonly used, but optical fields are also employed.

Magnetic traps depend on the intrinsic magnetism that many atoms have. If an atom has a magnetic moment, meaning that it acts as a tiny magnet, its energy is altered when it is put in a magnetic field. The

change in energy was first discovered by examining the spectra of atoms in magnetic fields and is called the Zeeman effect after its discoverer, the Dutch physicist Pieter Zeeman.

Because of the Zeeman effect, the ground state of alkali metal atoms, the most common atoms for ultracold atom research, is split into two states by a magnetic field. The energy of one state increases with the field, and the energy of the other decreases. Systems tend toward the configuration with the lowest accessible energy. Consequently, atoms in one state are repelled by a magnetic field, and atoms in the other state are attracted. These energy shifts can be used to confine the atoms in space.

The MOT

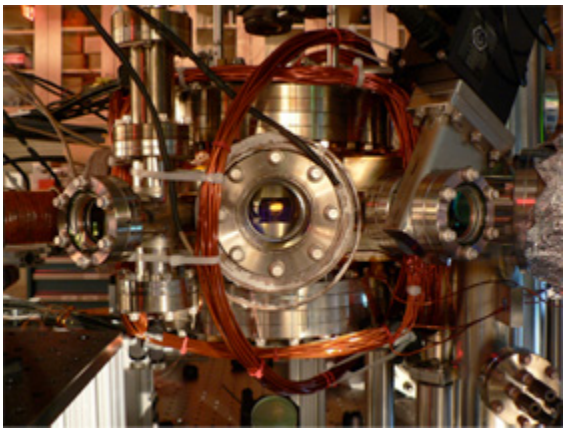


Figure 29: Atoms trapped in a MOT.
Source: © Martin Zwierlein.

The magneto-optical trap, or MOT, is the workhorse trap for cold atom research. In the MOT, a pair of coils with currents in opposite direction creates a magnetic field that vanishes at the center. The field points inward along the z-axis but outward along the x- and y-axes. Atoms in a vapor are cooled by laser beams in the same configuration as optical molasses, centered on the midpoint of the system. The arrangement by itself could not trap atoms because, if they were pushed inward along one axis, they would be pushed outward along another. However, by employing a trick with the laser polarization, it turns out that the atoms can be kept in a state that is pushed inward from every direction. Atoms that drift into the MOT are rapidly cooled and trapped, forming a small cloud near the center.

To measure the temperature of ultracold atoms, one turns off the trap, letting the small cloud of atoms drop. The speeds of the atoms can be found by taking photographs of the ball and measuring how rapidly it expands as it falls. Knowing the distribution of speeds gives the temperature. It was in similar

experiments that atoms were sometimes found to have temperatures below the Doppler cooling limit, not in the millikelvin regime, but in the microkelvin regime. The reason turned out to be an intricate interplay of the polarization of the light with the Zeeman states of the atom causing a situation known as the Sisyphus effect. The experimental discovery and the theoretical explanation of the "Sisyphus effect" were the basis of the Nobel Prize to Chu, Cohen-Tannoudji, and Phillips in 1997.

Evaporative cooling

When the limit of laser cooling is reached, the old-fashioned process of evaporation can cool a gas further. In thermal equilibrium, atoms in a gas have a broad range of speeds. At any instant, some atoms have speeds much higher than the average, and some are much slower. Atoms that are energetic enough to fly out of the trap escape from the system, carrying away their kinetic energy. As the remaining atoms collide and readjust their speeds, the temperature drops slightly. If the trap is slowly adjusted so that it gets weaker and weaker, the process continues and the temperature falls. This process has been used to reach the lowest kinetic temperatures yet achieved, a few hundred picokelvin. Evaporative cooling cannot take place in a MOT because the constant interaction between the atoms and laser beams keeps the temperature roughly constant. To use this process to reach temperatures less than a billionth of a degree above absolute zero, the atoms are typically transferred into a trap that is made purely of magnetic fields.

Optical traps

Atoms in light beams experience forces even if they don't actually absorb or radiate photons. The forces are attractive or repulsive depending on whether the laser frequency is below or above the transition frequency. These forces are much weaker than photon recoil forces, but if the atoms are cold enough, they can be large enough to confine them. For instance, if an intense light beam is turned on along the axis of a MOT that holds a cloud of cold atoms, the MOT can be turned off, leaving the atoms trapped in the light beam. Unperturbed by magnetic fields or by photon recoil, for many purposes, the environment is close to ideal. This kind of trap is called an optical dipole trap.

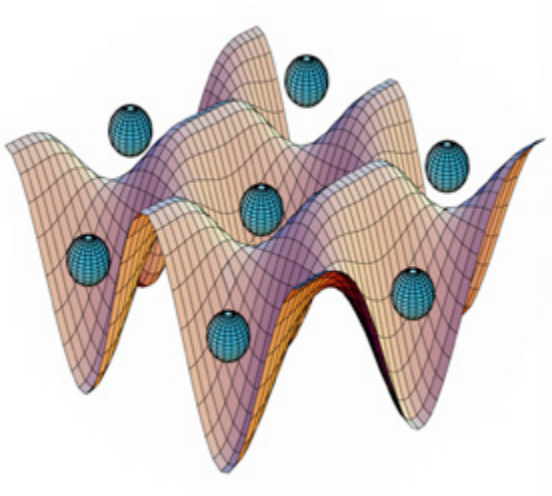


Figure 30: Atoms trapped in an optical lattice.
Source: © NIST.

If the laser beam is reflected back on itself to create a standing wave of laser light, the standing wave pattern creates a regular array of areas where the optical field is strong and weak known as an optical lattice. Atoms are trapped in the regions of the strong field. If the atoms are tightly confined in a strong lattice and the lattice is gradually made weaker, the atoms start to tunnel from one site to another. At some point the atoms move freely between the sites. The situation is similar to the phase transition in a material that abruptly turns from an insulator into a conductor. This is but one of many effects that are well known in materials and can now be studied using ultracold atoms that can be controlled and manipulated with a precision totally different from anything possible in the past.

Why the excitement?

The reason that ultracold atoms have generated enormous scientific excitement is that they make it possible to study basic properties of matter with almost unbelievable clarity and control. These include phase transitions to exotic states of matter such as superfluidity and superconductivity that we will learn about in Unit 8, and theories of quantum information and communication that are covered in Unit 7. There are methods for controlling the interactions between ultracold atoms so that they can repel or attract each other, causing quantum changes of state at will. These techniques offer new inroads to quantum entanglement—a fundamental behavior that lies beyond this discussion—and new possibilities for quantum computation. They are also finding applications in metrology, including atomic clocks.

Section 8: Atomic Clocks

The words "Atomic Clock" occasionally appear on wall clocks, wristwatches, and desk clocks, though in fact none of these devices are really atomic. They are, however, periodically synchronized to signals broadcast by the nation's timekeeper, the National Institute of Standards and Technology (NIST). The NIST signals are generated from a time scale controlled by the frequency of a transition between the energy states of an atom—a true atomic clock. In fact, the legal definition of the second is the time for 9,192,631,770 cycles of a particular transition in the atom ^{133}Cs.

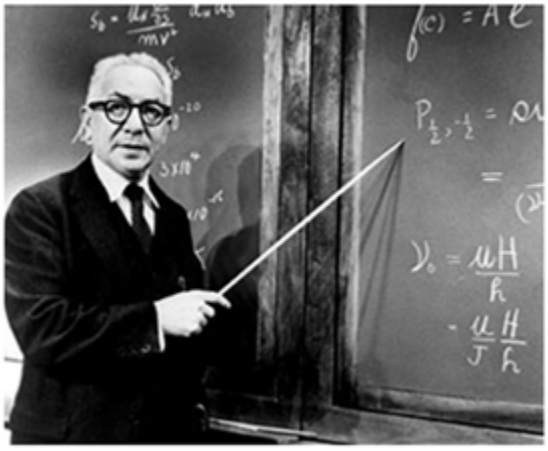


Figure 31: Isidor Isaac Rabi pioneered atomic physics in the U.S. during the 1930s, invented magnetic resonance, and first suggested the possibility of an atomic clock.
Source: © Trustees of Columbia University in the City of New York.

Columbia University physicist Isidor Isaac Rabi first suggested the possibility that atoms could be used for time keeping. Rabi's work with molecular beams in 1937 opened the way to broad progress in physics, including the creation of the laser as well as nuclear magnetic resonance, which led to the MRI imaging now used in hospitals. In 1944, the same year he received the Nobel Prize, he proposed employing a microwave transition in the cesium atom, and this system has been used ever since. The first atomic clocks achieved an accuracy of about 1 part in 10^{10}. Over the years, their accuracy has been steadily improved. Cesium-based clocks now achieve accuracy greater than 1 part in 10^{15}, 10,000 times more accurate than their predecessors, which is generally believed to be close to their ultimate limit. Happily, as will be described, a new technology for clocks based on optical transitions has opened a new frontier for precision.

A clock is a device in which a motion or event occurs repeatedly and which has a mechanism for keeping count of the repetitions. The number of counts between two events is a measure of the interval between them, in units of the period of the atomic transition frequency. If a clock is started at a given time—that is, synchronized with the time system—and kept going, then the accumulated counts define the time. This statement actually encapsulates the concept of time in physics.

In a pendulum clock, the motion is a swinging pendulum, and the counting device is an escapement and gear mechanism that converts the number of swings into the position of the hands on the clock face. In an atomic clock, the repetitious event is the quantum mechanical analogy to the physical motion of an atom: the frequency for a transition between two atomic energy states. An oscillator is adjusted so that its frequency matches the transition frequency, effectively making the atom the master of the oscillator. The number of oscillation cycles—the analogy to the number of swings of a pendulum—is counted electronically.

The quality of a clock—essentially its ability to agree with an identical clock—depends on the intrinsic reproducibility of the periodic event and the skill of the clock maker in counting the events. A cardinal principle in quantum mechanics is that all atoms of a given species are absolutely identical. Consequently, any transition frequency could form the basis for an atomic clock. The art lies in identifying the transition that can be measured with the greatest accuracy. For this, a high tick rate is desirable: It would be difficult to compare the rates of two clocks that ticked, for instance, only once a month. As the definition of the second reveals, cesium-based clocks tick almost 10 billion times per second.

Atomic clocks and the uncertainty principle

The precision with which a atomic transition can be measured is fundamentally governed by the uncertainty principle. As explained in Section 6, because of the time-energy uncertainty principle, there is an inherent uncertainty in the measurement of a frequency (which is essentially an energy) that depends on the length of the time interval during which the measurement is made. To reduce the uncertainty in the frequency measurement, the observation time should be as long as possible.

Controlling the frequency of an atomic clock

Four factors govern the quality of an atomic clock. They are:

1. The "tick rate," meaning the frequency of the transition. The higher the frequency, the larger the number of counts in a given interval, and the higher the precision. Cesium clocks tick almost 10 billion times per second.

2. The precision by which the transition frequency can be determined. This is governed fundamentally by the time-frequency uncertainty principle for a single measurement: $\tau\Delta f > 1$*. The fractional precision for a single measurement can be defined to be* $f/\Delta f = \tau f$*. Thus, the time interval during which the frequency of each atom is observed should be as long as possible. This depends on the art of the experimenter. In the most accurate cesium clocks, the observation time is close to one second.*

3. The rate at which the measurement can be repeated—that is, the number of atoms per second that are observed.

4. The ability to approach ideal measurement conditions by understanding and controlling the many sources of error that can affect a measurement. Those sources include noise in the measurement process, perturbations to the atomic system by magnetic fields and thermal (blackbody) radiation, energy level shifts due to interactions between the atoms, and distortions in the actual measurement process. The steady improvement in the precision of atomic clocks has come from progress in identifying and controlling these effects.

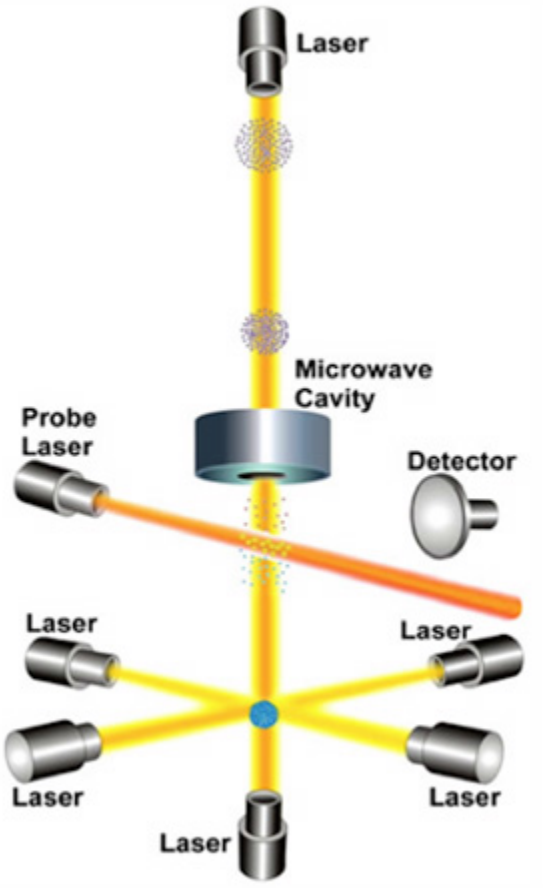


Figure 32: Schematic diagram of an atomic fountain clock.
Source: © NIST/Jeffrey Aarons.

In an atomic clock, the observation time is the time during which the atoms interact with the microwave radiation as they make the transition. Before the advent of ultracold atoms and atom trapping, this time was limited by the speed of the atoms as they flew through the apparatus. However, the slow speed of ultracold atoms opened the way for new strategies, including the possibility of an atomic fountain. In an atomic fountain, a cloud of cold atoms is thrust upward by a pulse of light. The atoms fly upward in the vacuum chamber, and then fall downward under the influence of gravity. The observation time is essentially the time for the atoms to make a roundtrip. For a meter-high fountain, the time is about one second.

The quality of an atomic clock depends on how well it can approach ideal measurement conditions. This requires understanding and controlling the many sources of error that can creep in. Errors arise from noise in the measurement process, perturbations to the atomic system by magnetic fields and thermal

(blackbody) radiation, energy level shifts due to interactions between the atoms, and distortions in the actual measurement process. The steady improvement in the precision of atomic clocks has come from incremental progress in identifying and controlling these effects.

The cesium fountain clock

The cesium clock operates on a transition between two energy states in the electronic ground state of the atom. As mentioned in Section 7, the ground state of an alkali metal atom is split into two separate energy levels in a magnetic field. Even in the absence of an external magnetic field, however, the ground state is split in two. This splitting arises from a magnetic interaction between the outermost in the atom electron and the atom's nucleus, known as the hyperfine interaction. The upper hyperfine state can in principle radiate to the lower state by spontaneous emission, but the lifetime for this is so long—thousands of years—that for all purposes, both states are stable. The transition between these two hyperfine states is the basis of the cesium clock that defines the second.

The cesium fountain clock operates in a high vacuum so that atoms move freely without colliding. Cesium atoms from a vapor are trapped and cooled in a magneto-optical trap. The trap lasers both cool the atoms and "pump" them into one of the hyperfine states, state A. Then, the wavelength of the trap laser beam pointing up is tuned to an optical transition in the atoms, giving the cloud a push by photon recoil. The push is just large enough to send the atoms up about one meter before they fall back down. The atoms ascend through a microwave cavity, a resonant chamber where the atoms pass through the microwave field from an oscillator. The field is carefully controlled to be just strong enough that the atoms make "half a transition," which is to say that if one observed the states of the atoms as they emerged from the cavity, half would be in hyperfine state A and half would be in state B. Then the atoms fly up, and fall back. If the frequency is just right, the atoms complete the transition as they pass through the cavity, so that they emerge in state B. The atoms then fall through a probe laser, which excites only those that are in state B. The fluorescence of the excited atoms is registered on a detector. The signal from the detector is fed back to control the frequency of the microwave oscillator, so as to continuously stay in tune with the atoms.

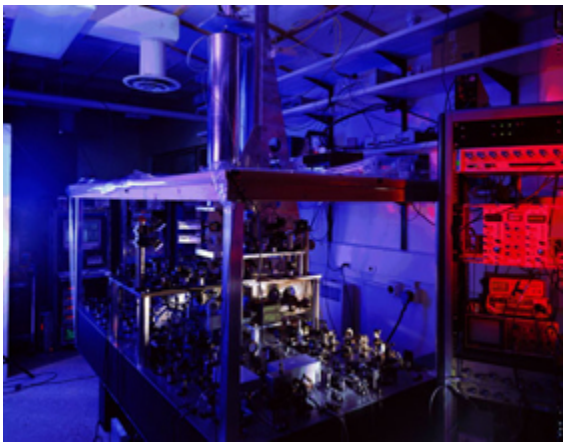


Figure 33: This apparatus houses the NIST F1 cesium fountain clock, which is the primary time and frequency standard of the United States.
Source: © NIST.

If we plot the signal on the detector against the frequency of the oscillator, we end up with what is known as a resonance curve. The pattern, called a Ramsey resonance curve, looks suspiciously like two-slit interference. In fact, it is an interference curve, but the sources interfere not in space but in time. There are two ways for an atom to go to state B from state A: by making the transition on the way up or on the way down. The final amplitude of the wavefunction has contributions from both paths, just as the wavefunction in two-slit interference has contributions from paths going through each of the slits. This method of observing the transition by passing the atom through a microwave field twice is called the "separated oscillatory field method" and its inventor, Norman F. Ramsey, received the Nobel Prize for it in 1989.

Optical clocks

A useful figure of quality for atomic clocks is the ratio of its frequency ν to the uncertainty in its frequency, $\Delta\nu$. For a given value of $\Delta\nu$, the higher the frequency, the better the clock. With atom-cooling techniques, there are many possibilities for keeping atoms close to rest so that $\Delta\nu$ is small. Consequently, clocks operating at optical frequencies, in the petahertz (10^{15} Hz) region, are potentially much more accurate than cesium-based clocks that operate in the gigahertz (10^{9} Hz) region. However, two impediments have delayed the advent of optical clocks. Fortunately, these have been overcome, and optical clock technology is moving forward rapidly.

The first impediment was the need for an incredibly stable laser to measure the atomic signal. In order to obtain a signal from the atoms, the laser must continue oscillating smoothly on its own during the entire time the atoms are being observed. The requirement is formidable: a laser oscillating at a frequency of close to 10^{15} Hz that fluctuates less than 1 Hz. Through a series of patient developments over many years, this challenge has been met.

The second impediment to optical clocks was the problem of counting cycles of light. Although counting cycles of an oscillating electric field is routine at microwave frequencies using electronic circuitry, until recently there was no way to count cycles at optical frequencies. Fortunately, a technology has been invented. Known as the "frequency comb," the invention was immediately recognized as revolutionary. The inventors, Theodor W. Hänsch and John L. Hall, were awarded the Nobel Prize in 2005 "for their contributions to the development of laser-based precision spectroscopy including the optical frequency-comb technique."

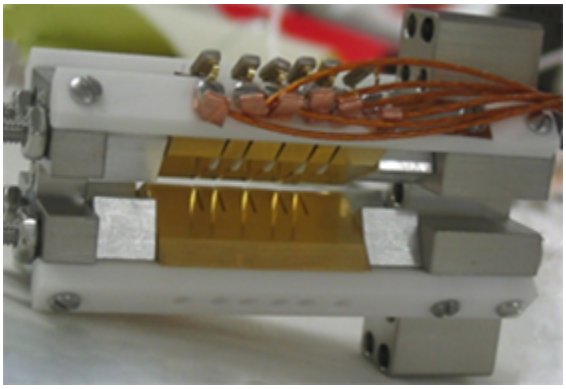


Figure 34: The heart of a next-generation optical clock.
Source: © Ion Storage Group, NIST.

Optical clocks are only in the laboratory stage but progress is rapid. One type of clock employs ions stored in electromagnetic traps, similar to the trap used in Figure 1; another employs neutral atoms confined in an optical lattice such as in Figure 2. Figure 34 shows a state-of-the-art ion-based clock at NIST. A pair of such clocks has recently demonstrated a relative accuracy greater than one part in 10^{17}. Making these clocks into practical devices is an interesting engineering challenge.

In the new world of precise clocks, transmitting timing signals and comparing clocks in different locations presents a major challenge. Transmissions through the atmosphere or by a satellite relay suffer bad atmospheric fluctuations. The signals can be transmitted over optical fibers, but fibers can introduce

timing jitter from vibrations and optical nonlinearities. These can be overcome for distances of tens of kilometers by using two-way monitoring techniques, but methods for extending the distances to thousands of kilometers have yet to be developed. However, there is an even more interesting impediment to comparing clocks at different locations. The gravitational redshift explained in Unit 3 changes the rates of clocks by 1 part in 10^{16} for each meter of altitude, near Earth's surface. Clocks are approaching the regime of parts in 10^{18}. To compare clocks in different locations, the relative altitudes would need to be known to centimeters. Earth's surface is constantly moving by tens of centimeters due to tides, weather, and geological processes. This presents not merely a practical problem but also a conceptual problem, for it forces us to realize that time and gravity are inextricably interlinked. Because of this, the view that time is essentially the clicks on a clock begins to seem inadequate.

Payoffs from basic research

When Isidor Isaac Rabi proposed the possibility of an atomic clock, he had a scientific goal in mind: to observe the effect of gravity on time—the gravitational redshift—predicted by Einstein's theory of general relativity. The quest to confirm Einstein's prediction motivated the field. Today, the gravitational redshift has not only been observed, but also measured to high precision. However, the biggest impacts of atomic clocks were totally unforeseen. Global Positioning System (GPS) is one of these.

The GPS is a network of satellites positioned so that several of them are essentially always in view. A receiver calculates its location from information transmitted by the satellites about their time and position at each instant. The satellites carry one or more atomic clocks whose times are periodically updated by a master atomic clock in a ground station. The GPS system is a miracle of engineering technology: sophisticated satellites, integrated electronics and advanced communications, information processing, geodesy, and orbital mechanics. But without atomic clocks, there would be no GPS. Furthermore, with the precision inherent in the GPS system, the gravitational redshift is not only detectable, but to overlook it would cause catastrophic navigational errors.

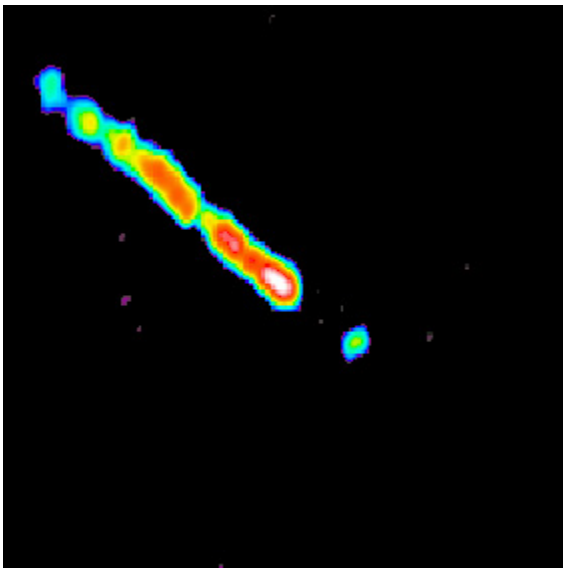


Figure 35: This VLBI image of jets from a black hole could not have been produced without atomic clocks.
Source: © Steven Tingay, Curtin University, Australia.

Atomic clocks have applications in fundamental science as well. The technique of very long baseline radio interferometry (VLBI) permits Earth to be converted to a giant radio telescope. Signals from radio observatories on different continents can be brought together and compared to provide the angular resolution of an Earth-sized dish. To do this, however, the astronomical radio signals must first be recorded against the signal from an atomic clock. The records are then brought together and their information is correlated. VLBI can reveal details less than a millionth of a degree, the highest resolution achieved in all of astronomy.

Although Einstein's theory of gravity is one of the most abstract subjects in science, the search to study it led to the invention of GPS and the creation of VLBI. This history illustrates, if illustration is needed, that the pursuit of basic knowledge is a worthy goal for scientists and a wise investment for society.

Section 9: Afterthoughts on a Paradox

The paradox of how a wave can be a particle and a particle can be a wave was brought up in Section 4, but not resolved. The issue is far from trivial and was fiercely debated in the early days of quantum mechanics. Niels Bohr even designed a hypothetical experiment to clarify the question of whether you could detect which slit a photon passed through in a two-slit interference experiment. For light to interfere, it must slightly change its direction as it passes through a slit in order to merge with the second beam.

Consequently, passing through a slit must slightly alter a photon's direction, which means that the slit has altered the photon's momentum. The photon must give an opposite momentum to the slit. Bohr's apparatus was designed to detect the recoil of the slit. If this were possible, an observer could decide which slit each photon passed through in creating an interference pattern, revealing both the particle and wave nature of light simultaneously. However, Bohr proved that detecting the photon would actually wipe out the interference pattern.

Thinking about waves passing through slits provides a different way to understand the situation. The waves might be light waves but they could just as well be matter waves. As the waves emerge from the slits, they diverge in a diffraction pattern. The wave intensity on the viewing screen might be registered on a camera, as in Figure 11, or measured by detections with particle counters, creating images similar to those in Figure 15. For the sake of discussion, we assume that the individual atoms or photons are detected with particle counters.

If the slits are close together, the diffraction patterns of particles coming through them overlap. In time the counts add to give a two-slit interference pattern, which is the signature of waves. What about the intermediate case? If the slits are far enough apart that the diffraction patterns only overlap a little bit, we should be able to place two detectors that only see particles passing through one or the other of the slits, and a detector in the center that sees two-slit interference. The conclusion is that if one knows from which of two slits the signal arises, one must ascribe the signal to the arrival of a particle. However, if there is no way to distinguish which of two possibilities gave rise to the signal, one must ascribe the signal to the arrival of waves.

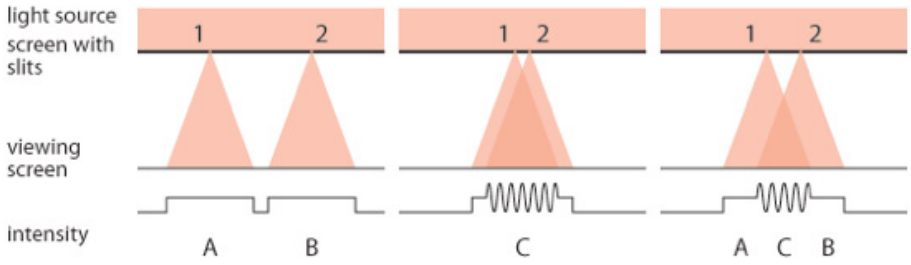


Figure 36: The two-slit interference pattern depends on the distance between the slits.
Source: © Daniel Kleppner.

The answer to the question, "Is light composed of waves or particles?" is "Both." If you search for light's wave properties, you will find them. If you search for light's particle properties, you will find them, too. However, you cannot see both properties at the same time. They are what Bohr called *complementary* properties. One needs both properties for a complete understanding of light, but they are fundamentally incompatible and cannot be observed at the same time. Thus, the wave-particle paradox presents a contradiction that is not really true, but merely apparent.

We have discussed the wave-particle paradox for light, but the same reasoning applies to atoms and matter waves. Atoms are waves and they are particles, but not at the same time. You will find what you look for.

Section 10: Further Reading

- Rainer Blatt and David Wineland, "Entangled states of trapped atomic ions," *Nature*, June 2008, p. 1008.
- Steven Chu, Claude Cohen-Tannoudji, and William D. Phillips: Nobel Prize Lectures, available at http://nobelprize.org/nobel_prizes/physics/laureates/1997/.
- Albert Einstein, "Einstein's Essays in Science," *Dover Publications* (2009).
- Tony Jones, "Splitting the Second: The Story of Atomic Time," *Taylor and Francis* (2008).

Glossary

evaporative cooling: Evaporative cooling is a process used in atomic physics experiments to cool atoms down to a few billionths of a degree above absolute zero. The way it works is similar to how a cup of hot coffee cools through evaporation. Atoms are pre-cooled, usually with some kind of laser cooling, and trapped in a manner that imparts no additional energy to the atoms. The warmest atoms are removed from the trap, and the remaining atoms reach a new, lower equilibrium temperature. This process is typically repeated many times, creating small clouds of very cold atoms.

alkali metals: The alkali metals are the chemical elements in the first column of the periodic table. They all have one valence electron. Alkali metals are commonly used atoms in atomic physics experiments for several reasons. Their structure is relatively simple and provides energy states that are convenient for laser cooling. Many of their transition frequencies match convenient laser sources. Also, the single valence electron's magnetic moment allows the atoms to be easily trapped using magnetic fields, which is convenient for the evaporative cooling process necessary to reach ultracold temperatures.

blackbody: A blackbody is an object that absorbs all incident electromagnetic radiation and re-radiates it after reaching thermal equilibrium. The spectrum of light emitted by a blackbody is smooth and continuous, and depends on the blackbody's temperature. The peak of the spectrum is higher and at a shorter wavelength as the temperature increases.

Bose-Einstein condensate: A Bose-Einstein condensate, or BEC, is a special phase of matter in which the quantum mechanical wavefunctions of a collection of particles line up and overlap in a manner that allows the particles to act as a single quantum object. The electrons in a superconductor form a BEC; superfluid helium is an example of a liquid BEC. BECs can also be created from dilute gases of ultracold atoms and molecules.

Bohr Correspondence Principle: The Bohr Correspondence Principle states that the predictions of quantum mechanics must match the predictions of classical physics in the physical situations that classical physics is intended to describe, and does describe very accurately. Mathematically, this means that the equations of quantum mechanics must smoothly turn into the equations of classical mechanics as the de Broglie wavelength of particles becomes very small, and the energy state quantum number gets very large.

cosmic microwave background: The cosmic microwave background (CMB) radiation is electromagnetic radiation left over from when atoms first formed in the early universe, according to our standard model of cosmology. Prior to that time, photons and the fundamental building blocks of matter formed a hot, dense soup, constantly interacting with one another. As the universe expanded and cooled, protons and neutrons formed atomic nuclei, which then combined with electrons to form neutral atoms. At this point, the photons effectively stopped interacting with them. These photons, which have stretched as the universe expanded, form the CMB. First observed by Penzias and Wilson in 1965, the CMB remains the focus of increasingly precise observations intended to provide insight into the composition and evolution of the universe.

diffraction: Diffraction is the spreading of a wave after it encounters an obstacle, a sharp corner, or emerges from a slit. If the slit is small, the spreading is large and is accompanied by an interference pattern with a central peak surrounded by weaker side lobes. In this context, "small" means comparable to the wavelength of the diffracting wave. The fact that light diffracts when passed through a small slit is evidence of its wave nature.

Doppler cooling: Doppler cooling is a technique that uses laser light to slow, and thus cool, moving atoms. An atom will absorb a photon that has an energy equal to the difference between two energy levels in the atom. When the atom absorbs a photon, it also absorbs the photon's momentum and gets a push in the direction that the photon was traveling. If the photon and atoms were traveling in opposite directions, the atom slows down. However, when the atom is moving relative to the laser, the laser light is Doppler shifted in the atom's reference frame. To cool moving atoms, the laser must be tuned slightly to the red to account for the Doppler shift of atoms moving toward the light source.

Doppler shift (Doppler effect): The Doppler shift is a shift in the wavelength of light or sound that depends on the relative motion of the source and the observer. A familiar example of a Doppler shift is the apparent change in pitch of an ambulance siren as it passes a stationary observer. When the ambulance is moving toward the observer, the observer hears a higher pitch because the wavelength of the sound waves is shortened. As the ambulance moves away from the observer, the wavelength is lengthened and the observer hears a lower pitch. Likewise, the wavelength of light emitted by an object moving toward an observer is shortened, and the observer will see a shift to blue. If the light-emitting object is moving away from the observer, the light will have a longer wavelength and the observer will see a shift to red. By observing this shift to red or blue, astronomers can determine the velocity of distant stars and galaxies

relative to the Earth. Atoms moving relative to a laser also experience a Doppler shift, which must be taken into account in atomic physics experiments that make use of laser cooling and trapping.

ground state: The ground state of a physical system is the lowest energy state it can occupy. For example, a hydrogen atom is in its ground state when its electron occupies the lowest available energy level.

harmonic oscillator: A harmonic oscillator is a physical system that, when displaced from equilibrium, experiences a restoring force proportional to the displacement. A harmonic oscillator that is displaced and then let go will oscillate sinusoidally. Examples from classical physics are a mass attached to a spring and a simple pendulum swinging through a small angle.

Heisenberg uncertainty principle: The Heisenberg uncertainty principle states that the values of certain pairs of observable quantities cannot be known with arbitrary precision. The most well-known variant states that the uncertainty in a particle's momentum multiplied by the uncertainty in a particle's position must be greater than or equal to Planck's constant divided by 4π. This means that if you measure a particle's position to better than Planck's constant divided by 4π, you know that there is a larger uncertainty in the particle's momentum. Energy and time are connected by the uncertainty principle in the same way as position and momentum. The uncertainty principle is responsible for numerous physical phenomena, including the size of atoms, the natural linewidth of transitions in atoms, and the amount of time virtual particles can last.

hyperfine interaction: When the nucleus of an atom has a non-zero magnetic moment, the magnetic field of the nucleus interacts with electrons in the atom. This interaction is called the hyperfine interaction, and leads to finely spaced atomic energy levels called hyperfine structure.

interference: Interference is an effect that occurs when two or more waves overlap. In general, the individual waves do not affect one another, and the total wave amplitude at any point in space is simply the sum of the amplitudes of the individual waves at that point. In some places, the two waves may add together, and in other places they may cancel each other out, creating an interference pattern that may look quite different than either of the original waves. Quantum mechanical wavefunctions can interfere, creating interference patterns that can only be observed in their corresponding probability distributions.

magnetic moment: The magnetic moment (or magnetic dipole moment) of an object is a measure of the object's tendency to align with a magnetic field. It is a vector quantity, with the positive direction defined by the way the object responds to a magnetic field: The object will tend to align itself so that its magnetic

moment vector is parallel to the magnetic field lines. There are two sources for a magnetic moment: the motion of electric charge and spin angular momentum. For example, a loop of wire with a current running through it will have a magnetic moment proportional to the current and area of the loop, pointing in the direction of your right thumb if your fingers are curling in the direction of the current. Alternatively, an electron, which is a spin-1/2 fermion, has an intrinsic magnetic moment proportional to its spin.

magneto-optical trap: A magneto-optical trap, or MOT, uses a combination of laser beams and magnetic fields to confine atoms at temperatures between a few millikelvin and a few microkelvin. Atoms in a MOT are constantly interacting with the laser beams, which cool them to the laser-cooling limit, but no further than that.

matrix mechanics: Matrix mechanics is the version of quantum mechanics formulated in the 1920s by Werner Heisenberg and his close colleagues Max Born and Pascual Jordan. It makes extensive use of matrices and linear algebra, which was relatively new mathematics at the time. Matrix mechanics, which is mathematically equivalent to Schrödinger's wave mechanics, greatly simplifies certain calculations.

natural linewidth: The natural linewidth of an atomic energy level is the intrinsic uncertainty in its energy due to the uncertainty principle.

optical dipole trap: An optical dipole trap is a type of atom trap that uses only laser light to trap atoms. The laser frequency is tuned to a frequency below an atomic resonance so the atoms do not absorb laser photons as they do in laser cooling or in a MOT. Instead, the electric field from the laser induces an electric dipole in the atoms that attracts them to regions of more intense laser light. Optical dipole traps are only strong enough to hold cold atoms, so atoms are typically cooled first and then transferred into the dipole trap.

optical lattice: An optical lattice is an optical dipole trap made from a standing wave laser beam, so there is a periodic array of regions with a strong and weak laser field. Atoms are attracted to regions of a strong field, so they are trapped in a lattice-like pattern.

optical molasses: Optical molasses is formed when laser beams for Doppler cooling are directed along each spatial axis so that atoms are laser cooled in every direction. Atoms can reach microkelvin temperatures in optical molasses. However, the molasses is not a trap, so the atoms can still, for example, fall under the influence of gravity.

photon: Photons can be thought of as particle-like carriers of electromagnetic energy, or as particles of light. In the Standard Model, the photon is the force-carrier of the electromagnetic force. Photons are massless bosons with integer spin, and travel through free space at the speed of light. Like material particles, photons possess energy and momentum.

Planck's constant: Planck's constant, denoted by the symbol h, has the value 6.626×10^{-34} m^2 kg/s. It sets the characteristic scale of quantum mechanics. For example, energy is quantized in units of h multiplied by a particle's characteristic frequency, and spin is quantized in units of $h/2\pi$. The quantity $h/2\pi$ appears so frequently in quantum mechanics that it has its own symbol: $\hbar$.

plane wave: A plane wave is a wave of constant frequency and amplitude with wavefronts that are an infinitely long straight line. Plane waves travel in the direction perpendicular to the wavefronts. Although they are a mathematical abstraction, many physical waves approximate plane waves far from their source.

polarization: The polarization of a wave is the direction in which it is oscillating. The simplest type of polarization is linear, transverse polarization. Linear means that the wave oscillation is confined along a single axis, and transverse means that the wave is oscillating in a direction perpendicular to its direction of travel. Laser light is most commonly a wave with linear, transverse polarization. If the laser beam travels along the x-axis, its electric field will oscillate either in the y-direction or in the z-direction. Gravitational waves also have transverse polarization, but have a more complicated oscillation pattern than laser light.

probability distribution: In quantum mechanics, the probability distribution is a mathematical function that gives the probability of finding a particle in any small region of space. The probability distribution for a quantum mechanical system is simply the square of the wavefunction.

quantized: Any quantum system in which a physical property can take on only discrete values is said to be quantized. For instance, the energy of a confined particle is quantized. This is in contrast to a situation in which the energy can vary continuously, which is the case for a free particle.

quantum number: A quantum number is a number that characterizes a particular property of a quantum mechanical state. For example, each atomic energy level is assigned a set of integers that is uniquely related to the quantized energy of that level.

resonance curve: A resonance curve is a graph of the response of an atomic system to electromagnetic radiation as a function of the frequency of the radiation. The simplest example of a resonance curve is the single peak that appears as a laser's frequency is scanned through the difference between two energy levels in the atoms.

spontaneous emission: An atom in an excited state can decay down to a lower state by emitting a photon with an energy equal to the difference between the initial, higher energy level and the final, lower energy level. When this process takes place naturally, rather than being initiated by disturbing the atom somehow, it is called spontaneous emission.

standing wave: A standing wave is a wave that does not travel or propagate: The troughs and crests of the wave are always in the same place. A familiar example of a standing wave is the motion of a plucked guitar string.

stationary states: In quantum mechanics, a stationary state is a state of a system that will always yield the same result when observed in an experiment. The allowed energy states of a harmonic oscillator (Unit 5, section 5) are an example, as are the allowed energy levels of an atom. Stationary states correspond to quantum wavefunctions that describe standing waves.

superposition principle: Both quantum and classical waves obey the superposition principle, which states that when two waves overlap, the resulting wave is the sum of the two individual waves.In quantum mechanics, it is possible for a particle or system of particles to be in a superposition state in which the outcome of a measurement is unknown until the measurement is actually made. For example, neutrinos can exist in a superposition of electron, muon, and tau flavors (Units 1 and 2). The outcome of a measurement of the neutrino's flavor will yield a definite result—electron, muon, or tau—but it is impossible to predict the outcome of an individual measurement. Quantum mechanics tells us only the probability of each outcome. Before the measurement is made, the neutrino's flavor is indeterminate, and the neutrino can be thought of as being all three flavors at once.

tunneling: Tunneling, or quantum tunneling, takes place when a particle travels through a region that would be forbidden according to the laws of classical physics. Tunneling occurs because quantum wavefunctions extend slightly past the boundaries that define where a particle is allowed to be. For example, in classical physics, an electron is allowed to move through a conductor but not through an insulator. However, if a thin layer of insulator is placed between two conductors, the electron can tunnel through from one conductor to the other because its wavefunction extends into the insulating layer.

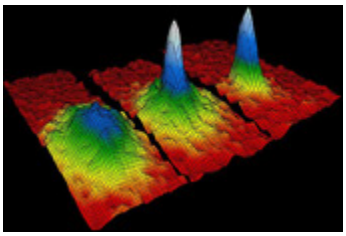
wave mechanics: Wave mechanics is the version of quantum mechanics formulated primarily by Erwin Schrödinger in the 1920s. Following de Broglie's hypothesis that particles can equally well be described as waves, Schrödinger set out to write down a wave equation for quantum systems and proceeded to solve it in many interesting examples. Wave mechanics is mathematically equivalent to Heisenberg's matrix mechanics.

Zeeman effect: Each atomic energy level in which an atom has a non-zero spin splits into two or more separate levels when the atom is placed in an external magnetic field. The splitting grows with the strength of the external field. This effect is named the Zeeman effect after the experimentalist who first studied it in the laboratory, Pieter Zeeman. He received the 1902 Nobel Prize for this work, along with Hendrik Lorentz, the theorist who explained the effect.

zero point energy: The zero point energy is the minimum energy a system can have based on the Heisenberg uncertainty principle.

Physics for the 21st Century

Unit 6: Macroscopic Quantum Systems



© Mike Matthews, JILA.

Unit Overview

The fundamentals of quantum mechanics that we met in Unit 5 characteristically appear on microscopic scales. Macroscopic quantum systems in which liquids, or electric currents, flow without friction or resistance have been known since the early part of the previous century: these are the superfluids and superconductors of traditional condensed matter physics that are discussed in Unit 8. In this unit we focus on an entirely new state of matter only recently created in the laboratory: this is the gaseous macroscopic quantum mechanical system known as a Bose-Einstein Condensate, or BEC. These quantum gases show the full panoply of quantum wave behavior as the individual particles of Unit 5, but now on a size scale visible to the naked eye because many millions, to many billions, of atoms occupy exactly the same quantum state, and thus form a coherent quantum whole. The quantum nature of a BEC can be directly visualized as the quantum effects are not hidden within liquids or solids as is the case with the more traditional superfluids and superconductors. Rather, they may be actually photographed, as the gas itself is a naked macroscopic quantum system. This unit starts by introducing the basic principles necessary to understand BECs, then details how the cooling and trapping introduced in Unit 5 led to the creation and subsequent manipulation of these quantum gases. Finally, we will see how atomic gases of ultra-cold fermions have evolved, in direct analogy to the Cooper paring needed to form bosonic pairs of electrons in superconductors, to molecular BECs, formed from pairs of the fermionic atoms.

Content for This Unit

Sections:

Section 1: Introduction

The quantum theory that evolved at the beginning of the 20th century is a strange creature: It assigns particulate properties to light—long thought to consist of waves—and, astonishingly, wave properties to individual fundamental particles that have definite masses. This seeming confusion between waves and particles never seems to bother us in ordinary life, where particles (such as grains of rice or baseballs) are simply particles, and what comes out of a flashlight or laser pointer can be focused or even diffracted, thereby revealing the wave-like properties we associate with light. Of course, large numbers of particles acting in collusion can create waves as a possible collective motion—on the surface of the ocean, for example. And a slinky or violin string can display wave properties, to the surprise of no one. But these are classical, not quantum, waves, a distinction we will make clear in this unit.

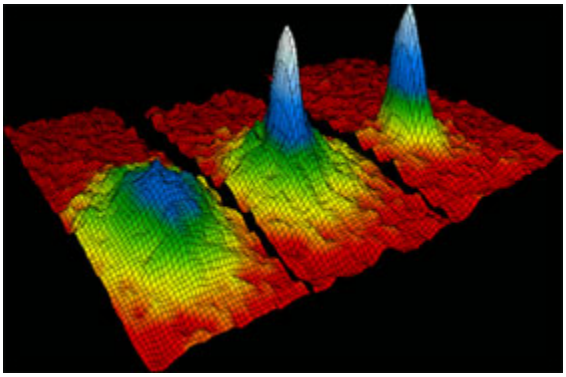


Figure 1: Some of the first experimental evidence for a gaseous macroscopic quantum state.
Source: © Mike Matthews, JILA.

We typically view quantum mechanics as applying only to the fundamental particles or fields of Units 1 and 2, and not to the objects we find in a grocery store or in our homes. If large objects like baseballs and our dining tables don't behave like waves, why do we bother about their possible quantum nature? More practically, we can ask: If we are supposed to believe that ordinary objects consist of wave-like quantum particles, how does that quantum nature disappear as larger objects are assembled from smaller ones? Or, more interestingly: Are there macroscopic objects large enough to be visible to the naked eye that still retain their quantum wave natures? To answer these questions, we need the rules for building larger objects out of smaller ones, and the means of applying them. The rules are of a highly quantum nature; that is, they have no counterpart at all in classical physics, nor are they suggested by the behavior of the ordinary material objects around us. But in surprising ways, they lead to macroscopic quantum behavior, as well as the classical world familiar to all of us.

To interpret these rules, we need to understand two critical factors. First, particles of all types, including subatomic particles, atoms, molecules, and light quanta, fall into one of two categories: fermions and bosons. As we shall see, the two play by very different rules. Equally important, we will introduce the concept of pairing. Invoked more than a century ago to explain the creation of molecules from atoms, this idea plays a key role in converting fermions to bosons, the process that enables macroscopic quantum behavior. This unit will explore all these themes and Unit 8 will expand upon them.

Building up atoms and molecules

Classical Pairing in DNA

The concept of pairing is not restricted to simple chemical bonds and the conversion of fermions to bosons. Pairs also occur at higher levels of molecule formation. In the Watson-Crick model of DNA, the molecules in each strand of the double helix are made up of four distinct groups of molecules denoted by A, G, C, and T; A and T form pairs, as do G and T. In cell division, one strand creates the template for correct formation of second strand. The proper pairings A, T and C, G allow transmission of genetic information when cells replicate.

DNA is tightly coiled in the nucleus of a cell. Stretched out, some chromosomes may contain DNA molecules whose lengths extend from ten centimeters to over one meter. These are indeed macroscopic single molecules. But at room temperature they have no special quantum properties, even though it is the quantum nature of the electrons in the A, G, C, and T that determine the forces between them. So although pairs are indeed ubiquitous, the chemical pairing of DNA is a "classical pairing," not the quantum pairing that can make bosons out of paired fermions.

We start with the empirically determined rules for building atoms out of electrons, protons, and neutrons, and then move on to the build-up of atoms into molecules. The quantum concept of the Pauli exclusion principle plays a key role in this build-up (or, in the original German, *aufbau*). This principle prevents more than one fermion of the same fundamental type from occupying the same quantum state, whether that particle be in a nucleus, an atom, a molecule, or even one of the atom traps discussed in Unit 5. But the exclusion principle applies only to *identical* fermions. Electrons, neutrons, and protons are all fermions; but, although all electrons are identical, and thus obey an exclusion principle, they certainly differ from neutrons or protons, and these non-identical fermions are not at all restricted by the presence of one another. Chemists and physicists understood the exclusion principle's empirical role in constructing the periodic table well before the discovery of the quantum theory of Unit 5; but they did not understand its mathematical formulation or its theoretical meaning.

Particles of light, or photons, are of a different type altogether: They are bosons, and do not obey the exclusion principle. Thus, identical photons can all join one another in precisely the same quantum state; once started, in fact, they are actually attracted to do so. This is the antithesis of the exclusion principle. Large numbers of such photons in the same quantum state bouncing back and forth between

two mirrors several inches or even many meters apart create an essential part of a laser that, in a sense, is a macroscopic quantum system with very special properties.

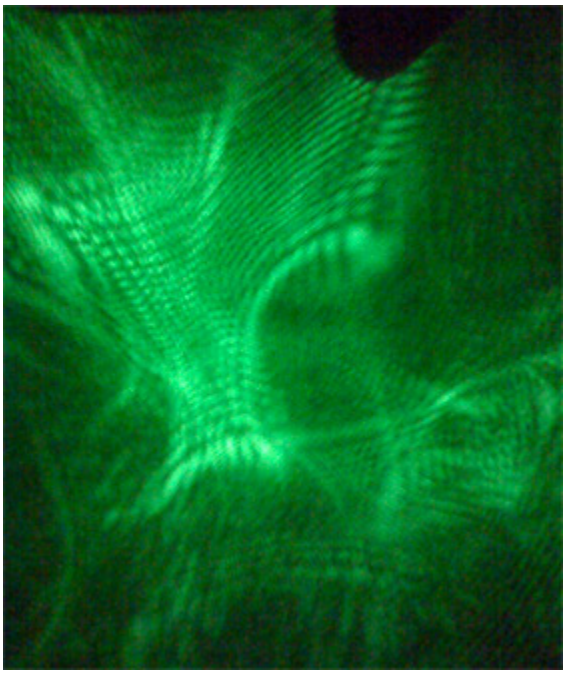


Figure 2: Diffraction of green laser light passing though a random medium.
Source: © William P. Reinhardt.

Lasers are everywhere in our technology, from laser pointers to surgical devices and surveying equipment, to the inner workings of CD and DVD players. Laser light sent from telescopes and reflected from mirrors on the Moon allows measurement of the distance between the Earth and Moon to better than one millimeter, allowing tests of gravitational theory, measurement of the slowly increasing radius of the moon's orbit around the Earth, and even allows geophysicists to observe the day-to-day motion of the continents (actually the relative motion of plate tectonics) with respect to one another: A bright light indeed. In addition to being bright and intense, laser light is *coherent*. Not only do all the particulate photons wave with the same wavelength (or color), but they are also shoulder to shoulder in phase with one another. This phase coherence of very large numbers of photons is a quantum property that follows from their bosonic nature. Yet, it persists up into the macroscopic world.

Superfluids and superconductors

Are there other macroscopic quantum systems in which actual particles (with mass) cooperate in the same coherent manner as the photons in a laser? We might use the molecules we will build up to make a

dining room table, which has no evident quantum wave properties; so perhaps quantum effects involving very large numbers of atoms simply don't persist at the macroscopic level or at room temperature. After all, light is quite special, as particles of light have no mass or weight. Of course, we fully expect light to act like a wave, it's the particulate nature of light that is the quantum surprise. Can the wave properties of massive particles appear on a macroscopic scale?

The surprising answer is yes; there are macroscopic quantum systems consisting of electrons, atoms, and even molecules. These are the superfluids and superconductors of condensed matter physics and more recently the Bose condensates of atomic and molecular physics. Their characteristics typically appear at low (about 1–20 K), or ultra-low (about 10^{-8} K and even lower) temperatures. But in some superconductors undamped currents persist at rather higher temperatures (about 80 K). These are discussed in Unit 8.

Superfluids are liquids or gases of uncharged particles that flow without friction; once started, their fluid motion will continue forever—not a familiar occurrence for anything at room temperature. Even super-balls eventually stop bouncing as their elasticity is not perfect; at each bounce, some of their energy is converted to heat, and eventually they lie still on the floor.

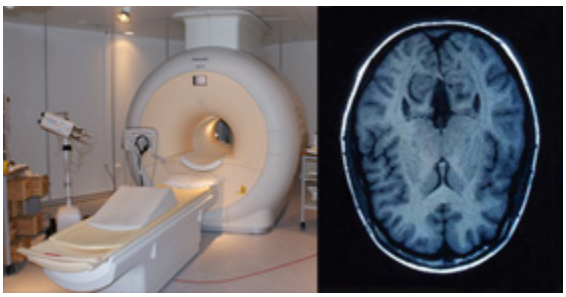


Figure 3: Superconducting magnets enable MRI machines to produce dramatic images.
Source: © Left: Wikimedia Commons, Creative Commons Attribution 3.0 Unported License. Author: Jan Ainali, 12 February 2008. Right: NIH.

Superconductors have this same property of flowing forever once started, except that they are streams of charged particles and therefore form electric currents that flow without resistance. As a flowing current generates a magnetic field, a superconducting flow around a circle or a current loop generates a possibly very high magnetic field. This field never dies, as the super-currents never slow down. Hospitals everywhere possess such superconducting magnets. They create the high magnetic fields that allow the diagnostic procedure called MRI (Magnetic Resonance Imaging) to produce images of the brain or other

soft tissues. This same type of powerful superconducting magnetic field is akin to those that might levitate trains moving along magnetic tracks without any friction apart from air resistance. And as we saw in Unit 1, physicists at CERN's Large Hadron Collider use superconducting magnets to guide beams of protons traveling at almost the speed of light.

Thus, there *are* large-scale systems with quantum behavior, and they have appropriately special uses in technology, engineering, biology and medicine, chemistry, and even in furthering basic physics itself. How does this come about? It arises when electrons, protons, and neutrons, all of which are fermions, manage to arrange themselves in such a way as to behave as bosons, and these composite bosons are then sucked into a single quantum state, like the photons in a laser.

Pairing and exclusion

What is required to create one of these composite quantum states, and what difference does it make whether the particles making up the state are fermions or bosons? Historically, the realization that particles of light behaved as bosons marked the empirical beginning of quantum mechanics, with the work of Planck, Einstein, and, of course, Bose, for whom the boson is named. In Unit 5, we encountered Planck's original (completely empirical) hypothesis that the energy at frequency in an equilibrium cavity filled with electromagnetic radiation at temperature T should be $nh\nu$, where h is Planck's constant, ν is the photon frequency, and n is a mathematical integer, say 0, 1, 2, 3.... Nowadays, we can restate the hypothesis by saying that n photons of energy $h\nu$ are all in the same quantum mode, which is the special property of bosons. This could not happen if photons were fermions.

Another, at first equally empirical, set of models building up larger pieces of matter from smaller came from Dmitri Mendeleev's construction and understanding of the periodic table of the chemical elements. American chemist G. N. Lewis pioneered the subsequent building up of molecules from atoms of the elements. His concept that pairs of electrons form chemical bonds and pairs and octets of electrons make especially stable chemical elements played a key role in his approach. This same concept of pairing arises in converting fermions into the bosons needed to become superconductors and superfluids, which will be our macroscopic quantum systems.

Group I	Group II	Group III	Group IV	Group V	Group VI	Group VII	Group VIII
H							
Li	Be	B	C	N	O	F	
Na	Mg	Al	Si	P	S	Cl	
K Cu	Ca Zn	‘ ‘	Ti	V As	Cr Se	Mn Br	Fe Co Ni
Rb Ag	Sr Cd	Y In	Zr Sn	Nb Sb	Mo Te	‘ I	Ru Rn Pd

Figure 4: Early periodic table.
Source:

Austrian theorist Wolfgang Pauli took the next step. The theoretical development of modern quantum mechanics by Heisenberg and Schrödinger allowed him to make a clear statement of the exclusion principle needed to build up the periodic table of elements. That principle also made clear the distinction between fermions and bosons as will be discussed in Section 4. But first, we will address how models of atomic structure were built, based on the discovery of the electron and the atomic nucleus. Then we will follow the quantum modeling of the atom, explaining these same empirical ideas once the quantum theory of Unit 5 is combined with Pauli's exclusion principle.

Section 2: Early Models of the Atom

The foundation for all that follows is the periodic table that Russian chemist Dmitri Mendeleev formulated by arranging the chemical elements in order of their known atomic weights. The table not only revealed the semi-periodic pattern of elements with similar properties, but also contained specific holes (missing elements, see Figure 4) that allowed Mendeleev to actually predict the existence and properties of new chemical elements yet to be discovered. The American chemist G. N. Lewis then created an electron shell model giving the first physical underpinning of both Mendeleev's table and of the patterns of chemical bonding. This is very much in the current spirit of particle physics and the Standard Model: symmetries, "magic numbers," and patterns of properties of existing particles that strongly suggest the existence of particles, such as the Higgs boson, yet to be discovered.

Figure 5: The father and son of chemical periodicity: Mendeleev and Lewis.
Source: © Left: Wikimedia Commons, Public Domain. Right: Lawrence Berkeley National Laboratory.

J. J. Thomson

When Mendeleev assembled his periodic table in 1869, he was the giant standing on the shoulders of those who, in the 6 decades before, had finally determined the relative masses of the chemical elements, and the formulae of those simple combinations of atoms that we call "molecules." In this arduous task, early 19th century chemists were greatly aided by using techniques developed over several millennia by the alchemists, in their attempts to, say, turn lead into gold. This latter idea, which nowadays seems either antiquated or just silly, did not seem so to the alchemists. Their view is reinforced by the realization that even as Mendeleev produced his table: *No one had any idea what the chemical elements, or atoms as we now call them, were made of.* Mendeleev's brilliant work was empiricism of the highest order.

John Dalton, following Democrates, had believed that they weren't made of anything: Atoms were the fundamental building blocks of nature, and that was all there was to it.

Figure 6: Plum pudding (or raisin scone) model of the atom.
Source: © William P. Reinhardt, 2010.

This all began to change with J. J. Thomson's discovery of the electron in 1897, and his proposed atomic model. Thomson found that atoms contained electrons, each with a single unit of negative charge. To his great surprise he also found that electrons were very light in comparison to the mass of the atoms from which they came. As atoms were known to be electrically neutral, the rest of the atom then had to be positively charged and contain most of the mass of the atom. Thomson thus proposed his plum pudding model of the atom.

Rutherford and Lewis

Thomson's model was completely overthrown, less than 15 years later, by Ernest Rutherford's discovery that the positive charge in an atom was concentrated in a very small volume which we now call the "atomic nucleus," rather than being spread out and determining the size of the atom, as suggested by Thomson's model. This momentous and unexpected discovery completely reversed Thomson's idea: Somehow, the negatively charged electrons were on the *outside* of the atom, and determined its volume, just the opposite of the picture in Figure 6. How can that be?

The fact that electrons determine the physical size of an atom suggested that they also determine the nature of atomic interactions, and thus the periodic nature of chemical and physical interactions as

summarized in Mendeleev's Table, a fact already intuited by Lewis, based on chemical evidence. At ordinary energies, and at room temperature and lower, nuclei never come into contact. But, then one had to ask: How do those very light electrons manage to fill up almost all of the volume of an atom, and how do they determine the atom's chemical and physical properties?

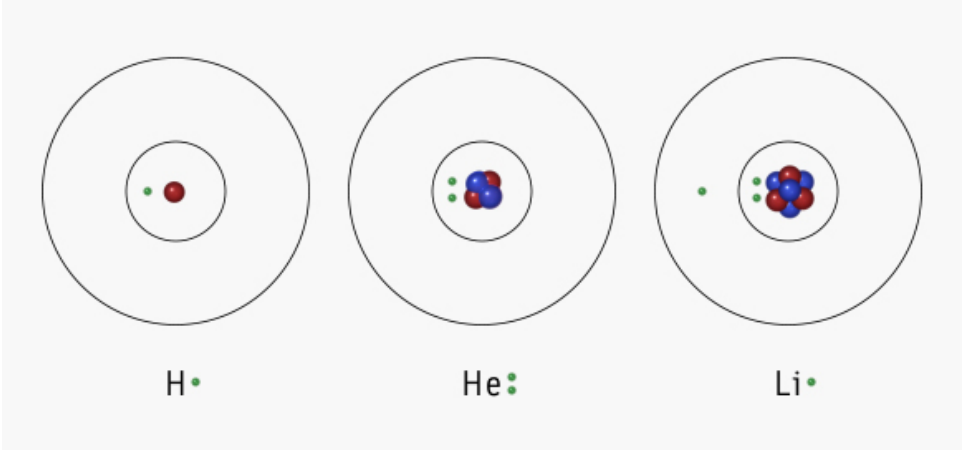


Figure 7: The Lewis shell model for the first three atoms in the modern periodic table.
Source:

A start had already been made. In 1902, when Rutherford's atomic nucleus was still almost a decade in the future, American chemist G. N. Lewis, had already proposed an empirically-developed shell model to explain how electrons could run the show, even if he was lacking in supplying a detailed model of how they actually did so.

He suggested that the first atomic shell (or *kernel* as he originally called it) held two electrons at maximum, the second and third shells a maximum of eight, and the fourth up to 18 additional electrons. Thus, for neutral atoms and their ion, the "magic numbers" 2, 8, 8, and 18 are associated with special chemical stability. Electrons in atoms or ions outside of these special closed shells are referred to as valence electrons, and determine much of the physical and chemical behavior of an atom. For example, in Unit 8, we will learn that atoms in metallic solids lose their valence electrons, and the remaining ionic cores form a metallic crystal, with the former valence electrons moving freely like water in a jar of beads, and not belonging to any specific ion. In doing so, they may freely conduct electrical currents (and heat), or under special circumstances may also become superconductors, allowing these free electrons to flow without resistance or energy dissipation.

Lewis also assumed that chemical bonding took place in such a way that stable molecules had fully filled shells, and that they formed these full shells by sharing electrons in pairs. The simplest example is the

formation of the hydrogen molecule. He denoted a hydrogen atom by H•, where the • is the unpaired electron in the atom's half-filled first shell. Why two atoms in a hydrogen molecule? The answer is easy in Lewis's picture: H:H. This pair of dots denotes that two shared electrons form the bond that holds the H_2 molecule together, where the subscript means that the molecule consists of two hydrogen atoms. As there are now no longer any unpaired electrons, we don't expect to form H_3 or H_4. Similarly, helium, which already has the filled shell structure He:, has no unpaired electrons to share, so does not bond to itself or to other atoms.

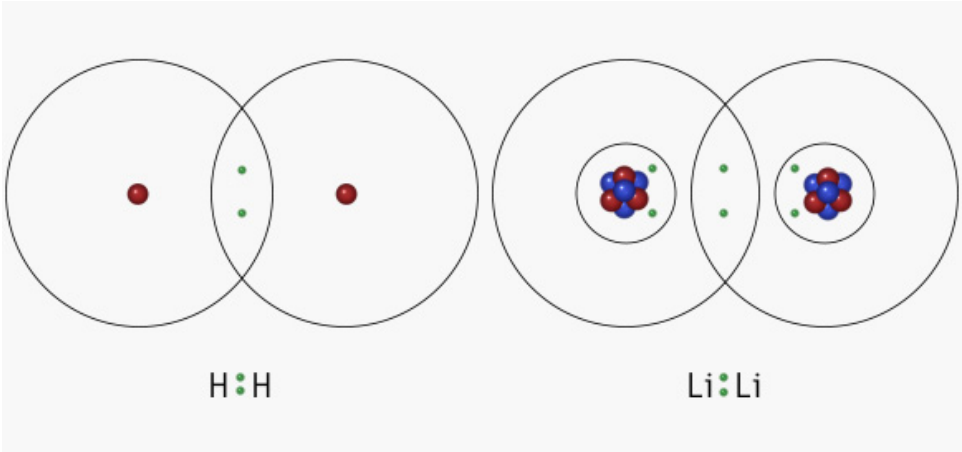


Figure 8: Molecules in the Lewis shell picture: the pair bond for H_2 and Li_2.
Source:

Next, Lewis introduced his famous counting rule (still used in models of bonding in much of organic chemistry and biochemistry): As the electrons are shared in overlapping shells, we count them twice in the dot picture for H_2, one pair for each H atom. Thus, in the Lewis manner of counting, each H atom in H_2 has a filled shell with two electrons just like Helium: He:. Here, we have the beginning of the concept of pairs or pairing, albeit in the form of a very simple empirical model. Such pairings will soon dominate all of our discussion.

How can those tiny electrons determine the size of an atom?

Lewis's model implies certain rules that allow us to understand how to build up an atom from its parts, and for building up molecules from atoms, at least for the electrons that determine the structure of the periodic table and chemical bonding. Another set of such rules tells us how to form the nucleus of an atom from its constituent neutrons and protons. The formation of nuclei is beyond the scope of this unit, but we should note that even these rules involve pairing.

We can take such a blithe view of the structure of the nucleus because this first discussion of atoms, molecules, solids, and macroscopic quantum systems involves energies far too low to cause us to worry about nuclear structure in any detail. It is the arrangement of and behavior of the electrons with respect to a given nucleus, or set of nuclei, that determines many of the properties of the systems of interest to us. However, we should not forget about the nucleus altogether. Once we have the basic ideas involving the structure of atoms and molecules at hand, we will ask whether these composite particles rather than their constituent parts are bosons or fermions. When we do this, we will suddenly become very interested in certain aspects of the nucleus of a given atom. But, the first critical point about the nucleus in our initial discussion involves its size in comparison with that of the atom of which it is, somehow, the smallest component.

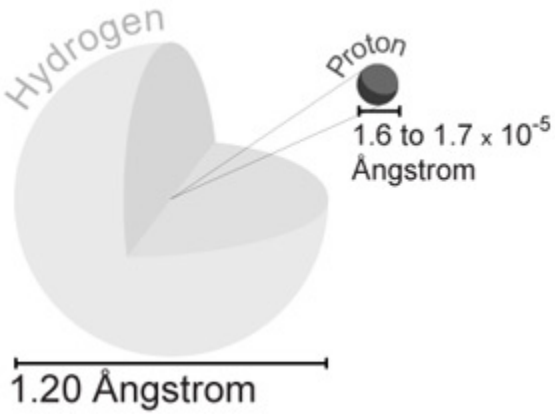


Figure 9: Rutherford's model of a hydrogen atom.
Source: © Wikimedia Commons, Public Domain. Author: Bensaccount, 10 July 2006.

Although it contains almost all of the atom's mass and its entire positive electrical charge, the nucleus is concentrated in a volume of 1 part in 10^{15} (that's one thousand trillion) of the physical volume of the atom. The size of the atom is determined by the electrons, which, in turn, determine the size of almost everything, be it solid or liquid, made of matter that we experience in daily life. For example, the volume of water in a glass of water is essentially the volume of the atoms comprising the water molecules that fill the glass. However, if the water evaporates or becomes steam when heated, its volume is determined by the size of the container holding the gaseous water. That also applies to the gaseous ultra-cold trapped atoms that we first met in Unit 5 and will encounter again later in this unit as Bose-Einstein condensates (BECs). They expand to fill the traps that confine them. How is it that these negatively charged and, in comparison to nuclei, relatively massless electrons manage to take up all that space?

Atoms and their nuclei

The atomic nucleus consists of neutrons and protons. The number of protons in a nucleus is called the atomic number, and is denoted by Z. The sum of the number of protons and neutrons is called the mass number, and is denoted by A. The relation between these two numbers Z and A will be crucial in determining whether the composite neutral atom is a boson or fermion. The volume of a nucleus is approximately the sum of the volumes of its constituent neutrons and protons, and the nuclear mass is approximately the sum of the masses of its constituent neutrons and protons. The rest of the atom consists of electrons, which have a mass of about 1/2,000 of that of a neutron or proton, and a negative charge of exactly the same magnitude as the positive charge of the proton. As suggested by its name, the neutron is electrically neutral. The apparently exact equality of the magnitudes of the electron and proton charges is a symmetry of the type we encountered in Units 1 and 2.

Do electrons have internal structure?

Now back to our earlier question: How can an atom be so big compared to its nucleus? One possibility is that electrons resemble big cotton balls of negative charge, each quite large although not very massive, as shown in Figure 10 below. As they pack around the nucleus, they take up lots of space, leading to the very much larger volume of the atom when compared to the volume of the nucleus.

Figure 10: A cotton ball model of an atom.
Source: © William P. Reinhardt.

However, there is a rather big problem with the simple cotton ball idea. When particle physicists try to measure the radius or volume of an individual electron, the best answer they get is zero. Said another way, no measurement yet made has a spatial resolution small enough to measure the size of an individual electron thought of as a particle. We know that electrons are definitely particles. With modern technology we can count them, even one at a time. We also know that each electron has a definite amount of mass, charge, and—last, but not at all least, as we will soon see—an additional quantity called spin. In spite of all this, physicists still have yet to observe any internal structure that accounts for these properties.

It is the province of string theory, or some yet-to-be created theory with the same goals, to attempt to account for these properties at length scales far too small for any current experiments to probe. Experimentalists could seek evidence that the electron has some internal structure by trying to determine whether it has a dipole moment. Proof of such a moment would mean that the electron's single unit of fundamental negative charge is not uniformly distributed within the electron itself. Since the Standard Model introduced in Unit 1 predicts that the electron doesn't have a dipole moment, the discovery of such an internal structure would greatly compromise the model.

For now, though, we can think of an electron as a mathematical point. So how do the electrons take up all the space in an atom? They certainly are not the large cotton balls we considered above; that would make everything too simple, and we wouldn't need quantum mechanics. In fact we do need quantum mechanics in many ways. Ironically, the picture that quantum mechanics gives, with its probability interpretation of an atomic wavefunction, will bring us right back to cotton balls, although not quite those of Figure 10.

Section 3: The Quantum Atom

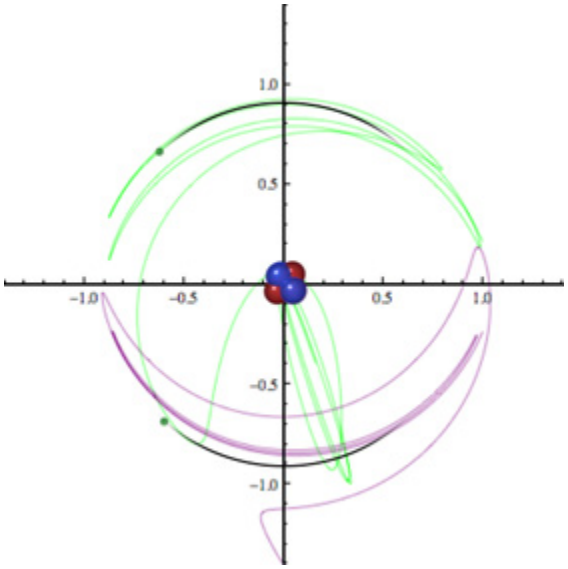


Figure 11: A simple classical model fails to explain the stability of the helium atom.
Source: © William P. Reinhardt, 2010.

As we learned in Unit 5, quantum theory replaces Bohr's orbits with standing, or stationary state, wavefunctions. Just as in the Bohr picture, each wavefunction corresponds to the possibility of having the system, such as an electron in an atom, in a specific and definite energy level. As to what would happen when an atom contained more than one electron, Bohr was mute. Astronomers do not need to consider the interactions between the different planets in their orbits around the Sun in setting up a first approximation to the dynamics of the solar system, as their gravitational interactions with each other are far weaker than with the Sun itself. The Sun is simply so massive that to a good first approximation it controls the motions of all the planets. Bohr recognized that the same does not apply to the classical motion of electrons in the next most complex atom: Helium, with a charge of +2 on its nucleus and two moving electrons. The electrons interact almost as strongly with each other as with the nucleus that holds the whole system together. In fact, most two-electron classical orbits are chaotically unstable, causing theoretical difficulties in reconciling the classical and quantum dynamics of helium that physicists have overcome only recently and with great cleverness.

Perhaps surprisingly, the energy levels described by Schrödinger's standing wave picture very quickly allowed the development of an approximate and qualitative understanding of not only helium, but most of the periodic table of chemical elements.

Energy levels for electrons in atoms

What is the origin of the Lewis filled shell and electron pair bonding pictures? It took the full quantum revolution described in Unit 5 to find the explanation—an explanation that not only qualitatively explains the periodic table and the pair bond, but also gives an actual theory that allows us to make quantitative computations and predictions.

When an atom contains more than one electron, it has different energies than the simple hydrogen atom; we must take both the quantum numbers n (from Unit 5) and l (which describes a particle's quantized angular momentum), into account. This is because the electrons do not move independently in a many-electron atom: They notice the presence of one another. They not only affect each other through their electrical repulsion, but also via a surprising and novel property of the electron, its spin, which appeared in Unit 5 as a quantum number with the values $\pm 1/2$, and which controls the hyperfine energies used in the construction of phenomenally accurate and precise atomic clocks.

The effect of electron spin on the hyperfine energy is tiny, as the magnetic moment of the electron is small. On the other hand, when two spin-1/2 electrons interact, something truly incredible happens if the two electrons try to occupy the same quantum state: In that case, one might say that their interaction becomes infinitely strong, as they simply cannot do it. So, if we like, we can think of the exclusion principle mentioned in the introduction to this unit as an extremely strong interaction between identical fermions.

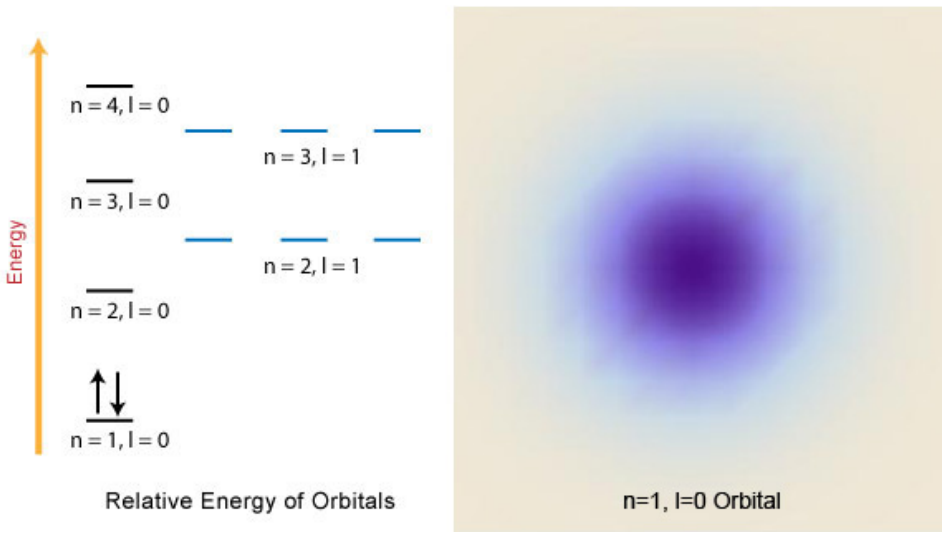


Figure 12: The ground state of helium: energy levels and electron probability distribution.
Source:

We are now ready to try building the periodic table using a simple recipe: To get the lowest energy, or ground state of an atom, place the electrons needed to make the appropriate atomic or ionic system in their lowest possible energy levels, noting that two parallel spins can never occupy the same quantum state. This is called the Pauli exclusion principle. Tradition has us write spin +1/2 as ↑ and spin -1/2 as ↓, and these are pronounced "spin up" and "spin down." In this notation, the ground state of the He atom would be represented as n = 1 and ↑↓, meaning that both electrons have the lowest energy principal quantum number n = 1, as in Unit 5, and must be put into that quantum state with opposite spin projections.

Quantum mechanics also allows us to understand the size of atoms, and how seemingly tiny electrons take up so much space. The probability density of an electron in a helium atom is a balance of three things: its electrical attraction to the nucleus, its electrical repulsion from the other electron, and the fact that the kinetic energy of an electron gets too large if its wavelength gets too small, as we learned in Unit 5. This is actually the same balance between confinement and kinetic energy that allowed Bohr, with his first circular orbit, to also correctly estimate the size of the hydrogen atom as being 10^5 times larger than the nucleus which confines it.

Assigning electrons to energy levels past H and He

Now, what happens if we have three electrons, as in the lithium (Li) atom? Lewis would write Li•, simply not showing the inert inner shell electrons. Where does this come from in the quantum mechanics of the *aufbau*? Examining the energy levels occupied by the two electrons in the He atom shown in Figure 12 and thinking about the Pauli principle make it clear that we cannot simply put the third electron in the n = 1 state. If we did, its spin would be parallel to the spin of one of the other two electrons, which is not allowed by the exclusion principle. Thus the ground state for the third electron must go into the next lowest unoccupied energy level, in this case n = 2, l = 0.

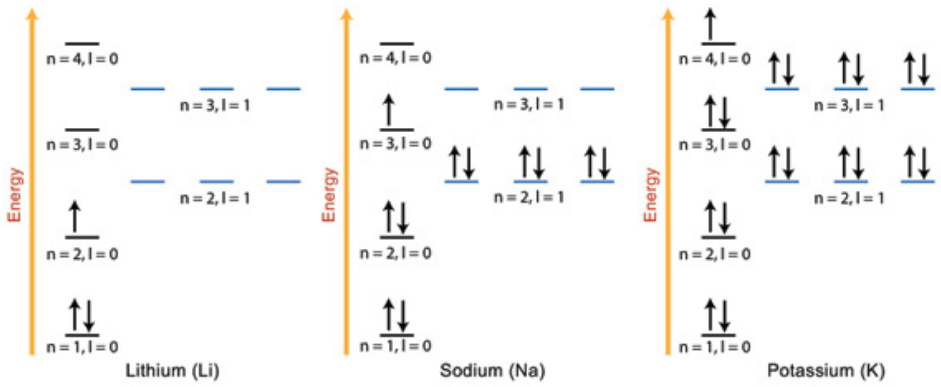


Figure 13: Electrons in atomic energy levels for Li, Na, and K
Source:

Using the exclusion principle to determine electron occupancy of energy levels up through the lithium (Li), sodium (Na), and potassium (K) atoms, vindicates the empirical shell structures implicit in the Mendeleev table and explicit in Lewis's dot diagrams. Namely Li, Na, and K all have a single unpaired electron outside of a filled (and thus inert and non-magnetic) shell. It is these single unpaired electrons that allow these alkali atoms to be great candidates for making atomic clocks, and for trapping and making ultra-cold gases, as the magnetic traps grab that magnetic moment of that unpaired electron.

Where did this magical Pauli exclusion principle come from? Here, as it turns out, we need an entirely new, unexpected, and not at all intuitive fundamental principle. With it, we will have our first go-around at distinguishing between fermions and bosons.

Section 4: Spin, Bosons, and Fermions

In spite of having no experimentally resolvable size, a single electron has an essential property in addition to its fixed charge and mass: a magnetic moment. It may line up, just like a dime store magnet, north-south or south-north in relation to a magnetic field in which it finds itself. These two possible orientations correspond to energy levels in a magnetic field determined by the magnet's orientation. This orientation is quantized in the magnetic field. It turns out, experimentally, that the electron has only these two possible orientations and energy levels in a magnetic field. The electron's magnetic moment is an internal and intrinsic property. For historical reasons physicists called it spin, in what turned out to be a bad analogy to the fact that in classical physics a rotating spherical charge distribution, or a current loop as in a superconducting magnet, gives rise to a magnetic moment. The fact that only two orientations of such a spin exist in a magnetic field implies that the quantum numbers that designate spin are not integers like the quantum numbers we are used to, but come in half-integral amounts, which in this case are $\pm 1/2$. This was a huge surprise when it was first discovered.

Figure 14: The Stern-Gerlach experiment demonstrated that spin orientation is quantized, and that "up" and "down" are the only possibilities.
Source: © Wikimedia Commons, GNU Free Documentation License, Version 1.2. Author: Peng, 1 July 2005.

What does this mean? A value of 1/2 for the spin of the electron pops out of British physicist Paul Dirac's relativistic theory of the electron; but even then, there is no simple physical picture of what the spin corresponds to. Ask a physicist what "space" spin lives in, and the answer will be simple: Spins are mathematical objects in "spin space." These spins, if unpaired, form little magnets that can be used to trap and manipulate the atoms, as we have seen in Unit 5, and will see again below. But spin itself has much far-reaching implications. The idea of "spin space" extends itself into "color space," "flavor space,"

"strangeness space," and other abstract (but physical, in the sense that they are absolutely necessary to describe what is observed as we probe more and more deeply into the nature of matter) dimensions needed to describe the nature of fundamental particles that we encountered in Units 1 and 2.

Spin probes of unusual environments

Atoms, Nuclei, and Nomenclature

The nuclei of atoms never come into contact with one another at the relatively low energies and moderate temperatures that correspond to our everyday experience. But those conditions do not apply to detonation of nuclear bombs or in nuclear reactors: here nuclei are in direct contact and either split or join together, releasing enormous amounts of energy. Atomic energy and atomic bombs involve the properties of nuclei rather their atoms. But in the 1940s, when they were being developed, scientists and politicians were afraid to use the word nucleus in public, fearing that nobody would know what that meant; so they stuck with the somewhat more familiar word atom. This type of fiction continues today. Because the phrases nuclear energy and nuclear reactors tend to make people nervous, hospitals using the diagnostic tool MRI (Magnetic Resonance Imaging) that involves direct probing of atomic nuclei choose not to mention that nuclei are involved. MRI is a technique that American physicists Ed Purcell and Bob Pound, who invented it, called NMR, for Nuclear Magnetic Resonance. Chemists, biophysicists, and biologists who use the technology to determine how proteins work, for example, call it NMR. But nonscientists who are afraid of nuclei don't want to hear anything about them, so NMR becomes MRI when doctors use it to image soft tissues in the body.

Is spin really important? Or might we just ignore it, as the magnetic moments of the proton and electron are small and greatly overwhelmed by their purely electrical interactions? In fact, spin has both straightforward implications and more subtle ones that give rise to the exclusion principle and determine whether composite particles are bosons or fermions.

The more direct implications of the magnetic moments associated with spin include two examples in which we can use spins to probe otherwise hard-to-reach places: inside our bodies and outer space. It turns out that the neutron and proton also have spin-1/2 and associated magnetic moments. As these magnetic moments may be oriented in only two ways in a magnetic field, they are usually denoted by the ideograms ↑ and ↓ for the spin projections +1/2 (spin up) and -1/2 (spin down). In a magnetic field, the protons in states ↑ and ↓ have different energies. In a strong external magnetic field, the spectroscopy of the transitions between these two levels gives rise to Nuclear Magnetic Resonance (NMR), a.k.a. MRI in medical practice (see Figure 3). Living matter contains lots of molecules with hydrogen atoms, whose nuclei can be flipped from spin up to spin down and vice versa via interaction with very low energy electromagnetic radiation, usually in the radiowave regime. The images of these atoms and their

interactions with nearby hydrogen atom nuclei provide crucial probes for medical diagnosis. They also support fundamental studies of chemical and biochemical structures and dynamics, in studies of the folding and unfolding of proteins, for example.

Another example of the direct role of the spins of the proton and electron arises in astrophysics. In a hydrogen atom, the spin of the proton and electron can be parallel or anti-parallel. And just as with real magnets, the configuration $p(\uparrow)e(\downarrow)$ has lower energy than $p(\uparrow)e(\uparrow)$. This small energy difference is due to the hyperfine structure of the spectrum of the hydrogen atom reviewed in Unit 5. The photon absorbed in the transition $p(\uparrow)e(\downarrow) \rightarrow p(\uparrow)e(\uparrow)$ or emitted in the transition $p(\uparrow)e(\uparrow) \rightarrow p(\uparrow)e(\downarrow)$ has a wavelength of 21 centimeters, in the microwave region of the electromagnetic spectrum. Astronomers have used this 21 centimeter radiation to map the density of hydrogen atoms in our home galaxy, the Milky Way, and many other galaxies.

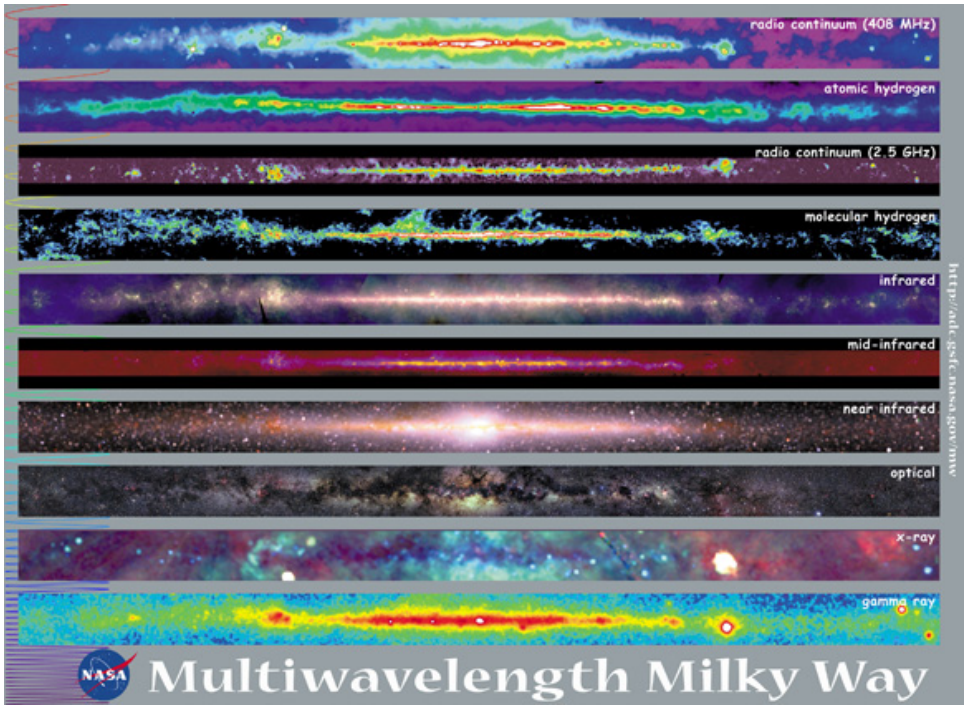


Figure 15: Our galaxy, imaged at many different wavelengths.
Source: © NASA.

Electrons are fermions

However, there is more: That a particle has spin 1/2 means more than that it has only two possible orientations in a magnetic field. Fundamental particles with intrinsic spin of 1/2 (or any other half-integer

spin, such as 3/2 or 5/2 or more whose numerators are odd numbers) share a specific characteristic: They are all fermions; thus electrons are fermions. In contrast, fundamental particles with intrinsic spin of 0, 1, 2, or any integral number are bosons; so far, the only boson we have met in this unit is the photon.

Is this a big deal? Yes, it is. Applying a combination of relativity and quantum theory, Wolfgang Pauli showed that identical fermions or bosons in groups have very different symmetry properties. No two identical fermions can be in the same quantum state in the same physical system, while as many identical bosons as one could wish can all be in exactly the same quantum state in a single quantum system. Because electrons are fermions, we now know the correct arrangement of electrons in the ground state of lithium. As electrons have only two spin orientations, ↑ or ↓, it is impossible to place all three electrons in the lowest quantum energy level; because at least two of the three electrons would have the same spin, the state is forbidden. Thus, the third electron in a lithium atom must occupy a higher energy level. In recognition of the importance of spin, the Lewis representation of the elements in the first column of the periodic table might well be H↑, Li↑, Na↑, K↑, and Rb↑, rather than his original H•, Li•, Na•, K•, and Rb•. The Pauli symmetry, which leads to the exclusion principle, gives rise to the necessity of the periodic table's shell structure. It is also responsible for the importance of Lewis's two electron chemical pair bond, as illustrated in Figure 16.

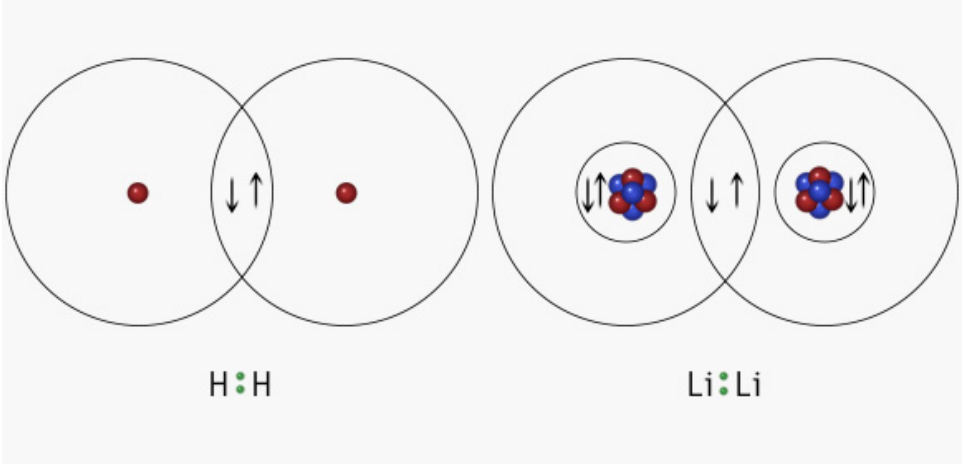


Figure 16: Spin pairing in the molecules H_2 and Li_2.
Source:

The Pauli rules apply not only to electrons in atoms or molecules, but also to bosons and fermions in the magnetic traps of Unit 5. These may be fundamental or composite particles, and the traps may be

macroscopic and created in the laboratory, rather than electrons attracted by atomic nuclei. The same rules apply.

Conversely, the fact that bosons such as photons have integral spins means that they can all occupy the same quantum state. That gives rise to the possibility of the laser, in which as many photons as we wish can bounce back and forth between two mirrors in precisely the same standing wave quantum state. We can think of this as a macroscopic quantum state. This is certainly another example of particles in an artificial trap created in the laboratory, and an example of a macroscopic quantum state. Lasers produce light with very special and unusual properties. Other bosons will, in fact, be responsible for all the known macroscopic quantum systems that are the real subject of this and several subsequent units.

Photons are bosons

When Max Planck introduced the new physical constant h that we now call Planck's constant, he used it as a proportionality constant to fit the data known about blackbody radiation, as we saw in Unit 5. It was Albert Einstein who noted that the counting number n that Planck used to derive his temperature dependent emission profiles was actually counting the number of light quanta, or photons, at frequency ν, and thus that the energy of one quantum of light was $h\nu$. If one photon has energy $h\nu$, then n photons would have energy $nh\nu$. What Planck had unwittingly done was to quantize electromagnetic radiation into energy packets. Einstein and Planck both won Nobel prizes for this work on quantization of the radiation field.

What neither Planck nor Einstein realized at the time, but which started to become clear with the work of the Indian physicist Satyendra Bose in 1923, was that Planck and Einstein had discovered that photons were a type of particle we call "bosons," named for Bose. That is, if we think of the frequency as describing a possible mode or quantum state of the Planck radiation, then what Planck's $nh\nu$ really stated was that any number, n = 0, 1, 2, 3..., of photons, each with its independent energy $h\nu$ could occupy the same quantum state.

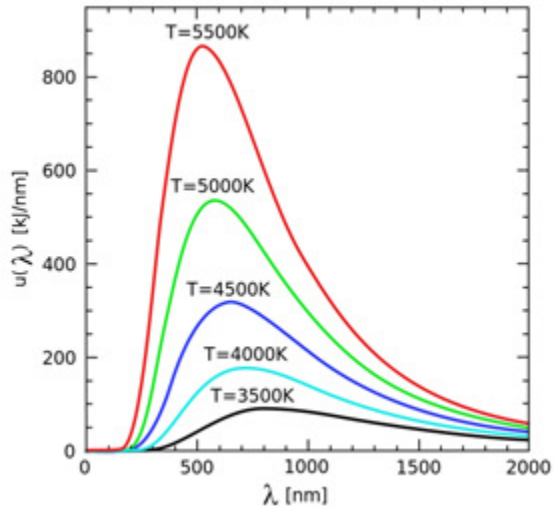


Figure 17: The spectrum of blackbody radiation at different temperatures.
Source: © Wikimedia Commons, GNU Free Documentation license, Version 1.2. Author: 4C, 04 August 2006.

In the following year, Einstein also suggested in a second paper following the work of Bose, that atoms or molecules might be able to behave in a similar manner: Under certain conditions, it might be possible to create a gas of particles all in the same quantum state. Such a quantum gas of massive particles came to be known as a Bose-Einstein condensate (BEC) well before the phenomenon was observed. Physicists observed the first gaseous BEC 70 years later, after decades of failed attempts. But, in 1924, even Einstein didn't understand that this would not happen for just any atom, but only for those atoms that we now refer to as bosons. Fermionic atoms, on the other hand, would obey their own exclusion principle with respect to their occupation of the energy levels of motion in the trap itself.

Section 5: Composite Bosons and Fermions

Gained in Translation

Source: © Wikimedia Commons, Public Domain.

In 1923 Satyendra Nath Bose, an Indian physicist working alone, outside of the European physics community, submitted a short article to a leading British physics journal. The article presented a novel derivation of the Planck distribution and implicitly introduced the concept of equal frequency photons as identical bosons. After the editors rejected the paper, Bose sent it to Einstein. Einstein translated the paper into German and submitted it for publication in a leading German physics journal with an accompanying paper of his own. These two papers set the basis for what has become called Bose-Einstein statistics of identical bosons. In his own accompanying paper, Einstein pointedly remarks that Bose's paper is of exceptional importance. But he then indicates that it contains a great mystery, which he himself did not understand. Perhaps the larger mystery is why Einstein, the first person to use the term quantum in his realization that Planck's E=nhν was the idea that light came in packets or quanta" of energy, didn't put the pieces together and discover modern quantum theory five years before Heisenberg and Schrödinger.

Whether atoms and molecules can condense into the same quantum state, as Einstein predicted in 1924, depends on whether they are bosons or fermions. We therefore have to extend Pauli's definition of bosons and fermions to composite particles before we can even talk about things as complex as atoms being either of these. As a start, let's consider a composite object made from two spin-1/2 fundamental particles. The fundamental particles are, of course, fermions; but when we combine them, the total spin is

either $1/2+1/2=1$ or $1/2-1/2=0$. This correctly suggests that two spin-1/2 fermions may well combine to form a total spin of 1 or 0. In either case, the integer spin implies that the composite object is a boson.

Why the caveat "may well"? As in many parts of physics, the answer to this question depends on what energies we are considering and what specific processes might occur. We can regard a star as a structureless point particle well described by its mass alone if we consider the collective rotation of billions of such point masses in a rotating galaxy. However, if two stars collide, we need to consider the details of the internal structure of both. In the laboratory, we can think of the proton (which is a composite spin-1/2 fermion) as a single, massive, but very small and inert particle with a positive unit charge, a fixed mass, and a spin of 1/2 at low energies. But the picture changes at the high energies created in the particle accelerators we first met in Unit 1. When protons moving in opposite directions at almost the speed of light collide, it is essential to consider their internal structures and the new particles that may be created by the conversion of enormous energies into mass. Similarly two electrons might act as a single boson if the relevant energies are low enough to allow them to do so, and if they additionally have some way to actually form that composite boson; this requires the presence of "other particles," such as an atomic nucleus, to hold them together. This also usually implies low temperatures. We will discuss electrons paired together as bosons below and in Unit 8. A hydrogen atom, meanwhile, is a perfectly good boson so long as the energies are low compared to those needed to disassemble the atom. This also suggests low temperatures. But, in fact, Einstein's conditions for Bose condensation require extraordinarily low temperatures.

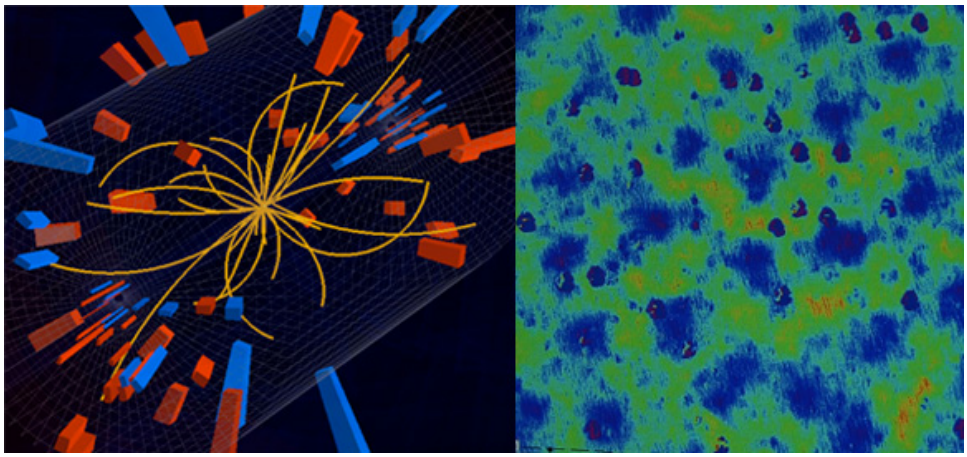


Figure 18: Protons in LHC collisions (left) and electrons in a superconductor (right) are examples of composite fermions and bosons.
Source: © CERN.

Helium as boson or fermion

When is helium a boson? This is a more complex issue, as the helium nucleus comes in two isotopes. Both have Z = 2, and thus two protons and two electrons. However, now we need to add the neutrons. The most abundant and stable form of helium has a nucleus with two protons and two neutrons. All four of these nucleons are spin-1/2 fermions, and the two protons pair up, as do the two neutrons. Thus, pairing is a key concept in the structure of atomic nuclei, as well as in the organization of electrons in the atom's outer reaches. So, in helium with mass number A = 4, the net nuclear spin is 0. Thus, the ^{4}He nucleus is a boson. Add the two paired electrons and the total atomic spin remains 0. So, both the nucleus and an atom of helium are bosons.

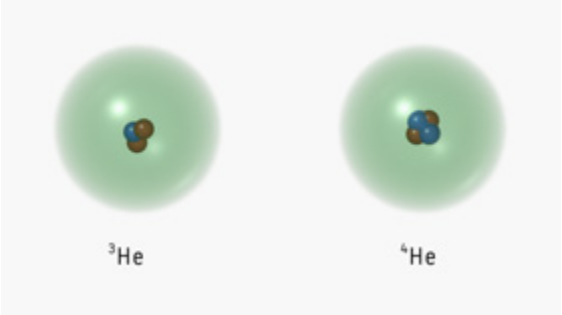


Figure 19: The two isotopes of helium: a fermion and a boson.
Source:

The chemical symbol He tells us that Z = 2. But we can add the additional information that $A = 4$ by writing the symbol for bosonic helium as ^{4}He, where the leading superscript 4 designates the atomic number. This extra notation is crucial, as helium has a second isotope. Because different isotopes of the same element differ only in the number of neutrons in the nucleus, they have different values of A. The second stable isotope of helium is ^{3}He, with only one neutron. The isotope has two paired protons and two paired electrons but one necessarily unpaired neutron. ^{3}He thus has spin-1/2 overall and is a composite fermion. Not surprisingly, ^{3}He and ^{4}He have very different properties at low temperatures, as we will discuss in the context of superfluids below.

Atomic bosons and fermions

More generally, as the number of spin-1/2 protons equals the number of spin-1/2 electrons in a neutral atom, any atom's identity as a composite fermion or boson depends entirely on whether it has an odd or

even number of neutrons, giving a Fermi atom and a Bose atom, respectively. We obtain that number by subtracting the atomic number, Z, from the mass number, A. Note that A is not entirely under our control. Experimentalists must work with the atomic isotopes that nature provides, or they must create novel ones, which are typically unstable; which isotopes these are depends on the rules for understanding the stability of nuclei. A difficult task as neutrons and protons interact in very complex ways, which befits their composite nature, and are not well represented by simple ideas like the electrical Coulomb's Law attraction of an electron for a proton.

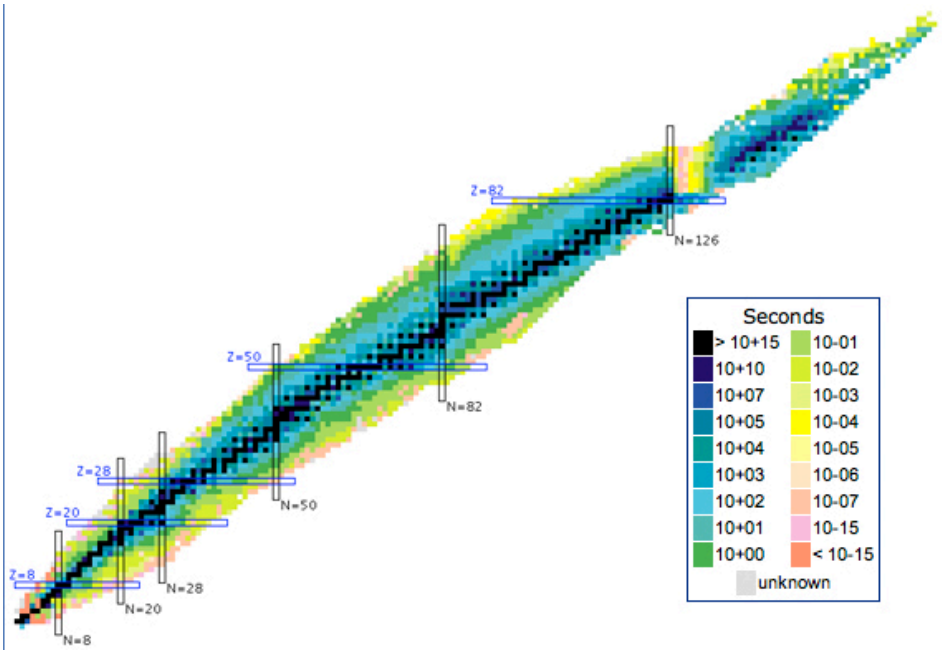


Figure 20: As this chart shows, not every imaginable nucleus is stable.
Source: © Courtesy of Brookhaven National Laboratory.

We are now in a position to understand why one would need bosonic atoms to make a BEC: Fermions cannot occupy the same macroscopic quantum state, but bosons can. And, we know how to recognize which atoms will actually be bosons. The first atomic gaseous Bose-Einstein condensates were made in Boulder, Colorado, using rubidium (Rb) atoms; in Cambridge, Massachusetts, using atomic sodium (Na); and in Houston, Texas, using atomic lithium (Li). These three elements are members of the alkali metal family, and share the property of having a single (and thus unpaired) electron outside a fully closed shell of electrons. These unpaired outer shell electrons with their magnetic moments allow them to be caught in a magnetic trap, as seen in Unit 5. If we want gases of these atoms to Bose condense, we must think counterintuitively: We need isotopes with odd values of A, so that the total number of spin-1/2 fermions —protons, electrons, and neutrons—is even. These alkali metals all have an odd number of protons, and

a matching odd number of electrons; therefore, we need an even number of neutrons. This all leads to A being an odd number and the alkali atom to being a boson. Thus the isotopes ^{7}Li, ^{23}Na, and ^{87}Rb are appropriate candidates, as they are composite bosons.

Section 6: Gaseous Bose-Einstein Condensates

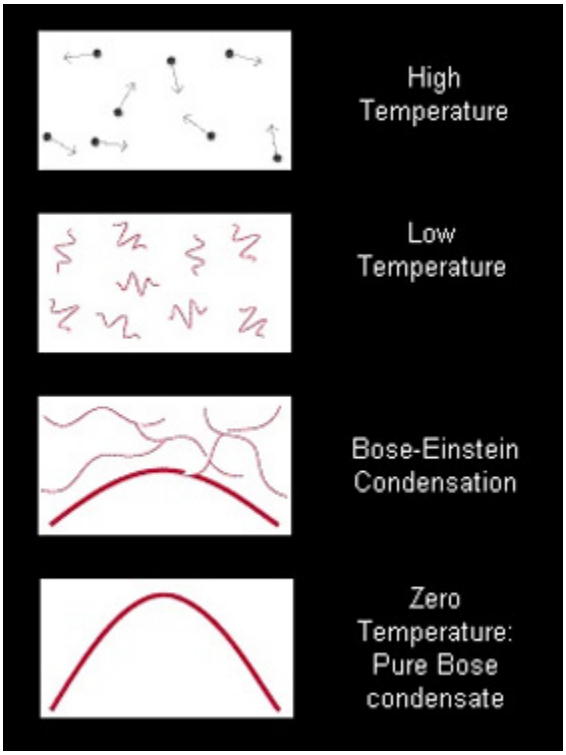


Figure 21: Atomic de Broglie waves overlap as temperatures are lowered.
Source: © Wolfgang Ketterle.

What does it take to form a gaseous macroscopic quantum system of bosonic atoms? This involves a set of very tricky hurdles to overcome. We want the atoms, trapped in engineered magnetic or optical fields, to be cooled to temperatures where, as we saw in Unit 5, their relative de Broglie wavelengths are large compared with the mean separation between the gaseous atoms themselves. As these de Broglie waves overlap, a single and coherent quantum object is formed. The ultimate wavelength of this (possibly) macroscopic quantum system is, at low enough temperatures, determined by the size of the trap, as that sets the maximum wavelength for both the individual particles and the whole BEC itself.

Einstein had established this condition in his 1924 paper, but it immediately raises a problem. If the atoms get too close, they may well form molecules: Li_2, Rb_2, and Na_2 molecules are all familiar characters in the laboratory. And, at low temperatures (and even room temperatures), Li, Na, and Rb are all soft metals: Cooling them turns them into dense hard metals, not quantum gases. Thus, experiments had to start with a hot gas (many hundreds of degrees K) and cool the gases in such a way that the atoms didn't simply

condense into liquids or solids. This requires keeping the density of the atomic gases very low—about a million times less dense than the density of air at the Earth's surface and more than a billion times less dense than the solid metallic forms of these elements.

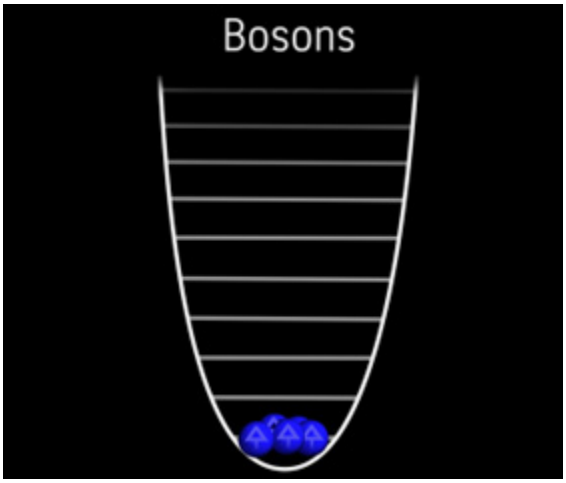


Figure 22: Atoms in a Bose condensate at 0 K.
Source:

Stating all this in terms of bosonic atoms in quantum energy levels, rather than overlapping de Broglie waves, leads to the picture shown in Figure 22. Here, we see the energy levels corresponding to the quantum motion of the atoms in the magnetic (or optical) trap made in the laboratory for collecting them. As the temperature cools, all of the bosonic atoms end up in the lowest energy level of the trap, just as all the photons in an ideal laser occupy a single quantum state in a trap made of mirrors.

The role of low temperature

The fact that gaseous atoms must be quite far apart to avoid condensing into liquids or solids, and yet closer than their relative de Broglie wavelengths, requires a very large wavelength, indeed. This in turn requires very slow atoms and ultra-cold temperatures. Laser cooling, as discussed in Unit 5, only takes us part of the way, down to about 10^{-6} (one one-millionth) of a degree Kelvin. To actually achieve the temperature of 10^{-8} K, or colder, needed to form a BEC, atoms undergo a second stage of cooling, ordinary evaporation, just like our bodies use to cool themselves by sweating. When the first condensates were made, no temperature this low had ever been created before, either in the laboratory or in nature. Figure 23 illustrates, using actual data taken as the first such condensate formed, the role of evaporative cooling and the formation of a BEC.

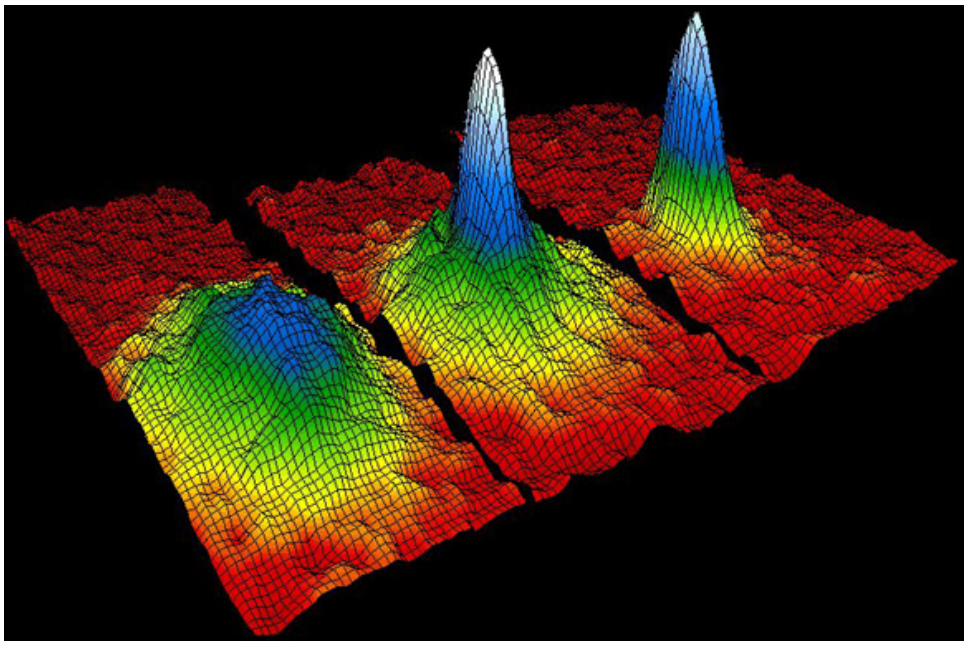


Figure 23: Three stages of cooling, and a quantum phase transition to a BEC.
Source: © Mike Matthews, JILA.

The process, reported in *Science* in July 1995, required both laser cooling and evaporative cooling of ^{87}Rb to produce a pretty pure condensate. Images revealed a sharp and smoothly defined Bose-Einstein condensate surrounded by many "thermal" atoms as the rubidium gas cooled to about 10^{-8} K. By "pretty pure," we mean that a cloud of uncondensed, and still thermal, atoms is still visible: These atoms are in many different quantum states, whereas those of the central peak of the velocity distribution shown in Figure 23 are in a single quantum state defined by the trap confining the atoms. Subsequent refinements have led to condensates with temperatures just above 10^{-12} K—cold, indeed, and with no noticeable cloud of uncondensed atoms. That tells us that all these many thousands to many millions of sodium or rubidium atoms are in a single quantum state.

The behavior of trapped atoms

How do such trapped atoms behave? Do they do anything special? In fact, yes, and quite special. Simulations and experimental observations show that the atoms behave very much like superfluids. An initial shaking of the trap starts the BEC sloshing back and forth, which it continues to do for ever longer-times as colder and colder temperatures are attained. This is just the behavior expected from such a gaseous superfluid, just as would be the case with liquid helium, ^{4}He.

The extent of the oscillating behavior depends on the temperature of the BEC. A theoretical computer simulation of such motion in a harmonic trap at absolute zero—a temperature that the laws of physics prevent us from ever reaching—shows that the oscillations would never damp out. But, if we add heat to the simulation, little by little as time passes, the behavior changes. The addition of increasing amounts of the random energy associated with heat causes the individual atoms to act as waves that get more and more out of phase with one another, and we see that the collective and undamped oscillations now begin to dissipate. Experimental observations readily reveal both the back-and-forth oscillations—collective and macroscopic quantum behavior taking place with tens of thousands or millions of atoms in the identical quantum state—and the dissipation caused by increased temperatures.

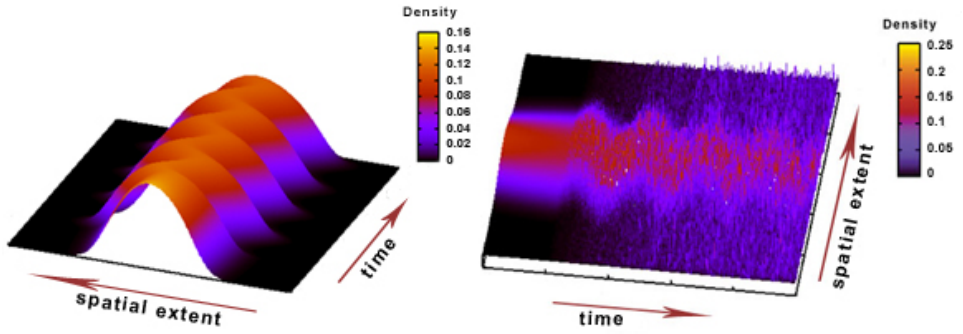


Figure 24: Super-oscillations of a quantum gas and their dissipation on heating.
Source: © William P. Reinhardt.

This loss of phase coherence leads to the dissipation that is part of the origins of our familiar classical world, where material objects seem to be in one place at a time, not spread out like waves, and certainly don't show interference effects. At high temperatures, quantum effects involving many particles "de-phase" or "de-cohere," and their macroscopic quantum properties simply vanish. Thus, baseballs don't diffract around a bat, much to the disappointment of the pitcher and delight of the batter. This also explains why a macroscopic strand of DNA is a de-cohered, and thus classical, macroscopic molecule, although made of fully quantum atoms.

Gaseous atomic BECs are thus large and fully coherent quantum objects, whereby coherent we imply that many millions of atoms act together as a single quantum system, with all atoms in step, or in phase, with one another. This coherence of phase is responsible for the uniformly spaced parallel interference patterns that we will meet in the next section, similar to the coherence seen in the interference of laser light shown in the introduction to this unit. The difference between these gaseous macroscopic quantum systems and liquid helium superfluids is that the quantum origins of the superfluidity of liquid helium are

hidden within the condensed matter structure of the liquid helium itself. So, while experiments like the superfluid sloshing are easily available, interference patterns are not, owing to the "hard core" interactions of the helium atoms at the high density of the liquid.

Section 7: Phase Control and New Structures

We have mentioned several times that laser light and the bosonic atoms in a gaseous BEC are coherent: The quantum phase of each particle is locked in phase with that of every other particle. As this phase coherence is lost, the condensate is lost, too. What is this quantum phase? Here, we must expand on our earlier discussion of standing waves and quantum wavefunctions. We will discuss what is actually waving in a quantum wave as opposed to a classical one, and acknowledge where this waving shows up in predictions made by quantum mechanics and in actual experimental data. Unlike poetry or pure mathematics, physics is always grounded when faced with experimental fact.

Imaginary Numbers and Quantum Waves

What do complex numbers have to do with waves? We can describe familiar waves such as those on the ocean in terms of trigonometric sines and cosines. But quantum waves are neither sines nor cosines; they are complex linear combinations of the two. We need this combination to account for the fact that, while wave amplitudes are definitely time-dependent, the probability densities of quantum mechanical stationary states don't depend on time at all, which is why we call them stationary states. How does this happen, and where do the complex numbers and functions come from? It turns out that the time-dependent part of the wave amplitude is a (complex) linear combination: cosine$(Et/\hbar)$ + i sine$(Et/\hbar)$, where i is $\sqrt{-1}$, t is the time, and E the energy of the quantum stationary state. A compact representation of this sum of real and imaginary sines and cosines is given by a remarkable, and intriguing formula due to the great Swiss mathematician Euler: $e^{i\varphi} = \cos(\varphi) + i \sin(\varphi)$. Physicist Richard Feynman often called this Euler formula one of the most unexpected and amazing results in all of mathematics; for example, if we choose $x = \pi$, we find that $e^{i\pi} + 1 = 0$, which, improbably at first sight, relates the five most fundamental numerical constants in all of mathematics: 0, 1, i, e and π. Here e is the base of natural logarithms, π the ratio of the circumference of a circle to its diameter, and i the square root of -1. Thus, we can now honestly state what is waving as time passes.

Like vibrating cello strings or ripples in the ocean, quantum wavefunctions are waves in both space and time. We learned in Unit 5 that quantum mechanics is a theory of probabilities. The full story is that quantum wavefunctions describe complex waves. These waves are sometimes also called "probability amplitudes," to make clear that it is not the wavefunction itself which is the probability. Mathematically, they are akin to the sines and cosines of high school trigonometry, but with the addition of i, the square root of -1, to the sine part. So, the wave has both a real part and an imaginary part. This is represented mathematically as $e^{i\varphi}$ (see sidebar), where the $i\varphi$ is called the "complex phase" of the wavefunction.

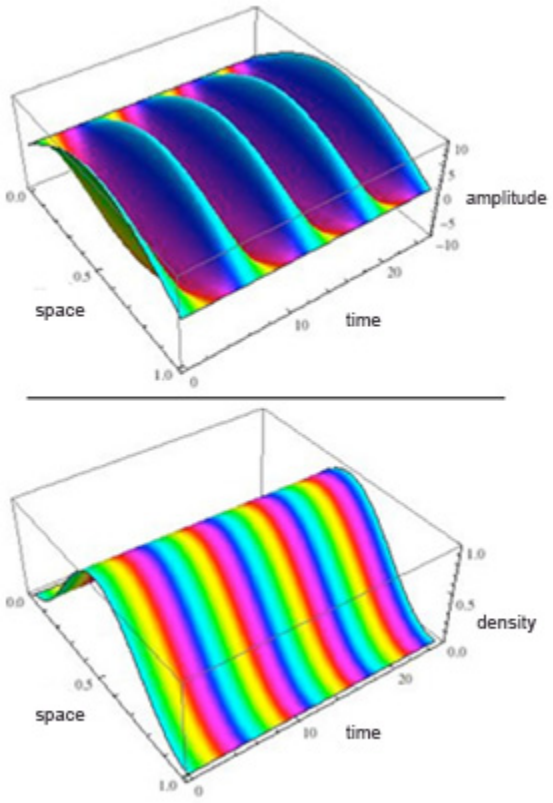


Figure 25: While the complex part of a quantum wavefunction "waves," the probability density does not.
Source: © William P. Reinhardt.

The probability density (or probability distribution), which tells us how likely we are to detect the particle in any location in space, is the absolute value squared of the complex probability amplitude, and as probabilities should be, they are real and positive. What is this odd-sounding phrase "absolute value squared"? A particle's probability density, proportional to the absolute value of the wavefunction squared, is something we can observe in experiments, and is always measured to be a positive real number. Our detectors cannot detect imaginary things. One of the cardinal rules of quantum mechanics is that although it can make predictions that seem strange, its mathematical description of physics must match what we observe, and the absolute value squared of a complex number is always positive and real.
See the math

What about the stationary states such as the atomic energy levels discussed both in this unit and in Unit 5? If a quantum system is in a stationary state, nothing happens, as time passes, to the observed

probability density: That's, of course, why it's called a stationary state. It might seem that nothing is waving at all, but we now know that what is waving is the complex phase.

For a stationary state, there is nothing left of the underlying complex waving after taking the absolute value squared of the wavefunction. But if two waves meet and are displaced from one another, the phases don't match. The result is quantum interference, as we have already seen for light waves, and will soon see for coherent matter waves. Experimentalists have learned to control the imaginary quantum phase of the wavefunction, which then determines the properties of the quantum probability densities. Phase control generates real time-dependent quantum phenomena that are characteristic of both BECs and superconductors, which we will explore below.

Phase imprinting and vortices

When does this hidden phase that we don't see in the experimentally measurable probability density become manifest? The simplest example is in interference patterns such as those shown in Figure 26. Here, the phase difference between two macroscopic quantum BECs determines the locations of the regions of positive and negative interference. So, situations do exist in which this mysterious phase becomes actual and measurable: when interference patterns are generated.

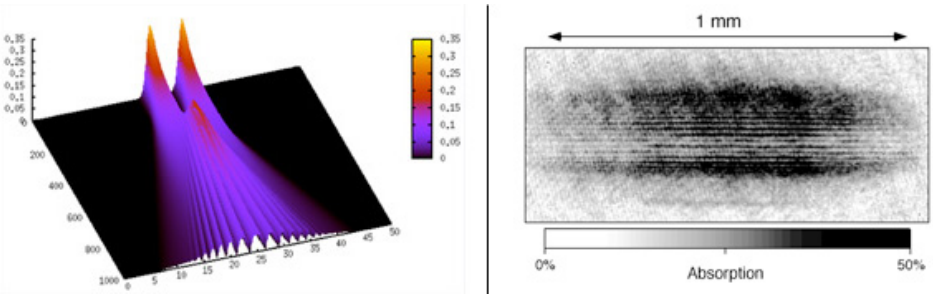


Figure 26: Interference of two coherent BECs, separated and allowed to recombine.
Source: © William P. Reinhardt (left image) and Wolfgang Ketterle (right image).

But, there is more if we fiddle with the overall phase of a whole superfluid or superconductor with its many-particle macroscopic wavefunction. Imprinting a phase on such a system can create moving dynamical structures called vortices and solitons. Both signify the existence of a fully coherent, macroscopic many-particle quantum system.

A phase can be "imprinted" by rotating such a macroscopic quantum system. This creates vortices as we see in the density profile calculated in Figure 27. Physicists have observed such vortices in liquid superfluids, gaseous BECs, and superconductors. All have the same origin.

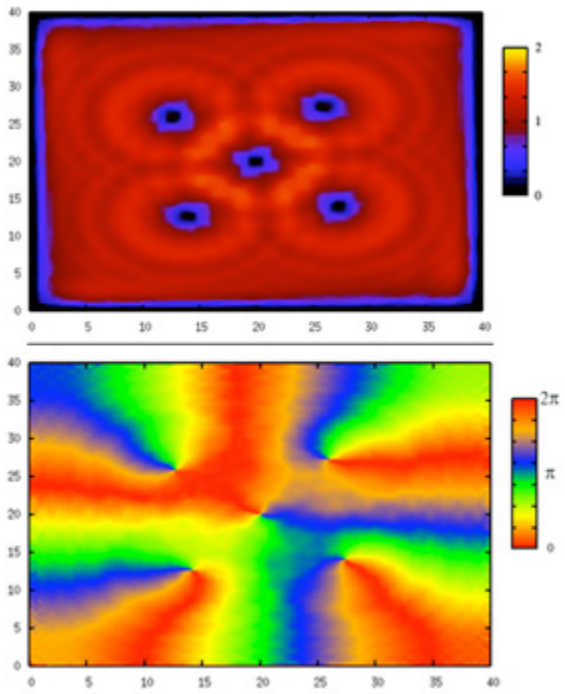


Figure 27: Quantum vortices in a BEC (top) and the corresponding phase of the quantum wavefunction (bottom).
Source: © William P. Reinhardt.

We can obtain a physical picture of what's happening if we examine the phase of the underlying macroscopic wavefunction. That's difficult to do experimentally, but easy if we use a computer calculation to generate the coherent many-body macroscopic wavefunction, as Figure 27 illustrates. Here, unlike in an experimental situation, we fully know the phase of the wavefunction because we have programmed it ourselves. We can then calculate its absolute value squared if we like, which gives the resulting probability density that an experiment would observe. If we calculate the phase in a model of a circulating BEC, superfluid, or superconducting current, and compare it to the observable probability density, we find the phase changes periodically as you trace a circle around what appear to be holes in the probability density, as you can see in Figure 27. Each hole is a tiny quantum whirlpool, or vortex; the lines of phase radiating outward indicate that the quantum fluid is circulating around these holes.

Solitons: waves that do not decay

Clearly, then, altering the phase of a macroscopic wavefunction may have dramatic effects. Another such effect is the creation of the special nonlinear waves called "solitons," which physicists have observed in BECs following a phase imprinting. Solitons are a very special type of wave. They can, and do, pass right though each other without dissipating. And if they are in a superfluid (a gaseous atomic BEC in this case), they will carry on this oscillatory motion forever. These are called "density notch solitons," and are created by phase imprinting on part of a condensate. The process both creates and drives the motion of the solitons, just as circular phase imprints both create and maintain superfluid vortices.

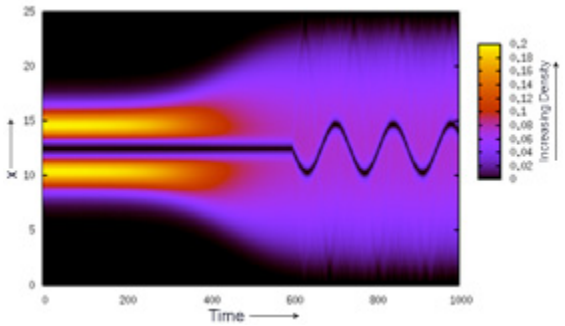


Figure 28: Density notch soliton.
Source: © William P. Reinhardt.

The vortices and solitons are both highly unusual persistent defects in the quantum wavefunction. Were these systems not macroscopic superfluids, such defects would just disappear and dissipate of their own accord. Try to make a hole in a tub of water and watch it fill in. (On the other hand, lumps of water can propagate without dispersion: Think of tsunamis, where, in exactly the right circumstances, lumps of water can move across a whole ocean.) The same would happen to a quantum bump or dip in the density corresponding to one or a few quantum particles in free space: The wavefunction would spontaneously disperse. It is the collective and coherent nature of the many-particle BEC that allows these nonlinear wave structures to persist. Of course, if you can create a whirlpool as the bath tub drains, it will persist; but such a classical whirlpool has nothing to do with the phase of the liquid water, phase being a purely quantum concept as we apply it here.

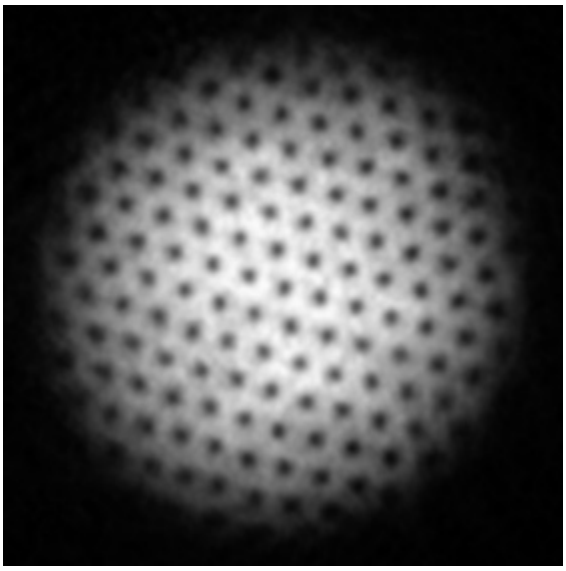


Figure 29: The formation of vortices in this BEC shows that it is a superfluid.
Source: © Martin Zwierlein.

The vortices and solitons we describe are all consistent with the idea that all the particles are in a single quantum state. These highly dilute gaseous condensates are essentially 100 percent pure, in the sense that all the particles occupy a single quantum state in the trap—although, to be sure, the quantum state wavefunction distorts as the trap is filled. The simple tools developed here have their limitations. They apply to dilute atomic gases, but cannot deal with a more traditional liquid such as ^{4}He in its superfluid state. Owing to the very strong interactions between the helium atoms at the much higher density of the liquid, the condensate is only about 10 percent pure condensate. This does not affect its remarkable properties as a superfluid, but certainly makes its theoretical description more difficult. In fact, it makes the quantitative description of the coherent quantum wavefunction a tremendously exciting exercise in computational physics.

Section 8: Making BECs from Fermi Gases

We have discussed the fact that a neutral atom of ^{7}Li is a boson. That's because the neutral atom's three protons, three electrons, and four neutrons add up to an even number of spin-1/2 fermions, thus giving the atom an integral total spin. ^{7}Li is the predominant isotope of lithium that occurs naturally on Earth. However, the isotope ^{6}Li accounts for about 7 percent of naturally occurring lithium. This isotope has an odd number of spin-1/2 fermion constituents: three protons, three neutrons, and three electrons, and is thus a fermionic atom. What happens to this fermionic isotope of Li if you trap it and cool it down? Rather than Bose condensing like ^{7}Li, it fills in the energy levels in the trap just like electrons fill in the energy levels of an atom. A fermionic gas in which the lowest energy levels are occupied with one particle in each level is called "degenerate." One can thus envisage making an ultra-cold degenerate Fermi gas from ^{6}Li, and researchers in the laboratory have actually done so.

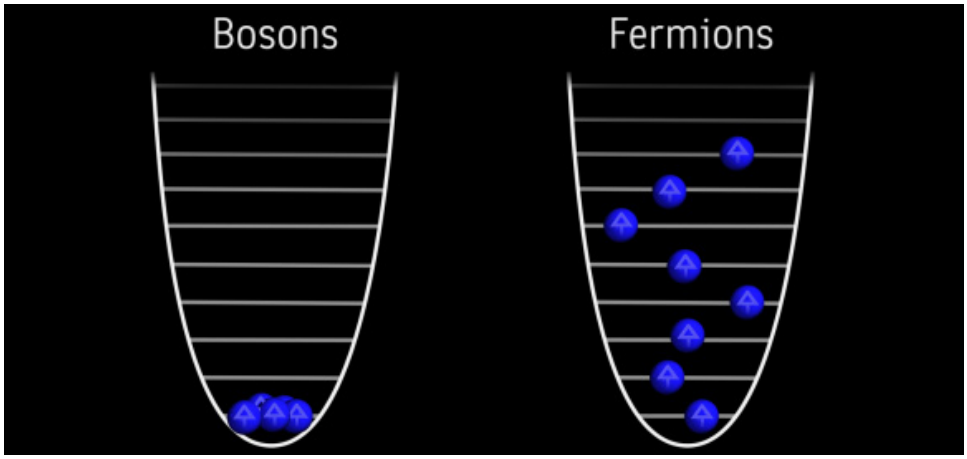


Figure 30: Atoms in a trap at 0 K: bosons form a BEC (left) and fermions form a degenerate Fermi gas (right).
Source:

Cooling Fermi gases in that way is far more difficult than making an atomic gaseous BEC. Once physicists learned to make them, gaseous BECs immediately became commonplace. Part of the advantage of making a BEC is the bosonic amplification—the effect of sucking all the bosons into the same quantum state that we encountered in our earlier discussion of lasers. The word "laser" is actually an acronym for "light amplification through stimulated emission of radiation." Once a single quantum state begins to fill up with bosons, others, miraculously, want to join them.

This is not at all the case with fermions, as their exclusion principle dictates entirely the opposite behavior. In fact, once a cold and partly degenerate Fermi gas begins to form, many of the energy levels are occupied but a (hopefully) small fraction are not. It then becomes very difficult to further cool the Fermi system. As most levels are already full, only the few empty ones are available to accept another atomic fermion. If these are few and far between, it takes a lot of time and luck to have a Fermi particle lose—through evaporative cooling, say—just the right amount of energy to land in one of the unoccupied energy levels. The other levels are blocked by the exclusion principle. Unsurprisingly called "Pauli blocking," this is a real impediment to making gaseous macroscopic, fully degenerate cold Fermi gases from fermionic atoms. Experimentalists often co-condense ^{6}Li with ^{7}Li and allow the ^{7}Li BEC to act as a refrigerator to cool the recalcitrant ^{6}Li atoms into behaving.

Pairing fermions to make bosons

In the style of earlier parts of this unit, we can represent fermionic ^{6}Li as ^{6}Li$\uparrow$, with the $\uparrow$ now representing the atom's outer unpaired electron. Here we can, at last, illustrate the pairing of fermions to give bosons—the analog of the Cooper pairing of electrons in a superconducting metal that Unit 8 will develop fully. Combine two fermions and you get a boson. This is simple numerics: Doubling an odd number produces an even number. So, for our example, the molecule $^{6}Li_2$ must be a boson whether it exists in the higher energy level ^{6}Li$\uparrow\uparrow^{6}$Li or the lower energy ^{6}Li$\downarrow\uparrow^{6}$Li. Now you can see why, in a chapter about macroscopic quantum states, we started off with a discussion of simple molecules. We should note, conversely, that combining two bosons just gives another boson, as the number of spin-1/2 particles is still even.

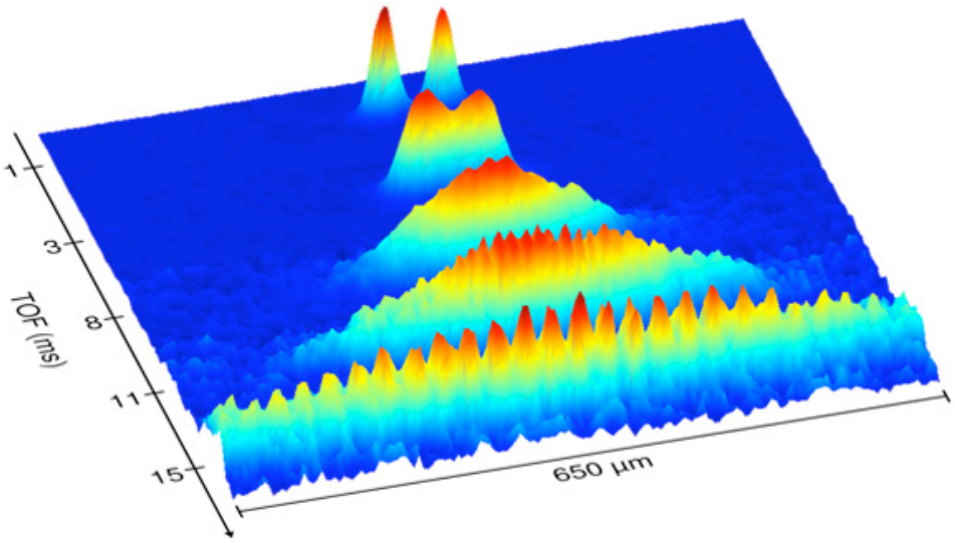


Figure 31: Interference pattern created by the overlap of two clouds of molecular BECs, each composed of $^{6}Li_{2}$ diatomic molecules.
Source: © C. Kohstall & R. Grimm.

Figure 31 indeed shows that fermionic atoms can pair and become molecular bosons, which can then condense into a molecular BEC. This is evident from the striking similarity of the interference patterns shown in Figures 26 and 31. Two overlapping degenerate Fermi gases would not create such macroscopic interference patterns because those wavefunctions have no intrinsic phase relationships, in contrast to the BEC wavefunctions. Molecular BECs, especially if the molecules are polar and can communicate with one anther through long-range forces, are of special interest as being quantum information storage devices, as individual molecules can potentially be addressed via their many internal degrees of freedom.

Section 9: Conclusions and a Look Ahead

Macroscopic quantum fluids, or superfluids, can be formed from cold bosons. Many types of composite entities fulfill the basic bosonic requirement of integral spin. So, the gaseous quantum superfluid can consist of bosonic atoms, and quite newly bosonic molecules, all of which contain an even overall number of spin-1/2 fermions, be they electrons, protons, or neutrons. All are uncharged superfluids. In the dilute gaseous superfluid phase, they display fully quantum properties, such as interference on the length scale of around a millimeter, which are visible to the naked eye. So, they are certainly macroscopic. Dense liquid superfluids such as ^{4}He, of course, contain many more atoms than gaseous BECs in laboratory traps, and have a myriad of unusual properties. But, they do not directly display their fundamental quantum nature quite so directly, even though it underlies their superfluid behavior.

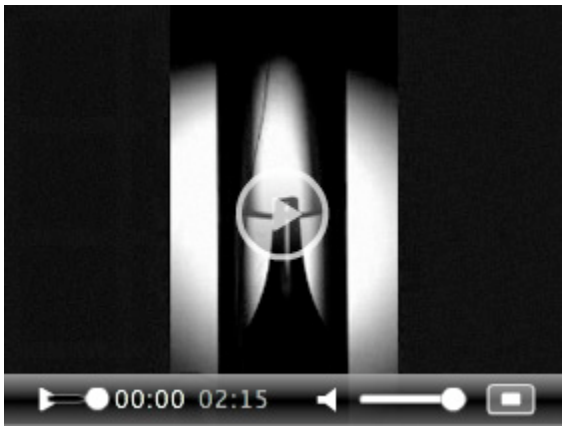


Figure 32: A fountain of superfluid ^{4}He
Source: © Peter Taborek, Low Temperature Materials Laboratory, University of California, Irvine.

Two fermions can pair to make a boson: This allows fermionic ^{3}He to become a superfluid, albeit at a much lower temperature than its bosonic neighbor ^{4}He. The pair interaction between these ^{3}He atoms is far too weak to allow formation of molecules in the liquid, so the pair formed is rather ephemeral, and is best described in analogy to what happens to electrons in a superconductor rather than atoms in ^{4}He.

Famously, and unexpectedly, even two electrons can pair in a metallic free-electron "sea" to make a composite boson inside a superconductor. This seems odd at first, as these electrons have like charges and thus repel each other. Thus, their pairing is not at all like making bosonic $^6Li_2$ from fermionic ^{6}Li, as those atoms actually attract and form molecules in physical space. The pairing of electrons to make

bosonic pairs of electrons, called "Cooper pairs," is indeed more complex. It does not even take place in the three-dimensional coordinate space in which we live. Rather, the pairing occurs in a more abstract "momentum" space. We leave that description to Unit 8, simply noting that it is yet another example of the pairing concepts that we have introduced here. And, because a bosonic pair of electrons carries two units of negative electrical charge, the Bose condensate of such paired electrons is not only a superfluid, but also a superconductor.

Section 10: Further Reading

- Eric A. Cornell and Carl E. Weimann, "The Bose-Einstein Condensate" *Scientific American*, March 1998.
- Eric A. Cornell, Carl E. Weimann, Wolfgang Ketterle: Nobel Prize Lectures, available at http://nobelprize.org/nobel_prizes/physics/laureates/2001/index.html.
- Wolfgang Ketterle, "How are temperatures close to absolute zero achieved and measured?" *Scientific American.com, Ask the Experts*, January 19, 2004, found here: http://www.scientificamerican.com/article.cfm?id=how-are-temperatures-clos.
- Carlos A.R. Sa de Melo, "When fermions become bosons: Pairing in ultracold gases" *Physics Today*, Oct. 2008, p. 45–51.
- Eric R. Scerri, "The Periodic Table, Its Story and Significance" *Oxford University Press*, 2007.
- Charles Townes, "How the Laser Happened: Adventures of a Scientist" *Oxford University Press*, 2002.

Glossary

alkali metals: The alkali metals are the chemical elements in the first column of the periodic table. They all have one valence electron. Alkali metals are commonly used atoms in atomic physics experiments for several reasons. Their structure is relatively simple and provides energy states that are convenient for laser cooling. Many of their transition frequencies match convenient laser sources. Also, the single valence electron's magnetic moment allows the atoms to be easily trapped using magnetic fields, which is convenient for the evaporative cooling process necessary to reach ultracold temperatures.

atomic number: The atomic number of an atom, denoted by Z, is the number of protons in its nucleus. The atomic number of an atom determines its place in the periodic table, and thus which chemical element it is.

blackbody: A blackbody is an object that absorbs all incident electromagnetic radiation and re-radiates it after reaching thermal equilibrium. The spectrum of light emitted by a blackbody is smooth and continuous, and depends on the blackbody's temperature. The peak of the spectrum is higher and at a shorter wavelength as the temperature increases.

Bose-Einstein condensate: A Bose-Einstein condensate, or BEC, is a special phase of matter in which the quantum mechanical wavefunctions of a collection of particles line up and overlap in a manner that allows the particles to act as a single quantum object. The electrons in a superconductor form a BEC; superfluid helium is an example of a liquid BEC. BECs can also be created from dilute gases of ultracold atoms and molecules.

boson: A boson is a particle with integer, rather than half-integer, spin. In the Standard Model, the force-carrying particles such as photons are bosons. Composite particles can also be bosons. Mesons such as pions are bosons, as are ^{4}He atoms. See: fermion, meson, spin.

complex: In the context of physics and math, the term complex refers to the presence of complex numbers and is not a synonym of complicated. Thus, a "complex wave" is a mathematical function that describes a wave that can take on complex number values.

complex number: A complex number is a composite of a real number and an imaginary number, and can be written in the form $a+bi$ where a and b are real numbers and i is the square root of -1.

counting number: The counting numbers are the integers greater than zero: 1, 2, 3 .

de Broglie wavelength: A particle's de Broglie wavelength, λ, is defined as Planck's constant divided by the particle's momentum, p: $\lambda = h/p$. The de Broglie wavelength is named after Louis de Broglie, the French physicist who first suggested that it might be useful to describe particles as waves. A relativistic electron has a de Broglie wavelength of around a nanometer, while a car driving down the highway has a de Broglie wavelength of around 10^{-38} meters. Quantum mechanical effects tend to be important at the scale of an object's de Broglie wavelength; thus we need to describe electrons quantum mechanically, but classical physics is adequate for cars and most other macroscopic objects.

electric dipole moment: The electric dipole moment of a system with two electric charges is defined as the product of the two charges divided by the distance between them. It is a vector quantity, with the positive direction defined as pointing from the (more) negative charge toward the (more) positive charge. The electric dipole moment of a more complicated system of charges is simply the sum of the moments of each pair of charges.

fermion: A fermion is a particle with half-integer spin. The quarks and leptons of the Standard Model are fermions with a spin of 1/2. Composite particles can also be fermions. Baryons, such as protons and neutrons, and atoms of the alkali metals are all fermions. See: alkali metal, baryon, boson, lepton, spin.

ground state: The ground state of a physical system is the lowest energy state it can occupy. For example, a hydrogen atom is in its ground state when its electron occupies the lowest available energy level.

harmonic trap: A harmonic trap is a trap in which the trapped objects (e.g., atoms) are pushed toward the center of the trap with a force proportional to their distance from the center of the trap. The motion of particles in a harmonic trap is analogous to the motion of a mass attached to a spring around the spring's equilibrium position. It is convenient to use harmonic traps in experiments because it is straightforward to calculate the motion of particles in analogy to the mass on a spring.

hyperfine structure: When the nucleus of an atom has a nonzero magnetic moment, some of the energy levels that electrons can occupy in the atom are very finely spaced. The arrangement of these finely spaced levels is called "hyperfine structure." The difference in energy between hyperfine levels typically

corresponds to a microwave photon frequency or light with a wavelength on the order of centimeters. The energy levels in the cesium atom used to define the second are hyperfine levels.

ion: An ion is an atom with nonzero electrical charge. A neutral atom becomes an ion when one or more electrons are removed, or if one or more extra electrons become bound to the atom's nucleus.

isotope: Different atoms of a chemical element in the periodic table all have the same number of protons, but may have a different number of neutrons in their nuclei. These different versions of the same element are called isotopes. The number of neutrons is not simply random, however—the nucleus will only be stable for combinations of protons and neutrons. Most chemical elements have several stable isotopes. For example, lithium (A=3) has two stable isotopes, one with three neutrons in the nucleus (^{6}Li) and one with four (^{7}Li). See: atomic number, mass number.

macroscopic: A macroscopic object, as opposed to a microscopic one, is large enough to be seen with the unaided eye. Often (but not always), classical physics is adequate to describe macroscopic objects, and a quantum mechanical description is unnecessary.

magnetic moment: The magnetic moment (or magnetic dipole moment) of an object is a measure of the object's tendency to align with a magnetic field. It is a vector quantity, with the positive direction defined by the way the object responds to a magnetic field: The object will tend to align itself so that its magnetic moment vector is parallel to the magnetic field lines. There are two sources for a magnetic moment: the motion of electric charge and spin angular momentum. For example, a loop of wire with a current running through it will have a magnetic moment proportional to the current and area of the loop, pointing in the direction of your right thumb if your fingers are curling in the direction of the current. Alternatively, an electron, which is a spin-1/2 fermion, has an intrinsic magnetic moment proportional to its spin.

mass number: The mass number (or atomic mass number) of an atom, denoted by A, is the total number of nucleons (protons+neutrons) in its nucleus. Sometimes, the mass number of an atom is written as a superscript to the left of its chemical symbol (e.g., ^{6}Li) to show which isotope is being discussed. See: atomic number, isotope.

Pauli exclusion principle: The Pauli exclusion principle states that no two identical fermions can occupy the same quantum state. It plays an important role in determining the structure of atoms and atomic nuclei, as well as how electrons behave in metals and semiconductors.

phase: In physics, the term phase has two distinct meanings. The first is a property of waves. If we think of a wave as having peaks and valleys with a zero-crossing between them, the phase of the wave is defined as the distance between the first zero-crossing and the point in space defined as the origin. Two waves with the same frequency are "in phase" if they have the same phase and therefore line up everywhere. Waves with the same frequency but different phases are "out of phase." The term phase also refers to states of matter. For example, water can exist in liquid, solid, and gas phases. In each phase, the water molecules interact differently, and the aggregate of many molecules has distinct physical properties. Condensed matter systems can have interesting and exotic phases, such as superfluid, superconducting, and quantum critical phases. Quantum fields such as the Higgs field can also exist in different phases.

phase coherence: If we think of a wave as having peaks and valleys with a zero-crossing between them, the *phase* of the wave is defined as the distance between the first zero-crossing and the point in space defined as the origin. Two waves are *phase coherent* (or simply *coherent*) if the distance between their respective peaks, valleys, and zero-crossings is the same everywhere.

photon: Photons can be thought of as particle-like carriers of electromagnetic energy, or as particles of light. In the Standard Model, the photon is the force-carrier of the electromagnetic force. Photons are massless bosons with integer spin, and travel through free space at the speed of light. Like material particles, photons possess energy and momentum.

Planck's constant: Planck's constant, denoted by the symbol h, has the value 6.626×10^{-34} m^2 kg/s. It sets the characteristic scale of quantum mechanics. For example, energy is quantized in units of h multiplied by a particle's characteristic frequency, and spin is quantized in units of $h/2\pi$. The quantity $h/2\pi$ appears so frequently in quantum mechanics that it has its own symbol: $\hbar$.

plum pudding model: The Plum Pudding Model is a model of atomic structure proposed by J.J. Thomson in the late 19th century. Thomson had discovered that atoms are composite objects, made of pieces with positive and negative charge, and that the negatively charged electrons within the atom were very small compared to the entire atom. He therefore proposed that atoms have structure similar to a plum pudding, with tiny, negatively charged electrons embedded in a positively charged substrate. This was later shown to be incorrect.

polar: A polar molecule has a nonzero electric dipole moment, so it has a side that is positively charged and a side that is negatively charged.

probability density: The exact location of a quantum mechanical particle is impossible to know because of the Heisenberg uncertainty principle. Rather than specifying the location of a particle such as an electron, quantum mechanics specifies a wavefunction. The probability density, which is a mathematical function that specifies the probability of finding the particle at any location in space, is the square of the wavefunction (technically, its absolute value squared).

real number: Real numbers are most easily defined by contrast to what they are not: imaginary numbers. The set of real numbers includes counting numbers, integers, rational numbers that can be written as fractions, and irrational numbers such as π. They can be thought of as all the points on a number line stretching from negative infinity to infinity.

shell model: The shell model of atomic structure is based on the notion that electrons in an atom occupy "shells" that can fill up, so only a certain number of electrons will fit in a given shell. G. N. Lewis found this idea useful in explaining the chemical properties of different elements. Lewis's shell model is consistent with the Bohr model of the atom in which electrons are thought of as orbiting the nucleus. The "shells" are three-dimensional counterparts of two-dimensional circular orbits with different radii. Although we now know that the Bohr model of the atom is not correct, the concept of shells is still sometimes used to describe the arrangement of electrons in atoms according to the Pauli exclusion principle.

soliton: A soliton is a stable, isolated wave that travels at a constant speed. As a soliton travels, its shape does not change and it does not dissipate. If it collides with another wave, it emerges from the collision unscathed. In a sense, it is a wave that behaves like a particle. Solitons have been predicted and observed in nearly every medium in which waves propagate, including fluids such as water, transparent solids such as optical fibers, and magnets.

standing wave: A standing wave is a wave that does not travel or propagate: The troughs and crests of the wave are always in the same place. A familiar example of a standing wave is the motion of a plucked guitar string.

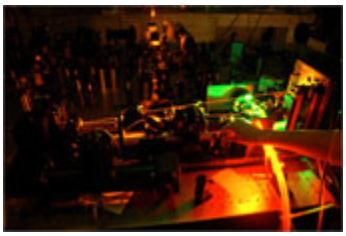
valence electron: A valence electron is an electron in the outermost shell of an atom in the Lewis model, or in the orbital with the highest value of the principal quantum number, *n*, in the quantum mechanical description of an atom. The valence electrons determine most of the chemical and physical properties of

the atom. It is the valence electrons that participate in ionic and covalent chemical bonds, and that make the primary contributions to an atom's magnetic moment.

vortex: A vortex is a region in which a fluid or gas flows in a spiral toward the vortex center. The speed of fluid flow is fastest at the center of the vortex, and decreases with distance from the vortex center. Tornados and whirlpools are examples of vortices. Quantized vortices will appear in a superfluid when it is rotated fast enough, and quantized vortices will form in the electron gas inside a type-II superconductor when it is placed in a strong enough magnetic field.

Physics for the 21st Century

Unit 7: Manipulating Light



Unit Overview

This unit continues to develop the theme of the practical and foundational effects of quantum mechanics. It focuses on the experimental achievements in reducing the speed of light by factors of tens of millions and covers some of the implications of that research. The first section emphasizes the critical role that the speed of light in a vacuum plays in our understanding of our universe. It also outlines the "natural" way of slowing light by small amounts by passing it through materials of different refractive indices. Section 3 then details the type of experimental setup used to slow down light "artificially" in the laboratory and analyzes the fundamental quantum processes that permit physicists to reduce light's speed to that of a cyclist—and even to stop light altogether and hold it in storage. Next, Section 7 covers methods of converting light into matter and back again. And finally, Section 8 points out various applications, real and potential, of the increasing ability to manipulate light.

Content for This Unit

Sections:

Section 1: Introduction

Figure 1: Saint Hans bonfire—Midsummer celebration in Skagen, Denmark.
Source: Painting by P.S. Krøyer: Sct. Hans-blus på Skagen, 1906; owned by Skagen Museum.

Light has fascinated humans for thousands of years. In ancient times, summer solstice was a celebrated event. Even to this day, the midsummer solstice is considered one of the most important events of the year, particularly in the northern-most countries where the contrast between the amount of light in summer and winter is huge. Visible daylight is intense due to the proximity of the Earth to the Sun. The Sun is essentially a huge nuclear reactor, heated by the energy released by nuclear fusion. When something is hot, it radiates. The surface of the Sun is at a temperature of roughly 6000 K (roughly 10,000°F). Our eyes have adapted to be highly sensitive to the visible wavelengths emitted by the Sun that can penetrate the atmosphere and reach us here on Earth. The energy carried in sunlight keeps the Earth at a comfortable temperature and provides the energy to power essentially everything we do through photosynthesis in plants, algae, and bacteria—both the processing happening right now as well as what took place millions of years ago to produce the energy stored in fossil fuels: oil, gas, and coal.

The interaction of light with different substances has long been a subject of great interest, both for basic science and for technological applications. The manipulation of light with engineered materials forms the basis for much of the technology around us, ranging from eyeglasses to fiber-optic cables. Many of these applications rely on the fact that light travels at different speeds in different materials.

The speed of light therefore plays a special role that spans many aspects of physics and engineering. Understanding and exploiting the interaction between light and matter, which govern the speed of light in different media, are topics at the frontier of modern physics and applied science. The tools and

techniques employed to explore the subtle and often surprising interplay between light and matter include lasers, low-temperature techniques (cryogenics), low-noise electronics, optics, and nanofabrication.

Figure 2: Flight controllers Charles Duke (Capcom), Jim Lovell (backup CDR), and Fred Haise (backup LMP) during lunar module descent.
Source: © NASA.

Light travels fast, but not infinitely fast. It takes about two and a half seconds for a pulse of light to make the roundtrip to the Moon and back. When NASA's ground control station was engaged in discussions with the Apollo astronauts at the Moon, the radio waves carrying the exchange traveled at light speed. The 2.5 second lag due to the radio waves moving at the finite speed of light produced noticeable delays in their conversations. Astronomers measure cosmic distances in terms of the time it takes light to traverse the cosmos. A light-year is a measure of distance, not time: it's the length that light travels in a year. The closest star to the Sun is about four light-years away. For more down-to-Earth separations, a roundtrip at light speed from London to Los Angeles would take about 0.06 seconds.

So, if the *upper* limit to the speed of light is c, an astonishing 300 million meters per second (186,000 miles per second), how much can we slow it down? Light traveling in transparent materials moves more slowly than it does in a vacuum, but not by much. In air, light travels about 99.97% as fast as it does in a vacuum. The index of refraction, n, of a material is the ratio between light's speed in empty space relative to its speed in the material. In a typical glass (n = 1.5), the light is traveling about 67% as fast as it does in a vacuum. But can we do better? Can we slow light down to human speeds? Yes!

This unit explores the fascinating quest for finding ways to slow down and even stop light in the laboratory. Perhaps this extreme limit of manipulating light in materials could give rise to novel and profound technological applications.

Section 2: Measuring and Manipulating the Speed of Light

Interest in studying light, and in measuring its speed in different circumstances, has a long history. Back in the 1600s, Isaac Newton thought of light as consisting of material particles (corpuscles). But Thomas Young's experiments showed that light can interfere like merging ripples on a water surface. This vindicated Christiaan Huygens' earlier theory that light is a wave. Quantum mechanics has shown us that light has the properties of both a particle and a wave, as described in Unit 5.

Scientists originally believed that light travels infinitely fast. But in 1676, Ole Rømer discovered that the time elapsed between eclipses of Jupiter's moons was not constant. Instead, the period varied with the distance between Jupiter and the Earth. Rømer could explain this by assuming that the speed of light was finite. Based on observations by Michael Faraday, Hans Christian Ørsted, André-Marie Ampère, and many others in the 19th century, James Clerk Maxwell explained light as a wave of oscillating electric and magnetic fields that travels at the finite speed of 186,000 miles per second, in agreement with Rømer's observations.

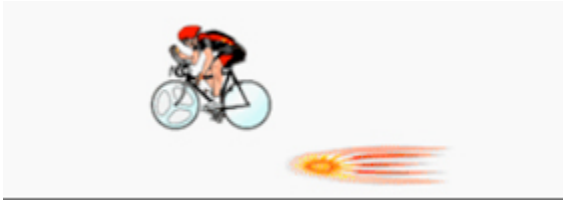


Figure 3: The biker wins the race against the light pulse!
Source: © John Newman and Lene V. Hau.

Experiments on the speed of light continued. Armand Hippolyte Fizeau was the first to determine the light speed in an Earth-based experiment by sending light through periodically spaced openings in a fast rotating wheel, then reflecting the light in a mirror placed almost 10 km away. When the wheel rotation was just right, the reflected light would again pass through a hole in the wheel. This allowed a measurement of the light speed. Fizeau's further experiments on the speed of light in flowing water, and other experiments with light by Augustin-Jean Fresnel, Albert Michelson, and Edward Morley, led Albert Einstein on the path to the theory of special relativity. The speed of light plays a special role in that theory, which states that particles and information can never move faster than the speed of light in a vacuum. In other words: the speed of light in a vacuum sets an absolute upper speed limit for everything.

Bending light with glass: classical optics

The fact that light travels more slowly in glass than in air is the basis for all lenses, ranging from contact lenses that correct vision to the telephoto lenses used to photograph sporting events. The phenomenon of refraction, where the direction that a ray of light travels is changed at an interface between different materials, depends on the ratio of light's speed in the two media. This is the basis for Snell's Law, a relationship between the incoming and outgoing ray angles at the interface and the ratio of the indices of refraction of the two substances.

By adjusting the curvature of the interface between two materials, we can produce converging and diverging lenses that manipulate the trajectory of light rays through an optical system. The formation of a focused image, on a detector or on a screen, is a common application. In any cell phone, the camera has a small lens that exploits variation in the speed of light to make an image, for example.

If different colors of light move at different speeds in a complex optical system, the different colors don't traverse the same path. This gives rise to "chromatic aberration," where the different colors don't all come to the same focus. The difference in speed with a wavelength of light also allows a prism to take an incoming ray of white light that contains many colors and produce beams of different colors that emerge at different angles.

Finding breaks in optical fibers: roundtrip timing

Figure 4: An optical time domain reflectometer in use.
Source: © Wikimedia Commons, Creative Commons Attribution-Share Alike 3.0. 25 September 2009.

A commercial application of the speed of light is the Optical Time Domain Reflectometer (OTDR). This device sends a short pulse of light into an optical fiber, and measures the elapsed time and intensity of any light that comes back from the fiber. A break in the fiber produces a lot of reflected light, and the elapsed time can help determine the location of the problem. Even bad joints or poor splices can produce measurable reflections, and can also be located with OTDR. Modern OTDR instruments can locate problems in optical fiber networks over distances of up to 100 km.

The examples given above involve instances where the light speeds in materials and in a vacuum aren't tremendously different, hardly changing by a factor of two. Bringing the speed of light down to a few meters per second requires a different approach.

Section 3: A Sound Design for Slowing Light

Our objective is to slow light from its usual 186,000 miles per second to the speed of a bicycle. Furthermore, we can show that light can be completely stopped, and even extinguished in one part of space and regenerated in a completely different location.

To manipulate light to this extreme, we first need to cool atoms down to very low temperatures, to a few billionths of a degree above absolute zero. As we cool atoms to such low temperatures, their quantum nature becomes apparent: We form Bose-Einstein condensates and can get millions of atoms to move in lock-step—all in the same quantum state as described in Units 5 and 6.

Atom refrigerator

Achieving these low temperatures requires more than the use of a standard household refrigerator: We need to build a special atom refrigerator.

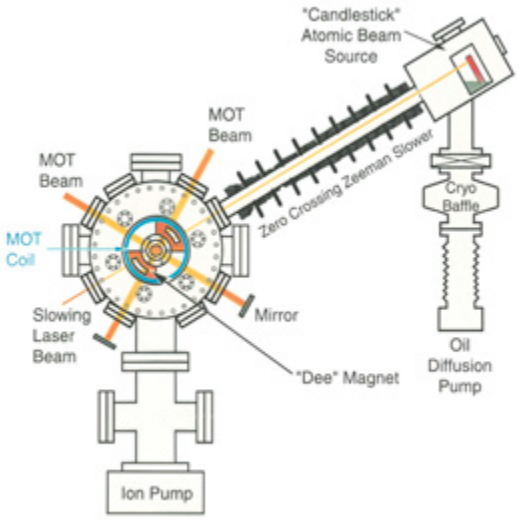


Figure 5: Sketch of experimental setup used to create ultracold atoms.
Source: © Brian Busch and Lene Hau.

In the setup we use in our laboratory, most parts have been specially designed and constructed. This includes the first stage of the cooling apparatus: the atom source. We have named it the "candlestick atomic beam source" because it works exactly like a candle. We melt a clump of sodium metal that we have picked from a jar where it is kept in mineral oil. Sodium from the melted puddle is then delivered by

wicking (as in a candle) to a localized hot spot that is heated to 650°F and where sodium is vaporized and emitted. To create a well-directed beam of sodium atoms, we send them through a collimation hole. The atoms that are not collimated are wicked back into the reservoir and can be recycled.

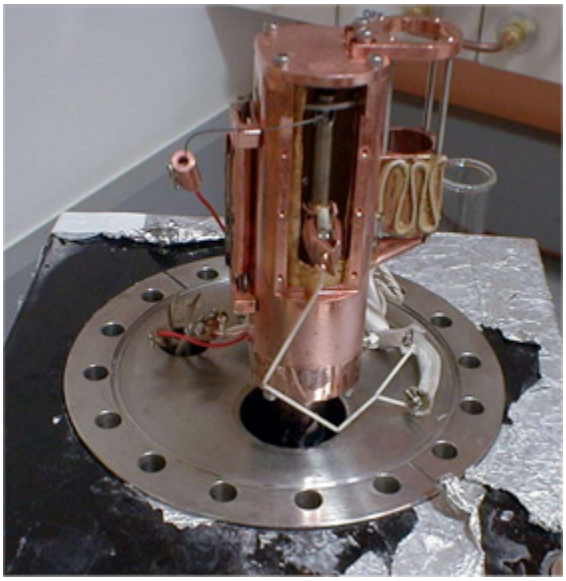


Figure 6: The atom source in the experimental setup is specially designed and called the "candlestick atomic beam source."
Source: © Hau Laboratory.

This atom source produces a high-flux collimated beam of atoms—there is just one problem: The atoms come out of the source with a velocity of roughly 1500 miles per hour, which is much too fast for us to catch them. Therefore, we hit the atoms head-on with a yellow laser beam and use radiation pressure from that laser beam to slow the atoms. By tuning the frequency of the laser to a characteristic frequency of the atoms, which corresponds to the energy difference between energy levels in the atom (we say the laser is *resonant* with the atoms), we can get strong interactions between the laser light and the atoms. We use a laser beam with a power of roughly 50 mW (a typical laser pointer has a power of 1 mW), which we focus to a small spot at the source. This generates a deceleration of the atoms of 100,000 g's, in other words 100,000 times more than the acceleration in the Earth's gravitational field. This is enormous; and in a matter of just a millisecond (one-thousandth of a second) —and over a length of one meter—we can slow the atoms to 100 miles per hour.

Optical molasses

The Concept of the Photon

According to quantum mechanics, the energy of a light field comes in packets—quanta—of finite energy. The energy of a photon increases in proportion to the frequency of the light field, and a light field can only have energies that are integer numbers of these energy quanta. A laser field is a very particular light field where all the quanta are similar or "indistinguishable" and are in lock-step.

At this point, we can load the atoms efficiently into an optical molasses created in the middle of a vacuum chamber by three pairs of counter-propagating laser beams. These laser beams are tuned to a frequency slightly lower than the resonance absorption frequency, and we make use of the Doppler effect to cool the atoms. The Doppler effect is familiar from everyday life: A passing ambulance will approach you with the siren sounding at a high pitch; and as the ambulance moves away, the siren is heard at a much lower pitch. Moving atoms bombarded from all sides with laser beams will likewise see the lasers' frequency shifted: If an atom moves toward the source of a laser beam, the frequency of the laser light will be shifted higher and into resonance with the atom; whereas for atoms moving in the same direction as the laser beam, the atoms will see a frequency that is shifted further from resonance. Therefore, atoms will absorb light—or photons—from the counter-propagating beams more than from the co-propagating beams; and since an atom gets a little momentum kick in the direction of the laser beam each time it absorbs a photon, the atoms will slow down. In this way, the laser beams create a very viscous medium—hence the name "optical molasses" in which the atoms will slow and cool to a temperature just a few millionths of a degree above absolute zero. We can never reach absolute zero (at -273°C or -460°F), but we can get infinitely close.

Figure 7: Running the experimental setup requires great focus.
Source:

It may be surprising that lasers can cool atoms since they are also being used for welding, in nuclear fusion research, and for other high-temperature applications. However, a laser beam consists of light that is incredibly ordered. It has a very particular wavelength or frequency and propagates in a very particular direction. In laser light, all the photons are in lock-step. By using laser beams, we can transfer heat and disorder from the atoms to the radiation field. In the process, the atoms absorb laser photons and spontaneously re-emit light in random directions: The atoms fluoresce. Figure 7 shows a view of the lab during the laser cooling process. Figure 8 shows the optical table in daylight: As seen, many manhours have gone into building this setup.

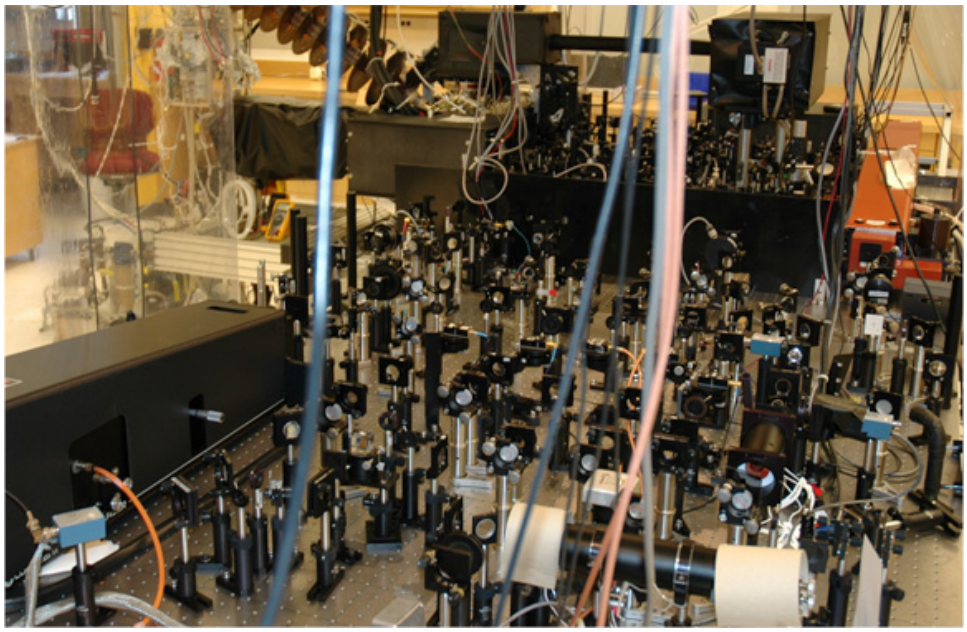


Figure 8: A view of the optics table.
Source: © Hau Laboratory.

Section 4: Making an Optical Molasses

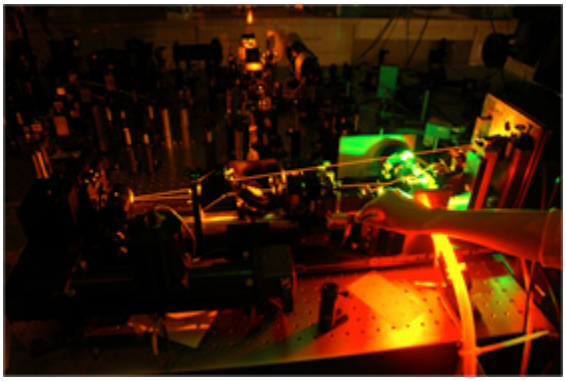


Figure 9: Adjustment of one of the dye lasers.
Source: © Hau Laboratory.

We have two laser systems on the table, each generating an intense yellow laser beam. The lasers are "dye lasers," pumped by green laser beams. A dye laser has a solution of dye molecules circulating at high pressure. This dye was originally developed as a clothing dye; but if illuminated by green light, the molecules emit light with wavelengths covering a good part of the visible spectrum: from red to greenish yellow.

In our apparatus, a narrow ribbon of dye solution is formed and squirted out at high velocity. We then hit the dye with a green laser beam inside a laser cavity where mirrors send light emitted by the dye molecules right back on the dye to stimulate more molecules to emit the same kind of light, at the same wavelength and in the same propagation direction (see Figure 9).

For the right length cavity, we can build up a large intensity of light at a particular wavelength. The cavity length should be an integer number of the desired wavelength. By intentionally leaking out a small fraction of the light, we generate the laser beam that we use in the experiments. How do we know that the wavelength is right for our experiments—that it is tuned on resonance with the atoms? Well, we just ask the atoms. We pick off a little bit of the emitted laser beam, and send it into a glass cell containing a small amount of sodium gas. By placing a light detector behind the cell, we can detect when the atoms start to absorb. If the laser starts to "walk off" (in other words if the wavelength changes), then we get less absorption. We then send an electronic error signal to the laser that causes the position of a mirror to adjust and thereby match the cavity length to the desired wavelength. Here—as in many other parts of the setup—we are using the idea of feedback: a very powerful concept.

So, perhaps you've started to get the sense that building an experiment like the one described here requires the development of many different skills: We deal with lasers, optics, plumbing (for chilled water), electronics, machining, computer programming, and vacuum technology. That is what makes the whole thing fun. Some believe that being a scientist is very lonely, but this is far from the case. To make this whole thing work requires great teamwork. And once you have built a setup and really understand everything inside out—no black boxes—that's when you can start to probe nature and be creative. When it is the most exciting, you set out to probe one thing, and then Nature responds in unanticipated ways. You, in turn, respond by tweaking the experiment to probe in new regimes where it wasn't initially designed to operate. And you might discover something really new.

Now, back to optical molasses: In a matter of a few seconds, we can collect 10 billion atoms and cool them to a temperature of 50 microkelvin (50 millionths of a degree above absolute zero). At this point in the cooling process, we turn off the lasers, turn on an electromagnet, and trap the atoms magnetically. Atoms act like small magnets: They have a magnetic dipole moment, and can be trapped in a tailored magnetic field. Once the laser-cooled atoms are magnetically trapped, we selectively kick out the most energetic atoms in the sample. This is called evaporative cooling, and we end up with an atom cloud that is cooled to a few nanoKelvin (a few billionths of a degree above absolute zero).

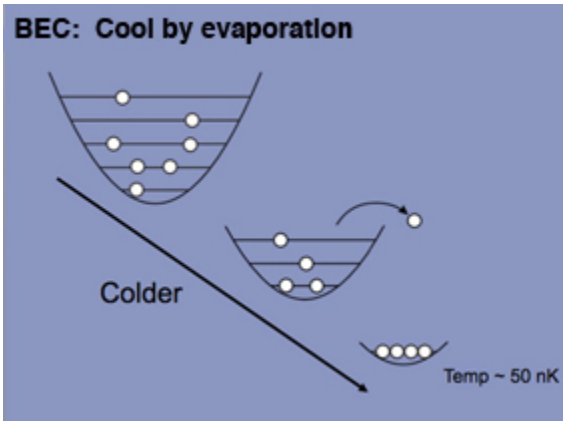


Figure 10: Evaporative cooling into the ground state.
Source: © Sean Garner and Lene V. Hau.

According to the rules of quantum mechanics, atoms trapped by the magnet can have only very particular energies—the energy is always the sum of kinetic energy (from an atom's motion) and of potential energy (from magnetic trapping). As the atoms are cooled, they start piling into the lowest possible energy state, the ground state (see Figure 10). As a matter of fact, because sodium atoms are bosons (there are just two types of particles in nature: bosons and fermions), once some atoms jump into the ground state,

the others want to follow; they are stimulated into the same state. This is an example of bosons living according to the maxim, the more the merrier. In this situation, pretty much all the atoms end up in exactly the same quantum state—we say they are described by the same quantum wavefunction, that the atoms are phase locked, and move in lock-step. In other words, we have created a Bose-Einstein condensate. Our condensates typically have 5–10 million atoms, a size of 0.004", and are held by the magnet in the middle of a vacuum chamber. (We pump the chamber out with pumps and create a vacuum because background atoms (at room temperature) might collide with the condensate and lead to heating and loss of atoms from the condensate).

It is interesting to notice that we can have these very cold atoms trapped in the middle of the vacuum chamber while the stainless-steel vacuum chamber itself is kept at room temperature. We have many vacuum-sealed windows on the chamber. During the laser cooling process, we can see these extremely cold atoms by eye. As described above, during the laser cooling process, the atoms absorb and reemit photons; the cold atom cloud looks like a little bright sun, about 1 cm in diameter, and freely suspended in the vacuum chamber.

Now, rather than just look at the atoms, we can send laser beams in through the windows, hit the atoms, manipulate them, and make them do exactly what we want...and this is precisely what we do when we create slow light.

Section 5: Slowing Light: Lasers Interacting with Cooled Atoms

So far, we have focused on the motion of atoms—how we damp their thermal motion by atom cooling, how this leads to phase locking of millions of atoms and to the formation of Bose-Einstein condensates.

For a moment, however, we will shift our attention to what happens internally within individual atoms. Sodium belongs to the family of alkali atoms, which have a single outermost, or valence, electron that orbits around both the nucleus and other more tightly bound electrons. The valence electron can have only discrete energies, which correspond to the atom's internal energy levels. Excited states of the atom correspond to the electron being promoted to larger orbits around the nucleus as compared to the lowest energy state, the (internal) ground state. These states determine how the atom interacts with light—and which frequencies it will absorb strongly. Under resonant conditions, when light has a frequency that matches the energy difference between two energy levels, very strong interactions between light and atoms can take place.

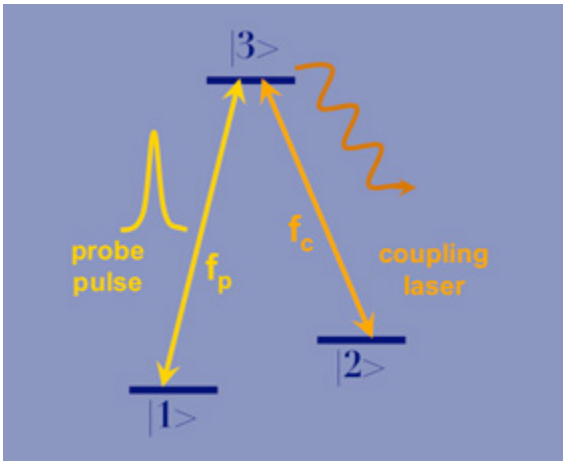


Figure 11: Internal, quantized energy levels of the atom.
Source: © Lene V. Hau.

When we are done with the cooling process, all the cooled atoms are found in the internal ground state, which we call 1 in Figure 11. An atom has other energy levels—for example state 2 corresponds to a slightly higher energy. With all the atoms in state 1, we illuminate the atom cloud with a yellow laser beam. We call this the "coupling" laser; and it has a frequency corresponding to the energy difference between states 2 and 3 (the latter is much higher in energy than either 1 or 2). If the atoms were actually in state 2, they would absorb coupling laser light, but since they are not, no absorption takes place.

Rather, with the coupling laser, we manipulate the optical properties of the cloud—its refractive index and opacity. We now send a laser pulse—the "probe" pulse into the system. The probe laser beam has a frequency corresponding roughly to the energy difference between states 1 and 3. It is this probe laser pulse that we slow down.

The presence of the coupling laser, and its interaction with the cooled atoms, generates a very strange refractive index for the probe laser pulse. Remember the notion of refractive index: Glass has a refractive index that is a little larger than that of free space (a vacuum). Therefore, light slows down a bit when it passes a window: by roughly 30%. Now we want light to slow down by factors of 10 to 100 million. You might think that we do this by creating a very large refractive index, but this is not at all the case. If it were, we would just create, with our atom cloud, the world's best mirror. The light pulse would reflect and no light would actually enter the cloud.

To slow the probe pulse dramatically, we manipulate the refractive index very differently. We make sure its average is very close to its value in free space—so no reflection takes place—and at the same time, we create a rapid variation of the index so it varies very rapidly with the probe laser frequency. A short pulse of light "sniffs out" this variation in the index because a pulse actually contains a small range of frequencies. Each of these frequency components sees a different refractive index and therefore travels at a different velocity. This velocity, that of a continuous beam of one pure frequency, is the phase velocity. The pulse of light is located where all the frequency components are precisely in sync (or, more technically, in phase). In an ordinary medium such as glass, all the components move at practically the same velocity, and the place where they are in sync—the location of the pulse—also travels at that speed. In the strange medium we are dealing with, the place where the components are in sync moves much slower than the phase velocity; and the light pulse slows dramatically. The velocity of the pulse is called the "group velocity," because the pulse consists of a group of beams of different frequencies.

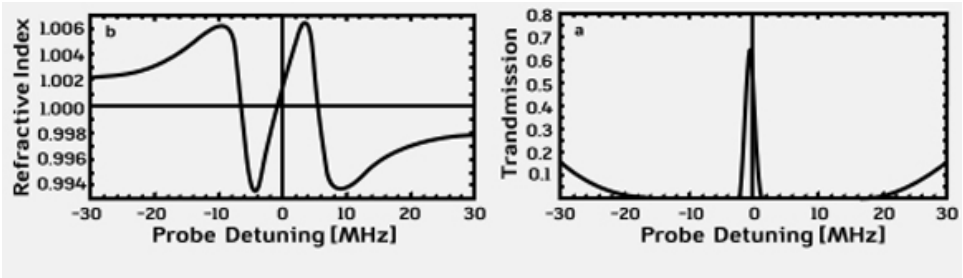


Figure 12: Refractive index variation with the frequency of a probe laser pulse.
Source: © Reprinted by permission from Macmillan Publishers Ltd: Nature 397, 594-598 (18 February 1999).

Another interesting thing happens. In the absence of the coupling laser beam, the "probe" laser pulse would be completely absorbed because the probe laser is tuned to the energy difference between states 1 and 3, and the atoms start out in state 1 as we discussed above. When the atoms absorb probe photons, they jump from state 1 to state 3; after a brief time, the excited atoms relax by reemitting light, but at random and in all directions. The cloud would glow bright yellow, but all information about the original light pulse would be obliterated. Since we instead first turn the coupling laser on and then send the probe laser pulse in, this absorption is prevented. The two laser beams shift the atoms into a quantum superposition of states 1 and 2, meaning that each atom is in both states at once. State 1 alone would absorb the probe light, and state 2 would absorb the coupling beam, each by moving atoms to state 3, which would then emit light at random. Together, however, the two processes cancel out, like evenly matched competitors in a tug of war—an effect called quantum interference.

The superposition state is called a dark state because the atoms in essence cannot see the laser beams (they remain "in the dark"). The atoms appear transparent to the probe beam because they cannot absorb it in the dark state, an effect called "electromagnetically induced transparency." Which superposition is dark—what ratio of states 1 and 2 is needed—varies according to the ratio of light in the coupling and probe beams at each location—more precisely, to the ratio of the electric fields of the probe pulse and coupling laser beam. Once the system starts in a dark state (as it does in this case: 100 percent coupling beam and 100 percent state 1), it adjusts to remain dark even when the probe beam lights up. The quantum interference effect is also responsible for the rapid variation of the refractive index that leads to slow light. The light speed can be controlled by simply controlling the coupling laser intensity: the lower the intensity, the steeper the slope, and the lower the light speed. In short, the light speed scales directly with the coupling intensity.

Section 6: Implementing These Ideas in the Laboratory

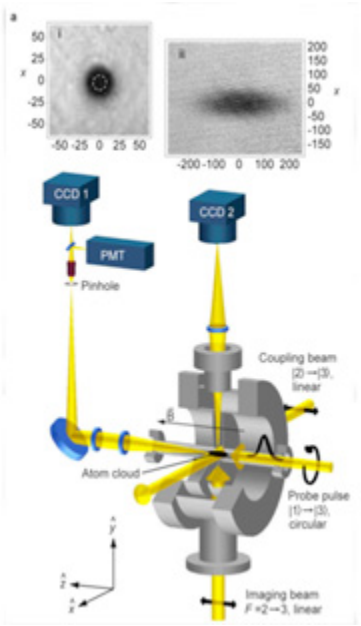


Figure 13: Experimental realization of slow light.
Source: © Reprinted by permission from Macmillan Publishers Ltd: Nature 397, 594-598 (18 February 1999).

Figure 13 shows a typical setup that we actually use for slowing light. We hold the atom cloud fixed in the middle of the vacuum chamber with use of the electromagnet and first illuminate the atom cloud from the side with the coupling laser. Then we send the probe light pulse into the cooled atoms. We can make a very direct measurement of the light speed: We simply sit and wait behind the vacuum chamber for the light pulse to come out, and measure the arrival time of the light pulse. To figure out the light speed, we just need to know the length of the cloud. For this purpose, we send a third "imaging" laser beam into the chamber from down below, after the probe pulse has propagated through the atom cloud and the coupling laser is turned off.

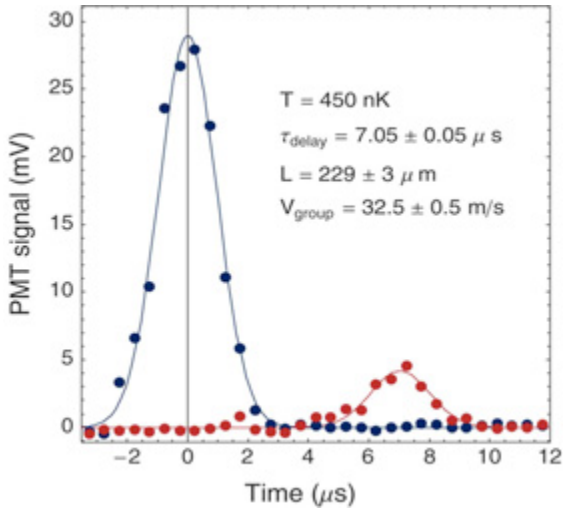


Figure 14: Measurement of ultra-slow light in a cold atom cloud.
Source: © Reprinted by permission from Macmillan Publishers Ltd: Nature 397, 594-598 (18 February 1999).

As the imaging laser is tuned on resonance with the atom's characteristic frequency, and there is only one laser beam present (i.e., there is no quantum interference), the atoms will absorb photons and create an absorption shadow in the imaging beam. By taking a photograph of this shadow with a camera, we can measure the length of the cloud. An example is seen in Figure 13 where the shadow (and the cloud) has a length of 0.008 inches. By sending a light pulse through precisely this cloud, we record the red light pulse in Figure 14. It takes the pulse 7 microseconds (7 millionths of a second) to get through the cloud. We now simply divide the cloud length by the propagation time to obtain the light speed of the light pulse: 71 miles/hour. So, we have already slowed light by a factor of 10 million. We can lower the intensity of the coupling beam further and propagate the pulse through a basically pure Bose-Einstein condensate to get to even lower light speeds of 15 miles/hour. At this point, you can easily beat light on your bicycle.

Figure 15 illustrates how the light pulse slows in the atom cloud: associated with the slowdown is a huge spatial compression of the light pulse. Before we send the pulse into the atom cloud, it has a length of roughly a mile [its duration is typically a few microseconds, and by multiplying the duration by the light speed in free space (186,000 miles per second), we get the length]. As the light pulse enters the cloud, the front edge slows down; but the back edge is still in free space, so that end will be running at the normal light speed. Hence, it will catch up to the front edge and the pulse will start to compress. The pulse is compressed by the same factor as it is slowed, so in the end, it is less than 0.001" long, or less than half the thickness of a hair. Even though our atom clouds are small, the pulse ends up being even smaller, small enough to fit inside an atom cloud. The light pulse also makes an imprint in the atom cloud,

really a little holographic imprint. Within the localized pulse region, the atoms are in these dark states we discussed earlier. The spatial modulation of the dark state mimics the shape of the light pulse: in the middle of the pulse where the electric field of the probe laser is high, a large fraction of an atom's state is transferred from 1 to 2. At the fringe of the pulse, only a small fraction is in 2; and outside the light pulse, an atom is entirely in the initial state 1. The imprint follows the light pulse as it slowly propagates through the cloud. Eventually, the light pulse reaches the end of the cloud and exits into free space where the front edge takes off as the pulse accelerates back up to its normal light speed. In the process, the light pulse stretches out and regains the length it had before it entered the medium.

It is interesting to note that when the light pulse has slowed down, only a small amount of the original energy remains in the light pulse. Some of the missing energy is stored in the atoms to form the holographic imprint, and most is sent out into the coupling laser field. (The coupling photons already there stimulate new photons to join and generate light with the same wavelength, direction, and phase). When the pulse leaves the medium, the energy is sucked back out of the atoms and the coupling beam, and it is put back into the probe pulse.

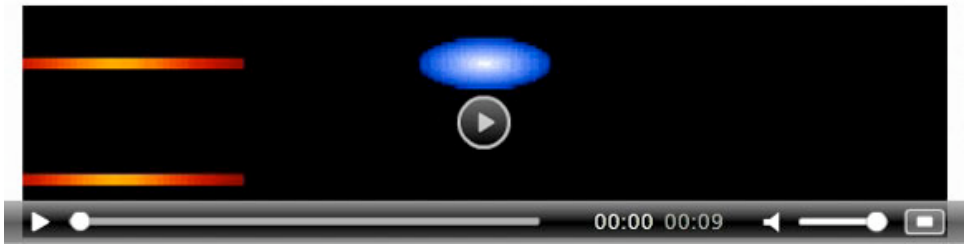


Figure 15: Light slows down in a Bose-Einstein condensate.
Source: © Chien Liu and Lene V. Hau.

Try to run the animation again and click on the screen when the pulse is slowed down, compressed, and contained in the atom cloud: The light pulse stops. You might ask: Could we do the same in the lab? The answer is: Yes, indeed, and it is just as easy—just block the coupling laser. As mentioned earlier, when the coupling intensity decreases, the light speed also decreases, and the light pulse comes to a halt. The atoms try to maintain the dark state. If the coupling laser turns off, the atoms will try to absorb probe light and emit some coupling light, but the light pulse runs empty before anything really changes. So, the light pulse turns off but the information that was in the pulse is not lost: it is preserved in the holographic imprint that was formed and that stays in the atom cloud.

Before we move on, it is interesting to pause for a moment. The slowing of light to bicycle speed in a Bose-Einstein condensate in 1999 stimulated a great many groups to push for achieving slow light.

In some cases, it also stopped light in all kinds of media, such as hot gases, cooled solids, room temperature solids, optical fibers, integrated resonators, photonic crystals, and quantum wells, and with classical or quantum states of light. The groups include those of Scully at Texas A&M, Budker at Berkeley, Lukin and Walsworth at Harvard and CFA, Hemmer at Hanscom AFB, Boyd at Rochester, Chuang at UIUC, Chang-Hasnain at Berkeley, Kimble at Caltech (2004), Kuzmich at Georgia Tech, Longdell et al. in Canberra, Lipson at Cornell, Gauthier at Duke, Gaeta at Cornell, Vuckovic at Stanford, Howell at Rochester, and Mørk in Copenhagen, to name just a few.

Section 7: Converting Light to Matter and Back: Carrying Around Light Pulses

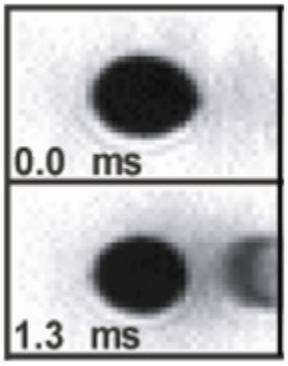


Figure 16: Turning light into matter with the creation of a perfect matter copy of the extinguished light pulse.
Source: © Reprinted by permission from Macmillan Publishers Ltd: Nature 445, 623-626 (8 February 2007).

Light and other electromagnetic waves carry energy, and that energy comes in quanta called "photons." The energy of a photon is proportional to the frequency of the light with a constant of proportionality called Planck's constant. We have already encountered photons on several occasions and seen that photons carry more than energy—they also carry momentum: an atom gets that little kick each time it absorbs or emits a photon.

When we slow, compress, stop, or extinguish a light pulse in a Bose-Einstein condensate, we end up with the holographic, dark state imprint where atoms are in states 1 and 2 at the same time. That part of an atom which is in state 2 got there by the absorption of a probe laser photon and the emission of a coupling laser photon. It has thus received two momentum kicks: one from each laser field. Therefore, the state 2 imprint starts to move slowly—at a few hundred feet per hour. It will eventually exit the condensate and move into free space.

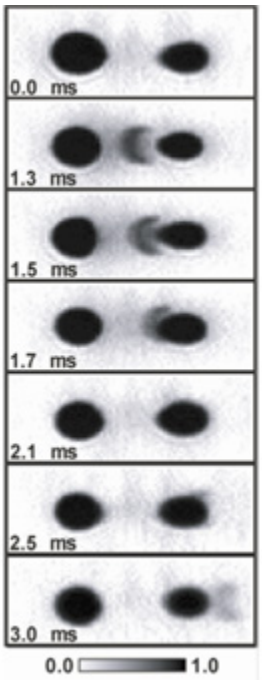


Figure 17: The matter copy continues the journey.
Source: © Reprinted by permission from Macmillan Publishers Ltd: Nature 445, 623-626 (8 February 2007).

At this point, we have created, out in free space, a matter copy of the light pulse that was extinguished. This matter copy can be photographed as shown in Figure 16. It takes the matter copy roughly one-half millisecond to exit the condensate, and we note that it has a boomerang shape. Why? Because when the light pulse slows down in the condensate, it slows most along the centerline of the condensate where the density of atoms is highest. Therefore, the light pulse itself develops a boomerang shape, which is clearly reflected in its flying imprint.

It is also interesting to remember that the atoms are in a superposition of quantum states: An atom in the matter copy is actually both in the matter copy traveling along and at the same time stuck back in the original condensate. When we photograph the matter copy, however, we force the atoms to make a decision as to "Am I here, or am I there?" The wavefunction collapses, and the atom is forced into a single definite quantum state. If we indeed have at hand a perfect matter copy of the light pulse, it should carry exactly the same information as the light pulse did before extinction; so could we possibly turn the matter copy back into light? Yes.

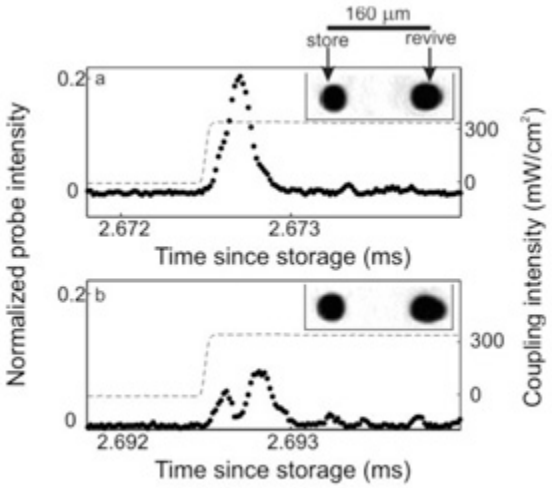


Figure 18: Turning light into matter in one BEC and then matter back into light in a second BEC in a different location.
Source: © Reprinted by permission from Macmillan Publishers Ltd: Nature 445, 623-626 (8 February 2007).

To test this, we form a whole different condensate of atoms in a different location, and we let the matter copy move into this second condensate. Two such condensates are shown in Figure 17. In these experiments, the light pulse comes in from the left and is stopped and extinguished in the first (leftmost) condensate. The generated matter copy travels into and across free space, and at some point it reaches the second (rightmost) condensate. If we don't do anything, the matter copy will just go through the condensate and come out on the other side at 2.5 ms and 3.0 ms (See Figure 18). On the other hand, once the matter copy is imbedded in the second condensate, if we turn on the coupling laser...out comes the light pulse! The regenerated light pulse is shown in Figure 18a.

So here, we have stopped and extinguished a light pulse in one condensate and revived it from a completely different condensate in a different location. Try to run the animation in Figure 19 and test this process for yourself.

Figure 19: A light pulse is extinguished in one part of space and then regenerated in a different location.
Source: © Sean Garner and Lene V. Hau.

Quantum puzzles

The more you think about this, the weirder it is. This is quantum mechanics seriously at play. The matter copy and the second receiver condensate consist of two sets of atoms that have never seen each other, so how can they together figure out to revive the light pulse? The secret is that we are dealing with Bose-Einstein condensates. When we illuminate the atoms in the matter copy with the coupling laser, they act as little radiating antennae. Under normal circumstances, these antennae would each do its own thing, and the emitted radiation would be completely random and contain no information. However, the lock-step nature of the receiver condensate will phase lock the antennae so they all act in unison, and together they regenerate the light pulse with its information preserved. When the light pulse is regenerated, the matter copy atoms are converted to state 1 and added as a bump on the receiver condensate at the revival location. The light pulse slowly leaves the region, exits the condensate, and speeds up.

This demonstrated ability to carry around light pulses in matter has many implications. When we have the matter copy isolated in free space, we can grab onto it—for example, with a laser beam—and put it "on the shelf" for a while. We can then bring it to a receiver condensate and convert it back into light. And while we are holding onto the matter copy, we can manipulate it, change its shape—its information content. Whatever changes we make to the matter copy will then be contained in the revived light pulse. In Figure 18b, you see an example of this: During the hold time, the pulse is changed from a single-hump to a double-hump pulse.

You might ask: How long can we hold on to a light pulse? The record so far is a few seconds. During this time, light can go from the Earth to the Moon! In these latest experiments, we let the probe and coupling

beams propagate in the same direction; so there is basically no momentum kick to the matter imprint, and it stays in the atom condensate where it was created. When atoms in the matter imprint collide with the host condensate, they can scatter to states other than 1 and 2. This will lead to loss of atoms from the imprint and therefore to loss of information. By exposing the atom cloud to a magnetic field of just the right magnitude, we can minimize such undesired interactions between the matter imprint and the host condensate. Even more, we can bring the system into a phase-separating regime where the matter imprint wants to separate itself from the host condensate, much like an oil drop in water. The matter imprint digs a hole for itself in the host, and the imprint can snugly nestle there for extended time scales without suffering damaging collisions. Also, we can move the matter imprint to the opposite tip of the condensate from where it came in, and the imprint can now be converted to light that can immediately exit the condensate without losses. This shows some of the possibilities we have for manipulating light pulses in matter form.

Section 8: Potential Long-Term Applications: Quantum Processing

The best and most efficient way to transport information is to encode it in light and send it down an optical fiber where it can propagate with low loss and at high data rates. This process is used for information traveling over the Internet. However to manipulate—or process—the information, it is much better to have it in matter form, as you can grab matter and easily change it. In current fiber optical networks, optical pulses are changed to electronic signals at nodes in the network. However, in these systems with existing conversion methods, the full information capacity of light pulses cannot be utilized and maintained. Using instead the methods described above, all information encoded in light—including its quantum character—can be preserved during conversion between light and matter, and the information can be powerfully processed. This is particularly important for creating quantum mechanical analogues of the Internet. Transport and processing of data encoded in quantum states can be used for teleportation and secure encryption. The ability to process quantum information is needed for the development of quantum computers.

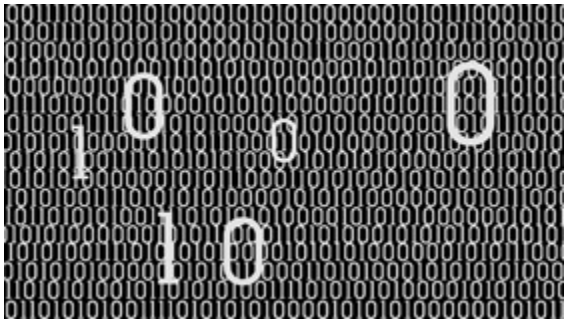


Figure 20: Classical computers use bits that can be valued either 0 or 1.
Source:

One possible application of the powerful methods for light manipulation described above is in quantum computing. In our current generation of computers, data are stored in memory as bit patterns of 0s and 1s (binary numbers). In a quantum computer, the 1s and 0s are replaced with quantum superpositions of 1's and 0's, called qubits, which can be 0 and 1 at the same time. Such computers, if they can be built, can solve a certain set of hard problems that it would take an ordinary computer a very long time (longer than the life of the universe) to crack. The trick is that because of quantum superposition, the input bit register in a quantum computer can hold all possible input values simultaneously, and the computer can

carry out a large number of calculations in parallel, storing the results in a single output register. Here we get to another important aspect of quantum computing: entanglement.

Let's look at a simple case of entanglement, involving a light pulse containing a single photon (i.e., we have a single energy quantum in the light pulse. There is some discussion as to whether this is a case of true entanglement, but it serves to illustrate the phenomenon in any case). If we send the photon onto a beamsplitter, it has to make a choice between two paths. Or rather, quantum mechanics allows the photon to not actually make up its mind: The photon takes both paths at the same time. We can see this through interference patterns that are made even when the photon rate is so low that less than one photon is traveling at a time. But if we set up a light detector to monitor one of the two paths, we will register a click (a photon hit) or no click, each with a 50% chance. Now say some other (possibly distant) observer sets up a detector to monitor the other path. In this case, if this second observer detects a click, that instantaneously affects our own measurement: We will detect "no click" with 100% probability (or vice versa, an absolutely certain click if the remote observer has detected no click), since each photon can be detected only once. And this happens even in cases where the detectors are at a very large distance from each other. This is entanglement: the quantum correlation of two spatially separated quantum states. The fact that a measurement in one location on part of an entangled state immediately affects the other part of the state in a different location—even if the two locations are separated by a very large distance—is a nonlocal aspect of quantum mechanics. Entangled states are hard to maintain, because interactions with the environment destroy the entanglement. They thus rarely appear in our macroscopic world, but that doesn’t stop them from being essential in the quantum world.

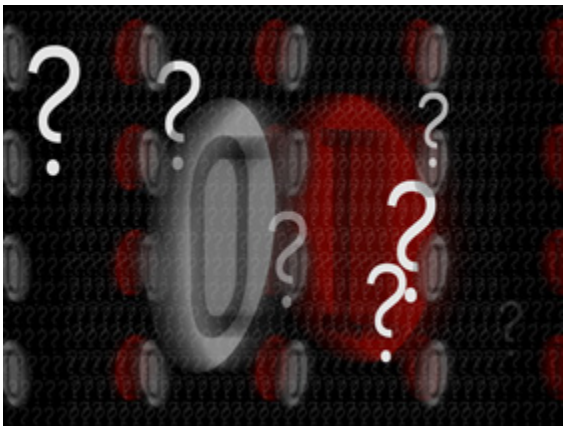


Figure 21: Quantum computers use qubits.
Source:

Now that we know what entanglement is, let's get back to the quantum computer. Let's say the quantum computer is set up to perform a particular calculation. One example would be to have the result presented at the output vary periodically with the input value (for example, if the computer calculates the value of the sine function). We wish to find the repetition period. The input and output registers of the computer each consist of a number of quantum bits, and the quantum computation leaves these registers in a superposition of matched pairs, where each pair has the output register holding the function value for the corresponding input register value. This is, in fact, an entangled state of the two registers, and the entanglement allows us to take advantage of the parallelism of the quantum computation. As in the beamsplitter example described above, a measurement on the output register will immediately affect the input register that will be left holding the value that corresponds to the measured output value. If the output repeats periodically as we vary the input, many inputs will correspond to a particular output, and the input register ends up in a superposition of precisely these inputs. This represents a great advantage over a classical calculation: After just one run-through of the calculation, global information—the repetition period—is contained in the input register. After a little "fiddling," this period can be found and read back out with many fewer operations than a classical computer requires. In a classical calculation, we would have to run the computation many times—one for each input value—to slowly, step by step, build up information about the periodicity.

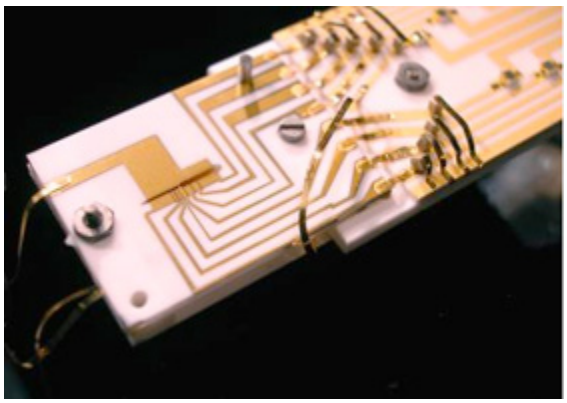


Figure 22: A two-bit quantum computer implemented with two beryllium ions trapped in a slit in an alumina substrate with gold electrodes.
Source: © J. Jost, NIST.

Whether quantum computers can be turned into practical devices remains to be seen. Some of the questions are: Can the technology be scaled? That is, can we generate systems with enough quantum bits to be interesting? Is the processing (the controlled change of quantum states) fast enough that it happens before coupling to the surroundings leads to "dephasing," that is to destruction of the

superposition states and the entanglement? Many important experiments aimed at the implementation of a quantum computer have been performed, and quantum computing with a few qubits has been successfully carried out, for example, with ions trapped in magnetic traps (See Figure 22). Here, each ion represents a quantum bit, and two internal states of the ion hold the superposition state corresponding to the quantum value of the bit. Encoding quantum bits—or more generally quantum states—via slow light-generated imprints in Bose-Einstein condensates presents a new and interesting avenue toward implementation of quantum computation schemes: The interactions between atoms can be strong, and processing can happen fast. At the same time, quantum dephasing mechanisms can be minimized (as seen in the experiments with storage times of seconds). Many light pulses can be sent into a Bose-Einstein condensate and the generated matter copies can be stored, individually manipulated, and led to interact. One matter copy thus gets affected by the presence of another, and these kinds of operations—called conditional operations—are of major importance as building blocks for quantum computers where generation of entanglement is essential. After a series of operations, the resulting quantum states can be read back out to the optical field and communicated over long distances in optical fibers.

The transfer of quantum information and correlations back and forth between light and matter in Bose-Einstein condensates may allow for whole new processing algorithms where the classical idea of bit-by-bit operations is replaced by global processing algorithms where operations are performed simultaneously on 3D input patterns, initially read into a condensate, and with use of the full coherence—phase-lock nature—of the condensate. Fundamental questions, regarding the information content of condensates, for example, will need to be addressed.

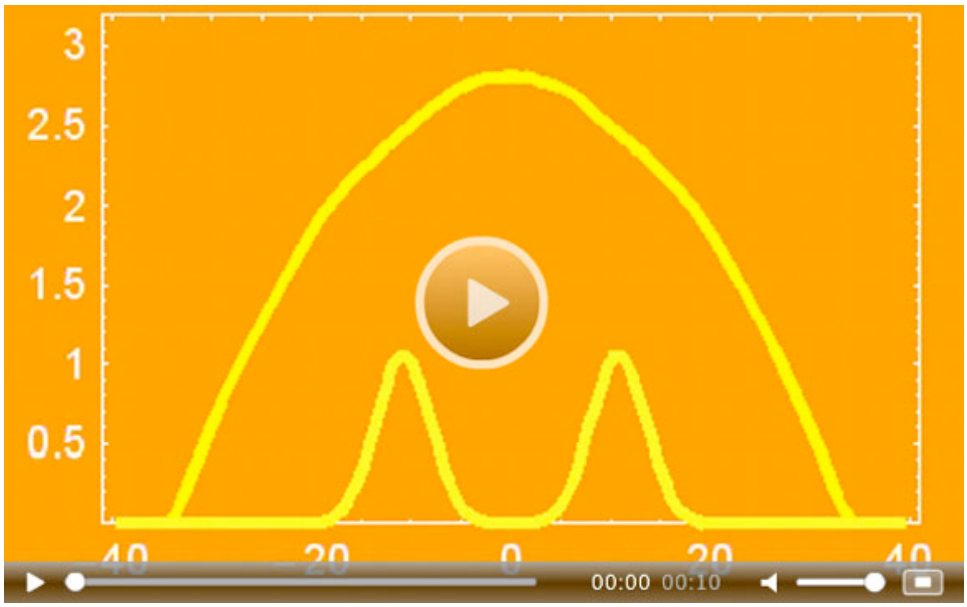


Figure 23: A Bose-Einstein condensate as a novel processor for quantum information.
Source: © Zachary Dutton and Lene V. Hau.

Another potential application of slow light-based schemes for quantum information processing, which has the promise to be of more immediate, practical use, is for secure encryption of data sent over a communication network (for example, for protecting your personal information when you send it over the Internet). Entangled quantum states can be used to generate secure encryption keys. As described above with the photon example, two distant observers can each make measurements on an entangled state. If we make a measurement we will immediately know what the other observer will measure. By generating and sharing several such entangled quantum states, a secure encryption key can be created. The special thing with quantum states is that if a spy listens in during transmission of the key, we would know: If a measurement is performed on a quantum state, it changes—once again: The wavefunction "collapses." So the two parties can use some of the shared quantum states as tests: They can communicate the result of their measurements using their cell phones. If there is not a complete agreement between expected and actual measurements at the two ends, it is a clear indication that somebody is listening in and the encryption key should not be used.

An efficient way to generate and transport entangled states can be achieved with the use of light (as in our photon example above) transmitted in optical fibers. However, the loss in fibers is not negligible, and entanglement distribution over long distances (above 100 km, say) has to be made in segments. Once the individual segments are entangled, they must then be connected in pairs to distribute the entanglement between more distant locations. For this to be possible, we must be able to sustain the

entanglement achieved in individual segments such that all the segments can eventually be connected. And here, the achievement of hold times for light of several seconds, as described above, is really important.

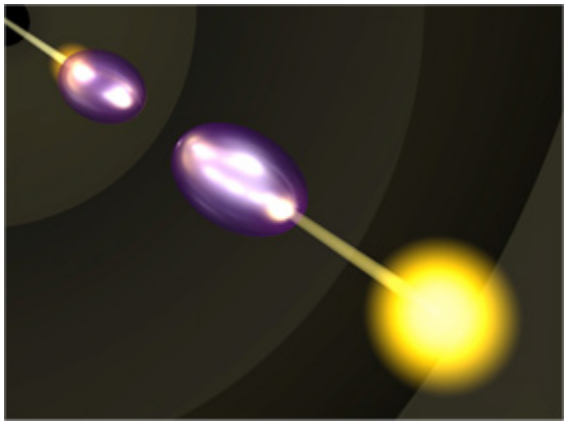


Figure 24: Was Newton right after all: "Are not gross Bodies and Light convertible into one another... ?"
Source: © Sean Garner and Lene V. Hau.

There are numerous applications for slow and stopped light, and we have explored just a few of them here. The important message is that we have achieved a complete symmetry between light and matter, and we get there by making use of both lasers and Bose-Einstein condensates. A light pulse is converted to matter form, and the created matter copy—a perfect imitation of the light pulse that is extinguished—can be manipulated: put on the shelf, moved, squeezed, and brought to interact with other matter. At the end of the process, we turn the matter copy back into light and beam it off at 186,000 miles per hour. During formation of the matter copy—by the slowing and imprinting of the input light pulse in a Bose-Einstein condensate—it is essential that the coupling laser field is present with its many photons that are all in lock-step. And when the manipulated matter copy is transformed back into light, the presence of many atoms in the receiver (or host) BEC that are all in lock-step is of the essence.

So, we have now come full circle: From Newton over Huygens, Young, and Maxwell, we are now back to Newton:

In *Opticks*, published in 1704, Newton theorized that light was made of subtle corpuscles and ordinary matter of grosser corpuscles. He speculated that through an alchemical transmutation, "Are not gross Bodies and Light convertible into one another, ...and may not Bodies receive much of their Activity from the Particles of Light which enter their Composition?"

Section 9: Further Reading

- Video lectures on quantum computing by Oxford Professor David Deutsch: http://www.quiprocone.org/Protected/DD_lectures.htm.
- Lene Vestergaard Hau,"Frozen Light," *Scientific American*, 285, 52-59 (July 2001), and Special Scientific American Issue entitled, "The Edge of Physics" (2003).
- Lene Vestergaard Hau, "Taming Light with Cold Atoms," *Physics World* 14, 35-40 (September 2001). Invited feature article. Published by Institute for Physics, UK.
- Simon Singh, "The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography," *Anchor*, reprint edition (August 29, 2000).

Glossary

Bose-Einstein condensate: A Bose-Einstein condensate, or BEC, is a special phase of matter in which the quantum mechanical wavefunctions of a collection of particles line up and overlap in a manner that allows the particles to act as a single quantum object. The electrons in a superconductor form a BEC; superfluid helium is an example of a liquid BEC. BECs can also be created from dilute gases of ultracold atoms and molecules.

Doppler shift (Doppler effect): The Doppler shift is a shift in the wavelength of light or sound that depends on the relative motion of the source and the observer. A familiar example of a Doppler shift is the apparent change in pitch of an ambulance siren as it passes a stationary observer. When the ambulance is moving toward the observer, the observer hears a higher pitch because the wavelength of the sound waves is shortened. As the ambulance moves away from the observer, the wavelength is lengthened and the observer hears a lower pitch. Likewise, the wavelength of light emitted by an object moving toward an observer is shortened, and the observer will see a shift to blue. If the light-emitting object is moving away from the observer, the light will have a longer wavelength and the observer will see a shift to red. By observing this shift to red or blue, astronomers can determine the velocity of distant stars and galaxies relative to the Earth. Atoms moving relative to a laser also experience a Doppler shift, which must be taken into account in atomic physics experiments that make use of laser cooling and trapping.

entanglement: In quantum mechanics, entanglement occurs when the quantum states of two particles that may be spatially separated are linked together. A measurement of one of the entangled particles implies the result of the same measurement made on the other entangled particle.

evaporative cooling: Evaporative cooling is a process used in atomic physics experiments to cool atoms down to a few billionths of a degree above absolute zero. The way it works is similar to how a cup of hot coffee cools through evaporation. Atoms are pre-cooled, usually with some kind of laser cooling, and trapped in a manner that imparts no additional energy to the atoms. The warmest atoms are removed from the trap, and the remaining atoms reach a new, lower equilibrium temperature. This process is typically repeated many times, creating small clouds of very cold atoms.

ground state: The ground state of a physical system is the lowest energy state it can occupy. For example, a hydrogen atom is in its ground state when its electron occupies the lowest available energy level.

index of refraction: A material's index of refraction is defined as the speed that light travels in the material divided by the speed of light in a vacuum. Therefore, the index of refraction of the vacuum is equal to one. Light slows down as it enters a material due to the interaction between the oscillating electric and magnetic fields of the light wave and the constituent parts of the material. The index of refraction of air is 1.003, and the index of refraction of water is 1.33.

phase: In physics, the term phase has two distinct meanings. The first is a property of waves. If we think of a wave as having peaks and valleys with a zero-crossing between them, the phase of the wave is defined as the distance between the first zero-crossing and the point in space defined as the origin. Two waves with the same frequency are "in phase" if they have the same phase and therefore line up everywhere. Waves with the same frequency but different phases are "out of phase." The term phase also refers to states of matter. For example, water can exist in liquid, solid, and gas phases. In each phase, the water molecules interact differently, and the aggregate of many molecules has distinct physical properties. Condensed matter systems can have interesting and exotic phases, such as superfluid, superconducting, and quantum critical phases. Quantum fields such as the Higgs field can also exist in different phases.

interference: Interference is an effect that occurs when two or more waves overlap. In general, the individual waves do not affect one another, and the total wave amplitude at any point in space is simply the sum of the amplitudes of the individual waves at that point. In some places, the two waves may add together, and in other places they may cancel each other out, creating an interference pattern that may look quite different than either of the original waves. Quantum mechanical wavefunctions can interfere, creating interference patterns that can only be observed in their corresponding probability distributions.

light-year: A light-year is the distance that light, which moves at a constant speed, travels in one year. One light-year is equivalent to 9.46×10^{15} meters, or 5,878 billion miles.

optical molasses: Optical molasses is formed when laser beams for Doppler cooling are directed along each spatial axis so that atoms are laser cooled in every direction. Atoms can reach microkelvin temperatures in optical molasses. However, the molasses is not a trap, so the atoms can still, for example, fall under the influence of gravity.

qubit: A qubit is the quantum counterpart to the bit for classical computing. A bit, which is short for binary digit, is the smallest unit of binary information and can assume two values: zero and one. A qubit, or quantum bit, is in a quantum superposition of the values zero and one. Because the qubit is in a

superposition and has no definite value until it is measured directly, quantum computers can operate exponentially faster than classical computers.

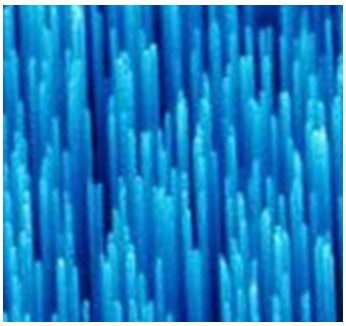
Snell's Law: Snell's law describes how the path of a light ray changes when it moves into a material with a different index of refraction. According to Snell's law, if a light ray traveling through a medium with index of refraction n_1 hits the boundary of a material with index n_2 at angle θ_1, the light path is bent and enters the new material at an angle θ_2 given by the relation $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

superposition principle: Both quantum and classical waves obey the superposition principle, which states that when two waves overlap, the resulting wave is the sum of the two individual waves.In quantum mechanics, it is possible for a particle or system of particles to be in a superposition state in which the outcome of a measurement is unknown until the measurement is actually made. For example, neutrinos can exist in a superposition of electron, muon, and tau flavors (Units 1 and 2). The outcome of a measurement of the neutrino's flavor will yield a definite result—electron, muon, or tau—but it is impossible to predict the outcome of an individual measurement. Quantum mechanics tells us only the probability of each outcome. Before the measurement is made, the neutrino's flavor is indeterminate, and the neutrino can be thought of as being all three flavors at once.

valence electron: A valence electron is an electron in the outermost shell of an atom in the Lewis model, or in the orbital with the highest value of the principal quantum number, *n*, in the quantum mechanical description of an atom. The valence electrons determine most of the chemical and physical properties of the atom. It is the valence electrons that participate in ionic and covalent chemical bonds, and that make the primary contributions to an atom's magnetic moment.

Physics for the 21st Century

Unit 8: Emergent Behavior in Quantum Matter



© Deli Wang Laboratory at UCSD.

Unit Overview

This unit takes an approach to physics that differs markedly from much of what we have encountered in previous units. Rather than cataloging the elementary components of matter, we look at what happens at the macroscopic scale when the interactions of these components with one another and their environment lead to entirely new—emergent—behavior. After introducing the concept of emergence, the unit examines emergent behavior in solid matter, quantum plasmas, and the very different behavior of the liquid forms of two different isotopes of helium (He). The next two sections cover the search for a microscopic theory of superconductivity and its culmination in Bardeen-Cooper-Schrieffer (BCS) theory, which triumphantly accounted for the emergent properties of conventional superconductors. The final three sections focus on efforts to understand emergence in new and different contexts, from freshly discovered forms of superconductivity on Earth to the cosmic superfluidity observed in pulsars —rotating stars made up primarily of neutrons.

Content for This Unit

Sections:

Section 1: Introduction

The term emergent behavior refers to the collective phenomena observed in macroscopic systems that are distinct from their microscopic constituents. It is brought about by the interaction of the microscopic constituents with one another and with their environment. Whereas the Standard Model of particle physics described in Units 1 and 2 has enjoyed great success by building up systems of particles and interactions from the ground up, nearly all complex and beautiful phenomena observed in the laboratory or in nature defy this type of reductionist explanation. Life is perhaps the ultimate example. In this unit, we explore the physics of emergent phenomena and learn a different approach to problem solving that helps scientists understand these systems.

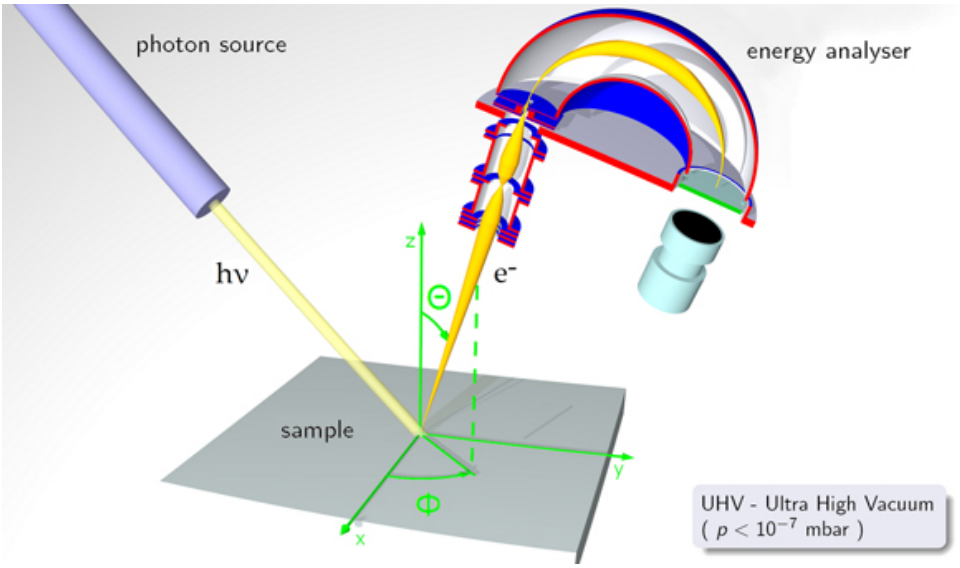


Figure 1: New experimental probes, such as angle-resolved photoemission spectroscopy (ARPES) enable us to study emergent behavior in quantum matter at an unprecedented level of detail. ARPES is an experimental technique to observe the distribution of the electrons (more precisely, the density of single-particle electronic excitations) in the reciprocal space of solids.
Source: © Wikimedia Commons, Public Domain. Author: Saiht, 15 June 2009.

Understanding emergent behavior requires a change of focus. Instead of adopting the traditional reductionist approach that begins by identifying the individual constituents and interactions of a system and then uses them as the basic building blocks for creating a model of a system's behavior, we must focus on identifying the origins of the emergent collective behavior characteristic of the system. Thus, in creating models of quantum matter, we use the organizing principles and concepts responsible for emergent quantum behavior as our basic building blocks. These new building blocks, the collective organizing principles, represent gateways to emergence.

Certain aspects of emergent behavior are considered "protected," in the sense that they are insensitive to the details of the underlying microscopic physical processes.

Much of this unit deals with the phenomenon of superconductivity, where electrical current can flow with absolutely no resistance whatsoever. Unraveling the mystery of how electrons (which experience a repulsive electrical interaction) can cooperate and produce coherent collective phenomena is a detective story that is still being written.

To build a little intuition about the subtle interactions that give rise to superconductivity (and the related phenomenon of superfluidity), and to set the stage for what follows, imagine two electrically charged bowling balls placed on a spring mattress. They experience a repulsive electrostatic interaction, but the indentation in the mattress made by one of the bowling balls can influence the other one, producing a net attractive force between them. This is crudely analogous to the direct interactions between electrons and their coupling to the excitations of the crystal lattice of a superconducting metal: an attractive interaction between electrons can result if the interaction between each electron and the excitations of the embedding lattice is strong enough to overcome the electrostatic repulsion and this interaction can give rise to the pairing condensate that characterizes the superconducting state.

Another concept that will arise in this unit is the notion of a "quasiparticle." The concept of a quasiparticle arises in a simplifying framework that ascribes the combined properties of an electron and its modified surroundings into a "virtual" equivalent composite object that we can treat as if it were a single particle. This allows us to use the theoretical toolkit that was built up for the analysis of single particles.

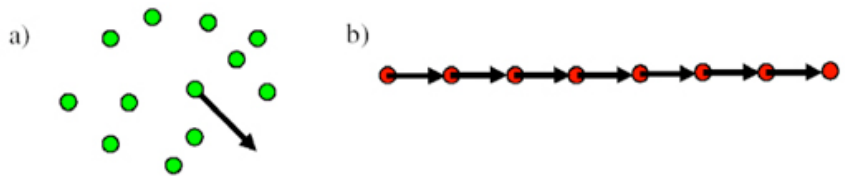


Figure 2: As shown in the figure, dimensionality can influence dramatically the behavior of quasiparticles in metals.
Source: © Joseph Sulpizio and David Goldhaber-Gordon, Stanford University.

As we will see below, both the superfluid state of ^{3}He and the superconducting states of metals come about because of quasiparticle pairing processes that transform a collection of fermions (the nuclei in the case of superfluid ^{3}He, and electrons in the case of superconductivity) into a collective, coherent single

quantum state, the superfluid condensate. This exhibits the macroscopic quantum mechanical effects of a superfluid flowing with no dissipation, and of the flow of electrical current with literally zero resistance, in a superconductor.

Understanding the subtle interactions that give rise to the pairing of fermions requires looking at the fluids and materials in a different way.

Our change in focus means, in general, that instead of following the motion of single particles in a material, we will focus on the behavior of the material as a whole—for example, the density fluctuations found in bulk matter. A simple example of a density fluctuation is a sound wave traveling through a medium such as air, water, or a solid crystal. Just as light can equally well be described as fluctuations of the electromagnetic field whose quantized particles are called "photons," the collective density fluctuations in crystalline solids can be described by quantized particles called phonons. Analogously to the particle interactions described in Unit 2, the electronic density fluctuations can couple to phonons and other fields, and the interaction between the density wave and various fields can represent both the influence of particle interactions and the external probes used to measure systems' behavior. We will also be interested in the spin fluctuations of fermions, which are particles with half-integer spin.

This unit will introduce the frontiers of research in the study of emergent behavior in quantum matter and call attention to the applicability of some key organizing principles to other subfields in physics. Sections 2 through 5 present what we might call "old wine in a new bottle"—an emergent perspective on subject matter described in many existing texts on quantum matter, while the last three sections highlight the frontiers of research in the field.

In the two sections devoted to superconductivity, I have gone to some length to sketch the immediate emergent developments that led up to the historic paper in which Bardeen, Cooper, and Schrieffer described their microscopic theory known as BCS. I have done so, in part, because I can write about these from personal experience. But I also believe that learning how a major problem in physics was finally solved after 45 years of trying might help nonscientists and would-be scientists appreciate the complex process of discovery and provide encouragement for young researchers seeking to solve some of the most challenging problems that our community faces today.

Section 2: Emergent Behavior in Crystalline Solids

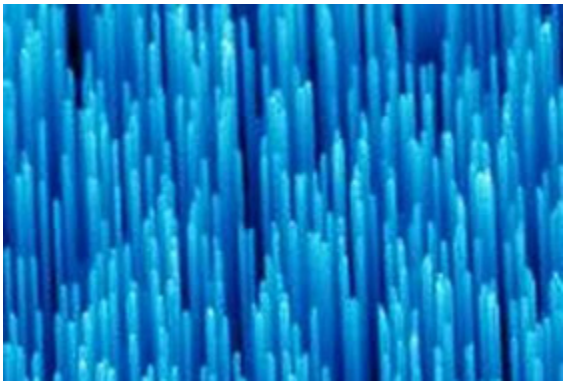


Figure 3: Nanowires are crystalline fibers with emergent behaviors expected to be used for nanoscale applications.
Source: © Deli Wang Laboratory at UCSD.

Crystalline solids provide familiar examples of emergent behavior. This section will outline the theoretical steps that have revealed the fundamental nature of solids and the ways in which such critical ideas as quantum statistics, excitations, energy bands, and collective modes have enabled theorists to understand how solids exhibit emergent behavior.

At high enough temperatures, any form of quantum electronic matter becomes a plasma—a gas of ions and electrons linked via their mutual electromagnetic interaction. As it cools down, a plasma will first become liquid and then, as the temperature falls further, a crystalline solid. For metals, that solid will contain a stable periodic array of ions along with electrons that are comparatively free to move under the application of external electric and magnetic fields.

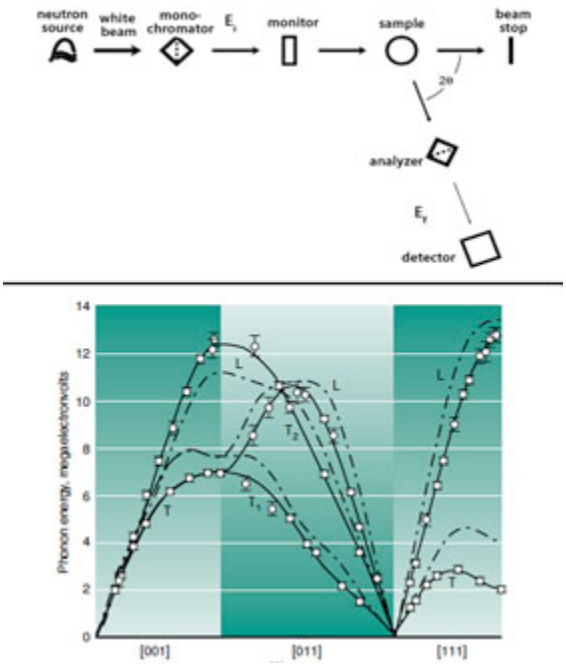


Figure 4: Left: Diagram showing experimental set-up for measurement of the energy loss spectrum of neutrons that are inelastically scattered by a crystal. Right: A typical phonon spectrum obtained through an inelastic neutron scattering (INS) experiment. **Source:** © Top: Wikimedia Commons, Public Domain. Author: Joachim Wuttke, 22 February 2007. Bottom: Courtesy of Joe Wong, Lawrence Livermore National Laboratory.

The crystallization process has broken a basic symmetry of the plasma: there are no preferred directions for the motion of its electrons. Broken symmetry is a key organizing concept in our understanding of quantum matter. The crystalline confinement of the ions to regular positions in the crystal leads to a quantum description of their behavior in terms of the ionic elementary excitations, called "phonons," which describe their vibrations about these equilibrium positions. Physicists can study phonons in detail through inelastic neutron scattering experiments (Figure 4) that fire neutrons at solid samples and measure the neutrons' energies after their collisions with the vibrating ions in the solid. Moreover, the solids' low-energy, long-wavelength behavior is protected: It is independent of details and describable in terms of a small number of parameters—in this case, the longitudinal and transverse sound velocities of their collective excitations, the quantized phonons.

Independent electrons in solids

What of the electrons? The interactions between the closely packed atoms in a periodic array in a crystal cause their outermost electrons to form the energy bands depicted in Figure 5. Here, the behavior of the

electrons is characterized by both a momentum and a band index, and their corresponding physical state depends to a first approximation on the valency of the ions and may be characterized by the material's response to an external electric field. Thus, the solid can take any one of three forms. It may be a metal in which the electrons can move in response to an external electric field; an insulator in which the electrons are localized and no current arises in response to the applied electric field; or a semi-conductor in which the valence band is sufficiently close to a conduction band that a small group of electrons near the top of the band are easily excited by heating the material and can move in response to an applied electric field.

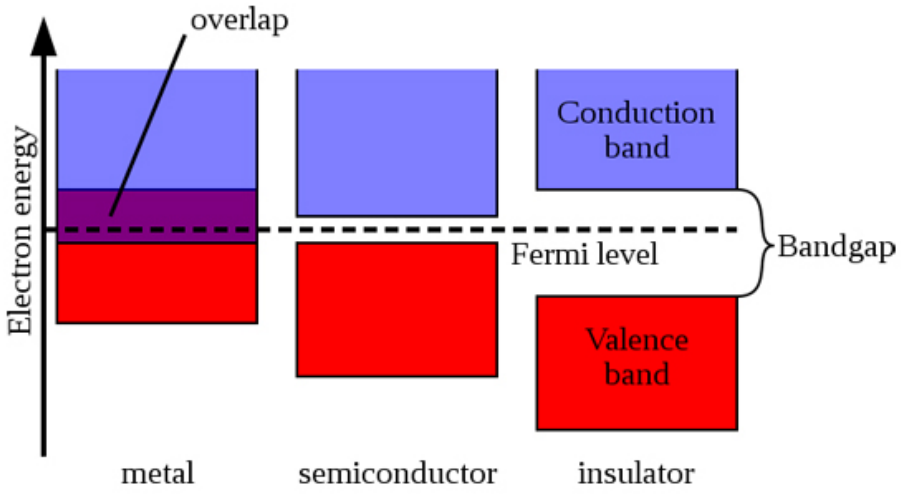


Figure 5: Comparison of the electronic band structures of metals, semiconductors, and insulators.
Source: © Wikimedia Commons, Public Domain. Author: Pieter Kuiper, 6 June 2007.

Physicists for many years used a simple model to describe conduction electrons in metals: the free electron gas or, taking the periodic field of ions into account through band theory, the independent electron model. The model is based on Wolfgang Pauli's exclusion principle that we met in Unit 2. Because electrons have an intrinsic spin of 1/2, no two can occupy the same quantum state. In the absence of the periodic field of ions, the quantum states of each electron can be labeled by its momentum, p, and its spin.

Specific Heat

The temperature dependence of the electronic specific heat in simple metals is easily explained; because of Pauli's principle that no two electrons can occupy the same quantum state, at a temperature T, only a fraction, kT/E_f, of the electrons inside the Fermi surface at which electrons possess the energy, E_f, can be excited. So, with each excited electron having an energy of about kT, the free energy of the system is roughly proportional to T^2 and its temperature derivative, the specific heat, varies linearly with T. Similar qualitative arguments explain why the electron spin susceptibility (its response to a uniform external magnetic field) first calculated by Wolfgang Pauli is independent of temperature.

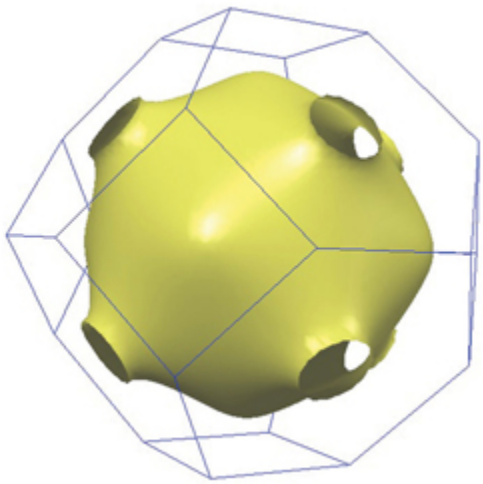


Figure 6: The Fermi surface reveals how the energy varies with momentum for the highest-energy electrons—those that have the Fermi energy.
Source: © Los Alamos National Laboratory.

Since the electrons can carry momentum in each of three independent spatial directions, it's useful to imagine a 3D coordinate system (which we call "momentum space") that characterizes the x, y, and z components of momentum.

The ground state of the electrons moving in a uniform background of positive charge would then be a simple sphere in momentum space bounded by its Fermi surface, a concept derived from the statistical work of Enrico Fermi and P.A.M. Dirac that defines the energies of electrons in a metal. When the uniform positive charge is replaced by the actual periodic array of the ions in a metallic lattice, the simple sphere

becomes a more complex geometric structure that reflects the nature of the underlying ionic periodic structure, an example of which is shown in Figure 6.

To calculate the ground state wavefunction of the electrons, physicists first applied a simple approach called the Hartree-Fock approximation. This neglects the influence of the electrostatic interaction between electrons on the electronic wavefunction, but takes into account the Pauli principle. The energy per electron consists of two terms: the electrons' average kinetic energy and an attractive exchange energy arising from the Pauli principle which keeps electrons of parallel spin apart.

Emergent concepts for a quantum plasma

Conscience and Contradictions

David Bohm.
Source: © Wikimedia Commons, Public Domain. Author: Karol Langner, 30 July 2005.

David Bohm's life involved a series of contradictions. Refused security clearance for work on the atom bomb during World War II, he made critical contributions to the development of the bomb. Ever suspicious of quantum mechanics, he wrote a classic textbook on the topic before attempting to develop a theory that explained all of quantum mechanics in terms of hidden variables whose statistical behavior produced the familiar quantum results. Widely regarded as one of the greatest physicists never to win the Nobel Prize, he spent the latter part of his career working primarly on philosophy and brain science.

In the early 1940s, Bohm performed theoretical calculations of collisions of deuterons and protons for his Ph.D. under Robert Oppenheimer at the University of California, Berkeley. But when Oppenheimer recruited him for the Manhattan Project, project head General Leslie Groves refused him security clearance because of his left-wing political associations. So, when his Ph.D. research was classified, he could not even finish his thesis; Oppenheimer had to certify that he had earned his Ph.D. Bohm then spent the years from 1942–45 working at Berkeley's Radiation laboratory on the classical plasmas found in the gas discharges associated with one of the methods used to

separate uranium isotopes and became famous there for his insight into the instabilities leading to plasma turbulence.

Politics reemerged in 1950, when the House Un-American Activities Committee cited Bohm for using the fifth amendment to refuse to answer its questions about his past political affiliations. By the time he was acquitted, the President of his university, Princeton, which had first suspended him, then refused to reappoint or promote him. Unable to obtain another scientific position in the United States because potential employers in both universities and the private sector feared being accused of appointing a communist, he became perhaps the most prominent scientific exile from the United States at a time when his scientific expertise was badly needed in the plasma-based efforts to develop a thermonuclear reactor. His exile first took him to professorships in Sao Paolo, Tel Aviv, and Bristol; he then spent the last 31 years of his life at the University of London's Birkbeck College and died in a London taxi in 1992.

The Russian physicist Lev Landau famously said, "You cannot repeal Coulomb's law." But until 1950, it appeared that the best way to deal with it was to ignore it, because microscopic attempts to include it had led to inconsistencies, or worse yet, divergent results. The breakthrough came with work carried out between 1949 and 1953 by quantum theorist David Bohm and myself, his Ph.D. student. Our research focused on the quantum plasma—electrons moving in a uniform background of positive charge, an idealized state of matter that solid-state physicist Conyers Herring called "jellium." Bohm and I discovered that when we viewed particle interactions as a coupling between density fluctuations, we could show, within an approximation we called "the random phase approximation (RPA)," that the major consequence of the long range electrostatic interaction between electrons was to produce an emergent collective mode: a plasma oscillation at a frequency, $\omega_p = (4\pi NE^2 / m)^{1/2}$, where N is the electron density and m its mass, whose quantized modes are known as plasmons. Once these had been introduced explicitly, we argued what was left was an effective short-range interaction between electrons that could be treated using perturbation-theoretic methods.

The plasma oscillation is an example of a "collisionless" collective mode, in which the restoring force is an effective field brought about by particle interaction; in this case, the fluctuations in density produce a fluctuating internal electric field. This is the first of many examples we will consider in which effective fields produced by particle interaction are responsible for emergent behavior. As we shall see later in this unit, the zero sound mode of ^{3}He furnishes another example, as does the existence of phonons in the

normal state of liquid ^{4}He. All such modes are distinct from the "emergent," but familiar, sound modes in ordinary liquids, in which the restoring forces originate in the frequent collisions between particles that make possible a "protected" long wavelength description using the familiar laws of hydrodynamics.

The importance of plasmons

Following their predicted existence, plasmons were identified as the causes of peaks that experimentalists had already seen in the inelastic scattering of fast electrons passing through or reflected from thin solid films. We now know that they are present in nearly all solids. Thus, plasmons have joined electrons, phonons, and magnons (collective waves of magnetization in ferromagnets) in the family of basic elementary excitations in solids. They are as well defined an elementary excitation for an insulator like silicon as for a metal like aluminum.

By the mid 1950s, it was possible to show that the explicit introduction of plasmons in a collective description of electron interactions resolved the difficulties that had arisen in previous efforts to deal in a consistent fashion with electron interaction in metals. After taking the zero point energy of plasmons into account in a calculation of the ground state energy, what remained was a screened electrostatic interaction between electrons of comparatively short range, which could be dealt with using perturbation theory. This work provided a microscopic justification of the independent electron model for metal, in which the effects of electron interaction on "single" electron properties had been neglected to first approximation. It also proved possible to include their influence on the cohesive energy of jellium with results that agreed well with earlier efforts by Eugene Wigner to estimate this quantity, and on the exchange and correlation corrections to the Pauli spin susceptibility, with results that agreed with its subsequent direct measurement by my Illinois colleague, C.P. Slichter.

It subsequently proved possible to establish that both plasma oscillations and screening in both quantum and classical plasmas are not simply mathematical artifacts of using the random phase approximation, but represent protected emergent behavior. Thus, in the limit of long wavelengths, plasma oscillations at ω_p are found at any density or temperature in a plasma, while the effective interaction at any temperature or density is always screened, with a screening length given by s / ω_p, where s is the isothermal sound velocity. Put another way, electrons in metals are never seen in isolation but always as "quasielectrons," each consisting of a bare electron and its accompanying screening cloud (a region in space in which there is an absence of other electrons). It is these quasielectrons that interact via a short range screened electrostatic interaction. For many metals, the behavior of these quasielectrons is likewise protected, in

this case by the adiabatic continuity that enables them to behave like the Landau Fermi liquids we will consider in the next section.

From plasmons to plasmonics

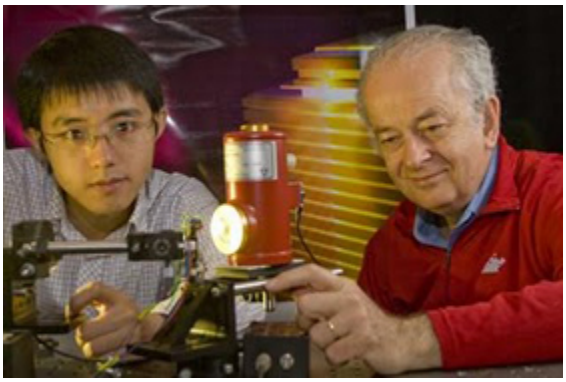


Figure 8: Harvard School of Engineering and Applied Sciences' researchers Federico Capasso (red shirt) and Nanfang Yu working on their nanoscale quantum clusters for plasmonic applications.
Source: © Eliza Grinnell.

When plasmons were proposed and subsequently identified, there seemed scant possibility that these would become a subject of practical interest. Unexpectedly, plasmons found at the surface of a metal, or at an interface between two solids, turn out to be sufficiently important in electronic applications at the nanoscale, that there now exists a distinct sub-field that marks an important intersection of physics and nanotechnology called "plasmonics." Indeed, beginning in 2006, there have been bi-annual Gordon research conferences devoted to the topic. To quote from the description of the founding conference: "Since 2001, there has been an explosive growth of scientific interest in the role of plasmons in optical phenomena, including guided-wave propagation and imaging at the subwavelength scale, nonlinear spectroscopy, and 'negative index' metamaterials. The unusual dispersion properties of metals near the plasmon resonance enables excitation of surface modes and resonant modes in nanostructures that access a very large range of wave vectors over a narrow frequency range, and, accordingly, resonant plasmon excitation allows for light localization in ultra-small volumes. This feature constitutes a critical design principle for light localization below the free space wavelength and opens the path to truly nanoscale plasmonic optical devices. This principle, combined with quantitative electromagnetic simulation methods and a broad portfolio of established and emerging nanofabrication methods, creates the conditions for dramatic scientific progress and a new class of subwavelength optical components." A description of the third such conference began with a description by Federico Capasso (Figure 8) of his bottom-up work on using self-assembled nanoclusters for plasmonic applications.

Section 3: Emergent Behavior in the Helium Liquids

The property that physicists call spin plays an essential role in the nature and emergent behavior of particles, atoms, and other units of matter. As we have noted previously, fermions have an intrinsic half-integer spin; no two fermions can occupy the same quantum state. And as we learned in Unit 6, because bosons have integral spin, any number of bosons can occupy the same quantum state. Those differences play out in the behavior at very low temperatures of the two isotopes of He—the fermion ^{3}He with spin of 1/2 owing to its single unpaired neutron and the boson ^{4}He with no net spin because it has two neutrons whose antiparallel spins sum to zero, as do the spins of the two protons in the He nucleus.

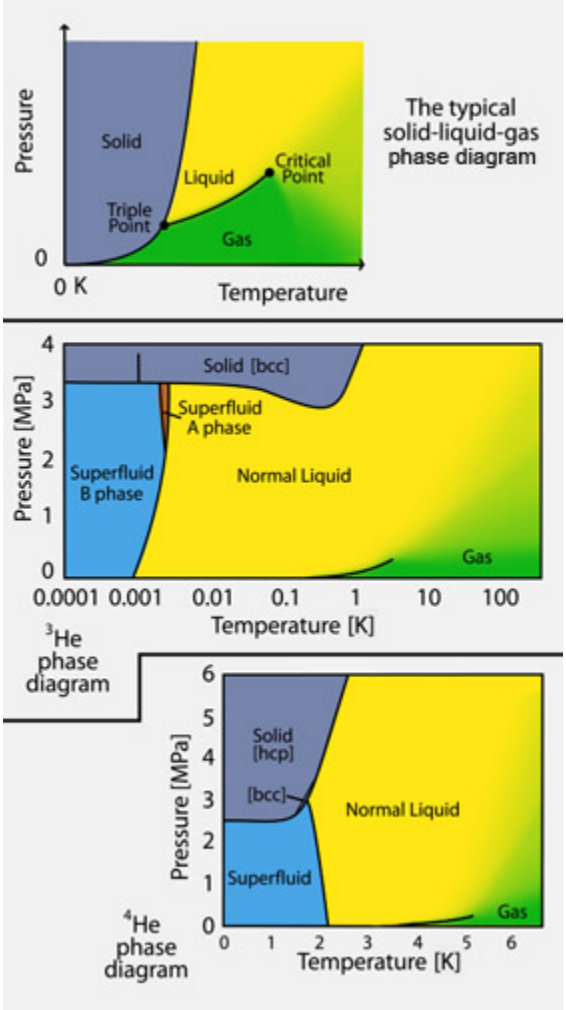


Figure 9: Temperature-pressure phase diagrams of the two quantum materials, ^{3}He and ^{4}He, that remain liquid down to the lowest temperatures in the absence of pressure compared to a typical liquid-solid phase diagram.
Source: Recreated from graphics by the Low Temperature Laboratory, Helsinki University of Technology.

The liquid forms of the two isotopes of He are the only two quantum liquids found in nature. Unlike all other atomic materials, because they are exceptionally light they do not freeze upon cooling; their zero point energy prevents them from freezing. As may be seen in Figure 9 (phase transitions) at low temperatures, the two isotopes of He exhibit remarkably different emergent behavior. Below 2.18 Kelvin (K), liquid ^{4}He becomes a superfluid that flows without appreciable resistance. Liquid ^{3}He behaves quite differently, flowing like a normal liquid down to temperatures in the millikelvin regime, some three orders of magnitude cooler, before it exhibits a transition to the superfluid state.

The reason is simple. Atoms of ^{4}He obey Bose-Einstein statistics. Below 2.18 K, a single quantum state, the Bose condensate, becomes macroscopically occupied; its coherent motion is responsible for its superfluid behavior. On the other hand, ^{3}He obeys Fermi-Dirac statistics, which specify that no two particles can occupy the same quantum state. While, as we shall see, its superfluidity also represents condensate motion, the condensate forms only as a result of a weak effective attraction between its quasiparticles—a bare particle plus its associated exchange and correlation cloud—rather than as an elementary consequence of its statistics.

Although physicists understood the properties of ^{3}He much later than those of ^{4}He, we shall begin by considering Landau's Fermi liquid theory that describes the emergent behavior displayed by the quasiparticles found in the normal state of liquid ^{3}He. We shall put off a consideration of their superfluid behavior until after we have discussed Bose liquid theory and its application to liquid ^{4}He, and explained, with the aid of the BCS theory that we will also meet later in this unit, how a net attractive interaction can bring about superconductivity in electronic matter and superfluidity in ^{3}He and other Fermi liquids, such as neutron matter.

Landau Fermi liquid theory

There are three gateways to the protected emergent behavior in the "Landau Fermi liquids" that include liquid ^{3}He and some simple metals: 1) adiabaticity; 2) effective fields to represent the influence of particle interactions; 3) a focus on long-wavelength, low-frequency, and low-temperature behavior. By incorporating these in his theory, Lev Landau was able to determine the compressibility, spin susceptibility, specific heat, and some transport properties of liquid ^{3}He at low temperatures.

Adiabaticity means that one can imagine turning on the interaction between particles gradually, in such a way that one can establish a one-to-one correspondence between the particle states of the noninteracting system and the quasiparticle states of the actual material. The principal effective fields introduced by Landau were scalar internal long-wavelength effective density fields, which determine the compressibility and spin susceptibility and can give rise to zero sound, and a vector effective field describing backflow that produces an increased quasiparticle mass. The focus on low-energy behavior then enabled him to determine the quasiparticle scattering amplitudes that specify its viscosity, thermal conductivity, and spin diffusion.

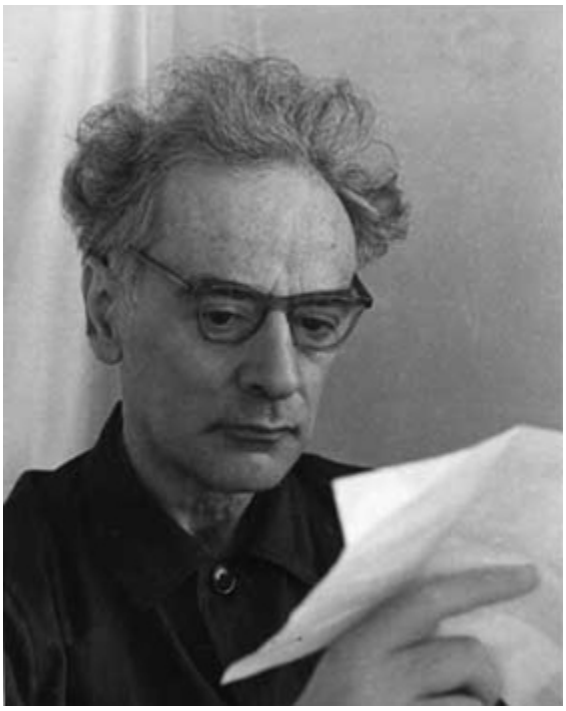


Figure 10: Landau impacted theoretical physics over much of the 20th century.
Source: © AIP Emilio Segrè Visual Archives, Physics Today Collection.

The restoring force for zero sound, a collective mode found in neutral Fermi liquids, is an internal density fluctuation field that is a generalization of that found in the random phase approximation (RPA). Its phenomenological strength is determined by the spin-symmetric spatial average of the effective interactions between parallel spin and antiparallel spin He atoms; the "Fermi liquid" correction to the Pauli spin susceptibility is determined by the corresponding spin antisymmetric average of the interactions between He atoms, i.e., the difference between the spatially averaged effective interactions between atoms of parallel and antiparallel spin. The vector field representing the strength of the effective current induced by a bare particle is just the backflow field familiar from a study of the motion of a sphere in an incompressible fluid. The collisions between quasiparticles produce an inverse quasiparticle lifetime that varies as the square of the temperature, and when modified by suitable geometric factors, give rise to the viscosity and thermal conductivity found experimentally in the normal state of liquid ^{3}He.

Fritz London

Source: © AIP Emilio Segrè Visual Archives, Physics Today Collection.

Fritz London was a seminal figure in the early days of quantum mechanics through his pioneering work on the chemical bond with Walter Heitler and his work with Edmond Bauer on the measurement problem. He was also the insufficiently celebrated heroic contributor to our understanding of superfluidity and superconductivity. As P. W. Anderson has emphasized, in his brilliant essay on London, "He was among the few pioneers who deliberately chose, once atoms and molecules were understood, not to focus his research on further subdividing the atom into its ultimate constituents, but on exploring how quantum theory could work, and be observed, on the macroscopic scale." With his younger brother Heinz, he proposed in 1934 the London equations that provided a phenomenological explanation for superconductivity, and a few years later was the first to recognize that the superfluidity of liquid ^{4}He was an intrinsic property of the Bose condensation of its atoms. In his two books on superfluids and superconductors, he set forth the basic physical picture for their remarkable quantum behavior on a macroscopic scale, but died in 1954 before he could see his ideas brought to fruition through microscopic theory. It is a tribute to both London and John Bardeen that with his share of his 1972 Nobel Prize, Bardeen endowed the Fritz London Memorial Lectures at Duke University, where London had spent the last 15 years of his life.

As we noted above, Landau's theory also works for electrons in comparatively simple metals, for which the adiabatic assumption is applicable. For these materials, Landau's quasiparticle interaction is the sum of the bare electrostatic interaction and a phenomenological interaction; in other words, it contains an add-on to the screening fields familiar to us from the RPA.

It is in the nonsimple metals capable of exhibiting the localized behavior predicted by Nevill Mott that is brought on by very strong electrostatic repulsion or magnetic coupling between their spins that one sees a breakdown of Landau's adiabatic assumption. This is accompanied by fluctuating fields and electronic scattering mechanisms that are much stronger than those considered by Landau. For these "non-Landau" Fermi liquids, the inverse quasiparticle lifetime may not vary as T^2, and the electron-electron interaction contribution to the resistivity will no longer vary as T^2.

We will consider Mott localization and some of the quantum states of matter that contain such non-Landau Fermi liquids in Section 7.

It is straightforward, but no longer exact, to extend Landau's picture of interacting quasiparticles to short-range behavior, and thereby obtain a physical picture of a quasiparticle in liquid ^{3}He. The theory that achieves this turns out to be equally applicable to ^{3}He and ^{4}He and provides insight into the relative importance of quantum statistics and the strong repulsive interaction between ^{3}He atoms.

The superfluid Bose liquid

While Heike Kamerlingh Onnes liquefied He in its natural ^{4}He form and then studied its properties in the 1920s, its superfluidity remained elusive until 1938. Jack Allen of Cambridge and Piotr Kapitsa in Moscow almost simultaneously found that, as the flowing liquid was cooled below 2.18 K, its viscosity suddenly dropped to an almost immeasurably low value. German-American physicist Fritz London quickly understood the gateway to the emergence of this remarkable new state of quantum matter. It was Bose condensation, the condensation of the ^{4}He atoms into a single quantum state that began at 2.18 K.

Superfluidity in liquid ^{4}He and other Bose liquids, such as those produced in the atomic condensates, is a simple consequence of statistics. Nothing prevents the particles from occupying the same momentum state. In fact, they prefer to do this, thereby creating a macroscopically occupied single quantum state, the condensate, that can move without friction at low velocities. On the other hand, the elementary

excitations of the condensate—phonons and rotons in the case of ^{4}He—can and do scatter against each other and against walls or obstacles, such as paddles, inserted in the liquid. In doing so, they resemble a normal fluid.

This is the microscopic basis for the two-fluid model of ^{4}He developed by MIT's Laszlo Tisza in 1940. This posits that liquid He consists of a superfluid and a normal fluid, whose ratio changes as the temperature falls through the transition point of 2.18 K. Tisza, who died in 2009 at the age of 101, showed that the model had a striking consequence; it predicted the existence of a temperature wave, which he called second sound, and which Kapitsa's student, Vasilii Peshkov subsequently found experimentally.

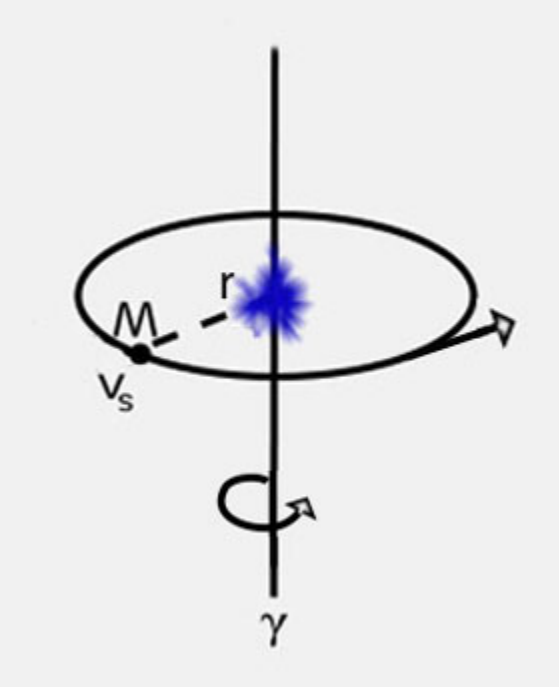


Figure 11: Geometry of a straight vortex line in a superfluid, showing how the superfluid velocity rotates about a normal core (shown in blue) whose size is the superfluid coherence length, ξ.
Source:

The superfluid flow of a condensate can also involve rotation. Norwegian-American scientist Lars Onsager and Richard Feynman independently realized that the rotational flow of the condensate of ^{4}He would be characterized by the presence of quantized vortex lines—singularities around which the liquid is rotating, whose motion describes the rotational flow. See the math

W.F. Vinen was subsequently able to detect a single vortex line, while Feynman showed that their production through friction between the superfluid flow and the pipe walls could be responsible for the existence of a critical velocity for superfluid flow in a pipe.

The introduction of rotons

It was Landau who proposed that the long wavelength elementary excitations in superfluid liquid He would be phonons; he initially expected additional phenomena at long length scales connected with vorticity, which he called rotons. He subsequently realized that to explain the experiment, his rotons must be part of the general density fluctuation spectrum, and would be located at a wave vector of the order of the inverse interatomic distance. Feynman then developed a microscopic theory of the phonon-roton spectrum, which was subsequently measured in detail by inelastic neutron scattering experiments. A key component of his work was the development with his student Michael Cohen of a ground state wavefunction that incorporated backflow, the current induced in the background liquid by the moving atom, thereby obtaining a spectrum closer to the experimental findings.

Figure 12: Lars Onsager, a physical chemist and theoretical physicist who possessed extraordinary mathematical talent and physical insight.
Source: © AIP Emilio Segrè Visual Archives, Segrè Collection.

By introducing response functions and making use of sum rules and simple physical arguments, it is possible to show that the long-wavelength behavior of a Bose liquid is protected, obtain simple quantitative expressions for the elementary excitation spectrum, and, since the superfluid cannot respond to a slowly rotating external probe, obtain an exact expression for the normal fluid density.

An elementary calculation shows that above about 1 K, the dominant excitations in liquid ^{4}He are rotons. Suggestions about their physical nature have ranged from Feynman's poetic tribute to Landau—"a roton

is the ghost of a vanishing vortex ring"—to the more prosaic arguments by Allen Miller, Nozières, and myself that we can best imagine a roton as a quasiparticle—a He atom plus its polarization and backflow cloud. The interaction between rotons can be described through roton liquid theory, a generalization of Fermi liquid theory. K.S. Bedell, A. Zawadowski, and I subsequently made a strong argument in favor of their quasiparticle-like nature. We described their effective interaction in terms of an effective quasiparticle interaction potential modeled after that used to obtain the phonon-roton spectrum. By doing so, we explained a number of remarkable effects associated with two-roton bound state effects found in Raman scattering experiments.

In conclusion we note that the extension of Landau's theory to finite wave vectors enables one to explain in detail the similarities and the differences between the excitation spectra of liquid ^{3}He and liquid ^{4}He in terms of modest changes in the pseudopotentials used to obtain the effective fields responsible for the zero sound spectrum found in both liquids. Thus, like zero sound, the phonon-roton spectrum represents a collisionless sound wave and the finding of well-defined phonons in the normal state of liquid ^{4}He in neutron scattering experiments confirms this perspective.

Section 4: Gateways to a Theory of Superconductivity

Superconductivity—the ability of some metals at very low temperatures to carry electrical current without any appreciable resistance and to screen out external magnetic fields—is in many ways the poster child for the emergence of new states of quantum matter in the laboratory at very low temperatures. Gilles Holst, an assistant in the Leiden laboratory of the premier low-temperature physicist of his time, Kamerlingh Onnes, made the initial discovery of superconductivity in 1911. Although he did not share the Nobel Prize for its discovery with Kamerlingh Onnes, he went on to become the first director of the Phillips Laboratories in Eindhoven. But physicists did not understand the extraordinary properties of superconductors until 1957, when Nobel Laureate John Bardeen, his postdoctoral research associate Leon Cooper, and his graduate student Robert Schrieffer published their historic paper (known as "BCS") describing a microscopic theory of superconductivity.

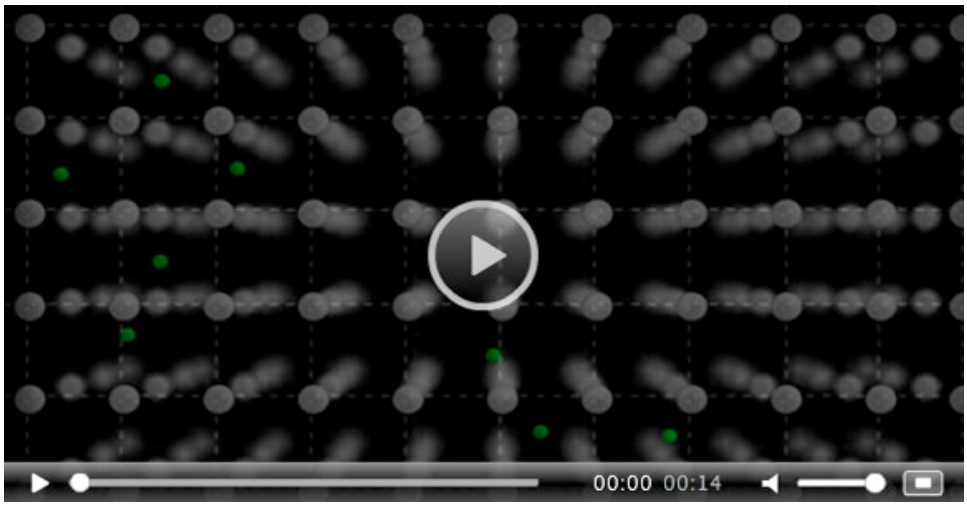


Figure 13: Superconductors carry electrical current without resistance and are almost perfect diamagnets (a more fundamental aspect of their behavior), in that they can screen out external magnetic fields within a short distance known as the "penetration depth."
Source:

We now recognize the two gateways to the emergence of the superconducting state: an effective attractive interaction between electrons (the quasiparticles of Landau's Fermi liquid theory), whose energies put them close to their Fermi surface; and the condensation of pairs of these quasiparticles of opposite spin and momentum into a macroscopically occupied single quantum state, the superfluid condensate.

BCS theory explains the superfluidity of quantum fermionic matter. It applies to conventional superconductors in which phonons, the quantized vibrations of the lattice, serve as the pairing glue that makes possible an attractive quasiparticle interaction and those discovered subsequently, such as superfluid pairing phenomena in atomic nuclei, superfluid ^{3}He, the cosmic superfluids of nuclear matter in the solid outer crust, and liquid interiors of rotating neutron stars. It also applies to the unconventional superconductors such as the cuprate, heavy electron, organic, and iron-based materials that take center stage for current work on superconductivity.

As we shall see, a remarkable feature of BCS theory is that, although it was based on an idealized model for quasiparticle behavior, it could explain all existing experiments and predict the results of many new ones. This occurs because the superconducting state is protected; its emergent behavior is independent of the details. As a result, a quite simple model that incorporates the "right stuff"—the gateways to superconducting behavior we noted above—can lead to a remarkably accurate description of its emergent behavior. In this section, we will trace the steps from 1950 to 1956 that led to the theory. The next section will outline the theory itself. And later in this unit, we will show how a simple extension of the BCS framework from the Standard Model considered in their original paper offers the prospect of explaining the properties of the unconventional superconductors at the center of current research on correlated electron matter.

Four decades of failed theories

In 1950, nearly 40 years after its discovery, the prospects for developing a microscopic theory of superconductivity still looked grim. Failed attempts to solve this outstanding physics challenge by the giants in the field, from Einstein, Bohr, Heisenberg, Bloch, and Landau to the young John Bardeen, led most theorists to look elsewhere for promising problems on which to work.

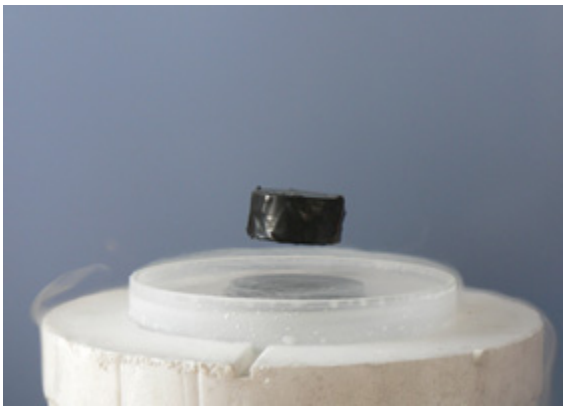


Figure 14: A photograph such as this of a levitating magnet is arguably the iconic image for superconductivity. It provides a vivid demonstration the way in which the near perfect diamagnetism of a superconducting material (the Meissner effect) makes it possible to levitate magnets above it.
Source: © Wikimedia Commons, GNU Free Documentation License, Version 1.2. Author: Mai-Linh Doan, 13 October 2007.

Despite that, experimentalists had made considerable progress on the properties of the superconducting state. They realized that a strong enough applied external magnetic field could destroy superconductivity and that a superconductor's almost perfect diamagnetism—its ability to shield out an external magnetic field within a short distance, known as the "penetration depth"—was key to an explanation. Theorists found that a two-fluid model analogous to that considered for superfluid ^{4}He in the previous section could connect many experimental results. Moreover, London had argued eloquently that the perfect diamagnetism could be explained by the rigidity of the superfluid wavefunction in the presence of an external magnetic field, while progress occurred on the "phase transition" front as Vitaly Ginzburg and Landau showed how to extend Landau's general theory of second-order phase transitions to superconductors to achieve an improved phenomenological understanding of emergent superconducting behavior.

Superconductivity was obviously an amazing emergent electronic phenomenon, in which the transition to the superconducting state must involve a fundamental change in the ground and excited states of electron matter. But efforts to understand how an electron interaction could bring this about had come to a standstill. A key reason was that the otherwise successful nearly free electron model offered no clues to how an electron interaction that seemed barely able to affect normal state properties could turn some metals into superconductors.

A promising new path

The Double Laureate and His Colleagues

Photo of John Bardeen.
Source: © Department of Physics, University of Illinois at Urbana-Champaign, courtesy AIP Emilio Segrè Visual Archives.

In 1951, after professional differences had undermined the relationship of John Bardeen and Walter Brattain with William Shockley, their team leader in the invention of the transistor at Bell Telephone Laboratories, Bardeen took a new job, as professor of electrical engineering and of physics in the University of Illinois at Urbana-Champaign. There, he was able to pursue freely his interest in superconductivity, setting out various lines of research to take on the immense challenge of understanding its nature. By 1957, when Bardeen, his postdoctoral research associate Leon Cooper, and his graduate assistant Robert Schrieffer developed what came to be known as the BCS theory, Bardeen had received the 1956 Nobel Prize in Physics for discovering the transistor. Bardeen knew that BCS was also worthy of the prize. But since no one had received a second Nobel Prize in the same subject, he reportedly worried that because of his earlier Nobel Prize, and his role in BCS, his two colleagues would not be eligible. Fortunately, the Nobel Committee eventually saw no reason to deny the prize to BCS: It awarded the three men the 1972 physics prize, in the process making Bardeen the first individual to become a double laureate in the same field.

Matters began to change in 1950 with the discovery of the isotope effect on the superconducting transition temperature, T_c, by Bernard Serin at Rutgers University, Emanuel Maxwell at the National Bureau of Standards, and their colleagues. They found that T_c for lead varied inversely as the square root of its isotopic mass. That indicated that quantized lattice vibrations must be playing a role in bringing about that transition, for their average energy was the only physical quantity displaying such a variation.

That discovery gave theorists a new route to follow. Herbert Fröhlich and Bardeen independently proposed theories in which superconductivity would arise through a change in the self-energy of individual electrons produced by the co-moving cloud of phonons that modify their mass. But, it soon became clear that efforts along these lines would not yield a satisfactory theory.

Frohlich then suggested in 1952 that perhaps the phonons played a role through their influence on the effective electron interaction. The problem with his proposal was that it was difficult to see how such an apparently weak phonon-induced interaction could play a more important role than the much stronger repulsive electrostatic interaction he had neglected. Two years later, Bardeen and his first postdoctoral research associate at the University of Illinois at Urbana-Champaign—myself—resolved that problem.

We did so by generalizing the collective coordinate formalism that David Bohm and I had developed for electron-electron interactions alone to derive their effective interaction between electrons when both the effects of the electrostatic interaction and the electron-phonon coupling are taken into account (Figure 15). Surprisingly, we found that, within the random phase approximation, the phonon-induced interaction could turn the net interaction between electrons lying within a characteristic phonon frequency of the Fermi surface from a screened repulsive interaction to an attractive one. We can imagine the phonon-induced interaction as the electronic equivalent of two children playing on a waterbed. One (the polarizer) makes a dent (a density wave) in the bed; this attracts the second child (the analyzer), so that the two wind up closer together.

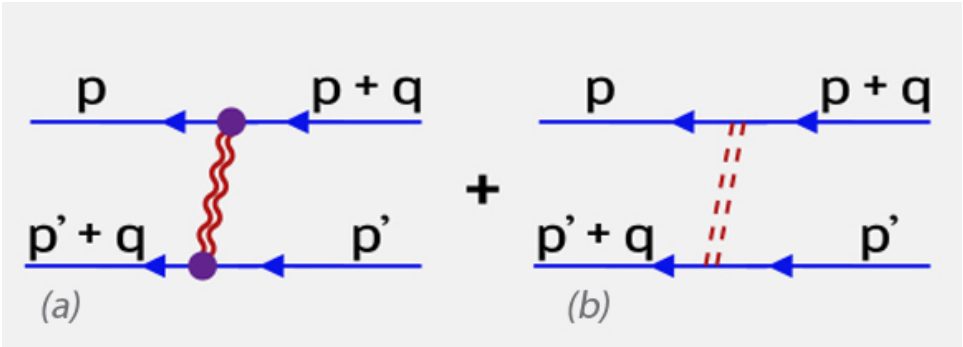


Figure 15: The net effective interaction between electrons in a metal, with (a) depicting the net phonon-induced interaction and (b) their intrinsic effective screened electrostatic interaction.
Source: © David Pines.

Leon Cooper, who replaced me in Bardeen's research group in 1955, then studied the behavior of two electrons of opposite spin and momentum near the Fermi surface using a simplified version of the Bardeen-Pines attractive interaction. In a 1956 calculation that allowed for the multiple scattering of the

pair above the Fermi surface, he showed that net attraction produced an energy gap in the form of a bound state for the electron pair.

During this period, in work completed just before he went to Stockholm to accept his 1956 Nobel Prize, Bardeen had showed that many of the key experiments on superconductivity could be explained if the "normal" elementary excitations of the two-fluid model were separated from the ground state by an energy gap. So the gap that Cooper found was intriguing. But there was no obvious way to go from a single bound pair to London's coherent ground state wavefunction that would be rigid against magnetic fields. The field awaited a breakthrough.

Section 5: The BCS Theory

The breakthrough came in January 1957, when Bardeen's graduate student, Robert Schrieffer, while riding a New York City subway train following a conference in Hoboken, NJ on The Many-Body Problem, wrote down a candidate wavefunction for the ground state and began to calculate its low-lying excited states. He based his wavefunction on the idea that the superconducting condensate consists of pairs of quasiparticles of opposite spin and momenta. This gateway to emergent superconducting behavior is a quite remarkable coherent state of matter; because the pairs in the condensate are not physically located close to one another, their condensation is not the Bose condensation of pairs that preform above the superconducting transition temperature in the normal state. Instead, the pairs condense only below the superconducting transition temperature. The typical distance between them, called the "coherence length," is some hundreds of times larger than the typical spacing between particles.

To visualize this condensate and its motion, imagine a dance floor in which one part is filled with couples (the pairs of opposite spin and momentum) who, before the music starts (that is, in the absence of an external electric field), are physically far apart (Figure 16). Instead of being distributed at random, each member of the couple is connected, as if by an invisible string, to his or her partner faraway. When the music begins (an electric field is applied), each couple responds by gliding effortlessly across the dance floor, moving coherently with the same velocity and never colliding with one another: the superfluid motion without resistance.

Figure 16: Like the condensate, these coupled dancers came together when the music started and continued in a fluid motion next to each other without bumping into each other or stepping on each other's toes.
Source: © Bruce Douglas.

Sorting out the details of the theory

Back in Urbana, Schrieffer, Bardeen, and Cooper quickly worked out the details of the microscopic theory that became known as BCS. A key feature was the character of the elementary excitations that comprise the normal fluid. We can describe these as quasiparticles. But in creating them, we have to break the pair bonds in the condensate. This requires a finite amount of energy—the energy gap. Moreover, each BCS quasiparticle carries with it a memory of its origin in the pair condensate; it is a mixture of a Landau quasiparticle and a Landau quasihole (an absence of a quasiparticle) of opposite spin.

Inspiration for the Critical Breakthrough

Schrieffer found the inspiration for his wavefunction in earlier work done by T. D. Lee, Francis Low, and myself on a quite different problem: that of understanding the behavior of a polaron—an electron moving in a polar crystal that is strongly coupled to its lattice vibrations. When I arrived in Urbana in 1952, John Bardeen suggested that insight into the next step on a microscopic, phonon-based theory of superconductivity might come from a solution of the polaron problem that went beyond perturbation theory. Lee and I found such a solution by adapting an approach developed by Sin-Itro Tomonaga for mesons coupled to nuclei. With the help of our Urbana colleague Francis Low, we then wrote the wavefunction for the solution in an especially simple way. The LLP wavefunction describes a coherent state in which the phonons in the cloud of strongly coupled phonons around the electron are emitted successively into the same momentum state; it took the form:

$$\Psi_{LLP} \sim \Pi_k \exp\left[\Sigma_k \left[f(k) \left[a_k^* + a_{-k} \right] \right] \Psi_0 \right]$$

where Ψ_0 was the ground state wavefunction, the operators a_k^* and a_k act to create or destroy phonons, and f(k) describes the phonon state. Schrieffer's brilliant insight was to try a ground state wavefunction for the superconductor in which the LLP phonon field was replaced by the pair field of the condensate, $b_k^* = c_{k\uparrow}^* c_{-k\downarrow}^*$, where the c are creation operators for a single electron (quasiparticle); his wavefunction thus took the form:*

$$\Psi \sim \Pi_k \exp\left[\Sigma_k b_k^* f(k)\right] \Psi_0$$

which reduces to the BCS wavefunction,

$$\Psi_{bcs} \sim \Pi_k \left[1 + \Sigma_k b_k^* f(k)\right] \Psi_0$$

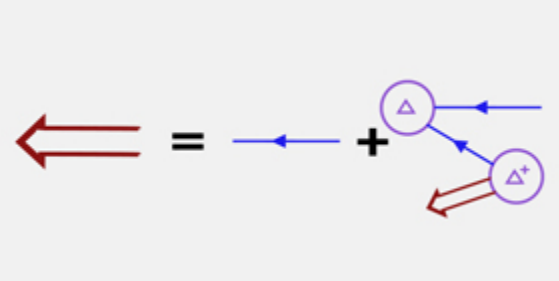


Figure 17: An illustration of the physical process by which a BCS quasiparticle becomes a mixture of a normal state quasiparticle and a quasihole and in so doing acquires an energy gap.
Source: © David Pines.

Figure 17 illustrates the concept. Because the key terms in the effective interaction of the BCS quasiparticle with its neighbors are those that couple it to the condensate, a quasiparticle that scatters against the condensate emerges as a quasihole of opposite spin. The admixture of Landau quasiholes with quasiparticles in a BCS quasiparticle gives rise to interference effects that lead the superconductor to respond differently to probes that measure its density response and probes that measure its spin response. Indeed, the fact that BCS could explain the major difference between the results of acoustic attenuation measurements that probe the density and nuclear spin relaxation measurements of the spin response of the superconducting state in this way provided definitive proof of the correctness of the theory.

The basic algebra that leads to the BCS results is easily worked out. It shows that the energy spectrum of the quasiparticles in the superconducting state takes an especially simple form:

$$E_p = [\varepsilon_p^2 + \Delta^2]^{1/2}$$

Here, ε_p is the normal state quasiparticle energy and Δ the temperature-dependent superconducting energy gap, which also serves as the order parameter that characterizes the presence of a superconducting condensate. When it vanishes at T_c, the quasiparticle and the metal revert to their normal state behavior.

The impact of BCS theory

The rapid acceptance by the experimental low-temperature community of the correctness of the BCS theory is perhaps best epitomized by a remark by David Shoenberg at the opening of a 1959 international conference on superconductivity in Cambridge: "Let us now see to what extent the experiments fit the

theoretical facts." Acceptance by the theoretical community came less rapidly. Some of those who had failed to devise a theory were particularly reluctant to recognize that BCS had solved the problem. (Their number did not include Feynman, who famously recognized at once that BCS had solved the problem to which he had just devoted two years of sustained effort, and reacted by throwing into the nearest wastebasket the journal containing their epochal result.) The objections of the BCS deniers initially centered on the somewhat arcane issue of gauge invariance. With the rapid resolution of that issue, the objections became more diffuse. Some critics persisted until they died, a situation not unlike the reaction of the physics community to Planck's discovery of the quantum.

Searching for Superfluid 3He

Not long after BCS published their epochal paper, both the theoretical and experimental low-temperature community recognized that ^{3}He would likely become a superfluid at some low temperature. But how low, and what form would that superfluidity take? Early on the community realized that because the spatial average of the effective interaction between the ^{3}He atoms measured by the dimensionless spin-symmetric Landau parameter, f_0^s, is positive (the strong short-range repulsion wins out over the weak long-range attraction), it was likely that the superfluid pairing would not be in the simple $^1S_0$ state found for conventional metallic superconductors. However, all attempts to predict the pairing state, much less the temperature at which superfluidity would be found, failed; while experimentalists searched for evidence for superfluid behavior at increasingly lower temperatures that were in the millikelvin range. Along the way, there were false sightings, notably a report by Peshkov at a low-temperature meeting in 1964 that was sharply and correctly criticized by John Wheatley at that same meeting. The discovery came in 1972 during an experimental study by Doug Osheroff, then a graduate student, David Lee, and Bob Richardson at Cornell University of the changes of the pressure as the volume of a sample of liquid ^{3}He that had been cooled to 2×10^{-3} K was slowly increased and then reduced. Tiny glitches that appeared in their results were at first attributed to solidification, but subsequent work when combined with a key interpretation of their results by Tony Leggett of Sussex University, who was visiting Cornell at the time, showed they had observed the onset of superfluid behavior, and that three different anisotropic superfluid phases could be identified. The coupling of the ^{3}He quasiparticles to the nearly antiferromagnetic spin fluctuations of the background liquid plays a key role in determining the anisotropic pairing states, which possess the common feature of being in a state in which the pairs have parallel spin and a p-wave relative orbital angular momentum, $l = 1$. Because the Nobel committee was reluctant to break its rule of awarding the prize to no more than three individuals, their work was recognized by separate Nobel Prizes, the first to Lee, Osheroff, and Richardson in 1996, and the second to Leggett in 2003.

For most physicists, however, the impact of BCS was rapid and immense. It led to the 1957 proposal of nuclear superfluidity, a 15-year search for superfluid ^{3}He, and to the exploration of the role played by pair condensation in particle physics, including the concept of the Higgs boson as a collective mode of a

quark-gluon condensate by Philip W. Anderson and Peter Higgs. It led as well to the suggestion of cosmic hadron superfluidity, subsequently observed in the behavior of pulsars following a sudden jump in their rotational frequency, as we will discuss in Section 8.

In addition, BCS gave rise to the discovery of emergent behavior associated with condensate motion. That began with the proposal by a young Cambridge graduate student, Brian Josephson, of the existence of currents associated with the quantum mechanical tunneling of the condensate wavefunction through thin films, called "tunnel junctions," that separate two superconductors. In retrospect, Josephson's 1962 idea was a natural one to explore. If particles could tunnel through a thin insulating barrier separating two normal metals, why couldn't the condensate do the same thing when one had a tunnel junction made up of two superconductors separated by a thin insulating barrier? The answer soon came that it could, and such superconducting-insulating-superconductor junctions are now known as "Josephson junctions." The jump from a fundamental discovery to application also came rapidly. Superconducting quantum interference devices (SQUIDS) (Figure 18), now use such tunneling to detect minute electromagnetic fields, including an application in magnetoencephalography—using SQUIDS to detect the minute magnetic fields produced by neurocurrents in the human brain.

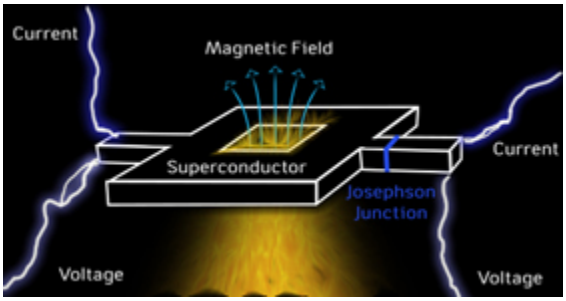


Figure 18: A Superconducting Quantum Interference Device (SQUID) is the most sensitive type of detector of magnetic fields known to science.
Source:

Still another fascinating property of a superconductor was discovered by a young Soviet theorist, Alexander Andreev, who realized that when an electron is reflected from the surface of a superconductor, because its wavefunction samples the condensate, it can break one of the pairs in the superconducting condensate, and so emerge as a hole. Measurements by point contact spectroscopy show this dramatic effect, known now as "Andreev reflection," which is the condensed matter equivalent of changing a particle into an antiparticle through a simple physical process.

Looking back at the steps that led to BCS as the Standard Model for what we now describe as conventional superconductors, a pattern emerges. Bardeen, who was key to the development of the theory at every stage from 1950 to 1957, consistently followed what we would now describe as the appropriate emergent strategy for dealing with any major unsolved problem in science:

- Focus first on the experimental results via reading and personal contact.
- Explore alternative physical pictures and mathematical descriptions without becoming wedded to any particular one.
- Thermodynamic and other macroscopic arguments have precedence over microscopic calculations.
- Aim for physical understanding, not mathematical elegance, and use the simplest possible mathematical description of system behavior.
- Keep up with new developments in theoretical techniques—for one of these may prove useful.
- Decide at a qualitative level on candidate organizing concepts that might be responsible for the most important aspect of the measured emergent behavior.
- Only then put on a "reductionist" hat, proposing and solving models that embody the candidate organizing principles.

Section 6: New Superconductors

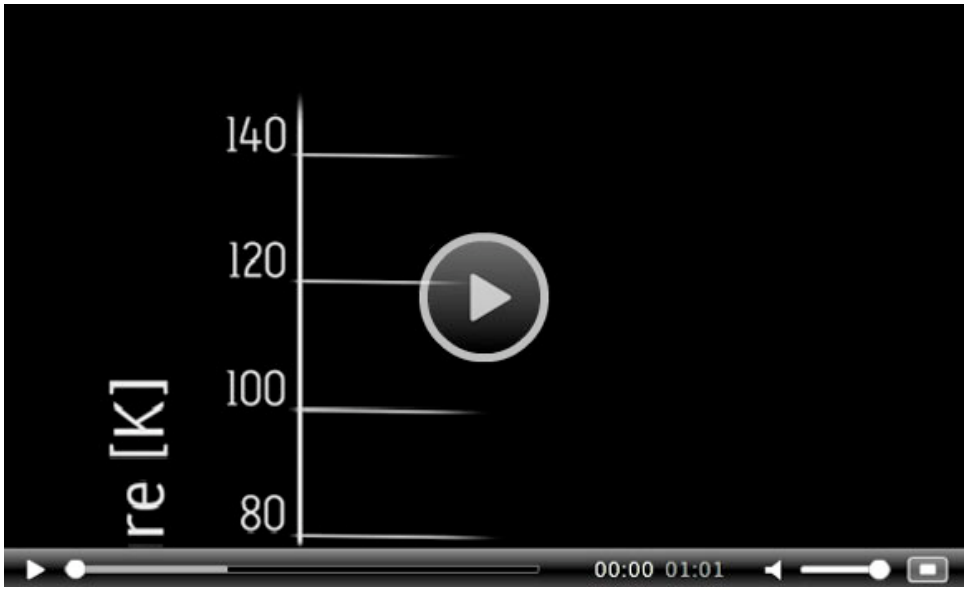


Figure 19: The first superconducting material was discovered in 1911 when mercury was cooled to 4 Kelvin (K). Seventy-five years later, thanks to the discovery of superconductivity in the family of cuprate materials by Bednorz and Mueller, scientists made a giant leap forward as they discovered many related materials that superconduct at temperatures well above 90 K.
Source:

Physics with the Whole World Watching

The discovery of materials exhibiting superconducting behavior above 23 K, the highest known transition temperature for traditional superconductors, created a huge, if delayed, response around the world. The reaction exploded at an event that became known as the "Woodstock of physics."

J. Georg Bednorz and K. Alex Müller of the IBM Zurich Research Laboratory had reported their discovery of a ceramic material with a superconducting transition temperature of 30 K in a German physics journal in April 1986. The news drew little immediate response. But excitement rose as other groups confirmed the find and discovered new high-T_c superconductors (including one with a T_c of 93 K reported by Paul Chu's team at the University of Houston). By 18 March 1987 the topic had gained so much traction that the American Physical Society (APS) added a last-minute session on it at its annual meeting in New York City. When the session started at 7:30 p.m., about 2,000 people filled the hall. Others watched the event on video monitors. And although organizers limited the 51 speakers' time on the podium, the meeting continued until 3:15 the next morning.

Characteristically, New York City embraced the event. That week, an APS badge guaranteed its wearer entry to several nightclubs without the need to pay a cover charge.

The past three decades have seen an outpouring of serendipitous discoveries of new quantum phases of matter. The most spectacular was the 1986 discovery by IBM scientists J. Georg Bednorz and K. Alex Müller of superconductivity at high temperatures in an obscure corner of the periodic table: a family of ceramic materials of which $La_xSr_{1-x}CuO_4$ (containing the elements lanthanum, strontium, copper, and oxygen) was a first example. By the American Physical Society meeting in March 1987 (often referred to as the "Woodstock of physics"), it was becoming clear that this was just the first of a large new family of cuprate superconductors that possess two factors in common. They have planes of cupric oxide (CuO_2) that can be doped with mobile electrons or holes. And the quasiparticles in the planes exhibit truly unusual behavior in their normal states while their superconducting behavior differs dramatically from that of the conventional superconductors in which phonons supply the pairing glue.

Over the past two decades, thanks to over 100,000 papers devoted to their study, we have begun to understand why the cuprate superconductors are so different. Moreover, it is now clear that they represent but one of an extended family of unconventional superconductors with three siblings: the heavy electron superconductors discovered in 1979; the organic superconducting materials discovered

in 1981; and the iron-based superconductors discovered in 2006. Although there is a considerable range in their maximum values of T_c—about 160 K for a member of the cuprate family, $HgBa_2Ca_2Cu_3O_x$, under pressure; roughly 56 K in the iron pnictides ($LnFeAsO_{1-x}$); and 18.5 K for $PuGaIn_5$, a member of the 115 ($RMIn_5$) family of heavy electron materials—they show remarkable similarities in both their transport and magnetic properties in the normal and superconducting states.

In particular, for all four siblings:

- Superconductivity usually occurs on the border of antiferromagnetic order at which the magnetic moments of atoms align.
- The behavior of the quasiparticles, density fluctuations, and spin fluctuations in their normal state is anomalous, in that it is quite different from that of the Landau Fermi liquids found in the normal state of liquid ^{3}He and conventional superconductors.
- The preferred superconducting pairing state is a singlet state formed by the condensation of pairs of quasiparticles of opposite spin in an orbital angular momentum, l, state, with l = 2; as a result, the superconducting order parameter and energy gap vary in configuration and momentum space.

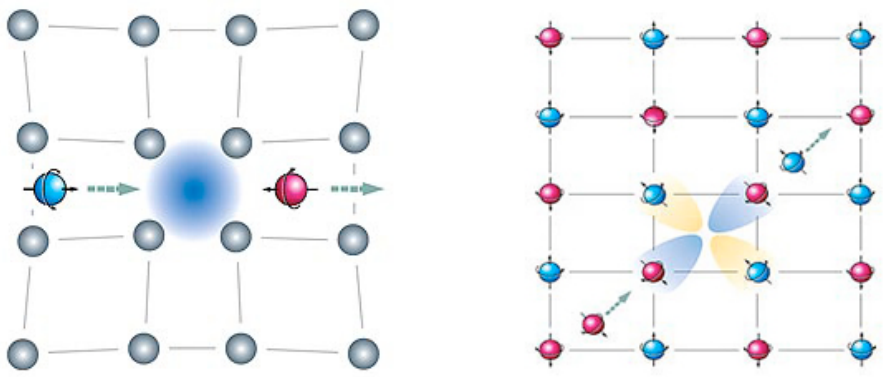


Figure 20: Left: Conventional superconductors and right: heavy-electron superconductors.
Source: © Los Alamos National Laboratory.

In this section, we can explore only a small corner of this marvelous variety of materials whose unexpected emergent properties continue to surprise and offer considerable promise for commercial application. To understand these, we must go beyond the standard model in which phonons provide the glue that leads to attraction. We explore the very real possibility that the net effective attraction between quasiparticles responsible for their superconductivity occurs without phonons and is of purely magnetic origin. In so doing, we enter territory that is still being explored, and in which consensus does not always exist on the gateways to the emergent behavior we find there.

Heavy electron materials

We begin with the heavy electron materials for three reasons. First, and importantly, they can easily be made in remarkably pure form, so that in assessing an experiment on them, one is not bedeviled by "dirt" that can make it difficult to obtain reliable results from sample to sample. Second, the candidate organizing concepts introduced to explain their behavior provide valuable insight into the unexpected emergent behavior seen in the cuprates and other families of unconventional superconductors. Third, these materials display fascinating behavior in their own right.

Heavy electron materials contain a lattice, called the "Kondo lattice," of localized outer f-electrons of cerium or uranium atoms that act like local magnetic moments magnetically coupled through their spins to a background sea of conduction electrons. It is called a Kondo lattice because in isolation each individual magnetic moment would give rise to a dramatic effect, first identified by Jun Kondo, in which the magnetic coupling of the conduction electrons to the moment acts to screen out the magnetic field produced by it, while changing the character of their resistivity, specific heat, and spin susceptibility. When one has a lattice of such magnetic moments, the results are even more dramatic; at low temperatures, their coupling to the conduction electrons produces an exotic new form of quantum matter in which some of the background conduction electrons form a heavy electron Fermi liquid for which specific heat measurements show that the average effective mass can be as large as that of a muon, some 200 or more bare electron masses. The gateway to the emergence of this new state of matter, the heavy electron non-Landau Fermi liquid, is the *collective* entanglement (hybridization) of the local moments with the conduction electrons.

Remarkably, the growth of the heavy electron Fermi liquid, which we may call a "Kondo liquid (KL)" to reflect its origin in the collective Kondo lattice, displays scaling behavior, in that its emergent coherent behavior can be characterized by the temperature, T*, at which collective hybridization begins.

Below T*, the specific heat and spin susceptibility of the emergent Kondo liquid display a logarithmic dependence on temperature that reflects their coherent behavior and collective origin. Its emergent behavior can be described by a two-fluid model that has itself emerged only recently as a candidate standard phenomenological model for understanding Kondo lattice materials. In it, the strength of the emergent KL is measured by an order parameter, f (T/T*), while the loss in strength of the second component, the local moments (LMs) that are collectively creating the KL, is measured by 1-f(T/T*). KL scaling behavior, in which the scale for the temperature dependence of all physical quantities is set by T*,

persists down to a temperature, T_0 close to those at which the KL or LM components begin to become ordered.

Phase diagram of a work in progress

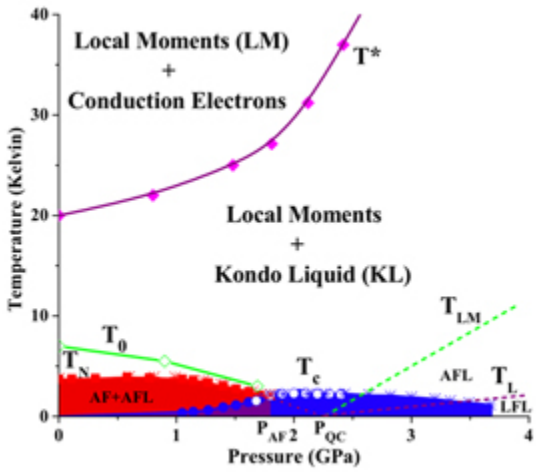


Figure 21: A candidate phase diagram for $CeRhIn_5$ depicting the changes in its emergent behavior and ordering temperatures as a function of pressure.
Source: © David Pines.

The accompanying candidate phase diagram (Figure 21) for the heavy electron material $CeRhIn_5$ (consisting of cerium, rhodium, and indium) gives a sense of the richness and complexity of the emergent phenomena encountered as a result of the magnetic couplings within and between the coexisting KL and LM components as the temperature and pressure are varied. It provides a snapshot of work in progress on these fascinating materials—work that will hopefully soon include developing a microscopic theory of emergent KL behavior.

Above T*, the local moments are found to be very weakly coupled to the conduction electrons. Below T*, as a result of an increasing collective entanglement of the local moments with the conduction electrons, a KL emerges from the latter that exhibits scaling behavior between T* and T_0; as it grows, the LM component loses strength. T_0 is the temperature below which the approach of antiferromagnetic or superconducting order influences the collective hybridization process and ends its scaling behavior.

Electronic order in metals

Electrons in a metal can become ordered through the magnetic coupling of their spins or the electrostatic interaction of their charges. The magnetic order can be ferromagnetic, as in iron, corresponding to a lattice of localized electron spins all of which point in the same direction, or antiferromagnetic, corresponding to a lattice of localized electron spins in which nearest neighbor spins point in opposite directions. The charge order can be localized, in which case electrons are no longer free to move throughout the metal, or coherent, in which case the electrons become superconducting. Interestingly, since a magnetic interaction between electron spins can bring about both superconductivity (as we discuss below) and antiferromagnetic order, one finds in some materials a competition between these two forms of order, a competition that is sometimes resolved, as is the case for $CeRhIn_5$, by both forms of competing order coexisting in a given material.

At T_N, the residual local moments begin to order antiferromagnetically, as do some, but not all, of the KL quasiparticles. The remaining KL quasiparticles become superconducting in a so-called d_{x2-y2} pairing state; as this state grows, the scale of antiferromagnetic order wanes, suggesting that superconductivity and antiferromagnetic order are competing to determine the low-temperature fate of the KL quasiparticles.

When the pressure, P, is greater than P_{AF}, superconductivity wins the competition, making long-range antiferromagnetic LM order impossible. The dotted line continuing TN toward zero for P greater than P_{AF} indicates that LM ordering is still possible if superconductivity is suppressed by application of a large enough external magnetic field. Experimentalists have not yet determined what the Kondo liquid is doing in this regime; one possibility is that it becomes a Landau Fermi liquid.

Starting from the high pressure side, P_{QC} denotes the point in the pressure phase diagram at which local moments reappear and a localized (AF) state of the quasiparticles first becomes possible. It is called a "quantum critical point" because in the absence of superconductivity, one would have a T = 0 quantum phase transition in the Kondo liquid from itinerant quasiparticle behavior to localized AF behavior.

Since spatial order is the enemy of superconductivity, it should not be surprising to find that in the vicinity of P_{QC}, the superconducting transition temperature reaches a maximum—a situation we will see replicated in the cuprates and one likely at work in all the unconventional superconducting materials. One explanation is that the disappearance of local moments at P_{QC} is accompanied by a jump in the

size of the conduction electron Fermi surface; conversely, as the pressure is reduced below P_{QC}, a smaller Fermi surface means fewer electrons are capable of becoming superconducting, and both the superconducting transition temperature and the condensation energy, the overall gain in energy from becoming superconducting, are reduced.

In the vicinity of a quantum critical point, one expects to find fluctuations that can influence the behavior of quasiparticles for a considerable range of temperatures and pressures. Such quantum critical (QC) behavior provides yet another gateway for emergent behavior in this and other heavy electron materials. It reveals itself in transport measurements. For example, in $CeRhIn_5$ at high pressures, one gets characteristic Landau Fermi liquid behavior (a resistivity varying as T^2) at very low temperatures; but as the temperature increases, one finds a new state of matter, quantum critical matter, in which the resistivity in the normal state displays anomalous behavior brought about by the scattering of KL quasiparticles against the QC fluctuations.

What else happens when the pressure is less than P_{QC}? We do not yet know whether, once superconductivity is suppressed, those KL quasiparticles that do not order antiferromagnetically exhibit Landau Fermi liquid behavior at low temperatures. But given their behavior for P less than P_{cr}, that seems a promising possibility. And their anomalous transport properties above T_N and T_c suggest that as the temperature is lowered below T_0, the heavy electron quasiparticles exhibit the anomalous transport behavior expected for quantum critical matter.

Superconductivity without phonons

We turn now to a promising candidate gateway for the unconventional superconducting behavior seen in this and other heavy electron materials—an enhanced magnetic interaction between quasiparticles brought about by their proximity to an antiferromagnetically ordered state. In so doing, we will continue to use BCS theory to describe the onset of superconductivity and the properties of the superconducting state. However, we will consider its generalization to superconducting states in which pairs of quasiparticles condense into states of higher relative angular momentum described by order parameters that vary in both configuration and momentum space.

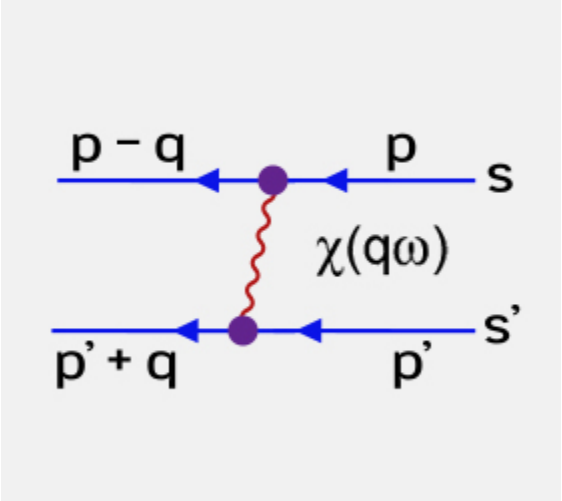


Figure 22: The magnetic quasiparticle interaction between spins s and s' induced by their coupling to the spin fluctuations, $\chi q\omega$, of the magnetic background material.
Source: © David Pines.

To see how the magnetic gateway operates, we turn to Figure 22 which illustrates how the magnetic interaction between two quasiparticles, whose spins are s and s', can be modified by their coupling to the spin fluctuations characteristic of the background magnetic behavior of the material in which they are located. Quite generally, spin, s, located, say, at the origin, acts to polarize the material by inducing a spin fluctuation; this induced magnetization, in turn, couples to the second spin, s', located a distance r away, producing an induced effective magnetic interaction which is analogous to the phonon-induced interaction responsible for superconductivity in ordinary BCS superconductors.

This induced effective magnetic interaction is highly sensitive to the magnetic properties of the background material. For an ordinary paramagnet exhibiting weakly magnetic behavior, the resulting magnetic interaction is quite weak and unremarkable. If, however, the background material is close to being antiferromagnetic, the spectrum of the spin fluctuations that provide the glue connecting the two spins becomes highly momentum dependent, exhibiting a significant peak for wave vectors that are close to those for which one finds a peak in the wave vector dependent magnetic susceptibility of the almost magnetically ordered material. As a result, the induced magnetic quasiparticle interaction will be strong and spatially varying.

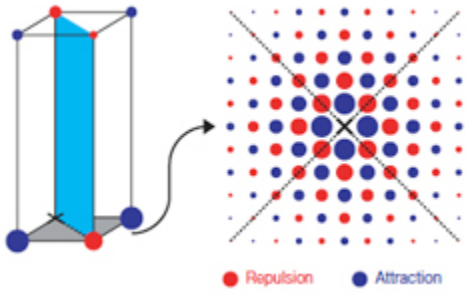


Figure 23: Magnetic interaction potential in a lattice.
Source: © Reprinted by permission from Macmillan Publishers Ltd: Nature 450, 1177–1183 (18 December 2007).

Consider, for example, a magnetic material that is at a pressure near the point at which the material exhibits simple two-dimensional planar commensurate AF order (in which the nearest neighbor spins point in opposite directions). Its momentum dependent susceptibility will then have a peak at the commensurate wave vector Q =[π/a, π/a] where a is the lattice spacing, as will its spin fluctuation spectrum. The corresponding induced magnetic quasiparticle interaction in configuration space will then be repulsive at the origin, attractive at its nearest neighbor sites, repulsive at next nearest neighbor sites, etc., as shown in Figure 23. See the math

Such an interaction, with its mixture of repulsion and attraction, does not give rise to the net attraction required for superconductivity in the conventional BCS singlet s-wave pairing state with an order parameter and energy gap that do not vary in space. The interaction can, however, be remarkably effective in bringing about superconductivity in a pairing state that varies in momentum and configuration space in such a way as to take maximum advantage of the attraction while possessing nodes (zeros) that minimize the repulsion. A dx2-y2 pairing state, the singlet d-wave pairing state characterized by an order parameter and energy gap Δ_{x2-y2} (k)= Δ [cos (k_xa)-cos (k_ya)], does just that, since it has nodes (zeroes) where the interaction is repulsive (at the origin or along the diagonals, for example) and is maximal where the interaction is maximally attractive (e.g., at the four nearest neighbor sites) as may also be seen in Figure 23.

We call such superconductors "gapless" because of the presence of these nodes in the gap function. Because it costs very little energy to excite quasiparticles whose position on the Fermi surface puts them at or near a node, it is the nodal quasiparticle excitations which play the lead role in determining the normal fluid density. Their presence is easily detected in experiments that measure it, such as the low-temperature specific heat and the temperature dependence of the London penetration depth. Nodal

quasiparticle excitations are also easily detected in NMR measurements of the uniform susceptibility and spin-lattice relaxation rate, and the latter measurements have verified that the pairing state found in the "high T_c" heavy electron family of $CeMIn_5$ materials, of which $CeRhIn_5$ is a member, is indeed d_{x2-y2}, the state expected from their proximity to antiferromagnetic order.

To summarize: In heavy electron materials, the coexistence of local moments and itinerant quasiparticles and their mutual interactions can lead to at least four distinct emergent states of matter: the Kondo heavy electron liquid, quantum critical matter, antiferromagnetic local moment order, and itinerant quasiparticle d_{x2-y2}, superconductivity, while the maximum superconducting transition temperature is found close to the pressure at which one finds a QCP that reflects the onset of local moment behavior. In the next section, we will consider the extent to which comparable emergent behavior is observed in the cuprate superconductors.

Section 7: Emergent Behavior in the Cuprate Superconductors

Nevill Mott and his exploits

Nevill Mott at University of Bristol; International Conference on the Physics of Metals, organized by N. F. Mott and A. M. Tyndall.
Source: © Archives of HH Wills Physics Laboratory, University of Bristol, courtesy AIP Emilio Segrè Visual Archives.

Nevill Mott was a world leader in atomic and solid-state physics who combined a keen interest in experiment with a gift for insight and exposition during a career in theoretical physics that spanned over 60 years.

We can best appreciate the remarkable properties of the cuprate superconductors by considering a candidate phase diagram (Figure 24) that has emerged following almost 25 years of experimental and theoretical study described in well over 100,000 papers. In it, we see how the introduction of holes in the CuO planes through chemical substitution in materials such corresponding to $La_{1-x}Sr_xCuO_4$ or $YBa_2Cu_3O_{6+x}$, the low-temperature phase is one of antiferromagnetic order. The gateway to this emergent behavior is the very strong electrostatic repulsion between the planar quasiparticles. This causes the planar Cu d electron spins to localize (a process called "Mott localization" in honor of its inventor, Nevill Mott rather than be itinerant, while an effective antiferromagnetic coupling between these spins causes them to order antiferromagnetically. The magnetic behavior of these localized spins is remarkably well described by a simple model of their nearly two-dimensional behavior, called the "two-dimensional Heisenberg model;" it assumes that the only interaction of importance is a nearest neighbor coupling between spins of strength J.

The impact of adding holes

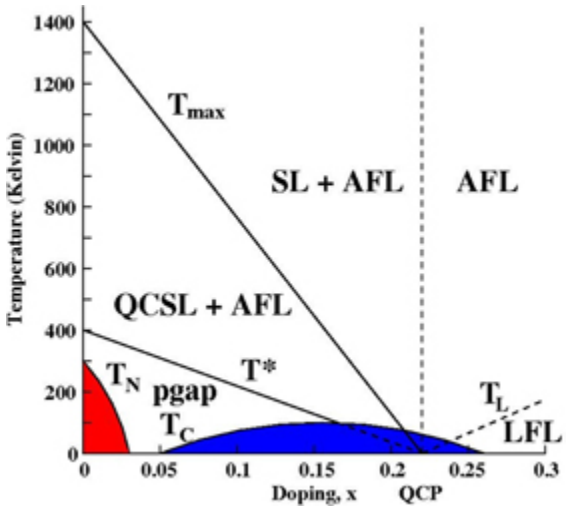


Figure 24: A candidate phase diagram based, in part, on magnetic measurements of normal state behavior, for the cuprate superconductors, in which are shown the changes in its emergent behavior and ordering temperatures as a function of the concentration of holes in the CuO planes.
Source: © David Pines.

When one adds holes to the plane, their presence has a number of interesting consequences for the localized Cu spins. Those, in turn, can markedly influence the behavior of the holes that coexist with them. The accompanying phase diagram (Figure 24) indicates some of the effects. Among them:

- Holes interfere with the long-range antiferromagnetic order of the localized spins, initially reducing its onset temperature, T_N, and then eliminating it altogether for hole doping levels $x > 0.03$.
- At higher hole doping levels, $0.03 < x < 0.22$, the local spins no longer exhibit long-range order. Instead they form a spin liquid (SL) that exhibits short-range spin order and scaling behavior controlled by their doping-dependent interaction. The measured scaling behavior of the SL can be probed in measurements using nuclear magnetic resonance to probe the temperature-dependent uniform magnetic susceptibility and measure the relaxation time of ^{63}Cu probe nuclei. These show that for temperatures above $T^*(x)$, the SL can still be described by the 2-d Heisenberg model, with a doping-dependent interaction, $J_{eff}(x)$, between nearest neighbor spins whose magnitude is close to the temperature, $T_{max}(x)$, at which the SL magnetic susceptibility reaches a maximum. As the density of holes increases, both quantities decrease linearly with x.
- $x = 0.22$ is a quantum critical point (QCP) in that, absent superconductivity, one would expect a quantum phase transition there from localized to itinerant behavior for the remaining Cu spins.
- Between T_{max} and T^*, the holes form an anomalous fermi liquid (AFL), whose anomalous transport properties are those expected for quantum critical matter in which the quasiparticles are scattered by the QC fluctuations emanating from the QCP at $x \sim 0.22$. Careful analysis of the nuclear spin-lattice relaxation rate shows that in this temperature range, the SL exhibits the dynamic quantum

critical behavior expected in the vicinity of 2d AF order, hence its designation as a quantum critical spin lLiquid, QCSL.

- Below T*, a quite unexpected new state of quantum matter emerges, pseudogap matter, so called because in it some parts of the quasihole Fermi surface become localized and develop an energy gap; the SL, which is strongly coupled to the holes, ceases to follow the two-dimensional Heisenberg scaling behavior found at higher temperatures.

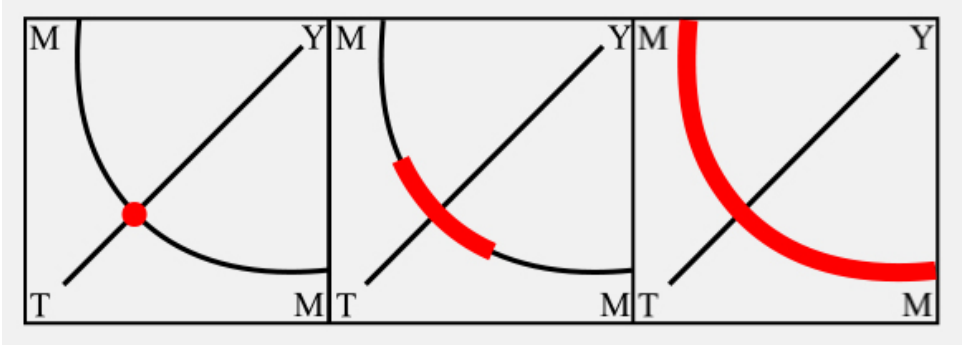


Figure 25: Schematic illustration of the temperature evolution of the Fermi surface in underdoped cuprates. The d-wave node below T_C (left panel) becomes a gapless arc in the pseudogap state above T_C (middle panel), which expands with increasing T to form the full Fermi surface at T* (right panel).
Source:

- For $0.05 < x < 0.17$, the hole concentration that marks the intersection of the T* line with T_c, the superconducting state that emerges from the pseudogap state is "weak"; some of the available quasiparticles have chosen, at a temperature higher than T_c, to become localized by condensing into the pseudogap state, and are therefore not available for condensation into the superconducting state. Their absence from itinerant behavior, illustrated in Figure 25, is seen, for example, in an ARPES (angle-resolved photoemission spectroscopy) probe of quasiparticles at the Fermi surface. Pseudogap matter and superconductivity thus compete for the low-temperature ordered state of the hole Fermi liquid in much the same way as antiferromagnetism and superconductivity compete in heavy electron materials.

- For $x > 0.17$, superconductivity wins the competition and is "strong," in that all available quasiparticles condense into the superconducting state. At these dopings, the pseudogap state does not form unless a magnetic field strong enough to destroy superconductivity is applied; when it is, the pseudogap state continues to form until one reaches the QCP at $x \sim 0.22$, behavior analogous to that found for the AF state in $CeRhIn_5$.

- Whether the superconductivity is weak or strong, the pairing state turns out to be the d_{x2-y2} state that, in the case of heavy electron materials, is the signature of a magnetic mechanism in which the magnetic quantum critical spin fluctuations provide the pairing glue. It is not unreasonable to conclude that the same physics is at work in the cuprates, with the nearly antiferromagnetic spin fluctuations playing a role for these unconventional superconductors that is analogous to that of phonons for conventional superconductors.

- The pseudogap state tends to form stripes. This tendency toward "inhomogeneous spatial ordering" reflects the competition between localization and itinerant behavior. It leads to the

formation of fluctuating spatial domains that have somewhat fewer holes than the average expected for their doping level that are separated by hole-rich domain walls.

- Scanning tunneling microscope experiments (STM) (Figure 26) on the BSCCO members of the cuprate family at low temperatures show that, for doping levels less than x ~ 0.22, even the samples least contaminated by impurities exhibit a substantial degree of spatial inhomogeneity, reflected in a distribution of superconducting and pseudogap matter energy gaps.

- Just as in the case of heavy electrons, the maximum T_c is not far from the doping level at which the spatial order manifested in pseudogap behavior enters.

Ingredients of a theory

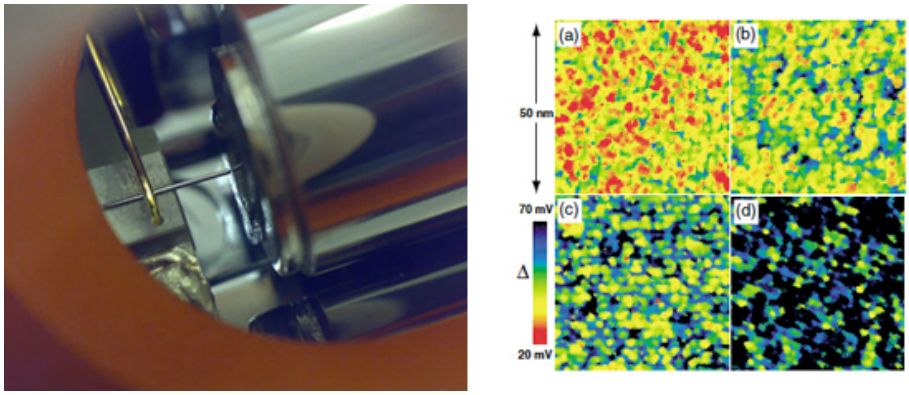


Figure 26: Left: A scanning tunneling microscope (STM) is a powerful instrument for imaging surfaces at the atomic level. Right: Inhomogeneous energy gaps measured in BSCCO; (a)-(d) correspond to doping levels that range from near optimal values of x = 0.19 seen in sample (a), for which T_C is 89 K, through levels of 0.15 (b), 0.13 (c) to the very underdoped material(d), for which x = 0.11 and T_C = 65 K; the color scales are identical.
Source: © Left: Wikimedia Commons, CC Share Alike 2.5 Generic License. Author: Royce Hunt, 5 March 2007. Right: K. McElroy, D.-H. Lee, J. E. Hoffman, K. M. Lang, J. Lee, E.W. Hudson, H. Eisaki, S. Uchida, and J. C. Davis.

We do not yet possess a full microscopic theory that explains these amazing emergent behaviors, but we see that the basic ingredients for developing such a theory are remarkably similar to those encountered in heavy electron materials. In both cuprates and heavy electron materials, local moments coexist with quasiparticles over a considerable portion of their generalized phase diagrams. Their mutual interaction and proximity to antiferromagnetism and a "delocalizing" quantum critical point lead to the emergence of quantum critical matter and d_{x2-y2} superconductivity, with the maximum T_c for the latter located not far from the QCP at which quasiparticle localization first becomes possible.

The principal differences are twofold: First, in the cuprates, the physical origin of the local moments is intrinsic, residing in the phenomenon of Mott localization brought about by strong electrostatic repulsion); second, in place of the AF order seen in heavy electron materials, one finds a novel ordered state, the

pseudogap, emerging from the coupling of quasiparticles to one another and to the spin liquid formed by the Cu spins. It is the task of theory to explain this last result.

We can qualitatively understand the much higher values of T_c found in the cuprates as resulting from a mix of their much higher intrinsic magnetic energy scales as measured by the nearest neighbor LM interaction—J ~1000 K compared to the 50 K typically found in heavy electron materials—and their increased two-dimensionality.

Theories in competition

Our present understanding of emergent behaviors in the cuprates would not have been possible without the continued improvement in sample preparation that has led to materials of remarkable purity; the substantive advances in the use of probes such as nuclear magnetic resonance and inelastic neutron scattering, to study static and dynamic magnetic behavior in these materials; and the development of probes such as ARPES, STM, and the de Haas von Alphen effect that enable one to track their quasiparticle behavior in unprecedented detail. The brief summary presented here has scarcely done justice to the much more detailed information that has emerged from these and other experiments, while it is even more difficult to present at a level appropriate for this unit an overview of the continued efforts by theorists to develop a microscopic explanation of this remarkable range of observed emergent behaviors.

The theoretical effort devoted to understanding the cuprate superconductors is some orders of magnitude greater than that which went into the search for a microscopic theory of conventional superconductors. Yet, as of this writing, it has not been crowned by comparable success. Part of the reason is that the rich variety of emergent behaviors found in these materials by a variety of different experimental probes are highly sample-dependent; it has not yet proved possible to study a sample of known concentration and purity using all the different probes of its behavior. This has made it difficult to reconcile the results of different probes and arrive at candidate phenomenological pictures such as that presented above, much less to arrive at a fundamental theory.

Another aspect is the existence of a large number of competing theories, each of which can claim success in explaining some aspect of the phase diagram shown in Figure 24. The proponents of each have been reluctant to abandon their approach, much less accept the possibility that another approach has been successful. Since none of these approaches can presently explain the complete candidate phase diagram discussed above, developing a microscopic theory that can achieve this goal continues to be a major challenge in condensed matter theory.

Still another challenge is finding new families of superconductors. Theory has not notably guided that quest in the past. However, the striking similarities in the families of novel unconventional superconductors thus far discovered suggest one strategy to pursue in searching for new families of unconventional (and possibly higher T_c) superconductors: Follow the antiferromagnetism, search for layered materials with high values of J, and pay attention to the role that magnetic order can play in maximizing T_c. In so doing, we may argue that we have learned enough to speculate that, just as there was a "phonon" ceiling of some 30 K for T_c in conventional superconductors, there may be a "magnetic" ceiling for T_c in unconventional superconductors. Both may be regarded as reflecting a tendency for strong quasiparticle interactions to produce localization rather than superconductivity. The question, then, is whether we have reached this ceiling with a T_c of about 160 K or whether new materials will yield higher transition temperatures using magnetic glues, and whether there are nonmagnetic electronic routes to achieving still higher values of T_c.

Section 8: Superfluidity on a Cosmic Scale

The Quick and the Dense: Pulsars and Neutron Stars

When a star with between four and eight times the mass of our Sun approaches the end of its life, it undergoes gravitational collapse, seen as a supernova, with the end-point being a comparatively tiny object—a star in which the inward pressure of gravity is balanced by the outward quantum pressure of the mostly neutrons it contains. These neutron-rich celestial objects arise from when the relentless gravitational pressure within its parent object exceeds the thermal pressure produced by its nuclear. The resulting stellar collapse drives a conversion in the supernova core of protons and electrons into neutrons, essentially compressing the atomic matter into nuclear material. This eliminates the empty space in atoms, and produces an object of extraordinarily high density. The typical neutron star packs a mass about 1.4 times that of the Sun into a diameter of order 10 kilometers so that its density is of the order of that found in atomic nuclei, equivalent to packing all the people on Earth into a single raindrop, so that a teaspoonful of the matter in a neutron star would weigh one billion tons on Earth.

Although the possibility of neutron stars was first suggested by Walter Baade and Fritz Zwicky in 1934, it was not until 1967 that astronomers took the proposal seriously. That year, a Cambridge University graduate student, Jocelyn Bell, convinced her thesis supervisor, Antony Hewish, that her observation of a radio source pulsing on and off with extraordinary regularity was not due to a system malfunction. The Cambridge radio astronomers first called the object LGM-1 (for little green men 1) because its precision seemed to indicate communication from intelligent life. But within a few months, theorists, led by Cornell astrophysicist Thomas Gold, persuaded the astronomical community that a pulsar had to be a rotating neutron star containing a giant residual magnetic field whose magnitude may be estimated by assuming that flux is conserved in the implosion which formed it, so that a progenitor core magnetic field of ~1 gauss becomes amplified into a field of ~10^{12} gauss, which could emit electron and electromagnetic beams of radiation as it spins.

So, while pulsars have not revealed anything about life elsewhere in the universe in the decades since their discovery, they have provided astrophysicists with a cosmic laboratory that can be used to study the behavior of matter at extraordinarily high densities, densities that are indeed the highest observable in our universe.

We conclude this unit with an introduction to some truly high T_c superfluids—the neutron superfluids found in the crust and core of a neutron star, for which T_c can be as large as 600,000 K. As we will see, the remarkable behavior of these superfluids not only enables us to study their emergent behavior by observing pulsars located many light-years away, but also establishes that these represent the most abundant superfluids in our universe.

Figure 27: The first director of the Los Alamos National Laboratory, Robert Oppenheimer (ca. 1944) was a brilliant theoretical physicist and inspired teacher who became famous for his remarkably effective leadership of the Manhattan Project.
Source: © Los Alamos National Laboratory.

Russian physicist Arkady Migdal suggested in the 1960s that cosmic hadron superfluids might exist in neutron stars. If the neutron stars studied by Robert Oppenheimer in 1939 existed, he reasoned, then in light of the fact that nuclei in the outer shells of terrestrial nuclei exhibited superfluid behavior, the neutrons these stars contained would surely be superfluid. Not long afterwards, Vitaly Ginzburg and David Kirshnitz argued that neutron stars, if indeed they existed, would surely rotate, in which case their rotation should be described in terms of the quantized vortex lines seen in liquid He.

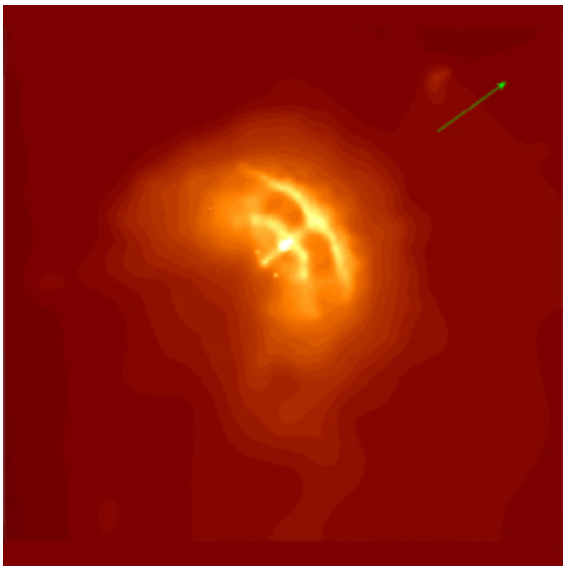


Figure 28: An image taken by the Chandra X-ray telescope of the Vela supernova remnant that shows dramatic bow-like structures produced by the interaction of radiation and electron beams coming from the rapidly rotating neutron star in its center with its immediate environment.
Source: © NASA, PSU, G. Pavlov et al., courtesy of Chandra X-ray Observatory.

The issue of such cosmic superfluidity remained a gleam in the theorist's eye until 1967. In that year, Cambridge graduate student Jocelyn Bell, who was studying atmospheric-produced scintillations of radio signals in Antony Hewish's radio astronomy laboratory, discovered pulsars. Astronomers soon identified these objects as rotating neutron stars which slowed down in remarkably regular fashion as they transferred their rotational energy into electromagnetic waves and accelerated electron beams.

Two years later, V. Radhakrishnan and Richard Manchester, using a radiotelescope in Australia, and Paul Reichley and George Downs, based at Caltech's Jet Propulsion Laboratory, independently observed that a comparatively young and fast pulsar, the Vela pulsar with an 89 ms period of rotation, "glitched." First, instead of continuing a remarkably regular spin-down produced by the transformation of its rotational energy into the beams of radio emission observed on Earth, it sped up by a few parts in a million.

Then, over some days to weeks, the sudden spin-up decayed. That a sudden spin-up of a tiny astronomical object with a radius of about 10 kilometers but a mass of the order of our Sun should occur at all is remarkable. Indeed, astronomers might have treated the glitch report as a malfunction of an observing radio telescope had not observers working independently in Australia and California both seen it. But perhaps more remarkable is the fact that astronomers could actually observe a glitch, since a

response time for ordinary neutron matter would be about 10^{-4} seconds, the time it takes a sound signal to cross the star. So, under normal circumstances, a glitch and its response would be gone in less time than the twinkling of an eye.

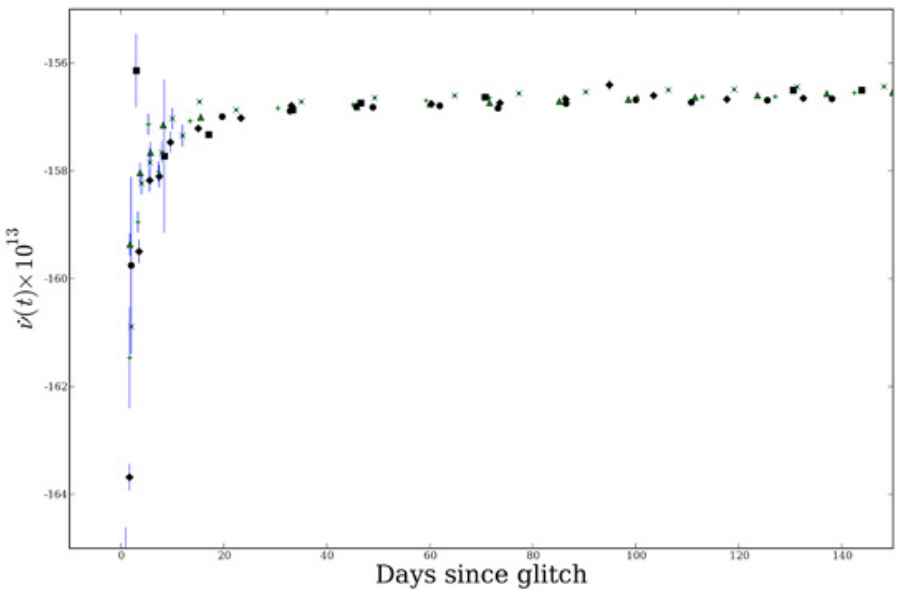


Figure 29: Radiotelescope observations of postglitch behavior in the Vela pulsar.
Source: © Sarah Buchner, Hartebeesthoek Radio Astronomy Observatory, South Africa.

The explanation was soon forthcoming: The slow decay time provided unambiguous evidence for the presence of superfluid neutrons in the pulsar. The reason was that these can change their angular momentum only by the postglitch motion of the vortex lines they carry and that process could easily be imagined to take days to months.

Why glitches occur

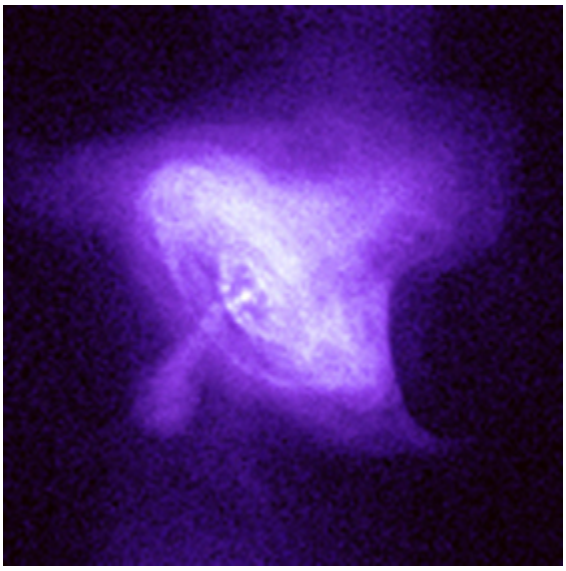


Figure 30: This recent image from the Chandra X-ray telescope shows the Crab Nebula, the remnant of a supernova explosion seen on Earth in 1054 AD that accompanied the formation of a rapidly rotating neutron star at its center.
Source: © NASA, CXC, and SAO.

Theorists initially thought that the origin of the glitch was extrinsic to the neutron superfluid. They envisioned a starquake in which part of the stellar crust crumbled suddenly in response to the forces produced by changes in the star's shape induced by pulsar spindown. But the observation of a second glitch in the Vela pulsar quickly ruled out that explanation for Vela pulsar glitches, since an elementary calculation showed that the expected time between such massive starquakes would be some thousands of years. Typical intervals between Vela pulsar glitches are some two years. It should be noted that for the much smaller glitches (a few parts in 100 million) seen in very young pulsars, such as that located in the Crab Nebula, whose age is less than 1,000 years, starquakes continue to provide a plausible explanation of their origin.

In 1975, Philip Anderson and Naoki Itoh came up with what astrophysicists now recognize as the correct explanation of the frequent pulsar glitches seen in the Vela and other older pulsars. Glitches, they argued, are an intrinsic property of the crustal neutron superfluid and come about because the vortex lines that carry the angular momentum of the crustal superfluid are pinned to the crustal nuclei with which they coexist. As a result, the superfluid's angular velocity, which can change only through the motion of its vortices, will lag that of the crust. The lag will persist until a sufficiently large number of vortices are pinned, at which point these unpin catastrophically, bringing about a sudden jump in the angular

momentum of the star—a glitch—while their subsequent motion determines the postglitch behavior produced by superfluid response to the glitch.

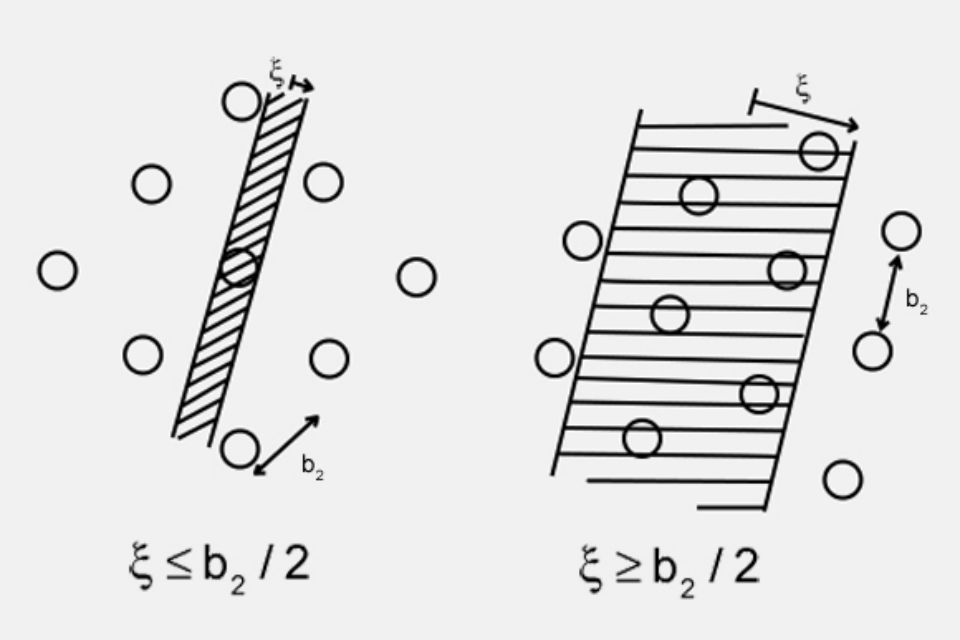


Figure 31: An illustration of two possible regimes of pinning for superfluid vortices in the crust of a neutron star. Left: the weak pinning expected when the size of the normal core, ξ of a vortex in the crustal neutron superfluid vortex is comparable to that of a crustal nucleus. Right the superweak pinning expected for a much larger vortex normal core. **Source:** © Reproduced by permission of the AAS. Courtesy of M.A. Alpar, P.W. Anderson, D. Pines, and J. Shaham, 1983.

Unexpectedly, perhaps, careful study of the range of pinning possibilities and the nonlinear response of the superfluid to unpinning events has made it possible to identify the distinct pinning regions in the stellar crust shown in Figure 30 by their different response to a glitch and to explain the characteristic response times identified in the postglitch behavior of the Vela and other pulsars. Still more surprising, a careful analysis of the glitch magnitude and the superfluid postglitch response of a given pulsar now makes it possible to predict with some accuracy (roughly tens of days, say, for glitches separated by intervals of order years) the time to its next glitch.

A summary of our present theoretical understanding of the components of a neutron star is given in Figure 32, while some recent observations of the two best-known neutron stars, those found in the Crab and Vela constellations, are illustrated in Figures 28 and 30.

Superfluids in the stars

Based on the roughly 25 glitching pulsars observed so far, we can easily establish that the amount of cosmic neutron superfluid observed thus far is several solar masses, or some 10^{34} grams. That far exceeds the quantity of terrestrial superfluid ever produced or observed. Further, the amount of cosmic neutron superfluid contained in neutron stars that have yet to be seen to glitch is likely an order of magnitude larger.

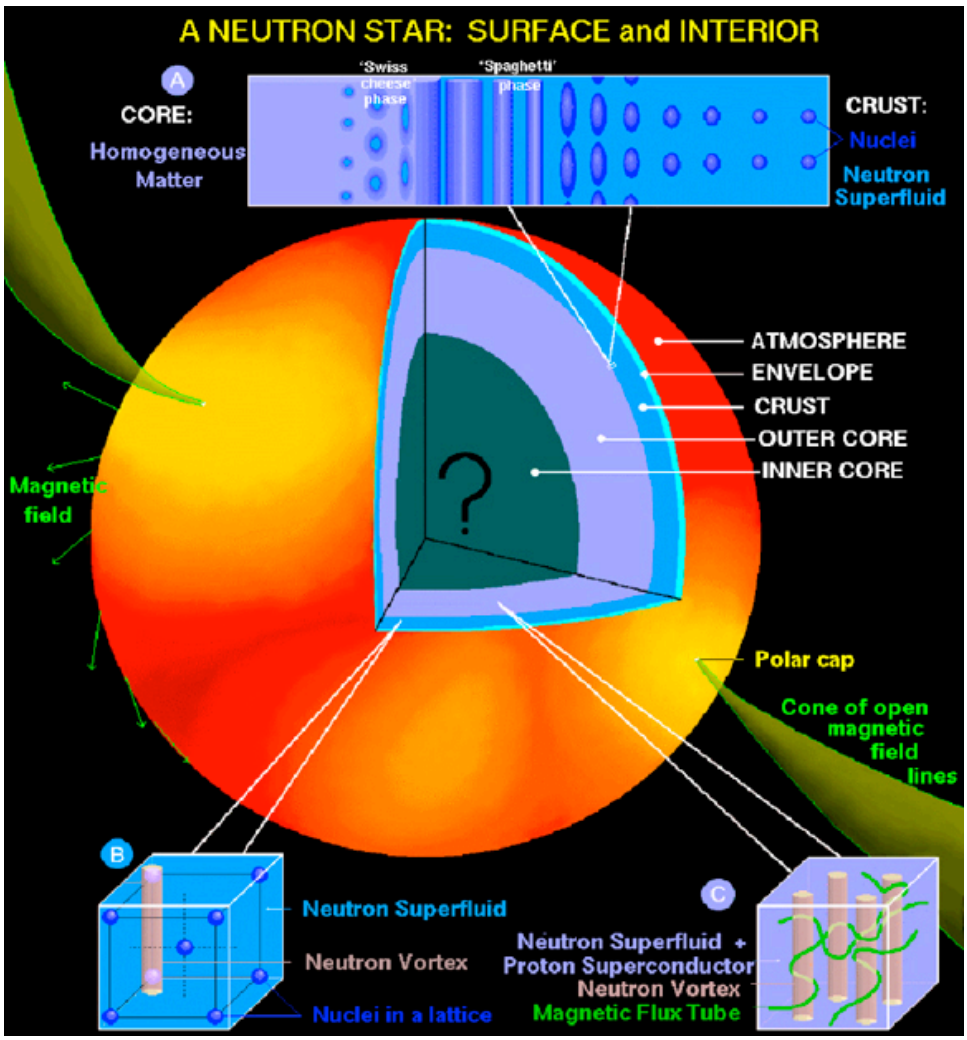


Figure 32: A cross section of a neutron star shows the rich variety of emergent quantum matter expected in its crust and core.
Source: © Dany Page.

Interestingly, glitch observations also provide us with important information on the hadron equation of state, since one that is too soft will not yield a crust sufficiently thick (~ 1 km) to support the region of pinned crustal superfluid we need to explain glitches. On combining this with information from direct

measurements of pulsar masses in binary systems, theorists now conclude that the hadron equation of state that describes the behavior of matter in the inner core of the star depicted in Figure 32, is sufficiently stiff that one will not find quark or other proposed exotic forms of matter there.

Developing an emergent perspective

While the author was receiving radiation treatment for prostate cancer at UCSF in San Francisco in the spring of 1999, with time on his hands following his early morning irradiations, he arranged to visit Stanford two days a week to discuss with his colleague Bob Laughlin various issues relating to the then newly formed Institute for Complex Adaptive Matter.

What emerged from those discussions was a paper, "The Theory of Everything." In it, we pointed out the obvious—that there can be no "theory of everything" in an emergent universe in which it is impossible to calculate with precision the result of bringing more than a dozen or so particles together, to say nothing of the difficulties in dealing with the living matter (discussed in Unit 9). We then called attention to the not so obvious; that despite this, one knows many examples of the existence of higher organizing principles in nature—gateways to emergence that lead to protected behavior in the form of exact descriptions of phenomena that are insensitive to microscopic details.

In this unit, we have considered a number of well-established quantum protectorates: the low-energy excitation spectrum of a conventional crystalline insulator, which consists of transverse and longitudinal sound, regardless of microscopic details; the low energy screening of electron interactions in quantum plasmas; the low-energy behavior of a Landau Fermi liquid; and the low-energy excitation spectrum of a conventional superconductor which is characterized by a handful of parameters that may be determined experimentally but cannot be computed from first principles. We have also considered a newly discovered candidate protectorate, the emergence of the Kondo liquid in heavy electron materials.

In "The Theory of Everything," we emphasized the importance of developing an emergent perspective on science, a perspective espoused years earlier by P. W. Anderson in his seminal article, "More is Different." The importance of acquiring and applying that emergent perspective—the realization that we have to study the system as a whole and search for the organizing principles that must be at work to bring about the observed emergent behavior—is arguably the most important takeaway message of this unit.

An emergent perspective is also needed as we confront emerging major societal challenges—human-induced climate change, terrorism, our current global economic meltdown. These are all caused by

humans; and in searching for an appropriate emergent response, we begin by seeking to identify their origins in societal behavior. But now there is a difference. Because these emerging challenges have no unique cause, it follows that there is no unique or even "best" solution. So we must try many different partial solutions, invent many new institutions, and, above all, experiment, experiment, experiment, as we address the various candidate causes, hoping (and expecting) that in the process some of these experiments will work. If all goes well, because everything is pretty much connected to everything else, a set of related solutions that begin to produce the desired result will emerge over time.

The selection of the examples of emergent behavior in quantum matter to be discussed in this unit has been a quite personal one. There are so many interesting examples of emergent behavior in quantum matter that the unit could easily have been 10 times its present length; in choosing which to present, the author decided to focus on examples drawn from his personal experience. He hopes the reader/viewer will be inspired to explore a number of other important examples on her/his own. Among those highly recommended are the discovery and explanation of quantum Hall states, metal-insulator transitions, dynamical mean field theory, quantum critical behavior, the recently discovered topological insulators, and the emerging fields of spintronics, nanoscience and nanotechnology, and quantum information.

Section 9: Further Reading

- M.Ali Alpar and Altan Baykal, "Pulsar Braking Indices, Glitches and Energy Dissipation in Neutron Stars," *M.N.R.A.S.* 372,489 (2006).
- M.A.Alpar, H.F.Chau, K.S.Cheng and D.Pines, "Postglitch Relaxation of the Vela Pulsar after its First Eight Glitches: A Re-evaluation with the Vortex Creep Model" *Ap. J.* 409,345 (1993).
- P.W. Anderson, "More is Different," *Science* 177, 393–396, 1972.
- P.W. Anderson, "Thinking Big," *Nature* 437, 625-628, 2005.
- Piers Coleman, "Quantum Criticality and Novel Phases: A panel discussion," *Physica Status Solidi* 247, 506-512, 2010.
- R.B. Laughlin and D. Pines, "The Theory of Everything," *PNAS* 97:28-31, 2000.
- P. Monthoux, D. Pines, and G.G. Lonzarich, "Superconductivity without phonons." *Nature* 450, 1177-1183, 2007.
- Philippe Nozieres and David Pines. "The Theory of Quantum Liquids," Vol 1. *Normal Fermi Liquids* and Vol. 2. *The Superfluid Bose Liquid*, Perseus Books, 1999.
- David Pines. *Elementary Excitations in Solids*, WA Benjamin, 1962.

Glossary

antiferromagnetic order: An antiferromagnet is a magnet in which the microscopic magnetic moments inside the material line up in a grid on which neighboring moments point in opposite directions. The interaction energy between two magnetic moments in an antiferromagnet is lower when the two moments point in opposite directions. This can lead to a frustrated system with multiple ground states.

BCS theory: BCS theory is the theory of superconductivity put forward in 1957 by John Bardeen, Leon Cooper, and John Schreiffer, who received the 1972 Nobel Prize for their effort. The basic premise of BCS theory is that under the right conditions inside a conductor, electrons can form weakly bound pairs called "Cooper pairs" that form a condensate. Pairs in the condensate experience no resistance as they travel through the conductor.

doping: In condensed matter physics, doping refers to the deliberate introduction of impurities into an extremely pure crystal. For example, a crystal of pure silicon might be doped with boron atoms that change the material's electrical properties, making it a more effective semiconductor.

emergent behavior: Emergent behavior is behavior of a complex system that is not easily predicted from a microscopic description of the system's constituent parts and the rules that govern them.

Fermi surface: According to the Pauli exclusion principle, it is not possible for identical fermions to occupy the same quantum state. In a system with many identical fermions, such as electrons in a metal, the fermions fill in the available quantum states in order of increasing energy. The energy of the highest occupied quantum state defines the energy of the Fermi surface, which is a surface of constant energy in momentum space.

ferromagnet: A ferromagnet is a magnet in which the microscopic magnetic moments inside the material all point in the same direction. Most magnetic materials we encounter in daily life are ferromagnets.

inelastic neutron scattering: Inelastic neutron scattering is an experimental technique for studying various properties of materials. A beam of neutrons of a particular energy is shot at a sample at a particular angle with respect to the crystal lattice. The energy of neutrons scattered by the sample is recorded, and the experiment is repeated at different angles and beam energies. The scattered neutrons lose some of their energy to the sample, so the scattering is inelastic. The results of inelastic neutron scattering are readily interpreted in terms of the wave nature of particles. The incident neutron beam is a wave with a

frequency proportional to the neutron energy. The crystal preferentially absorbs waves with frequencies that correspond to its natural modes of vibration. Note that the vibrations can be magnetic or acoustic. Thus, the modes of the sample can be inferred by mapping out how much energy is absorbed from the incident beam as a function of the incident beam energy. Inelastic neutron scattering has also been used to study acoustic oscillations and their corresponding quasiparticles in liquids.

itinerant: In condensed matter physics, the term itinerant is used to describe particles (or quasiparticles) that travel essentially freely through a material and are not bound to particular sites on the crystal lattice.

magnons: Magnons are the quasiparticles associated with spin waves in a crystal lattice.

phase: In physics, the term phase has two distinct meanings. The first is a property of waves. If we think of a wave as having peaks and valleys with a zero-crossing between them, the phase of the wave is defined as the distance between the first zero-crossing and the point in space defined as the origin. Two waves with the same frequency are "in phase" if they have the same phase and therefore line up everywhere. Waves with the same frequency but different phases are "out of phase." The term phase also refers to states of matter. For example, water can exist in liquid, solid, and gas phases. In each phase, the water molecules interact differently, and the aggregate of many molecules has distinct physical properties. Condensed matter systems can have interesting and exotic phases, such as superfluid, superconducting, and quantum critical phases. Quantum fields such as the Higgs field can also exist in different phases.

phonon: Phonons are the quasiparticles associated with acoustic waves, or vibrations, in a crystal lattice or other material.

plasma: A plasma is a gas of ionized (i.e., electrically charged) particles. It has distinctly different properties than a gas of neutral particles because it is electrically conductive, and responds strongly to electromagnetic fields. Plasmas are typically either very hot or very diffuse because in a cool, relatively dense gas the positively and negatively charged particles will bind into electrically neutral units. The early universe is thought to have passed through a stage in which it was a plasma of quarks and gluons, and then a stage in which it was a plasma of free protons and electrons. The electron gas inside a conductor is another example of a plasma. The intergalactic medium is an example of a cold, diffuse plasma. It is possible to create an ultracold plasma using the techniques of atom cooling and trapping.

plasmons: Plasmons are the quasiparticle associated with oscillations of charge density in a plasma.

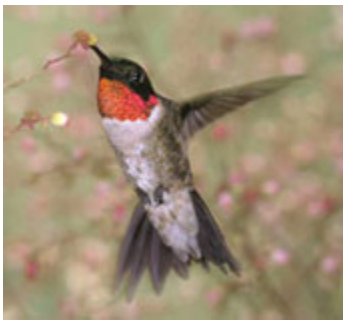
pulsar: A pulsar is a spinning neutron star with a strong magnetic field that emits electromagnetic radiation along its magnetic axis. Because the star's rotation axis is not aligned with its magnetic axis, we observe pulses of radiation as the star's magnetic axis passes through our line of sight. The time between pulses ranges from a few milliseconds to a few seconds, and tends to slow down over time.

quasiparticles: Just as particles can be described as waves through the wave-particle duality, waves can be described as particles. Quasiparticles are the quantized particles associated with various types of waves in condensed matter systems. They are similar to particles in that they have a well-defined set of quantum numbers and can be described using the same mathematical formalism as individual particles. They differ in that they are the result of the collective behavior of a physical system.

SQUID: A superconducting quantum interference device, or SQUID, is a tool used in laboratories to measure extremely small magnetic fields. It consists of two half-circles of a superconducting material separated by a small gap. The quantum mechanical properties of the superconductor make this arrangement exquisitely sensitive to tiny changes in the local magnetic field. A typical SQUID is sensitive to magnetic fields hundreds of trillions of times weaker than that of a simple refrigerator magnet.

Physics for the 21st Century

Unit 9: Biophysics



© Steve Maslowski, U.S. Fish and Wildlife Service.

Unit Overview

Following the example set in the previous unit, we now attempt to bring principles of physics to bear on the most complex systems of all: biological systems. Is it possible to describe living systems, or even small pieces of living systems, with the same concepts developed elsewhere in our ramble through physics? We begin with a discussion of whether physics can tell us if something is, in fact, alive. In the reductionist spirit, we then consider the physical principles that govern the constituent molecules of biological systems—and their emergent properties. From DNA and proteins, we move on to evolution and how it is physically possible for a species to genetically adapt to its environment quickly enough to survive. Finally, we seek to understand how the conscious mind can emerge from a network of communicating cells.

Content for This Unit

Sections:

Section 1: Introduction

Biology is complicated, really, really complicated. This should not surprise you if you think that ultimately the laws of physics explain how the world works, because biological activity is far beyond the usual realm of simple physical phenomena. It is easy to simply turn away from a true physical explanation of biological phenomena as simply hopeless. Perhaps it is hopelessly complex at some level of detail. The ghosts of biological "stamp collecting" are still alive and well and for a good reason.

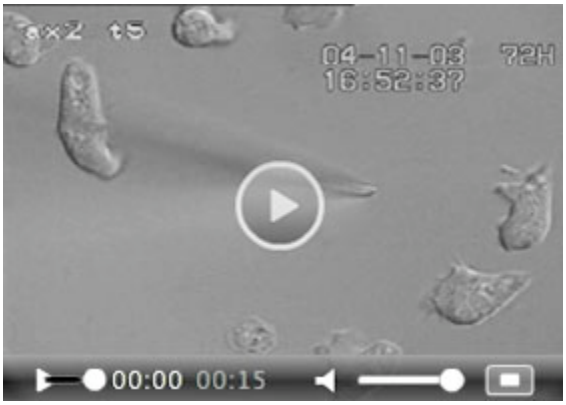


Figure 1: A population of slime-mold cells forms an aggregate in response to a signaling molecule.
Source: © Courtesy of BioMed Central Ltc. From: Maddelena Arigoni et al. "A novel Dictyostelium RasGEF required for chemotaxis and development," BMC Cell Biology, 7 December 2005.

However, it is possible that in spite of the seemingly hopeless complexity of biology, there are certain emergent properties that arise in ways that we can understand quantitatively. An emergent property is an unexpected collective phenomenon that arises from a system consisting of interacting parts. You could call the phenomenon of life itself an emergent property. Certainly no one would expect to see living systems arise directly from the fundamental laws of quantum mechanics and the Standard Model that have been discussed in the first seven units of this course.

The danger is that this concept of "emergent properties" is just some philosophical musing with no real deeper physics content, and it may be true that the emergent properties of life viewed "bottom up" are simply too complex in origin to understand at a quantitative level. It may not be possible to derive how emergent properties arise from microscopic physics. In his book *A Different Universe: Reinventing Physics from the Bottom Down*, the physicist Robert Laughlin compares the local movement of air molecules around an airplane wing to the large-scale turbulent hydrodynamic flow of air around the airfoil that gives rise to lift. Molecular motion is clearly the province of microscopic physics and

statistical mechanics, while turbulent flow is an emergent effect. As Laughlin puts it, if he were to discover that Boeing Aircraft began worrying about how the movement of air molecules collectively generates hydrodynamics, it would be time to divest himself of Boeing stock. Perhaps the same should have been said when banks started hiring theoretical physicists to run stock trading code.

In biology, we have a much greater problem than with the airplane, because the air molecules can be described pretty well with the elegant ideas of statistical mechanics. So, while it is a long stretch to derive the emergence of turbulence from atomic motion, no one would say it is impossible, just very hard.

Figure 2: Colored smoke marks the hydrodynamic flow around an aircraft, an emergent phenomenon.
Source: © NASA Langley Research Center (NASA-LaRC).

In biology, even the fundamentals at the bottom may be impossibly hard for physics to model adequately in the sense of having predictive power to show the pathways of emergent behavior. A classic example is the signaling that coordinates the collective aggregation of the slime-mold *Dictyostelium* cells in response to the signaling molecule cyclic AMP (cAMP). In the movie shown in Figure 1, the individual *Dictyostelium* cells signal to each other, and the cells stream to form a fruiting body in an emergent process called "chemotaxis." This fairly simple-looking yet spectacular process is a favorite of physicists and still not well understood after 100 years of work.

So, perhaps in a foolhardy manner, we will move forward to see how physics, in the discipline known as biological physics, can attack some of the greatest puzzles of them all. We will have to deal with the emergence of collective phenomena from an underlying complex set of interacting entities, like our *Dictyostelium* cells. But that seems still within the province of physics: really hard, but physics. But there are deeper questions that seem to almost be beyond physics.

Are there emergent physics rules in life?

The amazing array of knowledge in previous units contains little inkling of the complex, varied phenomena of life. Life is an astonishingly emergent property of matter, full-blown in its complexity today, some billions of years after it started out in presumably some very simple form. Although we have many physical ways to describe a living organism, quantifying its state of aliveness using the laws of physics seems a hopeless task. So, all our tools and ideas would seem to fail at the most basic level of describing what life is.

Biology has other incredible emergent behaviors that you can hardly anticipate from what you have learned so far. British physicist Paul Dirac famously said that, "The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved." Our question is: Is the biological physics of the emergent properties of life simply a matter of impossible complexity, or are there organizing principles that only appear at a higher level than the baseline quantum mechanics?

So far, we have talked about the emergent nature of life itself. The next astonishing emergent behavior we'll consider is the evolution of living organisms to ever-higher complexity over billions of years. It is strange enough that life developed at all out of inanimate matter, in apparent conflict with the Second Law of Thermodynamics. The original ur-cell, improbable as it is, proceeded to evolve to ever-greater levels of complexity, ultimately arriving at *Homo sapiens* several million years ago. Thanks to Darwin and Wallace and their concept of selection of the fittest, we have a rather vague hand-waving idea of how this has happened. But the quantitative modeling of evolution as an emergent property remains in its infancy.

The building blocks of evolution

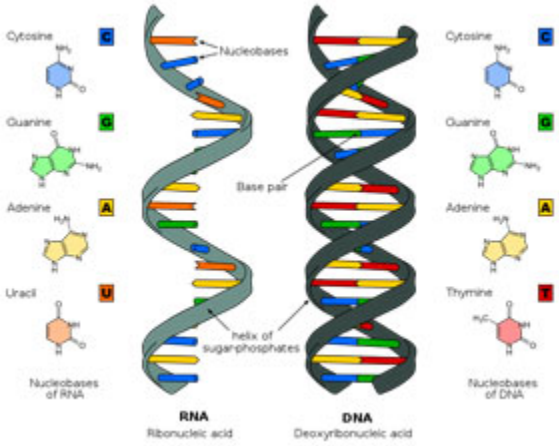


Figure 3: Molecules of life: RNA (left) and DNA (right).
Source: © Wikimedia Commons, Creative Commons Attribution-Share Alike 3.0 license. Author: Sponk, 23 March 2010.

Modeling evolution is a difficult task; but nevertheless, we can try. So let's start at the top and work down. Physicists believe (and it is somewhat more of a belief than a proven fact) that all life began with some ur-cell and that life evolved from that ur-cell into the remarkable complexity of living organisms we have today, including *Homo sapiens*. We will never know what path this evolution took. But the remarkable unity of life (common genetic code, common basic proteins, and common basic biological pathways) would indicate that, at its core, the phenomenon of life has been locked in to a basic set of physical modes and has not deviated from this basic set. At that core, lies a very long linear polymer, deoxyribonucleic acid (or DNA), which encodes the basic self-assembly information and control information. A related molecule, ribonucleic acid (RNA), has a different chemical group at one particular position, and that profoundly changes the three-dimensional structure that RNA takes in space and its chemical behavior.

Although evolution has played with the information content of DNA, its basic core content, in terms of how its constituent molecules form a string of pairs, has not obviously changed. And while there is a general relationship between the complexity of an organism and the length of its DNA that encodes the complexity, some decidedly simpler organisms than *Homo sapiens* have considerably longer genomes. So, from an information perspective, we really don't have any iron-clad way to go from genome to organismal complexity, nor do we understand how the complexity evolved. Ultimately, of course, life is matter. But, it is the evolution of information that really lies at the unknown heart of biological physics, and we can't avoid it.

The emergence of the mind in living systems

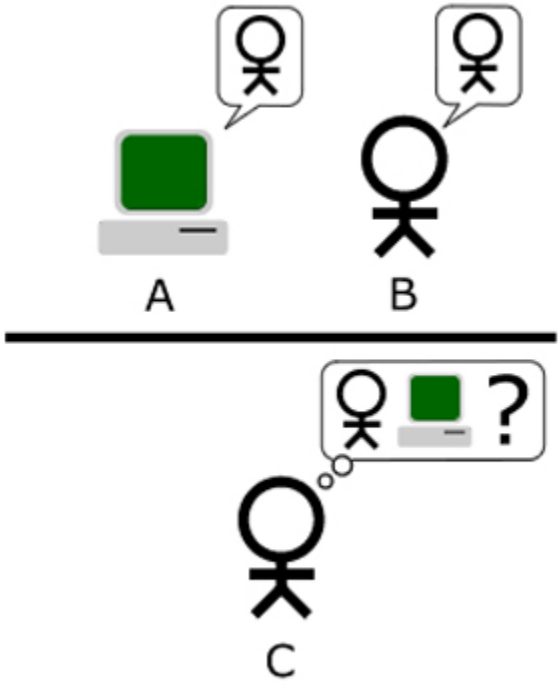


Figure 4: Player C is trying to determine which player—A or B—is a computer and which is human.
Source: © Wikimedia Commons, Public Domain. Author: Bilby, 25 March 2008.

Biology possesses even deeper emergent phenomena than the evolution of complexity. The writer and readers of this document are sentient beings with senses of identity and self and consciousness. Presumably, the laws of physics can explain the emergent behavior of consciousness, which certainly extends down from *Homo sapiens* into the "lower" forms of life (although those lower forms of life might object to that appellation). Perhaps the hardest and most impossible question in all of biological physics is: What is the physical basis behind consciousness? Unfortunately, that quest quickly veers into the realm of the philosophical and pure speculation; some would say it isn't even a legitimate physics question at all.

There is even an argument as to whether "machines," now considered to be computers running a program, will ever be able to show the same kind of intelligence that living systems such as human beings possess. Traditional reductionist physicists, I would imagine, simply view the human mind as some sort of a vastly complicated computational machine. But it is far from clear if this view is correct. The

mathematician Alan Turing, not only invented the Turing machine, the grandfather of all computers, but he also asked a curious question: Can machines think? To a physicist, that is a strange question for it implies that maybe the minds of living organisms somehow have emergent properties that are different from what a manmade computing machine could have. The answer to Turing's question rages on, and that tells us that biology has very deep questions still to be answered.

Section 2: Physics and Life

Here's a question from a biologist, Don Coffey at Johns Hopkins University: Is a chicken egg in your refrigerator alive? We face a problem right away: What does being alive actually mean from a physics perspective? Nothing. The concept of aliveness has played no role in anything you have been taught yet in this course. It is a perfectly valid biological question; yet physics would seem to have little to say about it. It is an emergent property arising from the laws of physics, which presumably are capable of explaining the physics of the egg.

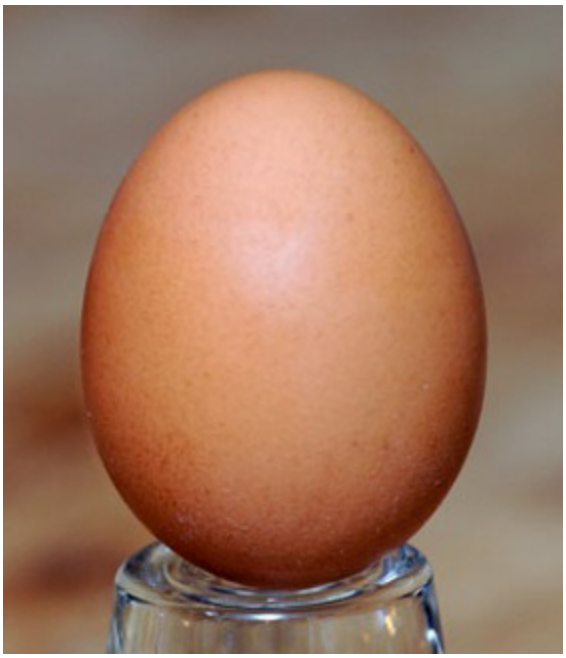


Figure 5: A chicken egg. Is it alive or dead?
Source:

The chicken egg is a thing of elegant geometric beauty. But its form is not critical to its state of aliveness (unless, of course, you smash it). However, you can ask pertinent physical questions about the state of the egg to determine whether it is alive: Has the egg been cooked? It's pretty easy to tell from a physics perspective: Spin the egg around the short axis of the ellipse rapidly, stop it suddenly, and then let it go. If it starts to spin again, it hasn't been cooked because the yolk proteins have not been denatured by heat and so remain as a viscous fluid. If your experiment indicates the egg hasn't been cooked it might be alive, but this biological physics experiment wouldn't take you much closer to an answer.

Assuming you haven't already broken the egg, you can now drop it. If you were right that it has not been cooked, the egg will shatter into hundreds of pieces. Is it dead now? If this were your laptop computer, you could pick up all the pieces and—if you are good enough—probably get it working again. However, all of the king's horses and all the king's men can't put Humpty Dumpty back together again and make him alive once more; we don't know how to do it. The egg's internal mechanical structure is very complex and rather important to the egg's future. It, too, is part of being alive, but surely rather ancillary to the main question of aliveness.

Aliveness is probably not a yes-no state of a system with a crisp binary answer, but rather a matter of degree. One qualitative parameter is the extent to which the egg is in thermodynamic equilibrium with its surroundings. If it is even slightly warmer, then I would guess that the egg is fertilized and alive, because it is out of thermodynamic equilibrium and radiating more energy than it absorbs. That would imply that chemical reactions are running inside the egg, maintaining the salt levels, pH, metabolites, signaling molecules, and other factors necessary to ensure that the egg has a future some day as a chicken.

Wait, the egg has a future? No proton has a future unless, as some theories suggest, it eventually decays. But if the egg is not dropped or cooked and is kept at exactly the right temperature for the right time, the miracle of embryonic development will occur: The fertilized nucleus within the egg will self-assemble in an intricate dance of physical forces and eventually put all the right cells into all the right places for a chick to emerge. Can the laws of physics ever hope to predict such complex emergent phenomena?

Emergent and adaptive behavior in bacteria

Here's an explicit example of what we are trying to say about emergent behavior in biology. Let's move from the complex egg where the chick embryo may be developing inside to the simple example of bacteria swimming around looking for food. It's possible that each bacterium follows a principle of every bug for itself: They do not interact with each other and simply try to eat as much food as possible in order to reproduce in an example of Darwinian competition at its most elemental level. But food comes and food goes at the bacterial level; and if there is no food, an individual bacterium will starve and not be able to survive. Thus, we should not be surprised that many bacteria do not exist at the level as rugged individuals but instead show quite startling collective behavior, just like people build churches.

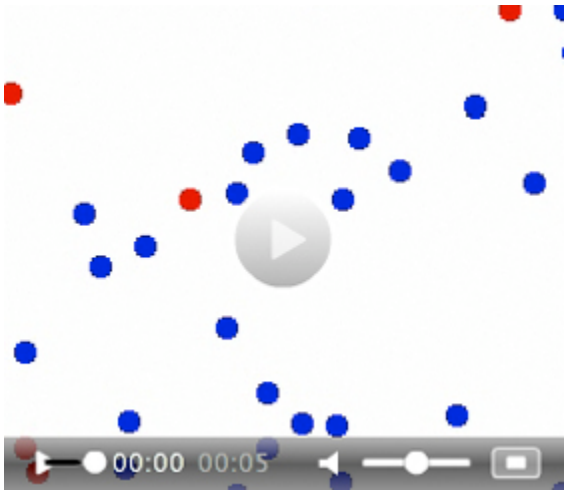


Figure 6: Random motion of gas molecules—bottom up.
Source: © Wikimedia Commons, GNU Free Documentation License.
Author: Greg L., 25 August 2006.

If bacteria acted as rugged individuals, then we would expect their movement through space looking for food to resemble what is called a random walk, which is different from the Brownian motion that occurs due to thermal fluctuations. In a random walk there is a characteristic step size L, which is how far the bacterium swims in one direction before it tumbles and goes off randomly in a new direction. Howard Berg at Harvard University has beautiful videos of this random movement of bacteria. The effect of this random motion is that we can view individual bacteria rather like the molecules of a gas, as shown in Figure 6. If that were all there is to bacterial motion, we would be basically done, and we could use the mathematics of the random walk to explain bacterial motion.

However, bacteria can be much more complicated than a gas when viewed collectively. In the Introduction, we discussed the chemotaxis of a population of individual *Dictyostelium* cells in response to a signal created and received by the collective population of the *Dictyostelium* cells. Bacteria do the same thing. Under stress, they also begin signaling to each other in various ways, some quite scary. For example, if one bacterium mutates and comes up with a solution to the present problem causing the stress, in a process called "horizontal gene transfer" they secrete the gene and transfer it to their buddies. Another response is to circle the wagons: The bacteria signal to each other and move together to form a complex community called a "biofilm." Figure 7 shows a dramatic example of the growth of a complex biofilm, which is truly a city of bacteria.

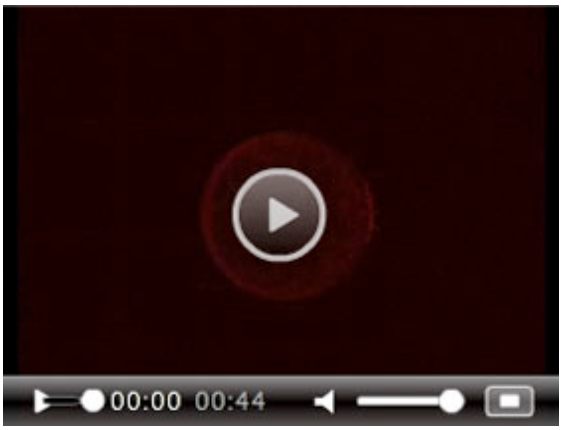


Figure 7: Growth of a biofilm of the bacteria *Bacillis subtilis* over four days.
Source: © Time-lapse movie by Dr. Remco Kort—published in J. Bacteroi. (2006) 188:3099-109.

The mystery is how the supposedly simple bacteria communicate with each other to form such a complex and adapted structure. There is a set of equations, called the "Keller-Segel equations," which are usually the first steps in trying to puzzle out emergent behavior in a collection of swimming agents such as bacteria. These equations are not too hard to understand, at least in principle. Basically, they take the random walk we discussed above and add in the generation and response of a chemoattractant molecule. A sobering aspect of these equations is that they are very difficult to solve exactly: They are nonlinear in the density of the bacteria, and one of the great secrets of physics is that we have a very hard time solving nonlinear equations.

Principles of a complex adaptive system

We are just skimming the surface of a monumental problem in biological physics: How agents that communicate with each other and adapt to the structures that they create can be understood. A biological system that communicates and adapts like the film-forming bacteria is an example of a complex adaptive system. In principle, a complex adaptive system could appear almost anywhere, but biological systems are the most extreme cases of this general phenomenon.

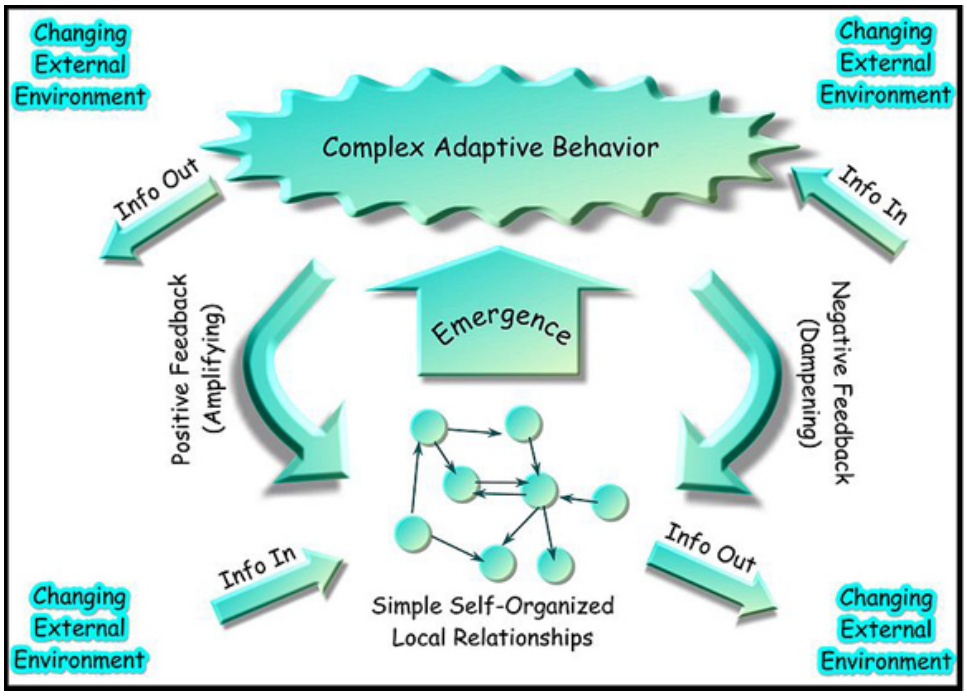


Figure 8: A schematic view of what constitutes a complex adaptive system.
Source: © Wikimedia Commons, Creative Commons Attribution ShareAlike 3.0.

The computer scientist John Holland and the physicist Murray Gell-Mann, who played a major role in the physics developments you read about in Units 1 through 4, have tried to define what makes a complex adaptive system. We can select a few of the key properties as presented by Peter Freyer that are most germane to biological systems:

1. Emergence: We have already discussed this concept, both in this unit and in Unit 8.
2. Co-evolution: We will talk about evolution later. Coevolution refers to how the evolution of one agent (say a species, or a virus, or a protein) affects the evolution of another related agent, and vice versa.
3. Connectivity: This is concerned with biological networks, which we will discuss later.
4. Iteration: As a system grows and evolves, the succeeding generations learn from the previous ones.
5. Nested Systems: There are multiple levels of control and feedback.

These properties will appear time and again throughout this unit as we tour various complex adaptive systems in biology, and ask how well we can understand them using the investigative tools of physics.

Section 3: The Emergent Genome

The challenge of biological physics is to find a set of organizing principles or physical laws that governs biological systems. It is natural to start by thinking about DNA, the master molecule of life. This super-molecule that apparently has the code for the enormous complexity seen in living systems is a rather simple molecule, at least in principle. It consists of two strands that wrap around each other in the famous double helix first clearly described by physicist Francis Crick and his biologist colleague James Watson. While the structure of DNA may be simple, understanding how its structure leads to a living organism is not.

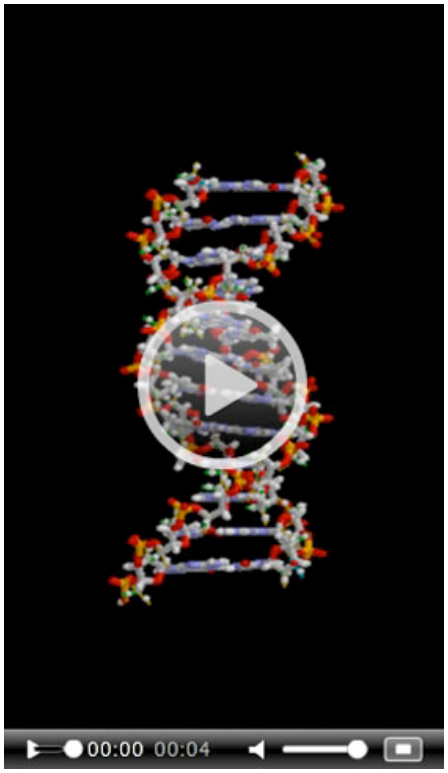


Figure 9: The double helix.
Source: © Wikimedia Commons, Public Domain. Author: brian0918, 22 November 2009.

We will use the word "emergent" here to discuss the genome in the following sense: If DNA simply had the codes for genes that are expressed in the organism, it would be a rather boring large table of data. But there is much more to the story than this: Simply knowing the list of genes does not explain the implicit emergence of the organism from this list. Not all the genes are expressed at one time. There is an

intricate program that expresses genes as a function of time and space as the organism develops. How this is controlled and manipulated still remains a great mystery.

As Figure 9 shows, the DNA molecule has a helicity, or twist, which arises from the fundamental handedness, or chirality, of biologically derived molecules. This handedness is preserved by the fact that the proteins that catalyze the chemical reactions are themselves handed and highly specific in preserving the symmetry of the molecules upon which they act. The ultimate origin of this handedness is a controversial issue. But we assume that a right-handed or left-handed world would work equally well, and that chiral symmetry breaking such as what we encountered in Unit 2 on the scale of fundamental particles is not present in these macroscopic biological molecules.

It is, however, a mistake to think that biological molecules have only one possible structure, or that somehow the right-handed form of the DNA double helix is the only kind of helix that DNA can form. It turns out that under certain salt conditions, DNA can form a left-handed double helix, as shown in Figure 10. In general, proteins are built out of molecules called "amino acids." DNA, itself a protein, contains the instructions for constructing many different proteins that are built from approximately 20 different amino acids. We will learn more about this later, when we discuss proteins. For now, we will stick to DNA, which is made of only four building blocks: the nitrogenous bases adenine (A), guanine (G), cytosine (C), and thymine (T). Adenine and guanine have a two-ring structure, and are classified as purines, while cytosine and thymine have a one-ring structure and are classified as pyrimidines. It was the genius of Watson and Crick to understand that the basic rules of stereochemistry enabled a structure in which the adenine (purine) interacts electrostatically with thymine (pyrimidine), and guanine (purine) interacts with cytosine (pyrimidine) under the salt and pH conditions that exist in most biological systems.

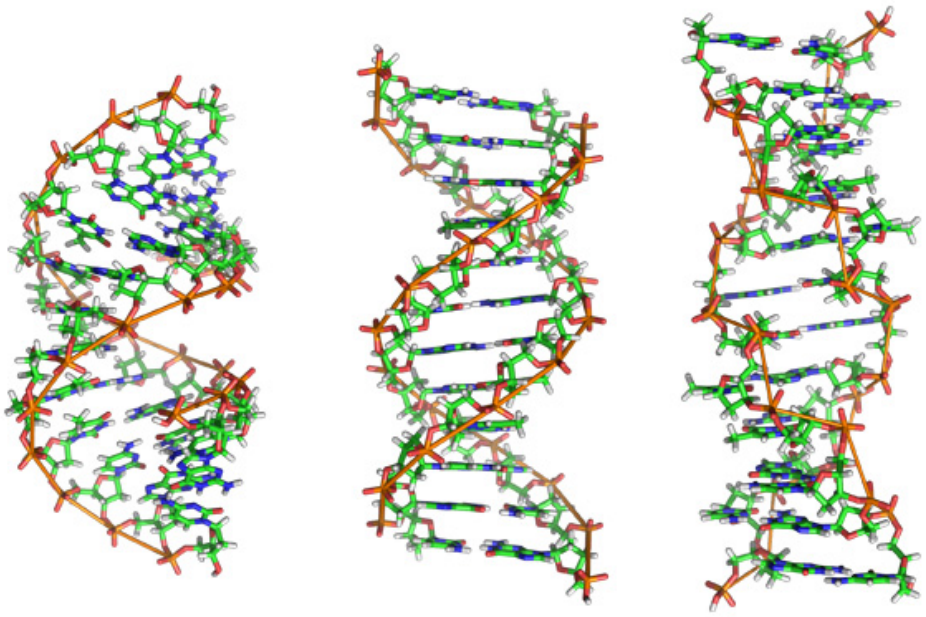


Figure 10: The DNA double helix, in three of its possible configurations.
Source: © Wikimedia Commons, GNU Free Documentation License, Version 1.2. Author: Zephyris (Richard Wheeler), 4 February 2007.

Not only does the single-stranded DNA (ssDNA) molecule like to form a double-stranded (dsDNA) complex, but the forces that bring the two strands together result in remarkably specific pairings of the base pairs: A with T, and G with C. The pyrimidine thymine base can form strong electrostatic links with the purine adenine base at two locations, while the (somewhat stronger) guanine-cytosine pair relies on three possible hydrogen bonds. The base pairs code for the construction of the organism. Since there are only bases in the DNA molecule, and there are about 20 different amino acids, the minimum number of bases that can uniquely code for an amino acid is three. This is called the triplet codon.

The remarkable specificity of molecular interactions in biology is actually a common and all-important theme. It is also a physics problem: How well do we have to understand the potentials of molecular interactions before we can begin to predict the structures that form? We will discuss this vexing problem a bit more in the protein section, but it remains a huge problem in biological physics. At present, we really cannot predict three-dimensional structures for biological structures, and it isn't clear if we ever will be able to given how sensitive the structures are to interaction energies and how complex they are.

An example of this extreme sensitivity to the potential functions and the composition of the polymer can be found in the difference between ribonucleic acids (RNA) and deoxyribonucleic acids (DNA). Structurally, the only difference between RNA and DNA is that at the 2' position of the ribose sugar, RNA

has a hydroxyl (OH) molecule—a molecule with one hydrogen and one oxygen atom—while DNA just has a hydrogen atom. Figure 11 shows what looks like the completely innocuous difference between the two fundamental units. From a physicist's bottom-up approach and lacking much knowledge of physical chemistry, how much difference can that lone oxygen atom matter?

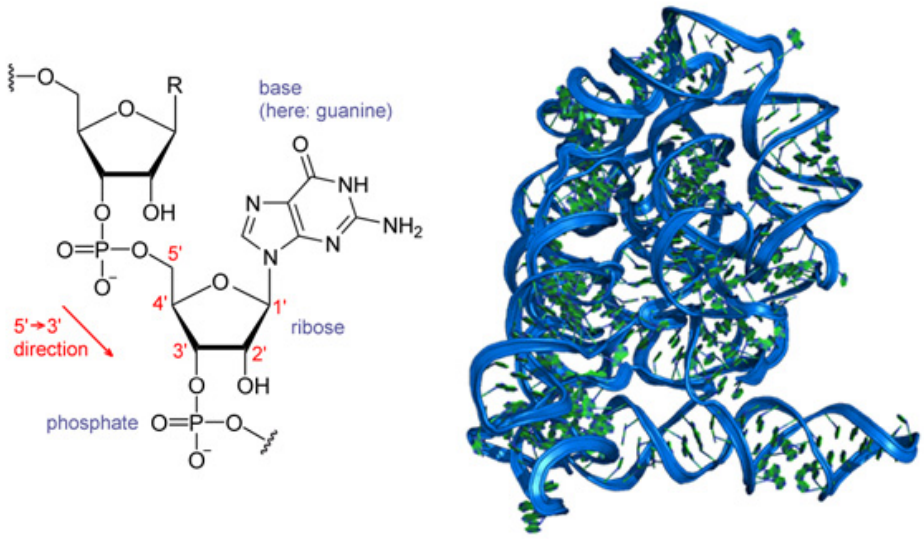


Figure 11: The chemical structure of RNA (left), and the form the folded molecule takes (right).
Source: © Left: Wikimedia Commons, Public Domain. Author: Narayanese, 27 December 2007. Right: Wikimedia Commons, Public Domain. Author: MirankerAD, 18 December 2009.

Unfortunately for the bottom-up physicist, the news is very bad. RNA molecules fold into a far more complex structure than DNA molecules do, even through the "alphabet," for the structures are just four letters: A,C, G, and bizarrely U, a uracil group that Nature for some reason has favored over the thymine group of DNA. An example of the complex structures that RNA molecules can form is shown in Figure 11. Although the folding rules for RNA are vastly simpler than those for DNA, we still cannot predict with certainty the three-dimensional structure an RNA molecule will form if we are given the sequence of bases as a starting point.

The puzzle of packing DNA: chromosomes

Mapping, Sequencing, and Controversy

Since the DNA molecules code for the proteins that are so critical to life, knowing the sequence of the base pairs is vital to knowing what proteins will be produced. This kind of single base pair resolution fractionation is key to "sequencing." At a much more coarse level, you might want to know the basic ordering of various proteins on a strand of DNA; we call this kind of low resolution "mapping" the DNA. The National Institutes of Health have established a National Center for Biotechnology Information with the express purpose of trying to centralize all the information pouring in from sequencing and mapping projects.

The sequencing and mapping of the human genome has been a huge national and private effort, and a very contentious one based upon the raft of ethical, legal, and social implications. The Human Genome Initiative was an astonishing success. However, one school of thought posits that the effort was partly (or maybe mostly) successful because of the efforts of a rival private program headed by entrepreneur Craig Venter, which used a different but complementary approach.

Let's consider a simpler problem than RNA folding: packaging DNA in the cell. A gene is the section of DNA that codes for a particular protein. Since an organism like the bacterium *Escherichia coli* contains roughly 4,000 different proteins and each protein is roughly 100 amino acids long, we would estimate that the length of DNA in *E. coli* must be about 2 million base pairs long. In fact, sequencing shows that the *E. coli* genome actually consists of 4,639,221 base pairs, so we are off by about a factor of two, not too bad. Still, this is an extraordinarily long molecule. If stretched out, it would be 1.2 mm in length, while the organism itself is only about 1 micron long.

The mathematics of how DNA actually gets packaged into small places, and how this highly packaged polymer gets read by proteins such as RNA polymerases or copied by DNA polymerases, is a fascinating exercise in topology. Those of you who are fishermen and have ever confronted a highly tangled fishing line can appreciate that the packaging of DNA in the cell is a very nontrivial problem.

The physics aspect to this problem is the stiffness of the double helix, and how the topology of the twisted and folded molecule affects its biological function. How much energy does it take to bend or twist the polymer into the complex shapes necessary for efficient packaging of DNA in a cell? And how does the intrinsic twist of the double helix translate into the necessity to break the double helix and reconnect it

when the code is read by proteins? In other words, biological physics is concerned with the energetics of bending DNA and the topological issues of how the DNA wraps around in space.

The incredible length of a DNA molecule, already bad enough for bacteria, gets more outrageous for higher organisms. Most mammals have roughly 3×10^9 base pairs wrapped up into chromosomes, which are very complex structures consisting of proteins and nucleic acids. However, although we view ourselves as being at the peak of the evolutionary ladder, there seems to be much more DNA in organisms we view as our intellectual inferiors: Some plants and amphibians have up to 10^{11} base pairs! If we laid out the DNA from our chromosomes in a line, it would have a length of approximately 1 meter; that of amphibians would stretch over 30 meters!

Dark matter in the genome

Why is the human DNA genome so long, and other genomes even longer still? We don't know exactly how many genes the human genome contains, but a reasonable guess seems to indicate about 30,000. If we imagine that each gene codes for a protein that has about 100 amino acids, and that three base pairs are required to specify each amino acid, the minimal size of the human genome would be about 10^7 base pairs. It would seem that we have at least 1,000 times as much DNA as is necessary for coding the genome. Clearly, the amount of "excess" DNA must be much higher for plants and amphibians. Apparently, the DNA is not efficiently coded in the cell, in the sense that lots of so-called "junk" DNA floats around in a chromosome. In fact, a large amount of noncoding DNA has a repeating motif. Despite some guesses about what role this DNA plays, its function remains a substantial puzzle. Perhaps the information content of the genome is not just the number of base pairs, but that there is much "hidden" information contained in this dark genome.

We have succeeded in sequencing the coding part of the human genome, but not the dark part. Are we done now that we know the coding sequence of one given individual? Hardly. We don't know how to extract the information content of the genome at many levels, or even how to define the genome's information quantitatively. The concept of "information" is not only a tricky concept, but also of immense importance in biological physics. Information is itself an emergent property in biology, and it is contextual: The environment gives meaning to the information, and the information itself means little without the context of the environment.

One problem is that we don't know how to measure information in the genome. Paradoxically, information to a physicist is related to entropy, which is a quantitative measure of disorder. The lower the entropy, the higher the information content. We do, however, need to be careful how we define entropy, because the standard equation in undergraduate physics courses does not apply to a string of base pairs.

Different meanings of information

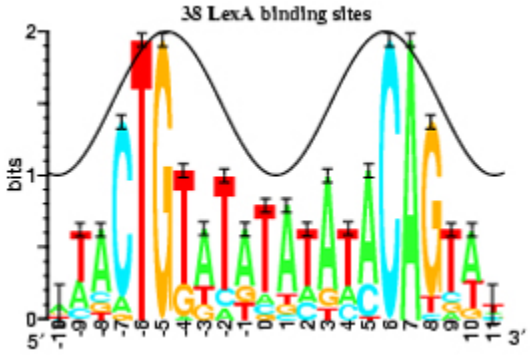


Figure 12: This sequence logo is a compact way of displaying information contained in a piece of genetic material.
Source: © P. P. Papp, D. K. Chattoraj, and T. D. Schneider, Information Analysis of Sequences that Bind the Replication Initiator RepA, J. Mol. Biol., 233, 219-230, 1993.

The introduction of entropy emphasizes a critical point: Information has a different meaning to a biologist than it does to a physicist. Suppose you look at some stretch of the genome, and you find that all four of the bases are present in roughly equal numbers—that is, a given base pair has a 25 percent chance to be present in the chain. To a biologist, this implies the sequence is coding for a protein and is information-rich. But a physicist would say it has high entropy and low information, somewhat like saying that it may or may not rain tomorrow. If you say it will rain tomorrow, you convey a lot of information and very little obvious entropy. The opposite is true in gene sequences. To a biologist, a long string of adenines, AAAAAAAAAAAA, is useless and conveys very little information; but in the physics definition of entropy, this is a very low entropy state. Obviously, the statistical concepts of entropy and the biological concepts of information density are rather different.

The dark matter that makes up a huge fraction of the total genome is still very much *terra incognito*. Entropy density maps indicate that it has a lower information density (in the biologist's use of the word) than the "bright matter" coding DNA. Unraveling the mysteries of the dark matter in the genome will challenge biologists just as much as exploring the cosmological variety challenges astrophysicists.

Section 4: Proteins

Having explored the emergent genome in the form of DNA from a structural and informational perspective, we now move on to the globular polymers called "proteins," the real molecular machines that make things tick. These proteins are the polymers that the DNA codes and are the business end of life. They regulate the highly specific chemical reactions that allow living organisms to live. At 300 K (80°F), the approximate temperature of most living organisms, life processes are characterized by tightly controlled, highly specific chemical reactions that take place at a very high rate. In nonliving matter, highly specific reactions tend to proceed extremely slowly. This slow reaction rate is another result of entropy, since going to a highly specific reaction out of many possible reactions is extremely unlikely. In living systems, these reactions proceed much faster because they are catalyzed by biological proteins called enzymes. It is the catalysis of very unlikely chemical reactions that is the hallmark of living systems.

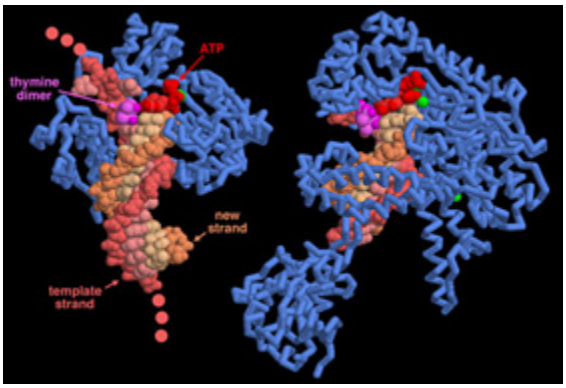


Figure 13: The enzyme on the left has a much easier time reading DNA than the enzyme on the right due to structural details that are difficult to predict from first principles.
Source: © RCSB Protein Data Bank.

The mystery of how these protein polymers do their magical chemical catalysis is basically the domain of chemistry, and we won't pursue it further here. As physicists, we will turn our attention to the emergent structure of biological molecules. We saw in the previous section how DNA, and its cousin RNA, have a relatively simple structure that leads, ultimately, to the most complex phenomena around. In this section, we will ask whether we can use the principles of physics to understand anything about how the folded structure of proteins, which is incredibly detailed and specific to biological processes, arises from their relatively simple chemical composition.

Proteins: the emergence of order from sequence

As polymers go, most proteins are relatively small but much bigger than you might expect is necessary. A typical protein consists of about 100 to 200 monomer links; larger polymers are typically constructed of subunits consisting of smaller balls of single chains. For example, the protein RNA polymerase, which binds to DNA and creates the single-strand polymer RNA, consists (in *E. coli*) of a huge protein with about 500,000 times the mass of a hydrogen atom, divided into five subunits. Despite their small size, folded proteins form exceedingly complex structures. This complexity originates from the large number of monomer units from which the polymers are formed: There are 21 different amino acids. We saw that RNA could form quite complex structures from a choice of four different bases. Imagine the complexity of the structures that can be formed in a protein if you are working with a choice of 21 of them.

Figure 14: As polarized light passes through corn syrup, which is full of right-handed sugar molecules, its plane of polarization is rotated.
Source: © Technical Services Group, MIT Department of Physics.

Note that you can assign a handedness to the bonding pattern within the protein: Some proteins are left-handed, and others are right-handed. Experimentally, it was observed that naturally occurring biological molecules (as opposed to molecules synthesized in the laboratory) could rotate the plane of polarization of light when a beam of light is passed through a solution of the molecule. It is easy to see this by getting some maple syrup from a store and observing what happens when a polarized laser beam passes through it. First, orient an "analyzing" polarizer so that no laser light passes through it. Then put the syrup in the laser's path before the analyzing polarizer. You will notice that some light now passes through the polarizer. The beam polarization (as you look at it propagating toward you) has rotated counterclockwise, or in a right-handed sense using the right-hand rule. The notation is that the sugar in the syrup is dextro-rotary (D-rotary), or right-handed. In the case of the amino acids, all but one are left-handed, or L-rotary. Glycine is the one exception. It has mirror symmetry.

We know how to denote the three-dimensional structure of a protein in a rather concise graphical form. But when you actually see the space-filling picture of a protein—what it would look like if you could see something that small—your physicist's heart must stop in horror. It looks like an ungodly tangled ball. Who in their right mind could possibly be interested in this unkempt beast?

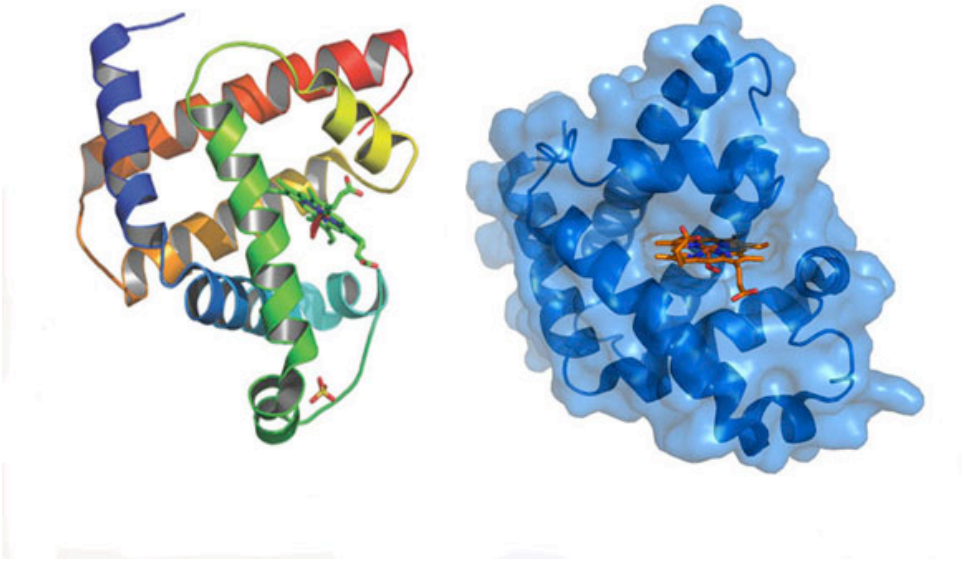


Figure 15: The structure of myoglobin (left) and the form it actually takes in space (right).
Source: © Left: Wikimedia Commons, Public Domain. Author: AzaToth, 27 February 2008; right: Wikimedia Commons Creative Commons Attribution-Share Alike 3.0 Unported License. Author: Thomas Splettstoesser, 10 July 2006.

We can say some general things about protein structure. First, the nearest-neighbor interactions are not totally random; they often show a fair amount of order. Experiments have revealed that nature uses several "motifs" in forming a globular protein, roughly specified by the choice of amino acids which naturally combine to form a structure of interest. These structures, determined primarily by nearest-neighbor interactions, are called "secondary structures." We commonly see three basic secondary structures: the α-helix, the β-strand (these combine into sheets), and the polyproline helix.

We can now begin to roughly build up protein structures, using the secondary structures as building blocks. For example, one of my favorite proteins is myoglobin because it is supposed to be simple. It is not. We can view it as basically a construction of several alpha helices which surround a "prosthetic group," the highly conjugated heme structure used extensively in biology. Biologists often regard

myoglobin as a simple protein. One possible function is to bind oxygen tightly as a storage reservoir in muscle cells. There may be much more to this molecule than meets the eye, however.

As their name indicates, globular proteins are rather spherical in shape. Also, the polarizability of the various amino acids covers quite a large range, and the protein is designed (unless it is membrane-bound) to exist in water, which is highly polar. As biologists see it, the polarizable amino acids are predominantly found in the outer layer of the globular protein, while the non-polar amino acids reside deep in the interior. This arrangement is not because the non-polar amino acids have a strong attraction for one another, but rather because the polar amino acids have strong interactions with water (the so-called hydrophilic effect) and because introducing non-polar residues into water gives rise to a large negative entropy change (the so-called hydrophobic effect). So, physics gives us some insight into structure, through electrostatic interactions and entropy.

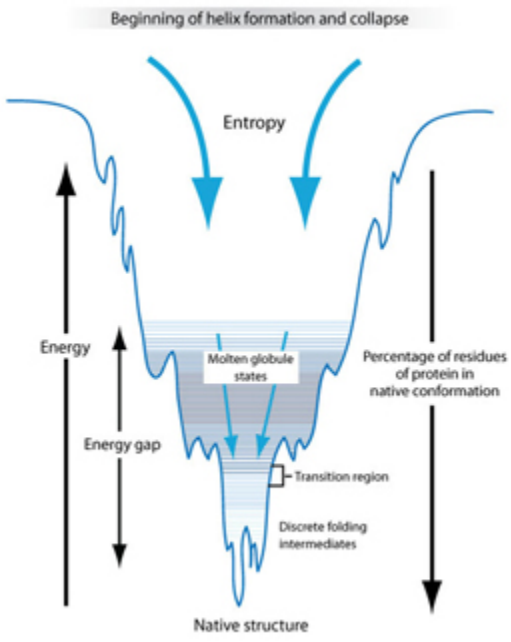


Figure 16: A schematic of how minimizing the free energy of a molecule could lead to protein folding.
Source:

One kind of emergence we wish to stress here is that, although you would think that a polymer consisting of potentially 21 different amino acids for each position would form some sort of a glue-ball, it doesn't. Many proteins in solution seem to fold into rather well-defined three-dimensional shapes. But can we

predict these shapes from the amino acid sequence? This question is known as the "protein-folding problem," and has occupied many physicists over the past 30 some years as they attempt with ever-increasingly powerful computers to solve it. While Peter Wolynes and Jose Onuchich have been able to sketch out some powerful ideas about the general path of the protein folding that make use of the physics concept of free energy, it could well be that solving the puzzle to a precise answer may be impossible.

There may well be a fundamental reason why a precise answer to the folding problem is impossible: because in fact there may be no precise answer! Experiments by Hans Frauenfelder have shown that even for a relatively simple protein like the myoglobin presented in Figure 16, there is not a unique ground state representing a single free energy minimum but rather a distribution of ground states with the same energy, also known as a conformation distribution, which are thermally accessible at 300 K. It is becoming clear that this distribution of states is of supreme importance in protein function, and that the distribution of conformations can be quite extreme; the "landscape" of conformations can be extremely rugged; and within a given local valley, the protein cannot easily move over the landscape to another state. Because of this rugged landscape, a protein might often be found in metastable states: trapped in a low-lying state that is low, but not the lowest, unable to reach the true ground state without climbing over a large energy barrier.

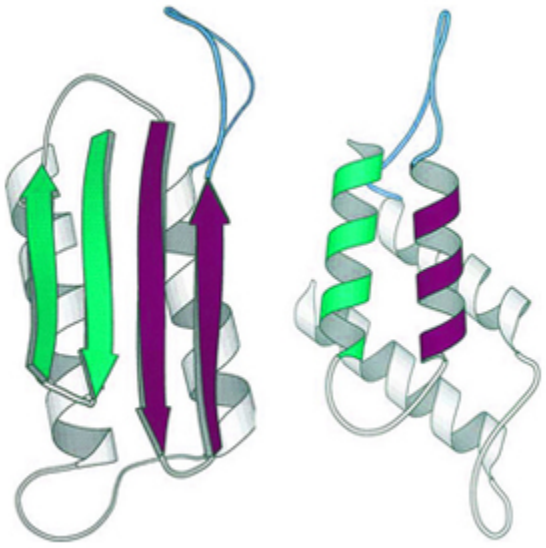


Figure 17: Two possible conformations of a prion protein: on the left as a beta sheet; on the right as an alpha helix.
Source: © Flickr, Creative Commons License. Author: AJC1, 18 April 2007.

An extreme example of this inherent metastability of many protein structures, and the implication to biology, is the class of proteins called "prions." These proteins can fold into two different deep valleys of free energy: as an alpha-helix protein rather like myoglobin, or as a beta-sheet protein. In the alpha-helix conformation, the prion is highly soluble in water; but in the beta-sheet conformation, it tends to aggregate and drop out of solution, forming what are called "amyloid plaques," which are involved with certain forms of dementia. One energy valley leads to a structure that leads to untreatable disease; the other is mostly harmless.

The apparent extreme roughness of biological landscapes, and the problems of ascertaining dynamics on such landscapes, will be one of the fundamental challenges for biological physics and the subject of the next section.

Section 5: Free Energy Landscapes

A certified unsolved problem in physics is why the fundamental physical constants have the values they do. One of the more radical ideas that has been put forward is that there is no deeper meaning. The numbers are what they are because an unaccountable number of alternate universes are forming a landscape of physical constants. We just happen to be in a particular universe where the physical constants have values conducive to form life and eventually evolve organisms who ask such a question.

This idea of a landscape of different universes actually came from biology, evolution theory, in fact, and was first applied to physics by Lee Smolin. Biology inherently deals with landscapes because the biological entities, whether they are molecules, cells, organisms, or ecologies, are inherently heterogeneous and complex. Trying to organize this complexity in a systematic way is beyond challenging. As you saw in our earlier discussion of the protein folding problem, it is easiest to view the folding process as movement on a free energy surface, a landscape of conformations.

Glasses, spin glasses, landscapes

There is a physical system in condensed matter physics that might provide a simpler example of the kind of landscape complexity that is characteristic of biological systems. Glasses are surprisingly interesting physical systems that do not go directly to the lowest free energy state as they cool. Instead, they remain frozen in a very high entropy state. For a physical glass like the windows of your house, the hand-waving explanation for this refusal to crystallize is that the viscosity becomes so large as the system cools, that there is not enough time in the history of the universe to reach the true ground state.

Figure 18: As a glass cools, the viscosity increases so rapidly that the atoms get frozen in a disordered state.
Source: © OHM Equipment, LLC.

A more interesting glass, and one more directly connected to biology, is the spin glass. It actually has no single ground state, which may be true for many proteins as well. The study of spin glasses in condensed matter physics naturally brings in the concepts of rough energy landscapes, similar to those we discussed in the previous section. The energy landscape of a spin glass is modified by interactions within the material. These interactions can be both random and frustrated, an important concept that we will introduce shortly. By drawing an analogy between spin glasses and biological systems, we can establish some overriding principles to help us understand these complex biological structures.

A spin glass is nothing more than a set of spins that interact with each other in a certain way. At the simplest level, a given spin can be pointing either up or down, as we saw in Unit 6; the interaction between two spins depends on their relative orientation. The interaction term J_{ij} specifies how spin i interacts with spin j. Magically, it is possible to arrange the interaction terms between the spins so that the system has a large set of almost equal energy levels, rather than one unique ground state. This phenomenon is called "frustration."

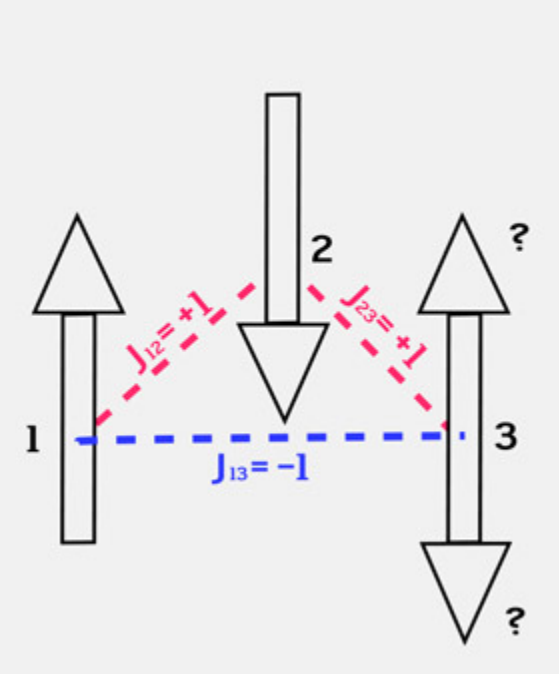


Figure 19: This simple system of three spins is frustrated, and has no clear ground state.
Source:

For a model spin glass, the rule that leads to frustration is very simple. We simply set the interaction term to be +1 if the two spins point in the same direction, and -1 if they point in different directions. If you go

around a closed path in a given arrangement of spins and multiply all the interaction terms together, you will find that if the number is +1, the spins have a unique ground state; and if it is -1, they do not. Figure 19 shows an example of a simple three-spin system that is frustrated. The third spin has contradictory commands to point up and point down. What to do? Note that this kind of a glass is different from the glass in your windows, which would find the true ground state if it just had the time. The spin glass has no ground state, and this is an emergent property.

Frustration arises when there are competing interactions of opposite signs at a site, and implies that there is no global ground energy state but rather a large number of states with nearly the same energy separated by large energy barriers. As an aside, we should note that this is not the first time we've encountered a system with no unique ground state. In Unit 2, systems with spontaneously broken symmetry also had many possible ground states. The difference here is that the ground states of the system with broken symmetry were all connected in field space—on the energy landscape, they are all in the same valley—whereas the nearly equal energy levels in a frustrated system are all isolated in separate valleys with big mountains in between them. The central concept of frustration is extremely important in understanding why a spin glass forms a disordered state at low temperatures, and must play a crucial role in the protein problem as well.

Hierarchical states

Figure 20: A Rubik's Cube is a familiar example of a hierarchical distribution of states.
Source: © Wikimedia Commons, GNU Free Documentation License 1.2. Author: Lars Karlsson (Keqs), 5 January 2007.

Take a look at a Rubik's cube. Suppose you have some random color distribution, and you'd like to go back to the ordered color state. If you could arbitrarily turn any of the colored squares, going back to the desired state would be trivial and exponentially quick. However, the construction of the cube creates large energy barriers between states that are not "close" to the one you are in; you must pass through many of the allowed states in some very slow process in order to arrive where you want to be. This distribution of allowed states that are close in "distance" and forbidden states separated by a large distance is called a hierarchical distribution of states. In biology, this distance can mean many things: how close two configurations of a protein are to each other, or in evolution how far two species are apart on the evolutionary tree. It is a powerful idea, and it came from physics.

To learn anything useful about a hierarchy, you must have some quantitative way to characterize the difference between states in the hierarchy. In a spin glass, we can do this by calculating the overlap between two states, counting up the number of spins that are pointing the same way, and dividing by the total number of spins. States that are similar to one another will have an overlap close to one, while those that are very different will have a value near zero. We can then define the "distance" between two states as one divided by the overlap; so states that are identical are separated by one unit of distance, and states that are completely different are infinitely far apart.

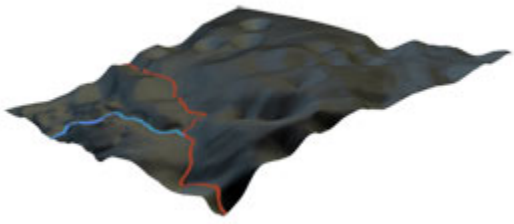


Figure 21: Here, we see two possible paths across an energy landscape strewn with local minima.
Source:

Knowing that the states of a spin glass form a hierarchy, we can ask what mathematical and biological consequences this hierarchy has. Suppose we ask how to pass from one spin state to another. Since the spins interact with one another, with attendant frustration "clashes" occurring between certain configurations, the process of randomly flipping the spins hoping to blunder into the desired final state is likely to be stymied by the high-energy barriers between some of the possible intermediate states. A consistent and logical approach would be to work through the hierarchical tree of states from one

state to another. In this way, one always goes through states that are closely related to one another and hence presumably travels over minimum energy routes. This travel over the space is movement over a landscape. In Figure 21, we show a simulated landscape, two different ways that system might pick its way down the landscape, and the local traps which can serve as metastable sticking points.

In some respects, this landscape picture of system dynamics is more descriptive than useful to the central problems in biological physics that we are discussing in this course. For example, in the protein section, we showed the staggering complexity of the multiple-component molecular machines that facilitate the chemical reactions taking place within our bodies, keeping us alive. The landscape movement we have described so far is driven by pre-existing gradients in free energy, not the time-dependent movement of large components. We believe that what we observe there is the result of billions of years of evolution and the output of complex biological networks.

Section 6: Evolution

Biological evolution remains one of the most contentious fields to the general public, and a dramatic example of emergent phenomena. We have been discussing the remarkable complexity of biology, and it is now natural to ask: How did this incredible complexity emerge on our planet? Perhaps a quote from French Nobel Laureate biologist Jacques Monod can put things into perspective: "Darwin's theory of evolution was the most important theory ever formulated because of its tremendous philosophical, ideological, and political implications." Today, over 150 years after the publication of "On the Origin of the Species," evolution remains hotly debated around the world, but not by most scientists. Even amongst the educated lay audience, except for some cranks, few have doubt about Newton's laws of motion or Einstein's theories of special and general relativity, but about half of the American public don't agree with Darwin's theory of evolution. Surely, physics should be able to clear this up to everybody's satisfaction.

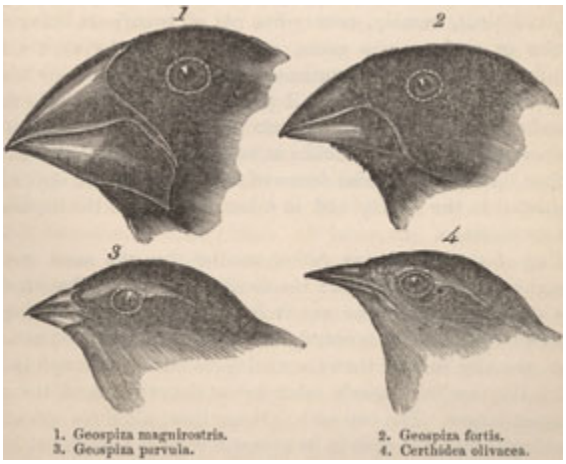


Figure 22: The variation in Galapagos finches inspired Charles Darwin's thinking on evolution, but may evolve too fast for his theory. **Source:** © Public Domain from The Zoology of the Voyage of H.M.S. Beagle, by John Gould. Edited and superintended by Charles Darwin. London: Smith Elder and Co., 1838.

Or maybe not. The problem is that simple theories of Darwinian evolution via random mutations and natural selection give rise to very slow change. Under laboratory conditions, mutations appear at the low rate of one mutated base pair per billion base pairs per generation. Given this low observed rate of mutations, it becomes somewhat problematic to envision evolution via natural selection moving forward to complex organisms such as humans. This became clear as evolution theories tried to move from Darwin's vague and descriptive anecdotes to a firmer mathematical foundation.

Recent work on the Galapagos Islands by the Princeton University biologists Peter and Rosemary Grant revealed something far more startling than the slow evolution of beak sizes. The Grants caught and banded thousands of finches and traced their elaborate lineage, enabling them to document the changes that individual species make in reaction to the environment. During prolonged drought, for instance, beaks may become longer and sharper, to reach the tiniest of seeds. Here is the problem: We are talking about thousands of birds, not millions. We are talking about beaks that change over periods of years, not thousands of years. How can evolution proceed so quickly?

Fitness landscapes and evolution

In our protein section, we discussed the concept of a free energy landscape. This indicates that proteins do not sit quietly in a single free energy minimum, but instead bounce around on a rough landscape of multiple local minima of different biological functional forms. But this idea of a complex energy landscape did not originate from proteins or spin glasses. It actually came from an American mathematical biologist named Sewall Wright who was trying to understand quantitatively how Darwinian evolution could give rise to higher complexity—exactly the problem that has vexed so many people.

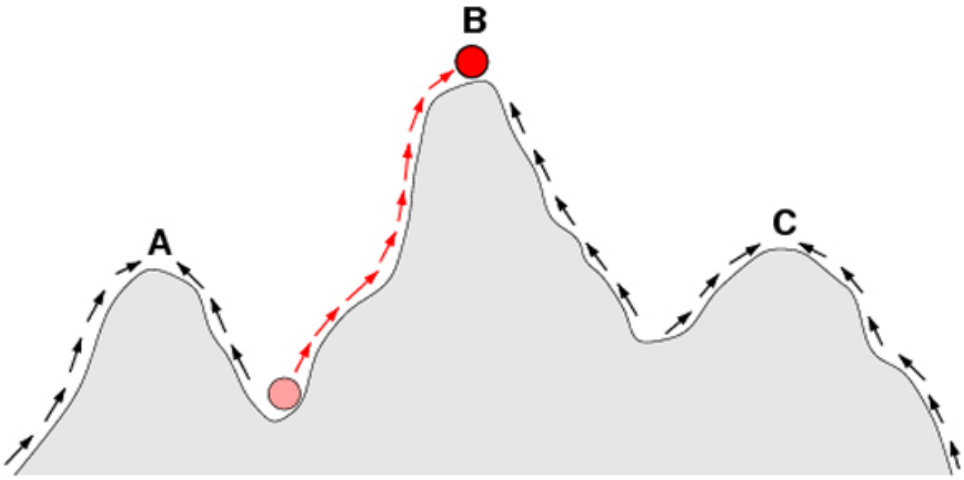


Figure 23: Natural selection can be viewed as movement on a fitness landscape.
Source: © Wikimedia Commons, Public Domain. Author: Wilke, 18 July 2004.

We can put the problem into simple mathematical form. Darwinian evolution is typically believed to be due to the random mutation of genes, which occurs at some very small rate of approximately 10^{-9} mutations/base pair-generation under laboratory conditions. At this rate, a given base pair would undergo a random mutation every billion generations or so. We also believe that the selection pressure—a quantitative measure of the environmental conditions driving evolution—is very small if we are dealing

with a highly optimized genome. The number of mutations that "fix," or are selected to enter the genome, is proportional to the mutation rate times the selection pressure. Thus, the number of "fixed" mutations is very small. A Galapagos finch, a highly evolved creature with a genome optimized for its environment, should not be evolving nearly as rapidly as it does by this formulation.

There is nothing wrong with Darwin's original idea of natural selection. What is wrong is our assumption that the mutation rate is fixed at 10^{-9} mutations/base-pair generation, and more controversially perhaps that the mutations occur at random on the genome, or that evolution proceeds by the accumulation of single base-pair mutations: Perhaps genomic rearrangements and basepair chemical modifications (a process called "epigenetics") are just as important. Further, we are beginning to understand the role of ecological complexity and the size of the populations. The simple fitness landscape of Figure 23 is a vast and misleading simplification. Even in the 1930s, Seawall Wright realized that the dynamics of evolution had to take into account rough fitness landscapes and multiple populations weakly interbreeding across a rough landscape. Figure 24 dating all the way back to 1932, is a remarkably prescient view of where evolution biological physics is heading in the 21st century.

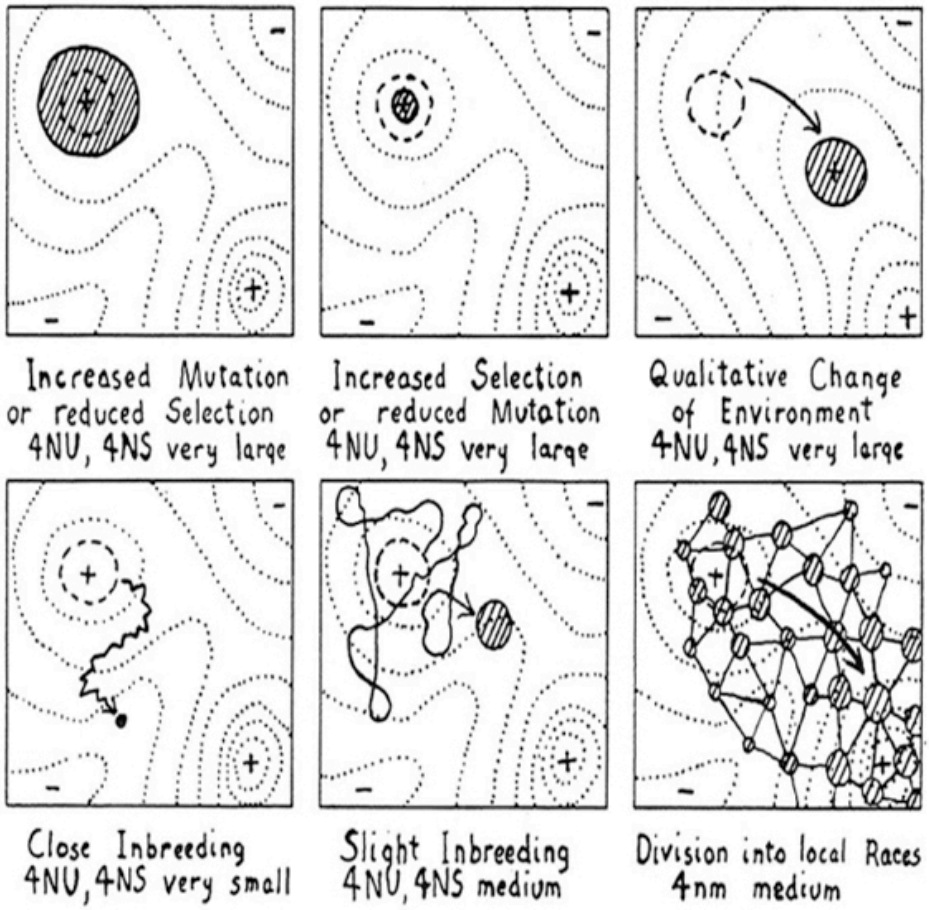


Figure 24: Sewall Wright sketched the path different populations might take on the fitness landscape.
Source: © Sewall Wright, "The Role of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution," Sixth International Congress of Genetics, Brooklyn, NY: Brooklyn Botanical Garden, 1932.

Darwinian evolution in a broader sense is also changing the face of physics as the fundamental concepts flow from biology to physics. Darwinian evolution as modified by recent theories teaches us that it is possible to come to local maxima in fitness in relatively short time frames through the use of deliberate error production and then natural selection amongst the errors (mutants) created. This seems somewhat counterintuitive, but the emergence of complexity from a few simple rules and the deliberate generation of mistakes can be powerfully applied to seemingly intractable problems in computational physics. Applications of Darwinian evolution in computational physics have given rise to the field of evolutionary computing. In evolutionary computing, principles taken from biology are explicitly used. Evolutionary computation uses the same iterative progress that occurs in biology as generations proceed, mutant individuals in the population compete with other members of the population in a guided random search

using parallel processing to achieve the increase in net fitness. To be more specific, the steps required for the digital realization of a genetic algorithm are:

1. A population of digital strings encode candidate solutions (for example, a long, sharp beak) to an optimization problem (needing to adapt to drought conditions).
2. In each generation, the fitness of every string is evaluated, and multiple strings are selected based on their fitness.
3. The strings are recombined and possibly randomly mutated to form a new population.
4. Re-iterate the next generation.

It is possible that by exploring artificial evolution, which came from biology and moved into physics, that we will learn something about the evolutionary algorithms running in biology and the information will flow back to biology.

Evolution and Understanding Disease in the 21st Century

The power influence of evolution is felt in many areas of biology, and we are beginning to understand that the origins of many diseases, most certainly cancer, may lie in evolution and will not be controlled until we understand evolution dynamics and history much better than we do today. For example, shark cartilage is one of the more common "alternative medicines" for cancer. Why? An urban legend suggests that sharks do not get cancers. Even if sharks have lower incidence rates of cancer than *Homo sapiens*, they possess no magic bullet to prevent the disease. However, sharks possess an important characteristic from an evolution perspective: They represent an evolutionary dead-end. Judging from the fossil record, they have evolved very little in 300 million years, and have not attempted to scale the fitness landscape peaks that the mammals eventually conquered.

Figure 25: Cartilage from the fin of the Mako shark.
Source: © www.OrangeBeach.ws.

We can ask two questions based on what we have developed here: Is cancer an inevitable consequence of rapid evolution, and in that sense not a disease at all but a necessary outlier tail of rapid evolution? And is cancer, then, inevitably connected with high evolution rates and high stress conditions and thus impossible to "cure"?

The Prisoner's Dilemma and Evolution

The prisoner's dilemma is a problem in game theory that links cooperation, competition, options, and decision-making in an uncertain environment. Devised by RAND staffers Merrill Flood and Melvin Dresher and formalized by Princeton mathematician Albert Tucker, it involves two suspects for a crime whom police are interrogating separately. Lacking evidence to convict the pair, the police use the incentive of getting out of jail free—or early—to persuade each prisoner to confess and implicate the other. If just one prisoner confesses, he goes free and his partner in crime receives the maximum sentence. If both confess, they will serve half the maximum time. But if both stay silent, each will serve a short stretch for a minor offense.

The dilemma stems from the fact that neither prisoner knows what option the other will choose. By confessing, a prisoner will definitely avoid the maximum sentence. He might avoid serving time altogether; but he might also spend half the maximum inside. If both prisoners say nothing, however, they would serve only minor time.

As with game theory prisoners, so it is with evolutionary biology. A species under stress can stand pat. Or it can mutate—a process that can lead either to death or vibrant new life.

Stress no doubt drives evolution forward, changing the fitness landscapes we have discussed from a basically smooth, flat, and boring plane into a rugged landscape of deep valleys and high peaks. Let us assume that in any local habitat or ecology is a distribution of genomes that includes some high-fitness genomes and some low-fitness genomes. The low-fitness genomes are under stress, but contain the seeds for evolution. We define stress here as something that either directly generates genomic damage, such as ionizing radiation and chemicals that directly attack DNA, viruses, or something that prevents replication of the genome, such as blockage of DNA polymerases or of the topological enzymes required for chromosome replication. Left unchallenged, all these stress inducers will result in the extinction of the quasi-species.

This is the business end of the grand experiment in exploring local fitness peaks and ultimately in generating resistance to stress. The system must evolve in response to the stress, and it must do this by deliberately generating genetic diversity to explore the fitness landscape—or not. Viewed in the perspective of game theory's prisoner's dilemma (see sidebar), the silent option under stress is not to evolve—to go down the senescent pathway and thus not attempt to propagate. Turning on mutational

mechanisms, in contrast, is a defection, in the sense that it leads potentially to genomes which can propagate even in what should be lethal conditions and could, in principle, lead to the destruction of the organism: disease followed by death, which would seem to be very counterproductive. But it may well be a risk that the system is willing to make. If ignition of mutator genes and evolution to a new local maximum of fitness increases the average fitness of the group, then the inevitable loss of some individuals whose genome is mutated into a fitness valley is an acceptable cost.

Section 7: Networks

The complex biological molecules we have spent the previous sections trying to understand are the building blocks of life, but it is far from obvious to put these building blocks together into a coherent whole. Biological molecules, as well as cells and complete organisms, are organized in complex networks. A network is defined as a system in which information flows into nodes, is processed, and then flows back out. The network's output is a function of both the inputs and a series of edges that are the bidirectional paths of information flow between the nodes. The theory and practice of networks is a vast subject, and with one ultimate goal of understanding that greatest of mysteries, the human brain. We will return to the brain and its neural networks in the next section. For now, we will discuss the more prosaic networks in living organisms, which are still complex enough to be very intimidating.

It isn't obvious when you look at a cell that a network exists there. The cytoplasm of a living cell is a very dynamic entity, but at least at first glance seems to basically be a bag of biological molecules mixed chaotically together. It is somewhat of a shock to realize that this bag of molecules actually contains a huge number of highly specific biological networks all operating under tight control. For example, when an epithelial cell moving across a substrate, patterns of specific molecules drive the motion of the cell's internal skeleton. When these molecules are tagged with a protein that glows red, displaying the collective molecular motion under a microscope, a very complex and interactive set of networks appears.

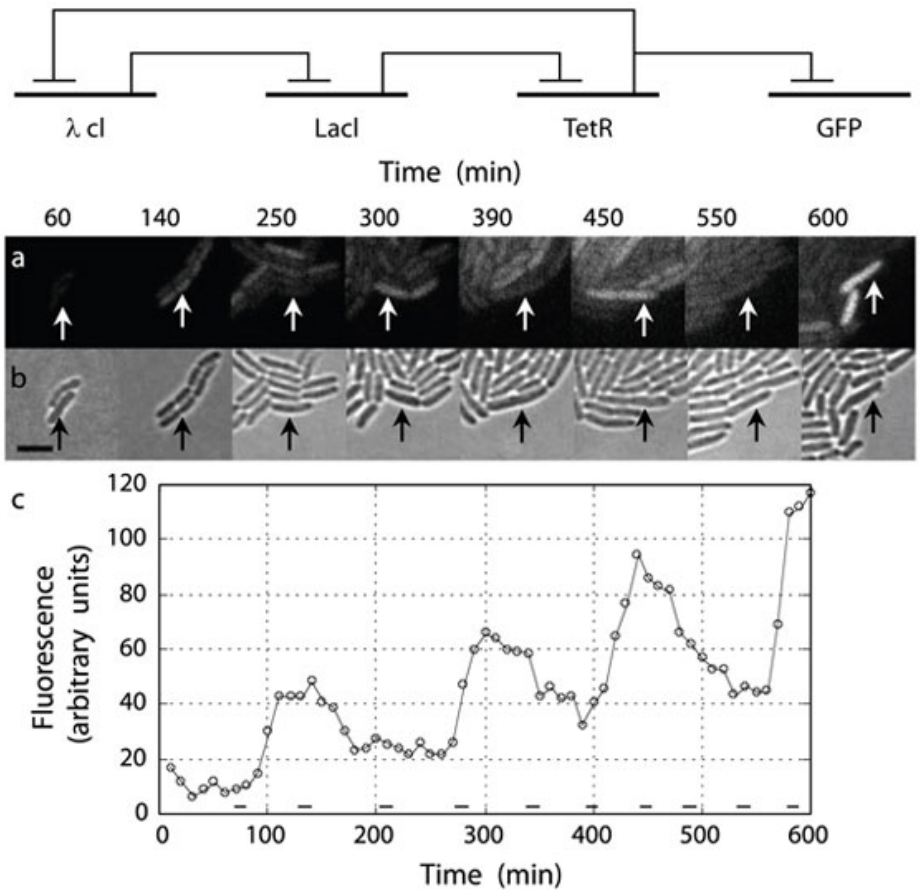


Figure 26: The circuit diagram (top), bacterial population (center), and plot of the dynamics (bottom) of the repressilator, an example of a simple synthetic biological network.
Source: © Top: Wikimedia Commons, Public Domain. Author: Timreid, 26 February 2007. Center and bottom: Reprinted by permission from Macmillan Publishers Ltd: Nature 403, 335-338 (20 January 2000).

The emergent network of the cytoplasm is a system of interconnected units. Each unit has at least an input and an output, and some sort of a control input which can modulate the relationship between the input and the output. Networks can be analog, which means that in principle the inputs and outputs are continuous functions of some variable; or they can be digital, which means that they have finite values, typically 1 or 0 for a binary system. The computer on which this text was typed is a digital network consisting of binary logic gates, while the person who typed the text is an analog network.

There are many different kinds of biological networks, and they cover a huge range of length scales, from the submicron (a micron is a millionth of a meter) to the scales spanning the Earth. Outside of neural networks, the most important ones are (roughly in order of increasing abstraction):

1. Metabolite networks: These networks control how a cell turns food (in the form of sugar) into energy that it can use to function. Enzymes (proteins) are the nodes, and the smaller molecules that represent the flow of chemical energy in the cell are the edges.

2. Signal transduction networks: These networks transfer information from outside the cell into its interior, and translate that information into a set of instructions for activity within the cell. Proteins are the nodes, typically proteins called kineases, and diffusible small signaling molecules which have been chemically modified are the edges. This is a huge class of networks, ranging from networks that process sensory stimuli to chemical inputs such as hormones.

3. Transcriptional regulatory networks: These networks determine how genes are turned on and off (or modulated).

4. Interorganism networks: This is a very broad term that encompasses everything from the coordinated behavior of a group of bacteria to complex ecologies. The nodes are individual cells, and the edges are the many different physical ways that cells can interact with each other.

There are probably fundamental network design principles that must be obeyed independent of their biological or manmade (which is still biological) origin if the network is to be stable to perturbations. Instability in a network is not generally viewed as a good thing, although there are exceptions to this rule. For example, the aerodynamics of most modern fighter jets makes the plane inherently unstable. This sounds like a very bad thing, except that it makes the fighter extremely adaptive to direction changes. Modern computers can constantly monitor and instantaneously correct the instability, so we end up with aircraft that are far more maneuverable—and thus more effective fighters—than the ordinary, stable variety.

The kind of stability issues the fighter jet and other similar networks face are deterministic, and can be modeled by ordinary differential equations that are straightforward to write down. One might then imagine designing a network based on a set of these equations. One of the pioneering exercises in designing "from scratch" was the work of Elowitz and Leibler of an oscillating gene expression pattern. It is sobering to understand the depth of understanding that was necessary to have made this simple oscillator work. For an electrical engineer, it is straightforward to design an oscillator following some basic rules of electromagnetism. However, in a biological network, the parameters are much less cleanly defined. Despite the inherent challenges, we now have a basic set of biological modules that is being developed in the new field of "synthetic biology," which is a mix of physics, engineering, and biology that exploits our knowledge of networks to design new functional biological "circuits." Figure 27 shows an example of a biological "film" consisting of bacteria. To do this, a gene was inserted into *E. coli* that coded for a protein that causes the bacteria to make a black pigment. The pigment production was coupled to a light sensor,

so that pigment would be made only in the dark. The group used stencils to pattern light exposure and produce bacterial photography.

Figure 27: Portrait of the University of Texas at Austin (UT Austin)/University of California, San Francisco (UCSF) Synthetic Biology team. The left panel shows the projected image of the students and professors from UT Austin and UCSF who participated in the project, and the right panel shows the resulting bacterial photo.
Source: © Aaron Chevalier and the University of Texas at Austin.

In addition to deterministic stability issues in biological networks, there is also the issue of stability in the presence of noise. For example, at the macro-scale, the sensory network of the dog's nose is about a million times more sensitive than a human noise. Despite this extreme sensitivity, the dog nose is not overwhelmed by the presence of an enormous background of other molecules. That is, the dog nose sensory network is extremely good at noise rejection. At the micro-scale of the single cell, the very finite number of molecules actually involved in the network node edges leads to statistical noise that can either confound the network's stability or increase its sensitivity. The dog has clearly resolved this issue to its benefit, but it remains a problem in the design of synthetic biological networks.

The obvious end goal of synthetic biology could be something truly astonishing: synthetic life. The key to synthetic life, if it is indeed achieved, will be our ability to harness biological networks. So far, scientists have synthesized genetic material and other important biological molecules from scratch, but have not put the pieces together into a complete living organism. The feat hailed by the media as the so-called creation of a "new form of life" by Craig Venter is something of a misnomer. While Venter's synthesis of a functional bacterial chromosome of one million base pairs was a fantastic technical achievement, it is very far from synthetic life, as the new chromosome was inserted into an already functioning cell consisting of a enormous collection of interacting networks, which we neither can understand nor can reproduce. Until we can understand the emergence of life from the networks of biology, we will remain very far from achieving synthetic life.

Section 8: The Emergence of the Mind

So far, we have wrestled with the structural diversity of proteins and its relationship to the free energy landscape, and we have tried to find some of the unifying and emergent properties of evolution that might explain the diversity of life and the increase in complexity. We have also taken a look at how the biological networks necessary to bind a collection of inanimate objects into a living system emerge. At the highest level lies the greatest mystery of biological physics: the emergence of the mind from a collection of communicating cells.

Figure 28: Ruby-throated hummingbird.
Source: © Steve Maslowski, U.S. Fish and Wildlife Service.

We started our discussion of biological physics by considering a chicken egg. Birds lay eggs, so let's consider a bird: the ruby-throated hummingbird presented in all its glory in Figure 28. About 7 cm long and weighing about 5 grams, this bird is capable of some rather amazing biophysical things. Its wings beat about 50 times each second, and they rotate around their central axis through almost 180 degrees, allowing the bird to fly backwards and forwards and hover. A lot of these fascinating mechanical properties can be considered the subject of biological physics.

But there is far more to these hummingbirds than just flying ability. They live for about nine years, spend their summers in the northern parts of North America and their winters in tropical Central America. So, this small animal can navigate in the fall over thousands of miles, including over hundreds of miles of open water, to certain locations and then return in the spring to the region it was born. How is it that a tiny hummingbird can do all this remarkable navigation? The advancement in capabilities of the digital

computer over the past 30 years has been truly staggering, yet it pales against what the hummingbird's brain can do. The human brain is far more impressive. Why can a couple of pounds of neurons drawing a few watts of chemical power with an apparent clock speed of maybe a kilohertz at best do certain tasks far better than a machine the size of a large truck running megawatts of power? And at a much more troubling level, why do we speak of the soul of a person when no one at this point would seriously ascribe any sense of self-recognition to one of our biggest computers? We seem to be missing something very fundamental.

Traditional computers vs. biology

We have moved into the computer age via the pathway pioneered by British mathematician Alan Turing, whom we first met in the introduction to this unit. Our modern-day computers all basically use the model described in Figure 29, coupled with the idea that any number is to be presented by bits in a binary representation. We have made things much faster than those early computers, but the basic idea has not changed. Even the quantum computers promised in Unit 7 keep the same basic design, replacing binary bits with more powerful qubits.

The Strange Case of Phineas Gage

Phineas Gage's brain injury.
Source: © Wikimedia Commons, Public Domain.

On September 13, 1848, a 25-year-old man named Phineas Gage was tamping a blasting charge with a steel rod about 2 cm in diameter and about 1 meter long. When he mistakenly ignited the charge, the rod shot through his left cheek, taking out his left eye in the process, went through his brain, exited through the top of his head, and landed some meters way. Amazingly, Phineas never really lost consciousness and lived another 13 years. His personality changed for a while into that of a "foul-mouthed, ill-mannered liar given to extravagant schemes that were never followed through." However, even that aberration stopped after a short time, and he lived a productive life and traveled widely.

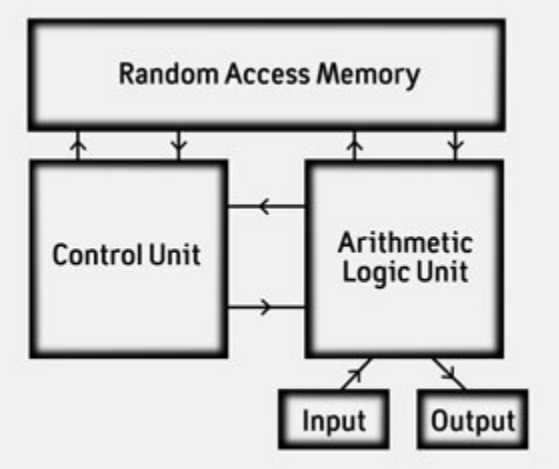


Figure 29: Schematic of a modern digital computer.
Source:

But this is not how biology has developed its own computers. The basic design has four major flaws as far as biology is concerned:

1. The machine must be told in advance, in great error-free detail, the steps needed to perform the algorithm.
2. Data must be clean; the potential loss of a single bit can crash the code.
3. The hardware must be protected and robust; one broken lead and the machine can crash.
4. There is an exact correspondence between a bit of data and a hardware location: The information in the machine is localized.

None of this is any good for a biological system. As far as biology is concerned, our computers are evolutionary dead-ends. We started this unit by considering the fragility of the egg in a large fall. Yet, as the example of Phineas Gage in the side bar shows, our brain can take enormous abuse and remain basically functional. I challenged you initially to drop the possibly cooked egg and see what happens. Now I challenge you to take a 2 cm diameter steel rod and thrust it through your laptop with great force, then try to surf the web.

The brain of a nematode

The human brain is probably the most complex structure in the universe that we know, but not only humans have brains. The adult hermaphrodite of the "lowly" nematode *C. elegans* consists of only 959 cells; yet when you watch it navigating around on an agar plate, it certainly seems to be computing

something based on its sensory input. The creature displays an astonishingly wide range of behavior: locomotion, foraging, feeding, defecation, egg laying, larva formation, and sensory responses to touch, smell, taste, and temperature, as well as some complex behaviors like mating, social behavior, and learning and memory. It would be quite hard to build a digital computer that could do all that, and certainly impossible to pack it into a tube about 1 mm long and a 100 microns in diameter that can reproduce itself.

The *C. elegans* doesn't have a brain per se, but it does have about 302 information-carrying neurons that form approximately 7,000 synapses. We believe that any real brain capable of making some sort of a computation, as opposed to the collective behavior seen in single-celled organisms, must consist of neurons that transfer information. That information is not transferred to some sort of a central processing unit. Biological computers are systems of interconnected cells that transfer and process information. The network of neurons in *C. elegans* displays the common feature in interconnectivity: the synaptic connections formed by the neurons.

The brain versus the computer

I want to concentrate on one thing here: how differently the brain, even the pseudo-brain of *C. elegans*, is "wired" from the computer that you're using to read this web page. Your computer has well-defined regions where critical functions take place: a section of random access memory (RAM) and a central processing unit (CPU). Each part is quite distinct, and buses transfer binary data between the different sections. Take out a single bus line or damage one of the RAM chips, and the system shuts down.

Brains in biology seem to have evolved in a different way. First, they are spatially diffuse. The computer is basically a two-dimensional device. Brains at every level seem to be basically three-dimensional. The interconnection takes place not via a bus, but rather through a vast network of input-output synaptic connections. For example, *C. elegans* has roughly 20 interconnects per neuron. In the human brain, we believe that the number is on the order of 10^3. Since the human brain has around 10^{12} neurons, the number of interconnects is on the order of 10^{15}—a huge number.

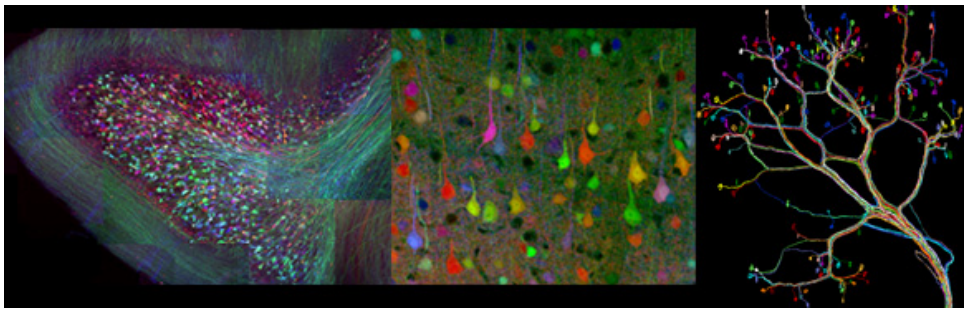


Figure 30: Rainbow images showing individual neurons fluorescing in different colors. By tracking the neurons through stacks of slices, we can follow each neuron's complex branching structure to create the treelike structures in the image on the right.
Source: © Jeff Lichtman, Center for Brain Science, Harvard University.

It would be a mistake to think that the 10^{12} neurons in the brain correspond to about 10^{12} bits of information, or about 100 Gigabytes. The number is much higher, because of the three-dimensional interconnections linking each neuron with about 10^3 other neurons. Returning to our theme of spin glasses, we can estimate the information capacity by making the simple assumption that each neuron can be like a spin which is either up or down depending on its storage of a bit of information. This means that the total number of differing configurations of the brain is on the order of $2^{10^{15}}$, an absurdly huge number, far greater than even the number of atoms in the universe. We can only assume that the brains of living organisms emerged as they exploited this immense 3-D information capacity owing to the ability of communities of cells to form neuronal interconnections throughout space.

How does the brain reason?

Given the large information capacity of even a small network of neurons and the fact that the human brain's capacity exceeds our ability to comprehend it, the next question is: How does a brain reason? As usual, we need to start by defining what we're talking about. According to the *Oxford English Dictionary*, "reasoning" is "find[ing] a solution to a problem by considering possible options." I suppose this dodges the question of the emergent property of consciousness, but I don't see this problem being solved any time soon, although I hope I am wrong.

The Dilemma of the Traveling Salesman—and the Hummingbird

Suppose a salesman must travel to N cities in the course of a trip. Naturally, he wants to travel through each city only once. In what order should he visit the cities? If N is some small number, the problem is trivial; but as N gets larger, the number of combinations to be considered blows up. To travel to the 15 cities shown in Figure 31, the salesman must consider 14!/2 or 43,589,145,600 different combinations. This is somewhat doable by brute force on a laptop computer; but if the number of cities, N, reaches 30, then the number of different combinations becomes about 10^{30}, clearly impossibly difficult to solve by brute force. As it navigates north in the spring, the hummingbird wants to pass through N locations where it will find flowers and to avoid traveling through a location again because all the flowers in that location have been drained of their nectar. In what order should it visit the locations?

The hummingbird has a big problem, essentially asking itself: How shall I fly back to a place I was at six months ago that is thousands of miles away from where I am now? Presumably, the bird uses different physics than that of a traditional computer, because the information content that the bird has to sort out would cause it to fail catastrophically. So, we finally have the problem that perhaps physics can attack and clarify in the 21st century: How can a set of interacting neurons with a deep level of interconnects take previously stored information and determine an optimal solution to a problem it has not yet seen?

The hummingbird faces a problem rather reminiscent of the traveling salesman problem, explained in the sidebar. To choose the correct locations to pass through on its springtime journey north, it must consider a number of combinations far beyond the power of any computer system to resolve. How does the hummingbird do it? Is it magic?

Physics shows that it isn't magic. As we have previously discussed, while a protein may fold or a species play with its genome in an almost uncountable number of ways, basic free energy minima schemes lead quite efficiently to a vastly smaller set of combinations that are roughly optimal. Nature doesn’t necessarily find the "best" solution, but it seems able to efficiently find a subset of solutions that works well enough. In the case of the traveling salesman problem, the vast combinatorics interconnects of a neural network of many neurons provides exactly the kind of search over a free energy surface that we need.

The "reasoning" ability of neural networks

We have discussed how landscapes—either fitness landscapes or free energy landscapes—can give rise to vastly complex surfaces with local minima representing some particular desired state. John Hopfield, a theoretical physicist at Princeton University, has explored ways for a system to find these minima. The three basic ideas below highlight how biological computers differ from their electronic counterparts:

1. Neural networks are highly interconnected. This interaction network can be characterized by a matrix, which tabulates the interaction between each pair of neurons.

2. Neurons interact in a nonlinear analog way. That is, the interconnection interaction is not an "all or nothing" matter, but a graded interaction where the firing rate of neurons varies smoothly with the input potential.

3. An "energy function" can be constructed that allows us to understand the collective (or emergent) dynamics of the neuron network as it moves over the information landscapes and finds local minima that represent effective solutions to the problem.

Figure 31: An optimal traveling salesman problem (TSP) tour through Germany's 15 largest cities. It is the shortest among 43,589,145,600 possible tours visiting each city exactly once.
Source: © Wikipedia, Creative Commons Attribution-ShareAlike License.

Hopfield and molecular biologist David Tank set out to make an analogy between neural networks and the energy network of a glassy system characterized by a large number of degrees of freedom. Following the three principles outlined above, they used this analogy to write an equation for the free energy of a neural network in terms of the interaction between each pair of neurons, the threshold for each neuron to self-

fire, and the potential for each of the neurons in the network. They also recognized that the interaction between pairs of neurons can change with time as the neural network learns.

The solution to a problem such as the traveling salesman problem emerges in the neural network as interaction strengths between the neutrons are adjusted to minimize the free energy equation. The flow of the neuron states during the computation can be mapped onto a flow on a free energy surface, similar to the flow of a spin glass toward its ground state or natural selection on a fitness landscape (but in the opposite direction). Clearly, quite complex and emergent neuronal dynamics can evolve with even the simple system we are considering here.

Hopfield and Tank showed that this neuronal map has quite impressive "reasoning" ability. A set of 900 neurons encoded to solve a 30-city traveling salesman problem was able to find 10^7 "best" solutions out of the 10^{30} possible solutions, a rejection ratio of 10^{23} in just a few clock cycles of the neural network.

Although we clearly are a long way from understanding the emergent nature of consciousness, this example reveals the immense computational power of neural networks. Surely, one of the grand challenges in 21st century physics will be to move from these simple physical models derived from very concrete physics concepts to the vastly more complex terrain of the human brain.

Section 9: Further Reading

- Howard Berg, movies of bacterial motion: http://webmac.rowland.org/labs/bacteria/index_movies.html.
- Michael B. Elowitz and Stanislas Leibler, "A Synthetic Oscillatory Network of Transcriptional Regulators," *Nature*, 2000 Jan. 20; 403(6767): 335-8.
- Robert Laughlin, "A Different Universe: Reinventing Physics from the Bottom Down," *Basic Books*, 2006.
- The National Human Genome Research Institute: http://www.genome.gov.

Glossary

Brownian motion: Brownian motion is the seemingly random motion that a small particle (say, a grain of pollen) undergoes when it is suspended in a liquid. First documented by Scottish botanist Robert Brown, it was explained by Einstein as the result of the pollen grain being buffeted by the random motion of molecules in the liquid. Brownian motion is similar to the random walk, and the equations governing Brownian motion can be derived from the random walk equations by making the step size infinitely small along with a few other mathematical assumptions.

catalyze: Some chemical reactions proceed much more quickly in the presence of a particular molecule than they do when that molecule is absent. The molecule, called a "catalyst," is said to catalyze the reaction.

complex adaptive system (CAM): A complex adaptive system, or CAM, is a population of individual components that react to both their environments and to one another. The state of the population is constantly evolving, and emergent behavior often appears. Biological and ecological systems are examples of complex adaptive systems, as are the Internet, human society, and the power grid.

conformation distribution: The internal potential energy that a molecule has depends on its physical structure, or conformation. Molecules tend toward structures that minimize their potential energy. Sometimes there is not a single, unique minimum energy conformation. In this case, the conformation distribution is the set of lowest energy states that a molecule can occupy.

energy landscape: The energy of a physical system can be represented by a mathematical function that depends on several variables. The energy landscape that the system occupies is this function plotted as a hypersurface in space that is one dimension higher than the relevant number of variables. If the energy depends on one variable, then the energy landscape is a line drawn in a two-dimensional plane. If the energy depends on two variables, the energy landscape is a two-dimensional surface embedded in three-dimensional space that can look like mountains and valleys in a real landscape that one might encounter on the Earth's surface. The ground state of a system is the lowest point on the energy landscape.

entropy: Entropy is a quantitative measure of the amount of order in a system. In statistical mechanics, a system's entropy is proportional to the logarithm of the number of states available to the system. If we

consider a collection of water molecules, its entropy is greater at room temperature, when the molecules are bouncing around in a gaseous phase, than at very low temperatures, when the molecules are lined up in a rigid crystal structure.

enzymes: Enzymes are proteins that catalyze chemical reactions in biological systems.

fitness landscape: The fitness landscape is a visual representation of how well adapted different genotypes are to a set of environmental conditions. Each possible genotype occupies a point on the landscape. The distance between each pair of genotypes is related to how similar they are, and the height of each point indicates how well adapted that genotype is.

frustrated: A physical system is frustrated if it has no well-defined ground state because there are competing interactions among the pieces of the system that cannot simultaneously be at an energy minimum. A simple example is a system of three spins. If the interaction energy between two spins is lowest when they point in opposite directions, the ground state of a pair of spins is clearly for the two spins to point in opposite directions. If a third spin is added, it is pulled in opposite directions attempting to minimize its interaction with the other two.

genome: An organism's genome is the complete set of genetic information required to reproduce and maintain that organism in a living state.

ground state: The ground state of a physical system is the lowest energy state it can occupy. For example, a hydrogen atom is in its ground state when its electron occupies the lowest available energy level.

handedness: Handedness, also called "chirality," is a directional property that physical systems may exhibit. A system is "right handed" if it twists in the direction in which the fingers of your right hand curl if your thumb is directed along the natural axis defined by the system. Most naturally occurring sugar molecules are right handed. Fundamental particles with spin also exhibit chirality. In this case, the twist is defined by the particle's spin, and the natural axis by the direction in which the particle is moving. Electrons produced in beta-decay are nearly always left handed.

metastable: A metastable state has a higher energy than the ground state that a physical system can become trapped in for some length of time. A simple example is a ball sitting on a hilltop. The ball's energy would be lower if it rolled down the hill; but unless something disturbs it, it will remain where it is. Metastable states of atoms are put to use in atomic clocks because they are long lived, and therefore

correspond to a clock frequency that can be known very precisely. In biological physics, valleys in the energy landscape correspond to metastable states, as do low-lying peaks in the fitness landscape.

monomer: A monomer is a small molecule that can bind to other like molecules to form a polymer. The amino acids that make up proteins are examples of monomers.

polar: A polar molecule has a nonzero electric dipole moment, so it has a side that is positively charged and a side that is negatively charged.

polarizability: Some atoms and molecules that have no electric dipole moment in an electrically neutral environment will develop one in an electric field. The polarizability of an atom or molecule is a quantity that describes how susceptible it is to this effect.

polarization: The polarization of a wave is the direction in which it is oscillating. The simplest type of polarization is linear, transverse polarization. Linear means that the wave oscillation is confined along a single axis, and transverse means that the wave is oscillating in a direction perpendicular to its direction of travel. Laser light is most commonly a wave with linear, transverse polarization. If the laser beam travels along the x-axis, its electric field will oscillate either in the y-direction or in the z-direction. Gravitational waves also have transverse polarization, but have a more complicated oscillation pattern than laser light.

polymer: A polymer is a large molecule that is made up of many repeating structural units, typically simple, light molecules, linked together. Proteins are polymers made up of amino acids. See: monomer.

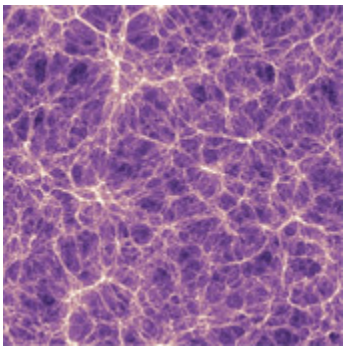
random walk: The random walk is the trajectory that arises when an object moves in steps that are all the same length, but in random directions. The path of a molecule in a gas follows a random walk, with the step size determined by how far (on average) the molecule can travel before it collides with something and changes direction. The behavior of many diverse systems can be modeled as a random walk, including the path of an animal searching for food, fluctuating stock prices, and the diffusion of a drop of food coloring placed in a bowl of water.

Second Law of Thermodynamics: The second law of thermodynamics states that the entropy of an isolated system will either increase or remain the same over time. This is why heat flows from a hot object to a cold object, but not the other way; and why it's easy to dissolve salt in water, but not so easy to get the salt back out again.

Turing machine: In 1937, Alan Turing outlined the details of the Turing machine in a paper investigating the possibilities and limits of machine computation. The machine is an idealized computing device that consists, in its simplest form, of a tape divided up into cells that are processed by an active element called a "head." The cells can be in one of two states. The head moves along the tape, changing the cells from one state to the other and moving either forward or backward according to a set of predetermined instructions. Turing machines can be described with a set of simple mathematical equations that allowed scientists to understand many of the basic properties of digital computing long before the first modern computer was built.

Physics for the 21st Century

Unit 10: Dark Matter



© Raul Angulo, Max Planck Institute for Astrophysics.

Unit Overview

Most of the mass in galaxies like our own Milky Way does not reside in the stars and gas that can be directly observed with telescopes. Rather, around 90 percent of the mass in a typical galaxy is "dark matter," a substance that has so far evaded direct detection. How can scientists make this astonishing claim? Because, although we have yet to detect dark matter, we can infer its existence from the gravitational pull it exerts on the luminous material we *can* see. Another facet of the dark matter problem comes at larger scales, where the total amount of mass in the universe exceeds the inventory of atoms we think were made in the Big Bang. A third indication of something missing comes from the evolution of the large-scale structure in the Universe, where fluctuations in the dark matter density are needed to seed the formation of the tendrils and filaments of galaxies we see in observations. So what is dark matter, and how might we find out? Determining the nature and distribution of dark matter is one of the most pressing (and most interesting!) open questions in modern science—it resides at the interface of particle physics, astrophysics, and gravity. Many candidates for dark matter have been suggested, from the ghostly axion (particles with a tiny amount of mass) to Weakly Interacting Massive Particles (WIMPs) that weigh in at 100 times the proton's mass. In this unit, we shall review the observational and theoretical evidence for dark matter, and describe the attempts that are under way to find it.

Content for This Unit

Sections:

Section 1: Introduction

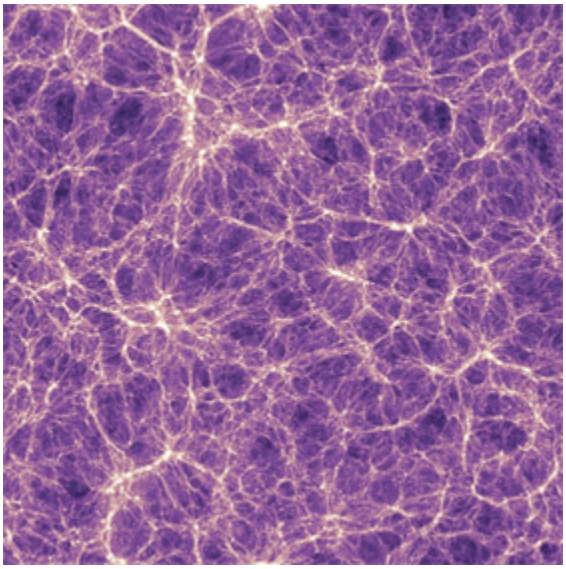


Figure 1: Distribution of dark matter in the universe.
Source: © Raul Angulo, Max Planck Institute for Astrophysics.

Dark matter is something beyond the stuff we encounter here on Earth. We all consist of neutrons, protons, and electrons, and our particle physics experiments with cosmic rays and accelerators tell us that a whole set of particles interact with each other to make up the world we see. As we learned in Units 1 and 2, the Standard Model describes these known particles and their interactions. But careful astronomical measurements, computer-based simulations, and nuclear theory calculations have all led us to believe that the particles described by the Standard Model account for only 4 percent of the mass of the universe. What makes up the missing 96 percent? Physicists believe, based on cosmological measurements described in this unit and Unit 11, that 23 percent is dark matter and 73 percent is dark energy. Dark energy and dark matter are very different. We shall learn about dark energy in Unit 11. Here, we focus on dark matter.

The first evidence of dark matter appeared in the 1930s, when astronomer Fritz Zwicky noticed that the motion of galaxies bound together by gravity was not consistent with the laws of gravity we learned about in Unit 3 unless there was a lot more matter in the galaxy cluster than he could see with his telescope. Development of more powerful and more precise theoretical and experimental tools in subsequent decades strengthened the case for dark matter. By the 1990s, dark matter was required to explain not just the motion of galaxies, but also how those galaxies and other large structures in the universe form,

and the detailed pattern of temperature fluctuations in the cosmic microwave background radiation left over from the early universe.

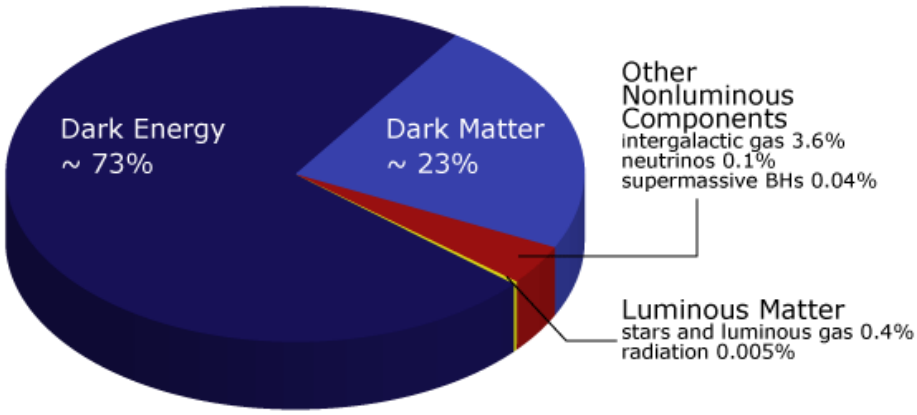


Figure 2: The composition of the universe, with 96 percent invisible and unfamiliar.
Source:

With these distinct reasons to believe that dark matter is a real part of our universe, scientists struggled to understand what comprises dark matter. Could it consist of familiar objects like brown dwarfs and large planets—made of the stuff of the Standard Model, but not emitting light and therefore invisible to astronomers? Both theory and experiment eventually pointed away from this simple explanation, strongly suggesting that dark matter is something entirely new and different. A generation of experiments was developed to look for new types of particles—beyond the Standard Model—that could account for some or all of dark matter. In parallel, theorists have developed creative visions of what new physics could explain about the motion of galaxies, large scale structure, and variations in the cosmic microwave background in one fell swoop.

The process of discovery has not run smoothly. It has survived successive periods of disinterest, progressing as new technologies developed, scientists made fresh observations in disparate fields, and general scientific interest in the topic increased. In this unit, we describe why we think dark matter exists, its role in determining the structure of galaxies and clusters of galaxies, and how it connects with particle physics. Finally, we discuss the ongoing quest to determine what dark matter is made of in both theory and experiment.

Dark matter and gravity

The connection between dark matter and gravity bears special mention because it is the one thing about dark matter of which physicists are certain. Everything we know about dark matter so far comes from astronomy. The astronomical measurements deal exclusively with the way in which dark matter interacts gravitationally. We have two ways of studying the effects of gravity on astronomical bodies: We can either see how a group of astronomical objects moves under the influence of gravity or measure how gravitation changes the way in which light travels. Experimentally, we have no reason to believe that dark matter interacts with normal matter or with itself in any way other than via gravitation, although there is a great deal of theoretical speculation to the contrary.

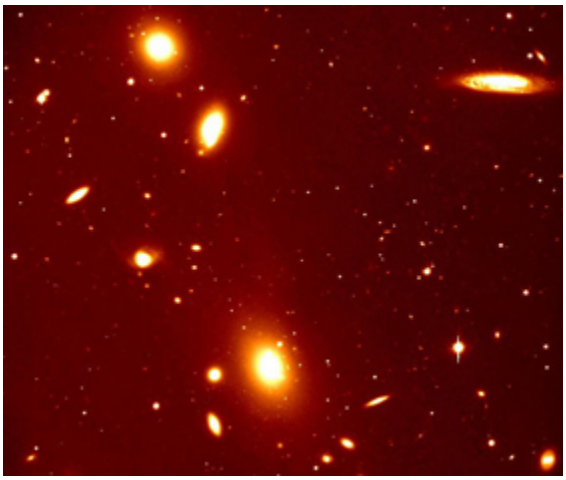


Figure 3: Clusters of galaxies contain significantly more dark matter than visible matter.
Source: © ESO.

The effects of dark matter did not become apparent until astronomers began to study the motion of galaxies and clusters of galaxies. Since a galaxy that measures 150,000 light-years across contains 2 to 10 times as much dark matter as normal matter, the gravity from the dark matter plays a large role in its movements. However, the normal matter is clumped in solar systems (stars and planets), while the dark matter is spread out. Typical solar systems are about 10 light-hours across and are separated from each other by about 2 light-years. So, in conventional terms, the galaxy consists of mostly empty space interspersed with very dense clumps of normal matter.

Since a solar system contains far more normal matter than dark matter (2×10^{30} kilograms vs. 9×10^{9} kilograms), dark matter plays an insignificant role in shaping our solar system. At the next level of size, observations indicate that normal and dark matter play roughly similar roles in determining the dynamics of galaxies. And at the largest-size scales, dark matter dominates the dynamics of galaxy clusters and

superclusters—clusters of clusters. To study dark matter, we need to investigate objects the size of a galaxy or larger.

Section 2: Initial Evidence of Dark Matter

Fritz Zwicky, an astronomer at the California Institute of Technology, stumbled across the gravitational effects of dark matter in the early 1930s while studying how galaxies move within the Coma Cluster. The Coma Cluster consists of approximately 1,000 galaxies spread over about two degrees on the sky —roughly the size of your thumb held at arm's length, and four times the size of the Sun and the Moon seen from Earth. Gravity binds the galaxies together into a cluster, known as a galaxy cluster. Unlike the gravitationally bound planets in our solar system, however, the galaxies do not orbit a central heavy object like the Sun and thus execute more complicated orbits.

The Father of Dark Matter—and More

Fritz Zwicky.
Source: © AIP Emilio Segrè Visual Archives, Physics Today Collection.

Bulgarian born and Swiss naturalized, Fritz Zwicky found his scientific home at the California Institute of Technology. From his perch on Caltech's Mount Wilson Observatory, Zwicky discovered more of the exploding stars known as "supernovae" than all his predecessors combined. But astrophysicists today admire him mostly for his theoretical insights into such phenomena as neutron stars, gravitational lenses, and—perhaps most important of all—dark matter.

Zwicky's observations of supernovae in distant galaxies laid the foundation of his theoretical work. As he detected supernovae in ever-more distant galaxies, he realized that most galaxies combined in clusters. Careful measurements of the light from clusters led him to suggest the existence of dark matter. That may represent his greatest legacy, but he made other key contributions to astrophysics. He predicted that galaxies could act as gravitational lenses, an effect first observed in 1979, five years after his death. And he and his colleague Walter Baade predicted the transition of ordinary stars into neutron stars, first observed in 1967.

To carry out his observations, Zwicky persuaded Caltech to build an 18-inch Schmidt telescope that could capture large numbers of galaxies in a single wide-angle photograph. He used the instrument to make a survey of all the galaxies in the cluster and used measurements of the Doppler shift of their spectra to determine their velocities. He then applied the virial theorem. A straightforward application of classical mechanics, the virial theorem relates the velocity of orbiting objects to the amount of gravitational force

acting on them. Isaac Newton's theory tells us that gravitational force is proportional to the masses of the objects involved, so Zwicky was able to calculate the total mass of the Coma Cluster from his measured galactic velocities. See the math

Figure 4: The Coma Cluster, which provided the first evidence for dark matter.
Source: © NASA, JPL-Caltech, SDSS, Leigh Jenkins, Ann Hornschemeier (Goddard Space Flight Center) et al.

Zwicky also measured the total light output of all the cluster's galaxies, which contain about a trillion stars altogether. When he compared the ratio of the total light output to the mass of the Coma Cluster with a similar ratio for the nearby Kapteyn stellar system, he found the light output per unit mass for the cluster fell short of that from a single Kapteyn star by a factor of over 100. He reasoned that the Coma Cluster must contain a large amount of matter not accounted for by the light of the stars. He called it "dark matter."

Zwicky's measurements took place just after astronomers had realized that galaxies are very large groups of stars. It took some time for dark matter to become the subject of active research it is today. When Zwicky first observed the Coma Cluster, tests of Einstein's theory were just starting, the first cosmological measurements were taking place, and nuclear physicists were only beginning to develop the theories that would explain the Big Bang and supernovae. Since galaxies are complex, distant objects, it is not surprising that astronomers did not immediately begin to worry about "the dark matter problem."

By the early 1970s, technology, astronomy, and particle physics had advanced enough that the dark matter problem seemed more tractable. General relativity and nuclear physics had come together in the Big Bang theory of the early universe, and the detection of microwave photons from the time when the first atoms formed from free electrons and protons had put the theory on a solid footing. Larger

telescopes and more precise and more sensitive light detectors made astronomical measurements quicker and better. Just as important, the emergence of affordable mini-computers allowed physics and astronomy departments to purchase their own high-performance computers for dedicated astronomical calculations. Every advance set the scene for a comprehensive study of dark matter, and two very important studies of dark matter soon appeared.

Dark matter appears in galactic simulations

Figure 5: James Peebles (left) and Jeremiah Ostriker (right) found evidence for dark matter in their computer simulations.
Source: © AIP, Physics Today Collection and Tenn Collection.

In 1973, Princeton University astronomers Jeremiah Ostriker and James Peebles used numerical simulation to study how galaxies evolve. Applying a technique called N-body simulation, they programmed 300 mass points into their computer to represent groups of stars in a galaxy rotating about a central point. Their simulated galaxy had more mass points, or stars, toward the center and fewer toward the edge. The simulation started by computing the gravitational force between each pair of mass points from Newton's law and working out how the mass points would move in a small interval of time. By repeating this calculation many times, Ostriker and Peebles were able to track the motion of all the mass points in the galaxy over a long period of time.

For a galaxy the size of the Milky Way ($4x10^{20}$ meters), a mass point about halfway out the edge moves at about 200 kilometers per second and orbits the center in about 50 million years. Ostriker and Peebles found that in a time less than an orbital period, most of the mass points would collapse to a bar-shaped, dense concentration close to the center of the galaxy with only a few mass points at larger radii. This looked nothing like the elegant spiral or elliptical shapes we are used to seeing. However, if they added a static, uniform distribution of mass three to 10 times the size of the total mass of the mass points, they

found a more recognizable structure would emerge. Ostriker and Peebles had solid numerical evidence that dark matter was necessary to form the types of galaxies we observe in our universe.

Fresh evidence from the Andromeda galaxy

At about the same time, astronomers Kent Ford and Vera Cooper Rubin at the Carnegie Institution of Washington began a detailed study of the motion of stars in the nearby galaxy of Andromeda. Galaxies are so large that even stars traveling at 200 kilometers per second appear stationary; astronomers must measure their Doppler shifts to obtain their velocities. However, early measurements of stellar velocities in different portions of Andromeda proved very difficult. Since the spectrometers used to measure the shift in frequency took a long time to accumulate enough light, observations of a given portion of Andromeda required several hours or even several nights of observing. Combining images from several observations was difficult and introduced errors into the measurement. However, new and more sensitive photon detectors developed in the early 1970s allowed much shorter measurement times and enabled measurements further out from the center of the galaxy.

From Controversy to Credibility

Vera Cooper Rubin at the Lowell Observatory. Kent Ford has his back to us.
Source: © Bob Rubin.

Vera Cooper Rubin faced several obstacles on her way to a career in astronomy. A high school physics teacher tried to steer her away from science. A college admissions officer suggested that she avoid majoring in astronomy. Princeton University did not grant her request for a graduate catalogue in 1948, because its graduate astronomy program did not accept women until 27 years later. Senior astronomers took a scornful view of her first paper, presented in 1950, on galactic motions independent of the classic expansion of the universe. And when she and collaborator Kent Ford expanded that research in the 1970s, they met so much dissent that they shifted to another field.

The shift proved providential. Rubin and Ford measured the rotational velocities of interstellar matter in orbit around the center of the nearby Andromeda galaxy. Their readings, confirmed by observations on other galaxies, led them to infer that the galaxies must contain dark matter. Confirmation of that fact sealed Rubin's reputation as an astronomer.

Rubin and Ford measured the velocity of hydrogen gas clouds in and near the Andromeda galaxy using the new detectors. These hydrogen clouds orbit the galaxy much as stars orbit within the galaxy. Rubin

and Ford expected to find that the hydrogen gas outside the visible edge of the galaxy would be moving slower than gas at the edge of the galaxy. This is what the virial theorem predicts if the mass in the galaxy is concentrated where the galaxy emits light. Instead, they found the opposite: the orbital velocity of the hydrogen clouds remained constant outside the visible edge of the galaxy. If the virial theorem is to be believed, there must be additional *dark* matter outside the visible edge of the galaxy. If Andromeda obeyed Newton's laws, Rubin reasoned, the galaxy must contain dark matter, in quantities that increased with increasing distance from the galactic center.

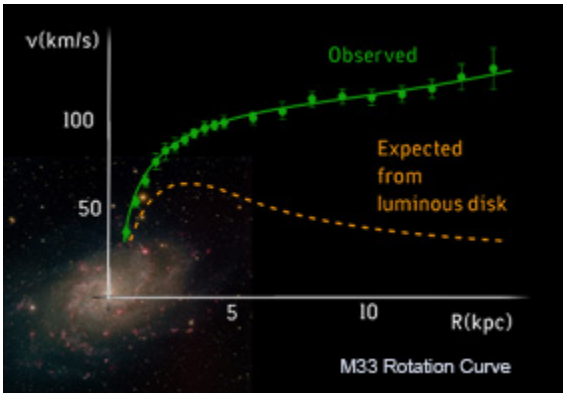


Figure 6: Observed and predicted rotation curves for the galaxy M33, also known as the "Triangulum Galaxy."
Source: © M33 Image: NOAO, AURA, NSF, T. A. Rector.

Alternative explanations of the Andromeda observations soon emerged. Theories of Modified Newtonian Dynamics (MOND), for example, aimed to explain the findings by modifying the gravitational interaction over galactic and larger distances. At very low accelerations, which correspond to galactic distances, the theories posit that the gravitational force varies inversely with the distance alone rather than the square of the distance. However, MOND would overturn Einstein's theory in an incredible way: General relativity is based on the simple idea of the equivalence principle. This states that there is no difference between gravitational mass (the mass that causes the gravitational force) and inertial mass (the mass that resists acceleration). There is no fundamental reason to expect these two masses to be the same, nor is there any reason to expect them to be different. But their equivalence forms the cornerstone of Einstein's general theory. MOND theories break that equivalence because they modify either gravity or inertia. If MOND were correct, a fundamental assumption underlying all of modern physics would be false.

Section 3: Dark Matter in the Early Universe

By the end of the 1970s, two compelling lines of evidence for dark matter had appeared. The motion of galaxies within clusters and the motion of gas clouds around individual galaxies strongly suggested that either our understanding of gravity is fundamentally wrong, or that there is far more matter in the galaxies and clusters than meets the eye. Further, simulations of galaxy formation showed that the spiral and elliptical galaxies we observe in the night sky cannot form without large amounts of dark matter in addition to the luminous stars. A third line of evidence developed in the 1990s, as radio telescopes above the atmosphere mapped the cosmic microwave background (CMB).

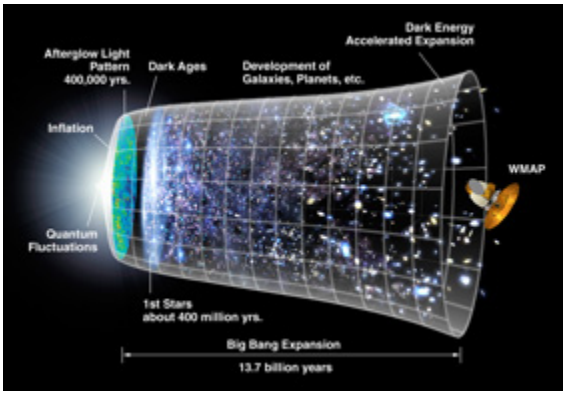


Figure 7: The history of the universe according to our standard model of cosmology.
Source: © NASA, WMAP Science Team.

This new evidence for dark matter has its origin in the early universe. About one second after the Big Bang, astrophysicists believe, a very dense mixture of protons, neutrons, photons, electrons, and other subatomic particles filled the universe. The temperature was so high that the electrons could not bind with the protons to form atoms. Instead, all the particles scattered off of each other at high rates, keeping all the different species at the same temperature—that is, in thermal equilibrium—with each other. The photons also scattered off of the electrically charged protons and electrons so much that they could not travel very far.

As the universe expanded, the temperature dropped to about one billion degrees Kelvin (K). At that point, the protons and neutrons began to bind together to form atomic nuclei. At roughly 390,000 years after the Big Bang, continued expansion and cooling had dropped the temperature of the universe to about 3000 K. By that point, all the electrons and protons had bound to form electrically neutral hydrogen atoms, and all the other charged particles had decayed. After the primordial hydrogen formed, the universe became

so transparent to photons that they have been traveling throughout it for the entire 13.7 billion years since then. These relic photons from the early universe have a microwave wavelength, and are known as the cosmic microwave background, or CMB.

Density fluctuations and dark matter

Before the neutral hydrogen formed, the matter was distributed almost uniformly in space—although small variations occurred in the density of both normal and dark matter owing to quantum mechanical fluctuations. Gravity pulled the normal and dark matter in toward the center of each fluctuation. While the dark matter continued to move inward, the normal matter fell in only until the pressure of photons pushed it back, causing it to flow outward until the gravitational pressure overcame the photon pressure and the matter began to fall in once more. Each fluctuation "rang" in this way with a frequency that depended on its size. The yo-yoing influenced the temperature of the normal matter. It heated up when it fell in and cooled off when it flowed out. The dark matter, which does not interact with photons, remained unaffected by this ringing effect.

When the neutral hydrogen formed, areas into which the matter had fallen were hotter than the surroundings. Areas from which matter had streamed out, by contrast, were cooler. The temperature of the matter in different regions of the sky—and the photons in thermal equilibrium with it—reflected the distribution of dark matter in the initial density fluctuations and the ringing normal matter. This pattern of temperature variations was frozen into the cosmic microwave background when the electrons and protons formed neutral hydrogen. So a map of the temperature variations in the CMB traces out the location and amount of different types of matter 390,000 years after the Big Bang.

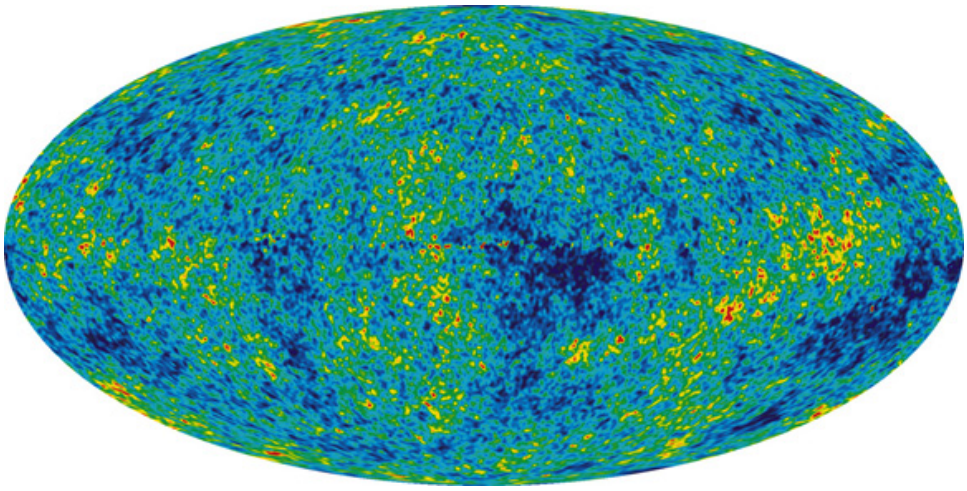


Figure 8: Map of the temperature variations in the cosmic microwave background measured by the WMAP satellite.
Source: © NASA, WMAP Science Team.

American physicists Ralph Alpher, Robert Herman, and George Gamow predicted the existence of the CMB in 1948. Seventeen years later, Bell Labs scientists Arno Penzias and Robert Wilson detected them. Initial measurements showed the intensity of the relic photons to be constant across the sky to a fraction of 1 percent. In the early 1990s, however, NASA's Cosmic Background Explorer (COBE) spacecraft used a pair of radio telescopes to measure differences among relic photons to one part per million between two points in the sky. A subsequent spacecraft, the Wilkinson Microwave Anisotropy Probe (WMAP), made an even more precise map. This revealed hot and cold spots about 1.8 degrees in size across the sky that vary in intensity by a few parts per million.

The angular size and the extent of variation indicate that the universe contained about five times as much dark matter as normal matter when the neutral hydrogen formed. Combined with measurements of supernovae and the clustering of galaxies, this indicates that dark energy comprises 73 percent of the universe, dark matter 23 percent, and normal matter just 4 percent.

Section 4: Dark Matter Bends Light

With three independent reasons to believe that dark matter existed—motion of galaxies, structure simulations, and temperature fluctuations in the cosmic microwave background—increasing numbers of physicists and astronomers turned their attention to trying to understand just what the dark matter is made of, and how it is distributed throughout the universe. Gravitational lensing proved a useful tool with which to probe the dark matter.

Quasars, lensing, and dark matter

Images of quasars gravitationally lensed by galaxies provide insight into the distribution of dark matter inside the lensing galaxies. Quasars are distant objects that emit huge amounts of light and other radiation. Since many quasars are visible behind galaxies, their light must pass through those intervening galaxies on the way to us. We know from general relativity theory that the matter in any galaxy—both normal and dark matter—bends space time. That bending distorts the image of any quasar whose light passes through a galaxy.

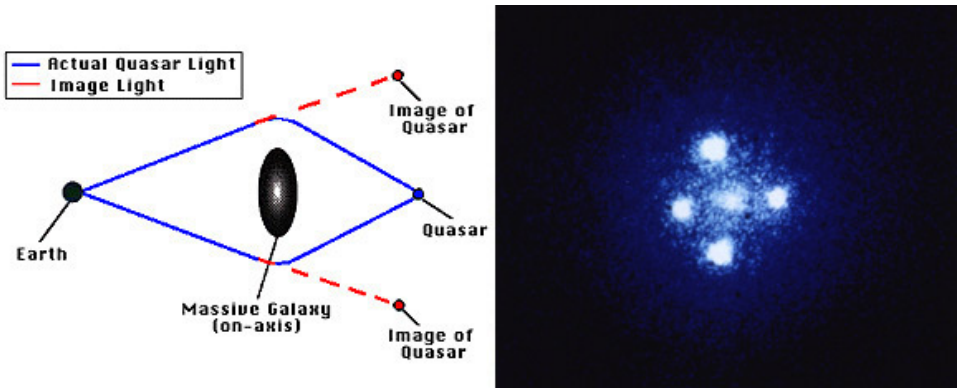


Figure 9: Gravitational lensing produces more than one image of distant quasars, as seen in this shot from the Hubble Space Telescope.
Source: © NASA/ESA Hubble Space Telescope, NASA/Goddard Space Flight Center.

In many cases, this lensing causes several images of the same quasar to appear in our telescopes. Careful measurements of the brightness of the different images of the quasar give hints about the distribution of the matter in the galaxy. Since the matter in each part of the galaxy determines the amount of bending of space time in that part of the galaxy, the brightness of the images tells us how matter, both normal and dark, is distributed. Optical measurements inform astronomers where the normal matter is. They can then use the brightness of the multiple quasar images to trace out the dark matter.

So far, astronomers have identified about 10 such lenses like this. Careful observations have shown that any clumps of dark matter in the galaxies must be smaller than about 3,000 light-years. More sensitive telescopes will find more lenses and will improve our understanding of how dark matter is distributed in galaxies.

Evidence from colliding clusters

Observing colliding galaxy clusters provides another useful way of understanding the nature of dark matter. When two clusters collide, the dark matter in one passes through the other unaffected; dark matter doesn't interact much with either itself or normal matter. But the normal matter in one cluster does interact with the dark matter and the normal matter in the other cluster, as well as with the dark matter in its own cluster. During the collision, the normal matter is dragged forward by the dark matter in its own cluster and dragged back by both the dark matter and normal matter in the other cluster. The net effect of the collision, therefore, is to cause the normal matter in each cluster to fall behind the dark matter in the same cluster.

Figure 10: X-ray and visible light images of the Bullet Cluster reveal strong evidence for the existence of dark matter.
Source: © X-ray: NASA, CXC, CfA, M. Markevitch et al.; Optical: NASA, STScI; Magellan, U. Arizona, D. Clowe et al.; Lensing Map: NASA, STScI; ESO WFI; Magellan, U. Arizona, D.Clowe et al.

Astronomers gained solid evidence of that scenario when they imaged a pair of colliding galaxy clusters named the Bullet Cluster in two ways: through its emission of visible light and x-rays. The collision between the normal matter in each subcluster heats up the normal matter, causing the colliding subclusters to emit x-rays. In 2004, NASA's orbiting Chandra x-ray observatory captured an x-ray image of the Bullet Cluster that gives the locations of the normal matter in the two subclusters. At the same time, the entire Bullet Cluster distorts the images of galaxies behind it through the gravitational lensing

effect that we reviewed above in the context of quasars. By carefully measuring the shape of the distorted background galaxies, astronomers could determine the average position and mass of each of the subclusters. Since galaxy clusters contain a few times as much dark matter as normal matter, the lensing measurement gives the location of the dark matter, while the x-rays locate the normal matter. The image that combines both measurements shows that the dark matter has run ahead of the normal matter in both subclusters, confirming expectation.

The measurements of the Bullet Cluster were a blow to the MOND theories that we encountered earlier in this unit. Those theories predict no difference between the x-ray and lensing images. Some theorists have tried to modify the MOND approach in such a way that it accommodates the evidence from the Bullet Cluster and other observations, but the clear consensus of astronomers is that dark matter is a reality.

Dark matter in our galaxy

With gravitational lensing successfully being used to "weigh" entire galaxy clusters, the question arose whether it could be brought to bear more locally, to search for dark matter objects in the outer regions of our own Milky Way galaxy. The answer is a resounding yes. A clever gravitational lensing survey to search for clumps of dark matter in the halo of our galaxy began in 1992. The survey was designed to find MACHOs, or massive compact halo objects, which is a fancy term for "chunks of dark matter." It was initially thought that MACHOs would be failed stars or large, drifting planets—familiar objects that don't emit light—but the MACHO project was designed to be sensitive to any lump of dark matter with a mass between the Earth's mass and 10 times the Sun's mass.

Figure 11: This Australian telescope unsuccessfully sought evidence for the existence of MACHOs based on their putative effect on starlight.
Source: © The Australian National University.

The MACHO project used a telescope to monitor the light from stars just outside the Milky Way in a very small satellite galaxy called the "Large Magellanic Cloud." If a MACHO passes in front of one of these stars, the gravitational lensing effect predicted by Einstein's general theory of relativity and confirmed in 1979 will increase the measured flux of the starlight by a tiny amount. The Anglo-American-Australian MACHO Project used an automated telescope at Australia's Mount Stromlo Observatory to observe transits. None showed anywhere near enough change in the starlight to account for dark matter as consisting of faint stars or large planets.

A similar project, named "EROS" and run by the European Organisation for Astronomical Research in the Southern Hemisphere at Chile's La Silla Observatory, has had the same negative result. For example, a study of 7 million stars revealed only one possible MACHO transit; in theory, MACHOs would have produced 42 events. But physicists refused to give up the hunt. The SuperMACHO survey, a successor to the MACHO Project, used the 4-meter Victor M. Blanco telescope in Chile's Cerro Tololo Inter-American Observatory to monitor tens of millions of stars in the Large Magellanic Cloud in search of evidence that

MACHOS exist. SuperMACHO also found that MACHOs cannot account for the vast amount of dark matter in the galaxy.

The astronomical evidence we have for dark matter ranges from within our galaxy to the farmost regions of space and time that we are able to probe. We now understand that dark matter dominates at the scale of galaxy clusters, normal matter dominates at the subgalactic scale, and they duke it out on the galactic scale. We know that dark matter gravitationally interacts with itself and normal matter, but we still do not know what the dark matter is.

Section 5: From Astronomy to Particle Physics

The abundance of astronomical evidence for dark matter in the early 1970s intrigued physicists working in other fields. Cosmologists and nuclear physicists were developing our current model of cosmology, trying to understand how the universe we live in—dark matter and all—formed. Concurrently, others wondered how the dark matter fit, if at all, into the Standard Model we learned about in Unit 1.

By the late 1970s, the Standard Model of particle interactions had gained a firm experimental footing. At the same time, physicists were refining their standard model of cosmology in which the universe began its existence when a singularity, a point of infinite density and infinite temperature, exploded in the Big Bang and began a process of expansion that continues today. Application of the Standard Model and nuclear theory to the Big Bang model allowed physicists to quantify nucleosynthesis, the process responsible for creating elements out of the protons, neutrons, electrons, and energy that suffused the infant universe.

$$
\begin{aligned}
&\mathrm{n}^0 \longrightarrow \mathrm{p}^+ + \mathrm{e}^- + \bar{\nu}_e && \mathrm{p}^+ + \mathrm{n}^0 \longrightarrow {}^2_1\mathrm{D} + \gamma \\
&{}^2_1\mathrm{D} + \mathrm{p}^+ \longrightarrow {}^3_2\mathrm{He} + \gamma && {}^2_1\mathrm{D} + {}^2_1\mathrm{D} \longrightarrow {}^3_2\mathrm{He} + \mathrm{n}^0 \\
&{}^2_1\mathrm{D} + {}^2_1\mathrm{D} \longrightarrow {}^3_1\mathrm{T} + \mathrm{p}^+ && {}^3_1\mathrm{T} + {}^2_1\mathrm{D} \longrightarrow {}^4_2\mathrm{He} + \mathrm{n}^0 \\
&{}^3_2\mathrm{He} + {}^4_2\mathrm{He} \longrightarrow {}^7_3\mathrm{Li} + \gamma && {}^3_2\mathrm{He} + \mathrm{n}^0 \longrightarrow {}^3_1\mathrm{T} + \mathrm{p}^+ \\
&{}^3_2\mathrm{He} + {}^2_1\mathrm{D} \longrightarrow {}^4_2\mathrm{He} + \mathrm{p}^+ && {}^3_2\mathrm{He} + {}^4_2\mathrm{He} \longrightarrow {}^7_4\mathrm{Be} + \gamma \\
&{}^7_3\mathrm{Li} + \mathrm{p}^+ \longrightarrow {}^4_2\mathrm{He} + {}^4_2\mathrm{He} && {}^7_4\mathrm{Be} + \mathrm{n}^0 \longrightarrow {}^7_3\mathrm{Li} + \mathrm{p}^+
\end{aligned}
$$

Figure 12: This series of reactions created the lightest elements in the infant universe.
Source: © Wikimedia Commons, Public Domain, Author: Lokal_Profil, 5 August 2008.

This model of Big Bang nucleosynthesis, supported by careful astronomical observations of the abundance of light elements in the universe, makes a particularly significant prediction about the density of baryons in the first few minutes: The Big Bang could not have created enough normal matter at the start of the universe to account for dark matter. Astrophysicists concluded that dark matter must be some new form of matter not yet observed, possibly even a new type of particle.

New dark matter particles

One of the first attempts to explain dark matter with new particles arose in a surprising place: the Homestake Gold Mine in South Dakota that we first encountered in Unit 1. The Homestake neutrino

detector was monitoring neutrinos thought to come from the Sun. In 1976, it became apparent that this experiment only counted about half the predicted number. One explanation was that some new form of heavy particles that did not interact much would collect in the center of the Sun, cooling it off very slightly. This new heavy particle would have the same properties required by dark matter: very weak interaction with other particles, copious in our solar system, and left over from the Big Bang.

We now know that the deficit of neutrinos is due to their oscillation; but at the time, it was an intriguing hint that dark matter could be made up of a new type of particle, possibly not included in the Standard Model. Heavy neutrinos were once considered a candidate for particle dark matter, but large-scale structure simulations of neutrino dark matter have ruled them out. The remainder of this unit will focus on particle dark matter in both theory and experiment. In section 8, we will explore the two leading non-Standard Model candidates for particle dark matter and experimental efforts to detect them. We also will examine how the constant theoretical effort to explain dark matter often generates new possibilities for particle dark matter. The table below summarizes all the possibilities for dark matter that appear in this unit.

Table 1. Possible candidates for Dark Matter

Candidate	Mass range	Pros	Cons
Astronomical objects (failed stars, black holes, MACHOs...)	10^{50}-10^{63} eV	Rather conservative scenario; lensing searches effective	Amount of ordinary matter made in Big Bang falls short of total dark matter we need; not detected via lensing searches
Neutrinos	< 2 eV	Known to exist, and have mass so a natural candidate	Tiny neutrino mass inhibits clumping on small scales needed to hold galaxies together
Axions	10^{-6} eV	Postulated to solve a different problem altogether; dark matter aspect comes for free	Tough to detect
Weakly Interacting Massive Particles (WIMPs)	10^{10} eV	Plausible class of new elementary particles that emerge from multiple theories beyond the Standard Model	Have evaded detection in accelerators to date
Alternative Gravity Scenarios	N/A	No mysterious new matter needed, but rather a modification of gravity	Hard to reconcile with Bullet Cluster observations; theories seen as "inelegant"
Dark Sector Interactions	N/A	Two new pieces of physics: exotic dark matter particles plus new interactions between them; might help reconcile experiments	Added complexity; wider range of phenomenology; tougher to rule out

Section 6: The Search for Particle Dark Matter

Starting in the late 1980s with the idea that dark matter could be a new kind of particle, nuclear and particle physicists began experiments to detect dark matter in the event that it interacts directly with normal matter. There are two main ideas about what these particles could be. One views the dark matter as a very light particle known as the axion. Hypothesized to explain a confusing property of the strong force that binds quarks together (see Unit 2), an axion would weigh about one-trillionth as much as a proton. The other idea comes from a very broad class of theories that predicts an electrically neutral particle weighing between 100 and 1,000 times as much as a proton. The general name of this kind of particle is a "weakly interacting massive particle" or WIMP. Physicists first introduced this concept to explain the problem of solar neutrinos that we met in Section 5.

Figure 13: If dark matter consists of axions, the Axion Dark Matter Experiment shown here could detect them in the next decade.
Source: © ADMX.

So far, physicists have found no evidence that axions or WIMPs actually exist; both particles remain in the realm of hypothesis. However, the physics community found the theoretical reasoning that led

to the hypotheses were compelling enough to mount experimental searches for them. Some of their experiments have provided fascinating hints of the presence of these peculiar particles.

The types of experiments differ considerably, based on which particle they aim to detect. In each case, they rely on the specific physical properties of the two proposed particles. Because axions are hypothesized to have no electric charge or spin, extremely small masses, and minimal interaction with ordinary matter, experimenters must use indirect methods to detect them. In contrast, theorists see WIMPs as not only possessing large masses but also interacting—although infrequently—with ordinary matter. Thus, it may be possible to detect them directly as well as indirectly.

The quest for axions

The concept of the axion emerged as a solution to the so-called strong-CP problem. We first encountered CP, the product of charge conjugation and parity, in Unit 1. There we discovered that CP violation occurs in weak interactions, but does not appear to occur in strong interactions. In 1977, theorists Roberto Peccei and Helen Quinn suggested that this difference between the strong and the weak force was due to a broken symmetry. In Unit 2, we learned that symmetry breaking is accompanied by a new particle called a "Nambu-Goldstone boson." The new particle associated with the broken Peccei-Quinn symmetry would interact with ordinary matter so weakly as to be virtually undetectable. MIT theorist Frank Wilczek named it the axion after a laundry detergent because, he said, it cleaned up the strong-CP problem. Later, the weakness of its interactions made it a strong candidate for dark matter.

Figure 14: Axion hunters: two Fermilab physicists with their experiment designed to detect axions.
Source: © Fermilab.

Experimentalists who want to detect the particle can choose either to make their own axions or to search for those that already exist. Many of these experiments attempt to detect axions as they interact with

photons. The basic idea is that when an axion collides with a photon, two photons are produced in the collision that have an energy proportional to the axion mass. Dark matter axions do not move very fast and are very light. Therefore, the photons produced would be low energy, with a wavelength roughly corresponding to radio waves. Axions are expected to interact with photons very weakly—much more weakly than electrons or protons—so the trick to detecting axions is to build a very sensitive radio antenna.

Trapping radio waves to identify axions

The process starts with a magnetic field about 200,000 times more powerful than Earth's field. When an axion interacts with the magnetic field, radio waves are generated. To capture the radio waves, experimentalists use a hollow superconducting cylinder called a "resonant cavity." The size and shape of the cavity are carefully selected to amplify radio waves of a particular frequency.

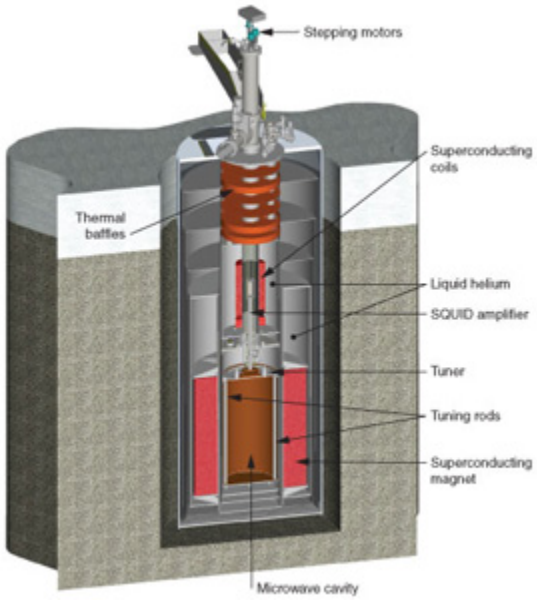


Figure 15: SQUID technology boosts the ability of the Axion Dark Matter Experiment to detect the faint signals that would indicate the presence of axions.
Source: © Lawrence Livermore National Laboratory.

For a typical mass of 2μeV, roughly 10^{30} axions would stream through the detector each second. Over time, the trapped radio waves would build up to a detectable amount. The radio waves built up in the resonant cavity are measured using a tool called a SQUID, for superconducting quantum interference device, which greatly improves the experiment's ability to detect faint signals. Since physicists do not

know the mass of the hypothetical axion, they would have to adjust the radio frequency of the cavity in small steps, like tuning a radio, to scan for a signal from dark matter axions.

The best-known experiment of this type, the Axion Dark Matter Experiment (ADMX), has operated since 1995 without detecting a signal. Physicists at Lawrence Livermore National Laboratory and collaborating institutions improved ADMX in 2008 by adding sensitive amplifiers to the apparatus. Further enhancements include adding a cooling system that will improve the system's sensitivity. The team will add more improvements and will continue to operate the experiment for many years before exhausting all its potential to hunt for axions.

Other searches for axions have started in recent years. A Japanese project, the Cosmic Axion Research with Rydberg Atoms in a Resonant Cavity (CARRAC) experiment, seeks axions in a range of masses similar to that sought by ADMX. An Italian group's PVLAS (for Polarizzazione del Vuoto con LASer) experiment looks for minute changes in the polarization of light that might stem from axions. And in contrast to those earthbound methods, the European Nuclear Research Center's Axion Solar Telescope (CAST) searches for axions produced in the Sun.

Seeking the elusive WIMPs

As theorized, WIMPs interact with normal matter in the simplest way, by colliding with it. They don't do that very often; they easily penetrate the Earth or Sun without interacting at all. But very occasionally a WIMP will hit an atomic nucleus and cause it to recoil. Theorists believe that 5 million dark matter particles will pass through a 2 kilogram piece of normal matter, containing roughly 10^{25} atoms, every second. In rough numbers, just one of the WIMPs will hit a nucleus in an entire year. The nucleus will recoil and deposit its energy in the surrounding matter in the form of ionization electrons, which can attach to ions to create neutral atoms, or heat. The amount of energy deposited in this way resembles that of an x-ray photon. Physicists searching for dark matter face the twin challenge of collecting this deposited energy and ensuring that the energy they collect came from a dark matter interaction and not from a conventional physics process.

Distinguishing between dark matter interactions and conventional interactions proves to be very difficult. At sea level, 100 cosmic rays pass through each square meter of the Earth's surface each second, along with 28 neutrons from cosmic ray interactions in the atmosphere and 10,000 x-rays from low-level contamination in normal materials. In addition, everything contains trace amounts of uranium and thorium,

both of which give rise to sequential radioactive decays. All these processes can mimic the scattering of dark matter off a nucleus.

Underground searches for WIMPs

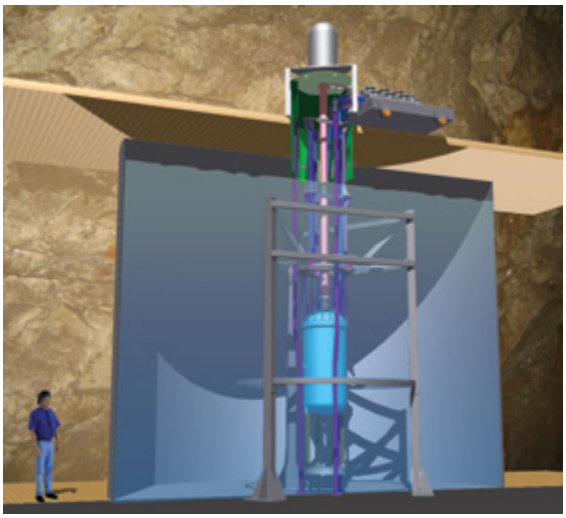


Figure 16: The Large Underground Xenon detector will have 100 times more sensitivity to WIMPs than previous detection methods.
Source: © The LUX Experiment.

Dark matter recoil experiments address these problems in several ways. Since few cosmic rays penetrate deep underground, experiments placed in tunnels and mines under a kilometer of rock remove that source of interference. The Large Underground Xenon (LUX) detector, which will operate 1,463 meters deep in the familiar Homestake Gold Mine in South Dakota, exemplifies this approach. As its detector, LUX will use a cylinder containing 350 kilograms of liquid and gaseous xenon, which scintillates and becomes ionized when struck by particles, including WIMPs. Several precautions will minimize the number of non-WIMP particles likely to impact the detector. Up to a meter of high-purity lead or copper shielding will absorb x-rays and gamma rays emitted by the walls of the mine. In future experiments, a meter or so of water will absorb neutrons from both cosmic rays and the cavern's walls. Finally, experimenters will use only tested, low-radioactivity materials to build the detector.

Other groups are also undertaking the underground route to detecting WIMPs. The international Xenon Dark Matter Project uses a xenon detector in a laboratory under Italy's Gran Sasso Mountain. The second Cryogenic Dark Matter Search (CDMSII) project relies on cryogenic germanium and silicon detectors in Minnesota's Soudan Mine, another location well used by scientists; the original experiment had taken

place in a tunnel under the Stanford University campus. And, the Italian-American WIMP Argon Program (WARP) uses argon in place of the more expensive xenon in its detector.

To clarify their results, the dark matter detectors measure the energy of the recoiling nucleus in two different ways. A neutron or dark matter interaction will divide its energy between heat and ionization electrons, while other radioactive decays will put virtually all their energy into ionization electrons. In the late 1980s, the first dark matter experiments were able to exclude neutrinos as dark matter by measuring the energy only one way. The two energy measurement techniques developed since then have led to an improvement of 10 million in sensitivity to dark matter interactions. Future detectors will have even greater sensitivity.

Monitoring the direction of dark matter

If dark matter WIMPs exist, we could learn more about them by measuring the direction from which they come toward Earth from space. A directional measurement would use gas molecules at about one-twentieth of an atmosphere pressure as targets for the dark matter particles to hit. Each nucleus struck by a WIMP would travel about 1 millimeter. That's a long enough distance for physicists to measure by collecting the ionization electrons created by the collisions directly or by converting them to scintillation light and using a charge-coupled device (CCD) camera to create an image. Since each struck nucleus will generally travel in the same direction as that in which the dark matter particle traveled before it hit the nucleus, measuring the direction of the recoiling nuclei will give experimenters critical details about dark matter in our galaxy.

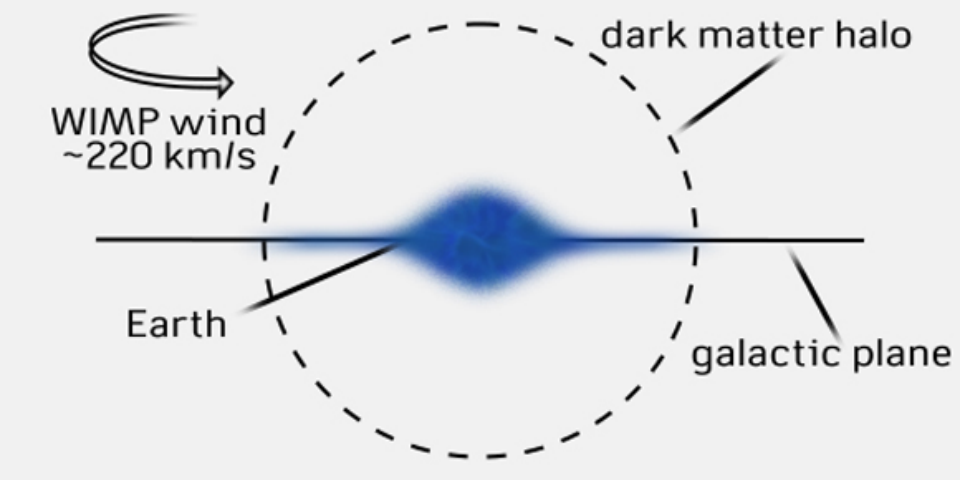


Figure 17: If they exist, WIMPs could stream toward Earth in a specific direction in a "WIMP wind" that might be experimentally detectable.
Source:

In the simplest picture, the normal matter in our Milky Way galaxy rotates through a stationary halo of dark matter. If we could easily detect dark matter on Earth, we would see a "wind" of dark matter coming from the direction in which our solar system is moving through the Milky Way. Since the constellation Cygnus orbits around the galactic center ahead of our solar system, the dark matter would appear to be streaming at us from Cygnus. Thus, a directional experiment would see nuclei recoiling away from Cygnus. Measuring direction in this way not only would yield information about dark matter, but it also would make the experiment more sensitive, since no background source of radiation would follow the trajectory of Cygnus. In addition, a detector able to measure direction would begin to explore the velocity distribution of dark matter in the Milky Way much more directly than ever before. A directional detector would work, in effect, as a dark matter telescope.

Collider and satellite searches for dark matter

If WIMPs comprise dark matter, high-energy collisions may also shed light on their nature. Both the Tevatron and the Large Hadron Collider (LHC) may be able to produce WIMPs by colliding protons and antiprotons or protons and protons at energies high enough to fuse the quarks inside those particles into WIMPs. Teams at both the Tevatron and LHC will continue sifting through vast amounts of data, hoping to find evidence of WIMPs in their detectors.

Figure 18: NASA's Fermi Gamma-ray Space Telescope has spotted an excess of normal matter particles that may have arisen when WIMPs annihilated each other.
Source: © NASA/Fermi Gamma-ray Space Telescope.

Finally, it may be that WIMP dark matter particles annihilate each other in the galaxy to produce extra amounts of normal matter (such as protons, electrons, antiprotons, positrons, neutrinos, or gamma rays), which could be detected from Earth or in space-borne experiments. Separating these extra normal particles from cosmic rays is difficult. But in the last year, two satellite experiments may have observed some hints of dark matter. NASA's Fermi Gamma-ray Space Telescope, launched in 2008, discovered evidence of more high-energy electrons and their antimatter positrons than anticipated. The excess could stem from WIMP annihilations. About the same time, the European Payload for Antimatter Matter Exploration and Light-nuclei Astrophysics (PAMELA) satellite, launched in 2006, detected more positrons than expected. However, it is much too early to tell whether either satellite has actually seen dark matter.

Section 7: Dark Forces

WIMPs and axions are compelling candidates for dark matter particles, but neither one has been detected experimentally. While ever-more sensitive laboratory experiments are conducted, theorists constantly develop new models, sometimes inventing new possibilities for dark matter. A plausible third candidate for dark matter has recently emerged, called dark forces. The dark forces theory is really an extension of the supersymmetry theory we first reviewed in Unit 2. In addition to the heavy WIMP particles, the latest version of supersymmetry theory posits the existence of light particles called φ, the Greek letter *phi*. If the φ exists, it is predicted to be more massive than two electrons, but less massive than 200 electrons. It would interact with other particles just like a photon, but with an interaction strength at least 1,000 times weaker.

The idea for dark forces arose when an Italian cosmic ray experiment called "DAMA/LIBRA" (DArk MAtter/Large sodium Iodide Bulk for RAre processes) observed energetic electrons and positrons unaccompanied by antiprotons. Ordinary WIMPs cannot explain this DAMA/LIBRA result, but in the dark forces version of supersymmetry, heavy WIMP particles would annihilate with one another and produce high-energy φ particles. The φ particles would then decay into energetic electron-positron pairs.

Figure 19: A technician works on detectors for the DAMA/LIBRA project, which stimulated the theory of dark forces.
Source: © DAMA/LIBRA.

The emergence of the dark forces theory has led to a series of new ideas for current and new experiments. If the theory is correct, WIMPs produced in high-energy collisions at the Tevatron and Large Hadron Collider would decay to several φ particles. Those particles would then decay to a large number

of electrons, positrons, or muons, giving a clear experimental signature. At low-energy colliders, the φ would manifest itself in rare decays of known particles. In lower-energy electron-proton collisions, extra electrons and positrons in the decay products would indicate that the collision produced φ particles. Physicists would need to gather a huge amount of data to test dark forces. Because the φ interacts with one-thousandth the strength of a photon, only one event in a million might contain a φ.

Although the dark forces theory arose to explain cosmic ray experiments and the DAMA/LIBRA results, it would still be viable even if the experimental basis were shown to be a fluctuation or result of a known process. Like axions and supersymmetry, the dark forces theory as yet has no solid experimental basis. However, it is a perfectly reasonable description of dark matter in every respect and should be experimentally pursued.

Supersymmetry theory has suggested other possible sources of dark matter. They include the gravitino, the supersymmetry partner of the graviton, and the electrically neutral neutralino, a particle with very small mass. Like other dark matter candidates, they have so far defied experimental efforts to detect them.

Section 8: The Search Continues

Dark matter remains an active area of research, with the results of current and planned experiments eagerly anticipated by the entire physics community. In coming years, large-scale studies of galaxies like the continuing Sloan Digital Sky Survey and the Anglo-Australian 2dF Galaxy Redshift Survey, supported by numerical simulations, will continue to develop our picture of the way in which dark matter is distributed over galactic and larger distances. Better cosmological measurements of supernovae and the cosmic microwave background will sharpen our knowledge of cosmological parameters, in particular the total amount of normal and dark matter. Detailed measurements of galaxies using gravitational lensing will tell us more about the distribution of dark matter within a galaxy. New space probes, nuclear recoil, and axion experiments will continue to hunt for evidence that dark matter interacts with normal matter in ways other than gravity. In addition, colliding beam accelerators, particularly the LHC, will try to make dark matter particles in the laboratory.

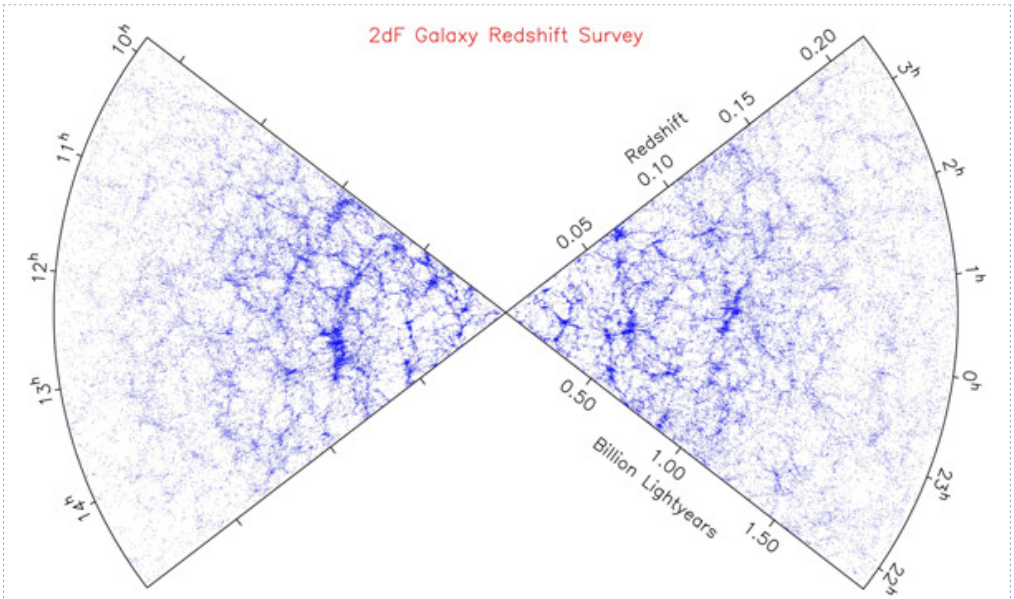


Figure 20: Galactic studies such as the "2dF Galaxy Redshift Survey" will improve our understanding of the distribution of dark and normal matter.
Source: © The 2dF Galaxy Redshift Team, 30 June 2003. Reproduced by permission of Commonwealth of Australia.

If some or all of the dark matter consists of WIMPs, the final picture will not emerge from any single endeavor. Rather, physicists will combine evidence produced by many different measurements to understand just what these new particles are. Even though the sensitivity of searches for dark matter on Earth has improved by about a factor of ten every few years over the past two decades, it might still take some time before the first convincing laboratory evidence for dark matter appears. Following first

indications, further measurements using different targets will sharpen the picture. But conclusive evidence will require a directional signal as well as consistency with cosmic ray experiments, astronomical observations, and exploration of the Terascale from collider experiments.

What if dark matter consists of axions? In that case, ADMX may make an observation in the next 10 years—if dark matter conforms to theory.

Of course, dark matter may be something completely new and unexpected. It may be a different manifestation of dark energy or it may be that we never find out. Dark matter raises the question of what it means to discover something. We already know what dark matter *does*: how it regulates the structure of galaxies and clusters of galaxies. This knowledge will certainly improve steadily as we make more astronomical observations. Learning what dark matter actually *is*, however, will take a big jump—one that we may never make. What does it mean for science if we find that we can't make this jump? Most likely, we will never have to answer that question. Physicists will continue to probe the universe in expectation of eventually unearthing its deepest secrets.

Section 9: Further Reading

- W. Goldstein, et al., "Neutrinos, Dark Matter and Nuclear Detection," *NATO Science for Peace and Security Series*, Part 2, 2007, p.117.
- Wayne Hu and Martin White, "The Cosmic Symphony," *Scientific American*, Feb. 2004, p. 44.
- Gordon Kane, "Supersymmetry: Unveiling the Ultimate Laws of Nature," *Basic Books*, 2001.
- Stacy McGaugh, "MOND over Matter," *Astronomy Now*, Nov./Jan. 2002, p. 63.
- Mordehai Milgrom, "Does Dark Matter Really Exist?," *Scientific American*, August 2002, p. 43.

Glossary

axion: The axion is a hypothetical particle that naturally arises in the solution to the strong-CP problem proposed by Peccei and Quinn in 1977. Axions are electrically neutral, and experiments have shown that their mass must be less than 1 eV. While they are relatively light particles, slow-moving axions could be produced in copious amounts in the early universe, and thus could be a significant component of the dark matter.

cosmic microwave background: The cosmic microwave background (CMB) radiation is electromagnetic radiation left over from when atoms first formed in the early universe, according to our standard model of cosmology. Prior to that time, photons and the fundamental building blocks of matter formed a hot, dense soup, constantly interacting with one another. As the universe expanded and cooled, protons and neutrons formed atomic nuclei, which then combined with electrons to form neutral atoms. At this point, the photons effectively stopped interacting with them. These photons, which have stretched as the universe expanded, form the CMB. First observed by Penzias and Wilson in 1965, the CMB remains the focus of increasingly precise observations intended to provide insight into the composition and evolution of the universe.

CP violation: The CP operation is a combination of charge conjugation (C) and parity (P). In most interactions, CP is conserved, which means that the interaction proceeds exactly the same way if the CP operation is performed on the interacting particles. If CP is conserved, particles with opposite charge and parity will interact in the same way as the original particles. CP violation occurs when an interaction proceeds differently when the CP operation is performed—particles with opposite charge and parity interact differently than the original particles. CP violation was first observed in neutral kaon systems.

dark forces: Dark forces arise in a 2009 theory to explain various experimental results in high-energy astrophysics. The theory proposes that dark matter WIMPs can decay into force-carrying particles, denoted by the Greek letter *phi* (φ). The φ particles would be associated with a new force of nature, distinct from the strong force, weak force, electromagnetism, and gravity.

Doppler shift (Doppler effect): The Doppler shift is a shift in the wavelength of light or sound that depends on the relative motion of the source and the observer. A familiar example of a Doppler shift is the apparent change in pitch of an ambulance siren as it passes a stationary observer. When the ambulance is moving toward the observer, the observer hears a higher pitch because the wavelength of the sound waves is shortened. As the ambulance moves away from the observer, the wavelength is lengthened and

the observer hears a lower pitch. Likewise, the wavelength of light emitted by an object moving toward an observer is shortened, and the observer will see a shift to blue. If the light-emitting object is moving away from the observer, the light will have a longer wavelength and the observer will see a shift to red. By observing this shift to red or blue, astronomers can determine the velocity of distant stars and galaxies relative to the Earth. Atoms moving relative to a laser also experience a Doppler shift, which must be taken into account in atomic physics experiments that make use of laser cooling and trapping.

equivalence principle: The equivalence principle is a basic premise that is essential to every experimentally verified physical theory, including General Relativity and the Standard Model. It states that an object's inertial mass is equivalent to its gravitational mass. The inertial mass of an object appears in Newton's second law: the force applied to the object is equal to its mass times its acceleration. The gravitational mass of an object is the gravitational equivalent of electric charge: the physical property of an object that causes it to interact with other objects through the gravitational force. There is no a priori reason to assume that these two types of "mass" are the same, but experiments have verified that the equivalence principle holds to a part in 10^{13}.

galaxy cluster: A galaxy cluster is a group of galaxies bound together by the force of gravity. Like the planets in our solar system, galaxies in a cluster orbit a common center of mass. However, galaxies execute more complicated orbits than the planets because there is no massive central body in the cluster playing the role of the Sun in our solar system. Galaxy clusters typically contain a few hundred galaxies, and are several megaparsecs (ten million light-years) in size. The orbital velocities of galaxies in clusters provide strong evidence for dark matter.

gravitational lensing: Gravitational lensing occurs when light travels past a very massive object. According to Einstein's theory of general relativity, mass shapes spacetime and space is curved by massive objects. Light traveling past a massive object follows a "straight" path in the curved space, and is deflected as if it had passed through a lens. Strong gravitational lensing can cause stars to appear as rings as their light travels in a curved path past a massive object along the line of sight. We observe microlensing when an object such as a MACHO moves between the Earth and a star. The gravitational lens associated with the MACHO focuses the star' light, so we observe the star grow brighter then dimmer as the MACHO moves across our line of sight to the star.

gravitino: The gravitino is the superpartner of the graviton. See: superpartner, supersymmetry.

graviton: The graviton is the postulated force carrier of the gravitational force in quantum theories of gravity that are analogous to the Standard Model. Gravitons have never been detected, nor is there a viable theory of quantum gravity, so gravitons are not on the same experimental or theoretical footing as the other force carrier particles.

ionization electron: An ionization electron is a free electron moving at high speed that knocks an electron off a neutral atom, turning the atom into an ion.

ion: An ion is an atom with nonzero electrical charge. A neutral atom becomes an ion when one or more electrons are removed, or if one or more extra electrons become bound to the atom's nucleus.

light-year: A light-year is the distance that light, which moves at a constant speed, travels in one year. One light-year is equivalent to 9.46×10^{15} meters, or 5,878 billion miles.

MACHO: A MACHO, or massive compact halo object, is a localized mass that has a gravitational influence on the matter around it but does not emit any light. Black holes and brown dwarf stars are examples of MACHOs. MACHOs were once thought to make a significant contribution to dark matter; however, gravitational lensing surveys have demonstrated that most of the dark matter must be something else.

mini-computer: The mini-computer was a precursor to the personal computers that are ubiquitous today. Prior to the development of the mini-computer, scientists doing computer-intensive calculations shared mainframe computers that were expensive multi-user facilities the size of small houses. Mini-computers cost ten times less than mainframe computers, fit into a single room, and had sufficient computing power to solve numerical problems in physics and astronomy when fully dedicated to that purpose. When mini-computers first became available, many areas of scientific research blossomed, including the study of how structure formed in the universe.

MOND: MOND, or Modified Newtonian Dynamics, is a theory that attempts to explain the evidence for dark matter as a modification to Newtonian gravity. There are many versions of the theory, all based on the premise that Newton's laws are slightly different at very small accelerations. A ball dropped above the surface of the Earth would not deviate noticeably from the path predicted by Newtonian physics, but the stars at the very edges of our galaxy would clearly demonstrate modified dynamics if MOND were correct.

N-body simulation: An N-body simulation is a computer simulation that involves a large number of particles interacting according to basic physical laws. N-body simulations are used to study how the structures in our universe may have evolved. Typically, many millions of particles are configured in an initial density distribution and allowed to interact according to the laws of gravity. The computer calculates how the particles will move under the influence of gravity in a small time step, and uses the resulting distribution of particles as the starting point for a new calculation. By calculating many time steps, the simulation can track the growth of structures in the model system. Depending on the initial density distribution and cosmological parameters selected, different structures appear at different stages of evolution. N-body simulations have provided strong support to the idea that our universe consists primarily of dark energy and dark matter. These simulations are resource intensive because the number of interactions the computer must calculate at each time step is proportional to the number of particles squared. A sophisticated N-body simulation can require tens of thousands of supercomputer hours.

neutralino: The neutralino is the superpartner of the neutrino. See: neutrino, superpartner, supersymmetry.

neutrinos: Neutrinos are fundamental particles in the lepton family of the Standard Model. Each generation of the lepton family includes a neutrino (see Unit 1, Fig. 18). Neutrinos are electrically neutral and nearly massless. When neutrinos are classified according to their lepton family generation, the three different types of neutrinos (electron, muon, and tau) are referred to as "neutrino flavors." While neutrinos are created as a well-defined flavor, the three different flavors mix together as the neutrinos travel through space, a phenomenon referred to as "flavor oscillation." Determining the exact neutrino masses and oscillation parameters is still an active area of research.

nucleosynthesis: The term "nucleosynthesis" refers either to the process of forming atomic nuclei from pre-existing protons and neutrons or to the process of adding nucleons to an existing atomic nucleus to form a heavier element. Nucleosynthesis occurs naturally inside stars and when stars explode as supernovae. In our standard model of cosmology, the first atomic nuclei formed minutes after the Big Bang, in the process termed "Big Bang nucleosynthesis."

parity: Parity is an operation that turns a particle or system of particles into its mirror image, reversing their direction of travel and physical positions.

singularity: Singularity is a mathematical term that refers to a point at which a mathematical object is undefined, either because it is infinite or degenerate. A simple example is the function 1/x. This function

has a singularity at x = 0 because the fraction 1/0 is undefined. Another example is the center of a black hole, which has infinite density. In our standard model of cosmology, the universe we live in began as a spacetime singularity with infinite temperature and density.

Sloan Digital Sky Survey: The Sloan Digital Sky Survey (SDSS) is one of the most extensive and ambitious astronomical surveys undertaken by modern astronomers. In its first two stages, lasting from 2000 to 2008, SDSS mapped almost 30 percent of the northern sky using a dedicated 2.5 meter telescope at the Apache Point Observatory in New Mexico. The survey used a 120-megapixel camera to image over 350 million objects, and collected the spectra of hundreds of thousands of galaxies, quasars, and stars. Notable SDSS discoveries include some of the oldest known quasars and stars moving fast enough to escape from our galaxy. SDSS data has also been used to map the distribution of dark matter around galaxies through observations of weak gravitational lensing and to study the evolution of structure in the universe through observations of how both galaxies and quasars are distributed at different redshifts. The third phase of the survey is scheduled to end in 2014, and is expected to yield many exciting scientific discoveries.

SQUID: A superconducting quantum interference device, or SQUID, is a tool used in laboratories to measure extremely small magnetic fields. It consists of two half-circles of a superconducting material separated by a small gap. The quantum mechanical properties of the superconductor make this arrangement exquisitely sensitive to tiny changes in the local magnetic field. A typical SQUID is sensitive to magnetic fields hundreds of trillions of times weaker than that of a simple refrigerator magnet.

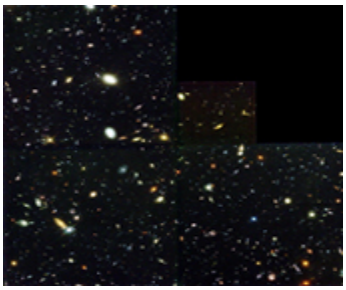
strong-CP problem: The strong-CP problem is a particular inconsistency between experimental observations of strong interactions and the theory that describes them. Unlike in weak interactions, CP violation has never been observed in strong interactions. However, the theory that describes strong interactions allows CP violation to occur. So why is it not observed? That is the strong-CP problem. Various attempts to resolve the strong-CP problem have been proposed, but none have been experimentally verified. See: axion, charge, CP violation, parity.

supersymmetry: Supersymmetry, or SUSY, is a proposed extension to the Standard Model that arose in the context of the search for a viable theory of quantum gravity. SUSY requires that every particle have a corresponding superpartner with a spin that differs by 1/2. While no superpartner particles have yet been detected, SUSY is favored by many theorists because it is required by string theory and addresses other outstanding problems in physics. For example, the lightest superpartner particle could comprise a significant portion of the dark matter.

WIMP: The WIMP, or weakly interacting massive particle, is a candidate for what may comprise the dark matter. WIMPs interact with other forms of matter through gravity and the weak nuclear force, but not through the electromagnetic force or the strong nuclear force. The lack of electromagnetic interactions means that WIMPs are nonluminous and therefore dark. They are assumed to be much heavier than the known Standard Model particles, with a mass greater than a few GeV. WIMP is a general term that can be applied to any particle fitting the above criteria. The neutralino, the supersymmetric partner of the neutrino, is an example of a WIMP candidate.

Physics for the 21st Century

Unit 11: Dark Energy



© NASA.

Unit Overview

This unit focuses on one of the biggest questions in 21st century physics: what is the fate of the universe? In recent years, astronomers have been surprised to discover that the expansion of the universe is speeding up. We attribute this to the influence of a "dark energy" that may have its origin in the microscopic properties of space itself. Even very simple questions about dark energy, like "has there always been the same amount?" are very difficult to answer. Observers are inventing programs to provide fresh clues to the nature of dark energy. Theorists hope to come up with a good new idea about gravity that will help us understand what we are seeing in the expansion causes the acceleration of the universe. Astronomers can observe the past but can only predict the future: if dark energy takes the simplest form we can think of, the universe will expand faster and faster, leaving our galaxy in a dark, cold, lonely place.

Content for This Unit

Sections:

Section 1: Introduction

We want to know what the universe is made of and how it has changed over time. Astronomers have been exploring these questions, and have discovered some surprising results. The universe is expanding, and the expansion is being accelerated by a "dark energy" that today apparently makes up more than 70 percent of the universe.

Figure 1: Type Ia supernovae, like the one shown here in the outskirts of galaxy NGC 4526, have been used to trace the influence of dark energy on the expansion of the universe.
Source: © High-Z Supernova Search Team, NASA/ESA Hubble Space Telescope.

The universe that contains our planet, our star, and our galaxy, as well as 10^{11} other galaxies and their stars and planets, obeys the same physical laws that we have uncovered in our exploration of nature here on Earth. By applying these laws, we have learned the scale of the universe, and the surprising fact that the other galaxies appear to be moving away from us as the universe stretches out in all directions. Astronomers detected this cosmic expansion in the 1920s. They understood it by applying Einstein's general relativity—the theory of gravity—to the universe as a whole. Recent work using exploding stars to measure the history of cosmic expansion shows that the universe is not slowing down due to the familiar braking action of gravity we know from everyday experience. The expansion of the universe has actually sped up in the last 5 billion years. In Einstein's theory, this can happen if there is another component to

the universe. We call that mysterious substance "dark energy," and we seek to discover its properties through astronomical observations.

Today's astronomical measurements show that dark energy makes up about 70 percent of the universe. So our ignorance is very substantial. This deep mystery lies at the heart of understanding gravity, which is no simple matter, as we saw in the first part of this course. Future observations will trace the growth of lumpy structures in the universe. These measurements can help distinguish between the effects of dark energy and possible imperfections in our understanding of gravity. We may face more surprises ahead.

This unit spells out how astronomers measure distance and velocity and describes how an expanding universe best explains these measurements. We show how recent observations of exploding stars, the glow from the Big Bang, and the clustering of galaxies make the case for dark energy as the largest, but least understood, component of the universe.

Section 2: Before the Beginning: A Static Universe

Our understanding of the universe has changed dramatically in the last century. One hundred years ago, astronomers thought the stars of our galaxy made up the whole universe, that the galaxy was nearly motionless, and that the universe was unchanging through time. Physicists trying to think of a model for the universe had to match these "facts."

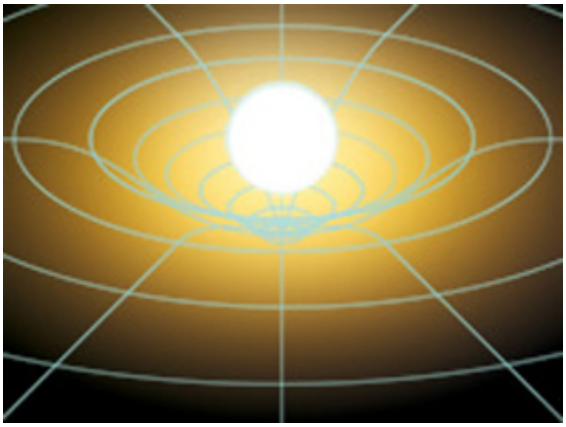


Figure 2: In Einstein's general relativity, mass warps the fabric of space.
Source: © NASA/STScI.

As Albert Einstein struggled to invent general relativity between 1914 and 1917, he was mindful of the possibilities for testing its predictions. Because gravity is so weak on the Earth, no laboratory measurements could test whether his ideas were right or wrong. So, Einstein looked to astronomers to help him find places to test his predictions, either because gravity was stronger (for example, near the edge of the Sun) or because the distances involved were large enough to show the cumulative effect of subtle differences from Newton's gravity. He posed three astronomical tests for the theory: the orbit of Mercury, the path of light near the limb of the Sun, and the effect of gravity on light from dense stars—tests that general relativity eventually passed *summa cum laude*.

Einstein's ideas about gravity were deeply original. He imagined that mass (and energy) would warp the fabric of space and time. Light or massive particles would then travel through this curved space. Einstein applied his equations to the universe as a whole, using the astronomical understanding of the day. In 1917, astronomers thought that our Milky Way galaxy, of which the Sun is an inconspicuous member, was, in fact, the whole universe.

Figure 3: Edge-on view of the Milky Way.
Source: © NASA Goddard Space Flight Center.

As far as astronomers knew at that time, the stars in the Milky Way were not moving in any systematic way. So when he wrote down an expression for the way gravity acts in the universe, Einstein added in an extra term to keep the universe static. This cosmological constant acted as a repulsive force that would balance out gravity and ensure that the universe would endure indefinitely without clumping together. Einstein found he could choose the value of the cosmological constant to produce just the right amount of curvature to make the universe "closed." This meant it behaved like the two-dimensional surface of a sphere, which has a finite surface area and has no edge, but Einstein was thinking of four dimensions of space and time. As Einstein apologized at the time, "...we admittedly had to introduce an extension of the field equations which is not justified by our actual knowledge of gravitation.... That [cosmological constant] term is necessary only for the purpose of making possible a quasi-static distribution of matter, as required by the fact of the small velocities of the stars."

Within a decade, astronomical observations showed that Einstein's picture of a static universe did not match the rapidly improving observational evidence of cosmic expansion. By 1931, Einstein considered the cosmological constant an unfortunate mistake. Today, however, careful measurements of distances and velocities from exploding stars observed halfway back to the Big Bang show that we need something very much like the cosmological constant to understand why the universe is speeding up.

Section 3: Discovery of the Expanding Universe

The surprising discovery that the universe is not static resulted from a long series of observational developments in astronomy. Establishing the distances to other galaxies and their recession from us was the work of many minds and hands. Building telescopes, making instruments to record and analyze the light they gather, deducing the properties of bright stars, applying measurements of those stars to measure distances, and heroic efforts to measure the spectra of galaxies all built the foundation for the discovery of cosmic expansion.

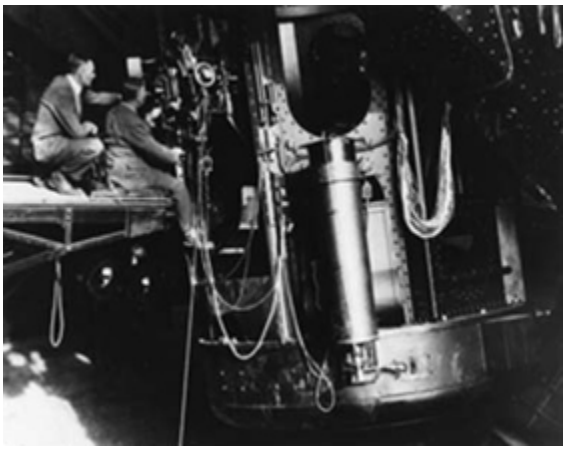


Figure 4: Edwin Hubble at the telescope.
Source: © AIP Emilio Segrè Visual Archives.

While Einstein was pioneering the theory of gravity, a technological revolution was under way in astronomy. Using photographs, astronomers began to measure the size of the Milky Way and began to study the fuzzy "nebulae" mixed in among the point-like images of stars. They found it difficult to determine what these pinwheel-like objects were because they did not know whether they were nearby small systems where one star was forming or distant large objects as big as the whole Milky Way. Distance measurements in astronomy are notoriously difficult and full of subtle errors. We can judge the distances to stars from their apparent brightness, but this can be deeply deceptive. If you look up on a summer night, you might see a firefly, a high-flying airplane, the planet Mars, the bright star Deneb, and M31, the Andromeda galaxy, all with about the same apparent brightness, even though it would take 10^{42} fireflies to emit as much light as a galaxy. To understand the real properties of these objects, we need to know how distance affects brightness.

In general, the brightness of an object falls off as the square of the distance to the object. This is called "the inverse square law." To use the inverse square law to determine the distance, however, you need to know how bright the object is to begin with. Stars come in an astonishing variety, from the dimmest brown dwarfs to the brightest blue supergiants. The power output, or luminosity, of stars whose distances we determine from geometry ranges over a factor of 10^{10}. If we were foolish enough to assume they were all the same as the Sun, we would introduce huge errors into our picture of how the Milky Way is organized. We need to find a way to identify stars that have the same luminosity—some stellar property that doesn't depend on the star's distance from us. See the math

Early Astronomical Computers

Henrietta Swan Leavitt
Source: © AIP Emilio Segrè Visual Archives, Physics Today Collection.

Before the emergence of the electronic version, the term "computer" referred to a person who carried out tedious, time-consuming measurements and calculations. Late in the 19th century, Edward Charles Pickering, director of the Harvard College Observatory, appointed a team of computers to measure the characteristics of stars in the observatory's 500,000 photographic plates. The team consisted entirely of women—mainly students or graduates of nearby Radcliffe College—and became known as "Pickering's Harem." Pickering was concerned about getting the most work done for the minimum expense. "A great observatory," he wrote, "should be as carefully organized and administered as a railroad." He noted that "a great savings may be effectuated by employing unskilled, and therefore inexpensive, labor, of course under careful supervision." However, the women who did these jobs turned out to have real talent for astronomy and published significant papers that eventually led to significant advances in many areas of astronomy, especially in understanding the physical nature of stars. The work of Henrietta Swan Leavitt on variable stars led to a revolution in understanding the scale of the universe.

A particularly helpful method came from studies of stars in the Magellanic Clouds, nearby satellites of our own Milky Way that we encountered in Unit 10. Careful studies of repeated photographic images revealed giant stars called Cepheid variables whose brightness increased and decreased in a rhythmic vibration repeated over a few days. Henrietta Swan Leavitt, a "computer" who studied the Cepheids, pointed out that "It is worthy of notice that... the brighter variables have the longer periods." Cepheids are what astronomers call standard candles, objects of known luminosity. If you find a Cepheid that has the same period as one of Henrietta Swan Leavitt's stars, no matter how bright it appears, you can assume

it has the same luminosity. If you know its intrinsic luminosity and how bright it appears, you can infer its distance.

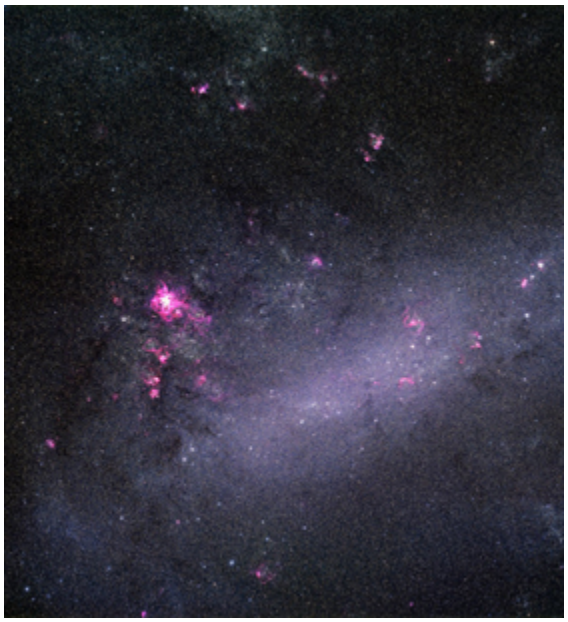


Figure 5: Large Magellanic Cloud.
Source: © ESO.

Measuring distances

Edwin Hubble, working at the Mount Wilson Observatory, home to the world's largest telescope, conducted a campaign to find Cepheid variables in the largest nebulae, M31, M33, and NGC 6822. His goal was to determine their distances and to find out whether they were small systems in the Milky Way or distant systems as big as the Milky Way. By 1925, he had a good sample of these vibrating stars. Like Henrietta Swan Leavitt, Hubble was able to measure their periods and their brightness from photographs. For Cepheid variable stars with the same periods, hence the same luminosities, as Leavitt's stars in the Magellanic Clouds, Hubble's stars appeared about 225 times dimmer. Since the brightness depends on the square of the distance, that meant that his stars were about $\sqrt{225}$ = 15 times more distant than the Magellanic Clouds. This placed these "nebulae" far outside the Milky Way. Einstein had tried to model a universe of stars assembled into the Milky Way, but Hubble glimpsed a much grander system. The universe was made up of galaxies as large as the whole Milky Way, separated by distances 10 times as big as their diameters.

To measure galaxy distances, astronomers use the light-year, the distance that light travels in one year. Since the speed of light is 3×10^8 meters per second and a year lasts about 3×10^7 seconds, a light year is about 10^{16} meters. A nearby star whose name we know, like Sirius, lies about eight light-years away; today, we see the light that left Sirius eight years ago, but the light from M31 takes about 2 million years to get here. So, the telescopes we use to view it act as no-nonsense time machines that allow us to record the earlier phases of cosmic evolution that occurred in the distant past. With modern equipment, we can easily detect galaxies billions of light-years away, and hence billions of years back in time.

The redshift and cosmic expansion

If Hubble had done nothing more than measure the true scale of the universe by measuring the period and brightness of Cepheids in the spiral nebulae, he would have sealed his place in the history of physics. However, he also used a different kind of measurement of the light from the nebulae to show that we live in a universe that is nothing like the static mathematical model that Einstein had constructed a few years earlier.

Hubble's Law in Different Dimensions

Imagine a classroom with students in rows of desks 1 meter apart. Suppose the whole array expands in 1 second; the desks are now 2 meters apart. The person who was next to you, 1 meter away, has moved away from you by 1 meter in 1 second—receding from you at 1 meter per second. But the next person over, in any direction, has gone from 2 meters away to 4, thus receding at 2 meters per second. Similarly, the next person beyond has moved away at 3 meters per second. Space that is stretching out in all directions will look just like Hubble's Law for everyone, with nearby objects moving away from you in all directions with velocities that are proportional to their distances.

That's just a two-dimensional example, but you could imagine a big 3D jungle gym made of live (and very fast-growing) bamboo. A monkey on this array would see the other monkeys receding as the bamboo grows, and would observe Hubble's Law in all directions.

Allusive talk about stretching classrooms and bamboo jungle gyms full of monkeys is not precise mathematical reasoning. However, the results are very similar when you formulate this problem more exactly and apply it to the universe.

Light from the galaxies would be stretched out by cosmic expansion, much as the sound from cars zooming by on a highway stretches out as they recede. For light, this means features in the spectrum of a receding galaxy are shifted to the red. For decades, measurements of galaxy velocities determined this way were accumulated at the Lowell Observatory by Vesto Melvin Slipher.

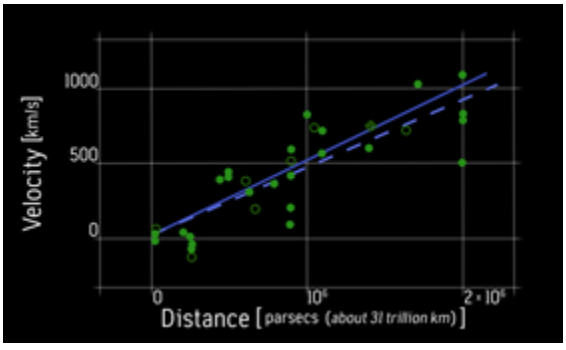


Figure 6: Hubble diagram, plotting velocity vs. distance for galaxies outside our own.
Source: Recreated from original plot by Edwin Hubble, March 15, 1929.

In 1929, Hubble combined his distance measurements from Cepheids with those velocity measurements in the form of a graph. Today, we call this plot the Hubble diagram. While the data were crude and limited, the relation between them was unmistakable: The velocity was proportional to the distance. Nearby galaxies are moving away from us slowly (and some of the very nearest, like M31, are approaching). As you look farther away, however, you see more rapid recession. You can describe the data by writing a simple equation v = H x d, where v is the velocity, d the distance, and H, the slope of the line is called the Hubble constant. We know the equation as Hubble's Law. See the math

The Hubble diagram shows a remarkable property of the universe. It isn't static as Einstein had assumed based on the small velocities of the stars in the Milky Way back in 1917; it is expanding. Those stars are not the markers that trace the universe; the galaxies are, and their velocities are not small. Even in this 1929 version of the Hubble diagram, the velocities for galaxies extend up to 1000 km/sec, much larger than the velocity of any star in the Milky Way.

Section 4: Mapping the Expansion with Exploding Stars

The galaxies we see with telescopes are massive objects, and the presence of mass in the universe should slow its expansion. Starting in the 1950s, astronomers saw their task in studying cosmology as measuring the current rate of cosmic expansion (the Hubble constant), and the rate at which gravity was slowing down that expansion (the deceleration.) With that information, they would be able to measure the age of the universe and predict its future. The deceleration would show up as a small but real departure from Hubble's Lbaw when the Hubble diagram is extended to very large distances, in the order of a few billion light-years. That's roughly 1,000 times farther away than the galaxies Hubble studied in the 1920s. The inverse square law tells us that galaxies that are 1,000 times farther away are $(1/1,000)^2$ times as bright. That's a million times dimmer, and it took several decades of technical advances in telescopes, detectors, and astronomical methods to compensate for this giant quantitative difference. Today's telescopes are 16 times bigger, our light detectors are 100 times as efficient, but we need to use much brighter stars than the Cepheids to look back far enough to see the effects of a changing rate of cosmic expansion.

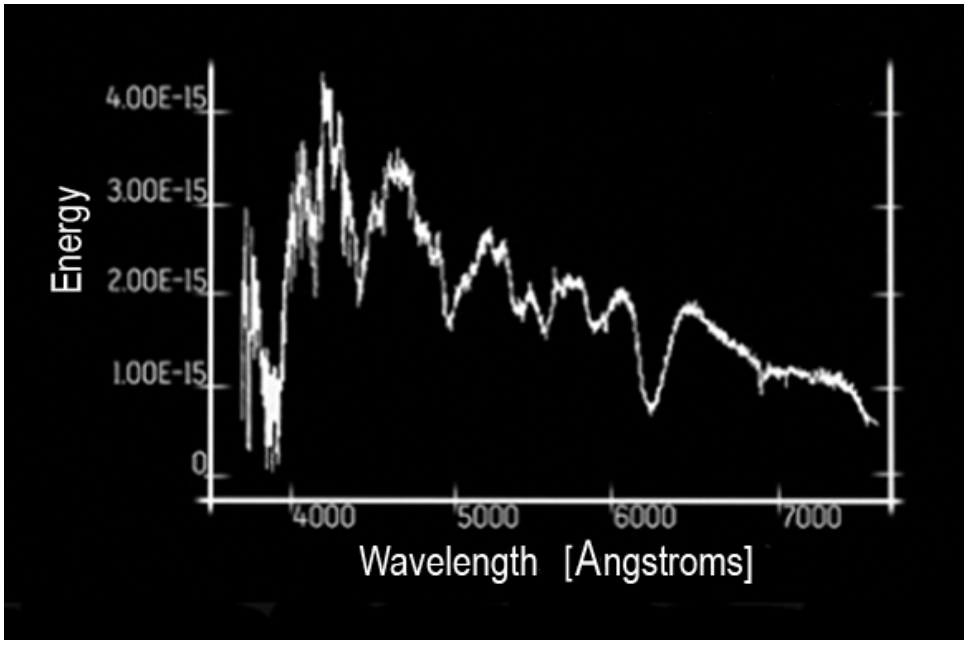


Figure 7: The spectrum of a Type Ia supernova, shown here, distinguishes it from other supernova types.
Source: Recreated from High-Z Supernova Team data, courtesy of Robert Kirshner.

Fortunately, nature has provided a brighter alternative. Some stars die a violent death in a blast of thermonuclear energy as a supernovae (SN) explosion. For a few weeks, a single star shines as brightly

as 4 billion stars like the sun. These thermonuclear supernovae, known as Type Ia supernovae, can be recognized from their spectra and, with some effort, used as standard candles for measuring cosmic distances.

The supernova challenge

Using Type Ia supernovae to measure cosmic acceleration or deceleration is no easy task. First, we have to find the supernovae; then we have to determine their distances and redshifts; and, finally, we have to find enough of them to make our measurement meaningful.

Although SN Ia are bright enough to see over distances of several billion light-years to detect the effects of cosmic acceleration or deceleration, they have one serious drawback: They are rare. Type Ia supernovae explosions take place about once per century in a typical galaxy. We cannot simply select the galaxies whose distances we wish to know. Instead, we have to inspect many galaxies to find the supernovae whose light has just arrived in the past week. Those are the ones we get to measure.

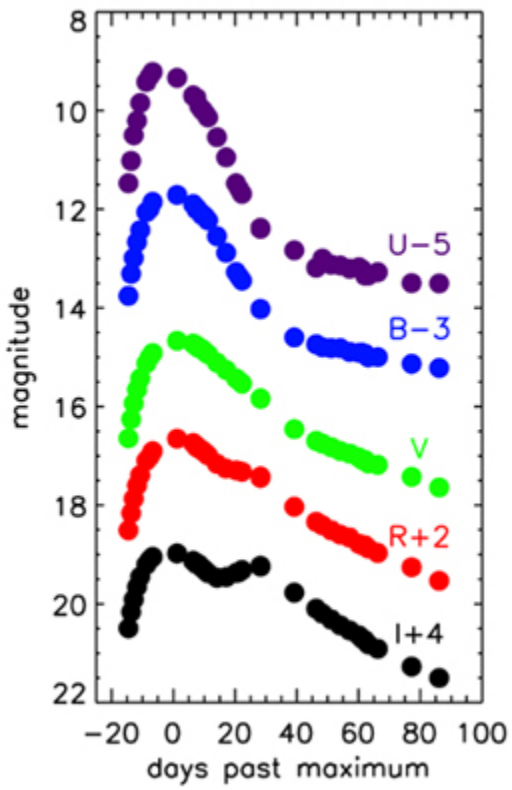


Figure 8: Light curve shape standardization.
Source: © Robert Kirshner.

How many galaxies do we need to search? Given that SN Ia appear about once per century per galaxy, and a year has about 50 weeks, we must search 5,000 galaxies to have a good chance of finding one within a week of its maximum light. This is a good job for a "computer"—but not the human kind.

Further, SN Ias don't possess the direct relation between brightness and vibration period exhibited by Cepheid variables. Fortunately, though, they have something similar. The light from an extra-bright SN Ia increases and decreases more slowly over time than that from a dimmer version. Careful study of the light curve can reveal which supernovae are extra bright and which are not so bright. Analyzing the light curve reduces errors in the distance to a single SN Ia to about 10 percent. This makes SN Ia plausible candidates for measuring the effect of cosmic acceleration or deceleration with a modest-sized sample, provided we look at SN Ia that are at a large enough distance that the effect of expansion, speeding up or slowing down, makes a 10 percent difference in the distance.

Finally, mathematical analysis shows that the uncertainty in the average value of something we're trying to measure becomes smaller as the square root of the number of times we repeat the measurement. In this case, a well-measured shape for the light curve of a single Type Ia supernova gave an uncertainty in a single distance of 10 percent. The effect due to cosmic deceleration was also expected to be about 10 percent for supernovae at a distance of a few billion light-years. So, an astronomer who wanted to make the uncertainty in the mean significantly smaller than the signal that would show evidence for cosmic deceleration would need at least nine objects to push the error down to about 3 percent (10 percent/$\sqrt{9}$) and about 100 to push the uncertainty in the mean down to 1 percent (10 percent/$\sqrt{100}$). Somewhere in that range, where the ratio of the expected signal to the uncertainty in the measurement is a factor of 3 to 10, astronomers might begin to believe that they have really measured something.

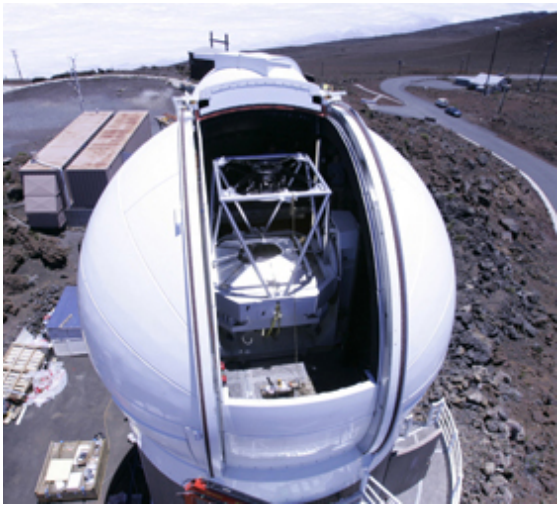


Figure 9: This Pan-STARRS telescope will find thousands of supernovae in the next decade.
Source: © Pan-STARRS, University of Hawaii.

If we need to search 5,000 galaxies to find one supernova, we'll need independent searches of 50,000 galaxies to find 10 and 500,000 galaxies to find 100. This became practical in the 1990s, when large digital cameras with tens of megapixels were mounted on 4-meter telescopes. In the coming decades, as the detector arrays grow, the hunt will reel in dozens of supernovae every night. Of course, this is still a small fraction of all the supernovae; we estimate that 30 explode somewhere in the universe every second. We have a long way to go before we will see them all.

The Astonishing Discovery of Cosmic Acceleration

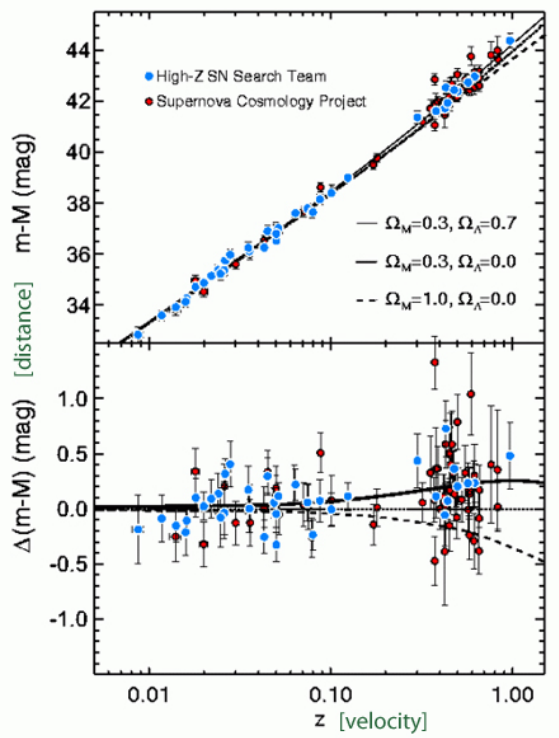


Figure 10: Adding high-redshift supernovae to the Hubble diagram revealed the effects of dark energy.
Source: © High-Z Supernova Search Team.

Starting in the mid-1990s, the pieces were in place for a direct approach to measuring cosmic deceleration; and by the end of the decade, two international teams were ready to publish their results. In September 1998, the High-Z Supernova Team published measurements of 16 distant and 34 nearby supernovae. To their surprise, the data pointed firmly toward *acceleration*. Instead of averaging a little brighter than expected for their redshift, this sample showed that the distant supernovae were about 25 percent dimmer than expected. This meant that the expansion of the universe was speeding up. Nine months later, the Supernova Cosmology Project reported on a larger sample of 42 distant supernovae. It agreed with the High-Z Team's results. It looked as if the universe was not decelerating, but accelerating. What could cause this? One possibility was that old bugaboo of cosmological theory, the cosmological constant.

Scientists are naturally suspicious of measurements and interpretations in new areas. However, subsequent work at large and small telescopes has augmented the samples of low redshift and high redshift supernovae into the hundreds. There's no escaping it now: The evidence from SN Ia shows that the universe is speeding up.

Section 5: Beyond Hubble's Law

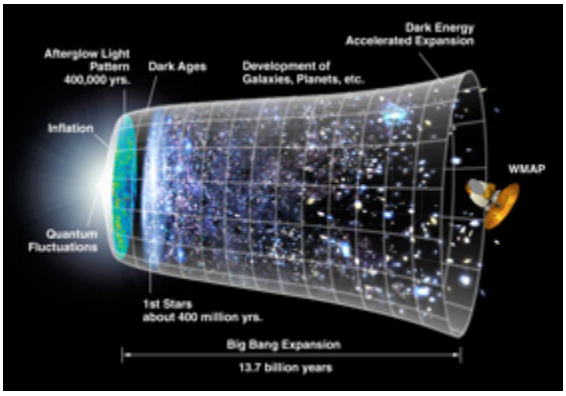


Figure 11: Dark energy is now the dominant factor pushing the universe to expand.
Source: © NASA, WMAP Science Team.

The supernovae results by themselves show that the universe is accelerating, but they don't say exactly what is causing it. Nor, by themselves, do they tell us how much of the universe is matter and how much is the agent causing the acceleration. The supernovae measure the acceleration in cosmic expansion, which stems from the difference between the component of the universe that makes things speed up (Ω_Λ) and the component that makes them slow down (Ω_M). Apparently, Ω_Λ now has the upper hand in the cosmic tug-of-war, but we'd like to know how much of the universe is in each form. We can obtain a much more complete idea of the contents of the universe by including other strands of evidence.

To determine the cosmology of the universe we live in, the two most effective pieces of information are the geometry of the universe, which tells us about the sum of the amount of matter and the cosmological constant (or whatever it truly is) driving the acceleration, and direct measurements of the amount of matter. In Unit 10, we learned that the cosmic microwave background (CMB) gives us direct information on the total amount of matter in the universe. It turns out that the CMB also gives us an excellent measurement of the geometry.

Density, geometry, and the fate of the universe

Einstein's theory of general relativity describes the connection between the density of the universe and its geometry. A particular value of the cosmic density, called the critical density, corresponds to flat space—the geometry of Euclid that is taught in schools. For cosmic expansion, this term depends on the current

rate of expansion—the Hubble constant. More precisely, the critical density is given by $\rho_{crit} = 8\pi H_0/3G$, where G is the gravitational constant that we first met in Unit 3.

The arithmetic shows that, for a Hubble constant of 70 km/sec/Mpc, $\rho_{crit} = 9 \times 10^{-27}$ kg/m^3. This is a significantly small number compared with the emptiest spaces we can contrive in a laboratory on Earth. It corresponds to about five hydrogen atoms in a cubic meter. Modern evidence, especially from the analysis of the cosmic microwave background that we encountered in Unit 10, shows that our universe has the geometry of flat space, but the sum of all the forms of gravitational matter is too small to supply the needed density.

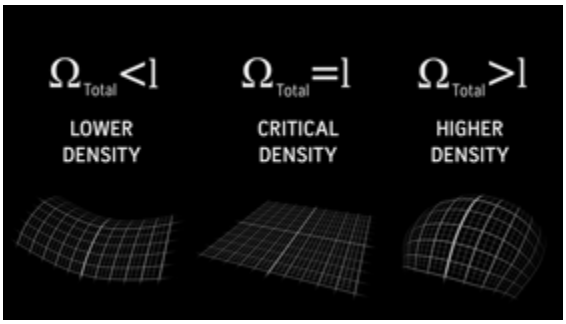


Figure 12: The geometry of the universe depends on its density.
Source:

Astronomers usually compare any density ρ they are measuring to the critical density (ρ_{crit}). We call this ratio omega (Ω) after the last letter in the Greek alphabet. (We should use the last letter to describe something that tells us how the world will end.) So $\Omega = \rho / \rho_{crit}$—a pure number with no units. The total density of the universe is simply the sum of the densities of all its constituents, so Ω_{total} is equal to the matter density Ω_M that we discussed in unit 10 plus the density of anything else out there, including the energy density associated with the cosmological constant, Ω_Λ. A value of Ω_{total} less than one means that the universe has an "open" geometry, and space is negatively curved like the surface of a saddle. If Ω_{total} is greater than one, the universe has a "closed" geometry, and space is positively curved like the surface of a sphere. And if Ω_{total} equals one, the geometry of the universe is that of flat space.

Modern evidence, especially from the analysis of the cosmic microwave background, shows that our universe has the geometry of flat space, but the sum of all the forms of gravitating matter is too small to

supply the needed density. The energy density associated with the cosmological constant, we believe, makes up the difference.

The anisotropic glow of the cosmic microwave background

The measurement of cosmic geometry comes from observations of the glow of the Big Bang, the cosmic microwave background (CMB). As we saw in Unit 10, Bell Labs scientists Arno Penzias and Robert Wilson observed this cosmic glow to be nearly the same brightness in all directions (isotropic), with a temperature that we now measure to be 2.7 Kelvin.

Although the CMB is isotropic on large scales, theory predicts that it should have some subtle texture from point to point, like the skin of an orange rather than a plum. The size of those patches would correspond to the size of the universe at a redshift of 1,000 when the universe turned transparent, and the radiant glow we see today was released. We know the distance to these patches, and we know their size: The angle they cover depends on how the universe is curved. By measuring the angular scale of this roughness in the CMB, we can infer the geometry of the universe.

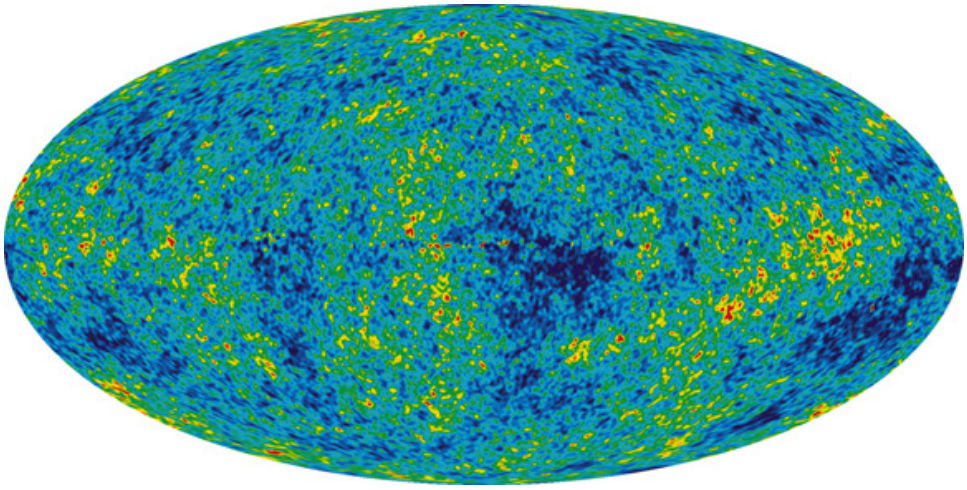


Figure 13: The cosmic microwave background, shown here, has subtle texture from point to point.
Source: © NASA, WMAP Science Team.

The technical difficulty of this measurement is impressive: The variations in temperature that we must measure correspond to about 0.001 percent of the signal. However, in 2000, astrophysicists measured the angular scale well from Earth and better from the WMAP satellite three years later. The angular scale is just about 1 degree, which corresponds with amazing precision to the angle we would measure if the

geometry of the universe as a whole were Euclidean. To an uncertainty of just a few percent, the universe is flat, and $\Omega_M + \Omega_\Lambda$ is 1.

The results from Unit 10 are consistent with Ω_M of about 1/3. If the total of Ω_M and Ω_Λ is 1, as the angular scale of the temperature fluctuations in the CMB strongly suggest, this suggests that 2/3 of the energy density in the universe is made up of Ω_Λ. Not only is there something driving cosmic acceleration, as the supernovae show, but the CMB observations also require a lot of it.

Summing up

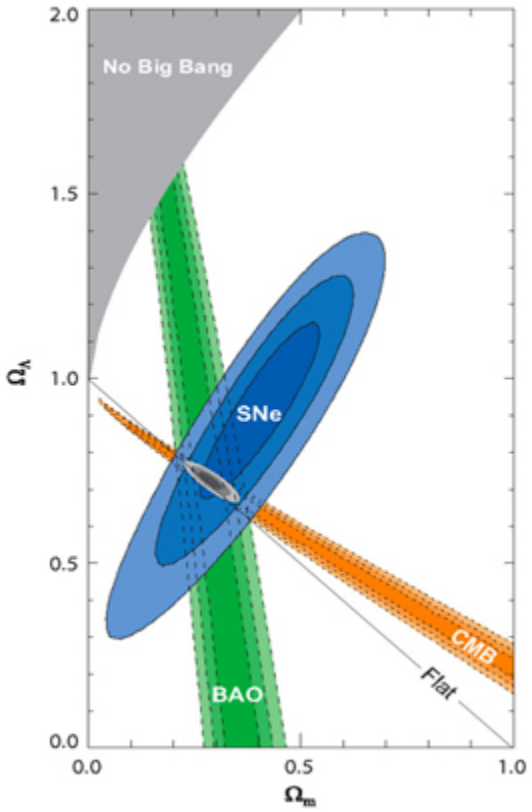


Figure 14: Combining evidence from supernova and the CMB makes a strong case for dark energy.
Source: © Supernova Cosmology Project, Lawrence Berkeley National Laboratory.

The CMB gives us a way to measure the geometry of the universe, which tells us about the sum Ω_M and Ω_Λ, the CMB as well as measurements of galaxy clustering tell us about Ω_M, and the supernova results

tell us about Ω_Λ independently. We can compare the components best in a graph that plots the value of Ω_M on one axis and Ω_Λ on the other. The supernova results (shown in blue) are broadly consistent with a line that looks like a constant value of $\Omega_M - \Omega_\Lambda$. The CMB results (shown in orange) are approximately what we would get for a constant value (of 1) for the sum of these two components of the universe. On the plot, these represent two lines that are very nearly perpendicular. They cross at a value of about 1/3 for Ω_M and 2/3 for Ω_Λ. Of course, they have to cross somewhere. So how do we know how much to trust this result?

A separate line of evidence employs data on the way that the action of gravity clusters galaxies and the imprint of the very subtle early density fluctuations on today's pattern of galaxy locations. Both depend chiefly on the value of Ω_M, and they give another constraint in the diagram (shown in green), as a nearly vertical line at Ω_M equaling about 0.3. While two lines in a plane have to cross, there's no guarantee that three lines will cross in the same place, but they do. We take this as a sign that the picture we are developing for the universe is a reliable one and that a mixed dark matter and dark energy universe with 1/3 dark matter and 2/3 dark energy is about right.

Section 6: The Concept of Dark Energy

It's a good thing that we have such confidence in the concordance of the measurements described in this unit, because the overall picture they indicate is completely astonishing. If we take the measurements seriously—and we should—they point to a universe dominated by something that acts to make the universe speed up, partially balanced by matter that makes galaxies clump together. However, two large and deep mysteries surround this view.

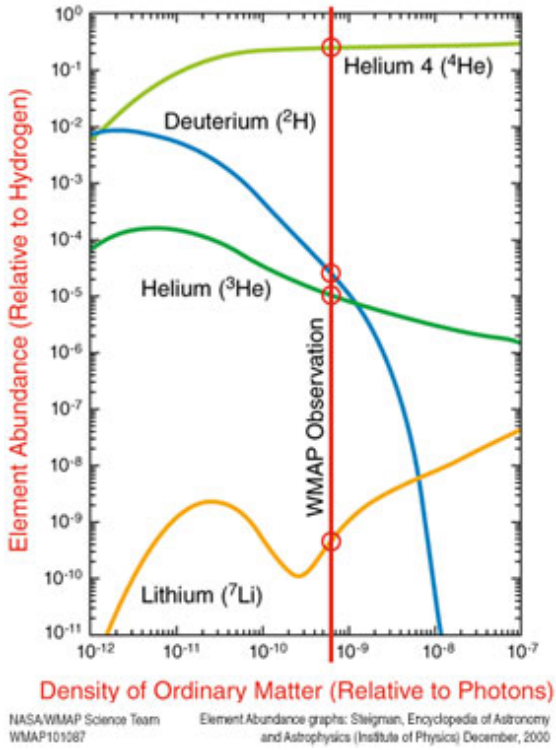


Figure 15: The abundance of light elements indicates that most of the universe is not protons, neutrons, or electrons.
Source: © NASA, WMAP Science Team.

One is that most of the matter in the universe cannot take the form that makes up galaxies, stars, and people: the familiar elements of the periodic table and their subatomic constituents of protons, neutrons, and electrons. Based on our understanding of the way light elements such as helium would form in the hot Big Bang, we know that this nuclear cooking can agree with the data only if most of the universe consists of something that is not protons, neutrons, or electrons. We call this dark matter, but, as we saw in Unit 10, we don't know what it is.

The other mystery stems from the fact that the cosmological constant is not the only candidate for the component of the universe that makes expansion speed up. If we try to estimate the energy density of the cosmological constant from basic quantum mechanical principles, we get a terrible quantitative disagreement with the observations. The computed number is at least 10^{60} times too large. Before the discovery of cosmic acceleration, physicists took this disagreement to mean that somehow nature covers up for our ignorance, and the real value is zero, but now we know that can't be right.

A dense web of evidence tells us that the energy associated with gravity acting in empty space is not exactly zero, but it isn't the gigantic value computed from theory, either. Clearly, something is missing. That something is a deeper understanding of how to bridge the two great pillars of modern physics: quantum mechanics and general relativity. If we had a good physical theory for that combination, presumably the value of the cosmological constant would be something we could predict. Whether that hope is valid or vain remains to be seen. In the meantime, we need a language for talking about the agent that causes cosmic acceleration.

Dark energy, dark matter, and the cosmological constant

To cover the possibility that the thing that causes the acceleration is not the cosmological constant, we use a more general term and call it dark energy. For example, dark energy might change over time, whereas the cosmological constant would not. To express our combination of confidence that dark energy is real and our ignorance of its precise properties, we describe those properties by talking about dark energy's equation of state—the relation between the pressure of a gas and its density.

The cosmological constant doesn't act like any gas we have ever used to squirt paint from an aerosol can. We're used to pressure going down as the gas expands. If dark energy is really a constant energy density, as it would be if it were identical to the cosmological constant, then the vacuum energy in each cubic meter would remain the same as the universe expands. But if dark energy behaves slightly differently from the cosmological constant, that energy density could go up or down; this would have important, and possibly observable, consequences for the history of cosmic expansion.

Taking our best estimates for the fraction of the gravitating matter that is dark matter and the fraction associated with the glowing matter we see, and assigning the convergence value to dark energy yields the amazing pie chart diagram for the universe that we encountered in Unit 10. To our credit, the diagram

is full. However, it is full of ignorance: We don't know what dark matter is, and we don't know what dark energy is. In science, this is not such a bad situation to be in. It leaves plenty of room for future discovery.

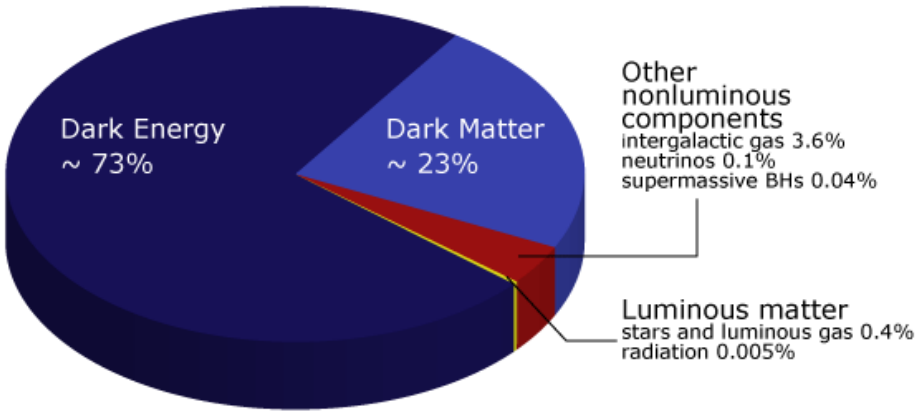


Figure 16 : The composition of the universe, with 96 percent invisible and unfamiliar.
Source:

We have some ideas of how to proceed in learning more about the nature of both dark matter and dark energy. As we saw in Unit 10, physicists have embarked on several efforts to determine the identity of dark matter, using beams in accelerators, detectors in underground laboratories, and instruments in space. All of this is quite speculative, but very exciting. We could well be on the threshold of making a crucial measurement about dark matter.

Our expectations for learning more about dark energy are more modest. Larger, more sophisticated, and more certain surveys of supernovae, galaxy clustering, and other tracers of the way the universe has changed over time provide the most promising path forward.

Section 7: From Deceleration to Acceleration

We have learned that the universe—far from being small and static as was thought in Einstein's day—is large and dynamic. At present, it is expanding at an ever-faster rate. As it does so, the relative amounts of matter and dark energy change. If this version of cosmology is correct, we can use it to reconstruct the past and predict the future. Whenever possible, we test these predictions against observation.

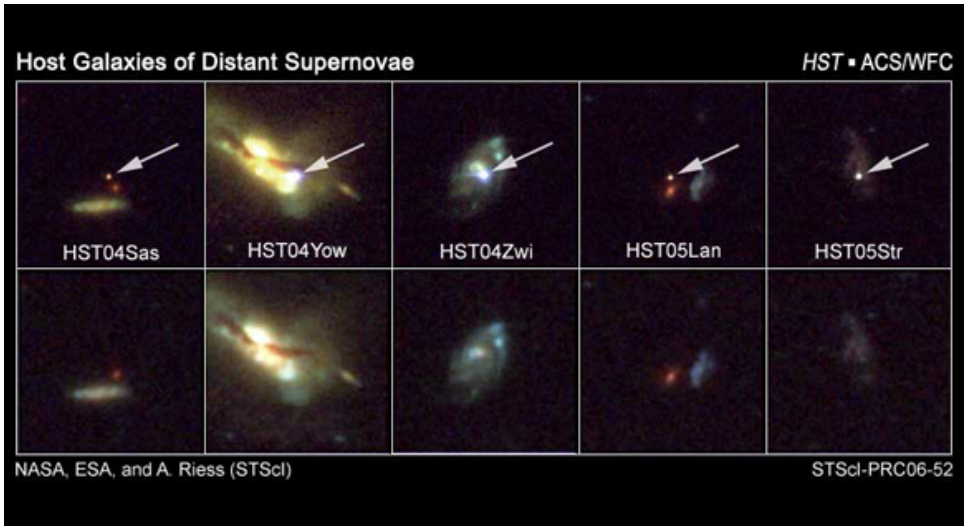


Figure 17: High-redshift supernovae, together with their host galaxies.
Source: © NASA, ESA, A. Riess (STScI).

SN Ia provide a proven method for tracing the history of cosmic expansion. One direction that has proven fruitful has been to press the search for supernovae to fainter fluxes and hence larger distances. That allows us to probe the accumulated effect of cosmic expansion over ever-longer stretches of time. The initial measurements from 1998 had their strongest information at distances that correspond to about 5 billion light-years. But it is possible, with great effort, to push these measurements to earlier epochs. Astrophysicists have done this most effectively using the Hubble Space Telescope (HST), which is no larger than the telescope Hubble himself used at Mount Wilson, but observes from a much better site, above the blurring caused by the Earth's atmosphere.

Looking back at the past

Hubble in—and out of—Trouble

Hubble spacecraft.
Source: © NASA/STScI.

When NASA launched the Hubble Space Telescope into an orbit about 600 kilometers above Earth in April 1990, astronomers anticipated a cornucopia of new observations. But, within a few weeks, a serious problem emerged: The telescope's primary mirror had been precisely figured, but to the wrong shape. Fortunately, STScI engineer Jim Crocker devised a clever correction inspired by the folding spray head of the shower at the Hoyacker Hof in Garching. His plan involved adding two small new mirrors that unfolded into the light path and corrected for the telescope's main mirror. In December 1993, a team of astronauts aboard the space shuttle installed the "corrective optics space telescope axial replacement" package. The fix worked perfectly as a stopgap, and all the instruments brought up to HST since then have had their own corrections built in.

A decade after the successful repair, the HST faced another crisis. Sean O'Keefe, NASA's administrator at the time, canceled the final servicing mission to the telescope, scheduled for early 2005, asserting that the mission involved too much risk for the astronauts. However, astronomers were not willing to abandon one of NASA's most productive scientific satellites without a fight. They mounted a campaign that persuaded O'Keefe's successor, Michael Griffin, to reschedule the mission. Carried out in May 2009, it left the Hubble in good health. Astronomers hope that HST will be pouring out valuable data on cosmic questions for many years to come.

If the universe consists of a combination of dark energy and dark matter, adding up to a Ω_{total} of 1, the equations of cosmic expansion guarantee that at earlier times (and later) the Ω_{total} will remain 1. But the

balance of dark energy and dark matter is expected to change with cosmic epoch. As we look into the past, we see that the density of dark matter must have been higher than it is now. After all, every cubic meter of the universe we see today has stretched out from a smaller region. Compared to redshift 1, the length of any piece of the universe has doubled; so the volume of space that occupies one cubic meter today took up only 1/2 x 1/2 x 1/2, or 1/8, of a cubic meter back then. If you had the same matter in that volume at redshift 1 as we do today, then the density would have been higher in the past, by a factor of 8.

How this affects cosmic acceleration depends on the difference between dark matter and dark energy. If dark energy is identical to the cosmological constant, it lives up to its name by having a constant value of energy density. In that case, we expect that, when we look back to redshift 1, we would find dark matter to be eight times more important than it is today, while dark energy would show no change. This means that dark matter, the gravitating material, would have had the upper hand, and the universe should have been decelerating.

This is not just a matter for speculation. Using the HST, we can find and measure supernovae at this epoch, halfway back to the Big Bang. Measurements of 23 supernovae with large redshifts discovered and measured with HST were reported in 2004 and 2007 by Adam Riess and his colleagues. The data show that we live in a universe that was slowing down about 7 billion years ago. The balance shifted somewhere around 5 billion years ago, and we now live in an era of acceleration. In more detail, the data now in hand from 400 supernovae near and far allow us to trace the history of cosmic expansion with some precision. A dark energy that acts like the cosmological constant would be adequate to fit all the facts we have today. That doesn't mean this is the final answer, but a better sample will be needed to find out if the universe has a more complicated form of dark energy.

The prospect of a lonely future

A telescope lets us collect light emitted in the distant past. Looking into the future is more difficult, since we have only human minds as our tools, and our present understanding of dark energy is incomplete. But if dark energy acts just like the cosmological constant, the future of cosmic expansion is quite dramatic.

A constant dark energy produces exponential expansion. The larger the universe becomes, the faster it will expand. The expansion can become so fast that light from distant galaxies will never reach us. Even galaxies we see now will be redshifted right out of our view; so as the universe ages, an observer at our location will see fewer galaxies than we can see today.

Figure 18: Hubble Space Telescope panoramic view of thousands of galaxies in various stages of evolution.
Source: © NASA, ESA, R. Windhorst, S. Cohen, M. Mechtley, and M. Rutkowski (ASU, Tempe), R. O'Connell (UVA), P. McCarthy (Carnegie Observatories), N. Hathi (UC, Riverside), R. Ryan (UC, Davis), H. Yan (OSU), and A. Koekemoer (STScI).

If we follow the logic (and assume that our present understanding is perfect), eventually our Milky Way Galaxy and nearby Andromeda will be separated from this outwardly accelerating scene, and Andromeda and our other local neighbors will be the only galaxies we can see. Worse, we will eventually collide with Andromeda, leaving just one big galaxy in an otherwise empty universe. In a strange way, if this prophecy for the long-term future of our galaxy comes true, it will produce a situation much like the picture of the Milky Way as the entire universe that prevailed in Einstein's time.

Section 8: Dark Energy Theory

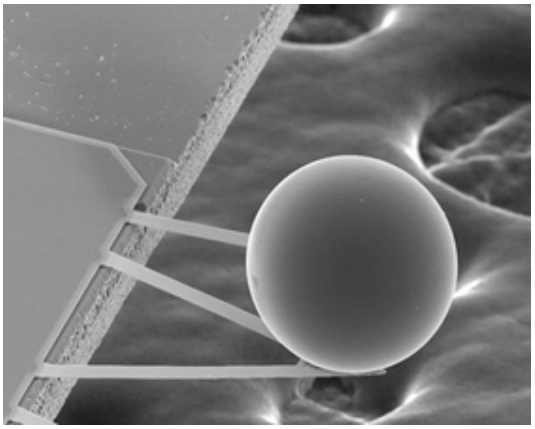


Figure 19: Precise laboratory experiments like the one shown here measure the energy of empty space.
Source: © Umar Mohideen, University of California at Riverside.

The advent of cosmic acceleration about 5 billion years ago looks just like the change in the expansion rate that the cosmological constant would produce. Does this prove that dark energy is a modern version of the cosmological constant? Not exactly. The modern view of the cosmological constant is that the vacuum—empty space itself—has some properties that we can understand only by using quantum mechanical ideas. In the case of electromagnetism, for example, the quantum picture views the vacuum not as an inert background on which the electrical and magnetic forces act, but on the submicroscopic scale, as a seething froth of particles and their antiparticles that are being created and annihilated all the time.

One way to think of this busy scene on very small scales involves the Heisenberg uncertainty principle that we encountered in Units 2 and 5. This tells us that the better we know the location, the more uncertain is our knowledge of the energy at that place. If we insist on looking on very fine spatial scales (much smaller than an atom), the energy could be large enough to create a particle and its antiparticle. These particles would find each other soon enough and annihilate. If we look on a big scale, the average value of their density would be zero, but on a very small the fluctuations about zero would be quite large.

For electromagnetism, this picture of the vacuum makes a subtle difference to the forces we predict between charged particles. Physicists can test these predictions in high-precision laboratory experiments. The measurements agree better with this picture than with one in which the vacuum is a featureless

mathematical space for electric and magnetic fields. So, this seems like the right way to proceed for other forces.

An appalling disagreement on vacuum energy

If we do the same kind of analysis of the vacuum for the gravitational force, we find that, because of that force's extreme weakness, the appropriate length scale is much smaller than for electromagnetism and the resulting energy fluctuations are much larger. In units where the observed value of the cosmological constant is 0.7, the calculated value for the vacuum energy associated with gravity is not 1 (which would be close enough) or 10, 100 (10^2), 1000 (10^3), or even 10^4. It's at least 10^{60}.

This is not just a small numerical puzzle: It is the worst quantitative disagreement in all of physical science. For decades, physicists have swept this strange result under the rug. But now that we have a real astronomical measurement of the effects of vacuum energy, it seems to demand an explanation. Why is the energy of the vacuum so small?

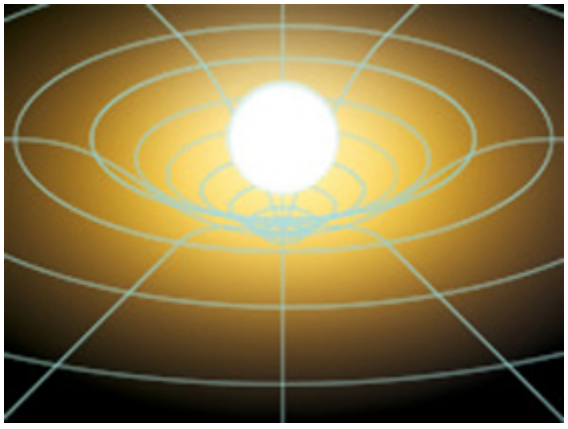


Figure 20: In Einstein's theory of gravity, space is warped but featureless.
Source: © NASA/STScI.

The honest answer is that we don't know. That's why the discovery of cosmic acceleration points directly to a problem at the heart of physics: What, exactly, is gravity? Or, more specifically, what is the right way to incorporate quantum ideas into the theory of gravity? Einstein's gravity is not a quantum theory. It is one in which a featureless mathematical space is warped by the presence of mass and energy and through which massive particles and photons travel. The appalling discrepancy between the predictions of theory and the astronomical observations has led to some novel ideas that seem a bit odd.

The anthropic principle: We're here because we're here

One novel idea posits many possible universes that make up a more varied "multiverse" of more or less unrelated regions. Each universe might have its own set of physical constants that governs such factors as the energy of the vacuum. If we happen to be in a region of the multiverse where that value is big, the acceleration sets in immediately, gravity never gets a chance to pull matter together, galaxies never form, stars never form, and interesting chemical elements like carbon are never produced. This boring universe contains no life and nobody to say proudly that "the energy of the vacuum is just as we predicted". Even though it could be that this large value for vacuum energy is the norm and our patch of the multiverse has an extremely low and unlikely value for the vacuum energy, we can't be completely surprised that a place also exists in which galaxies did form, stars did form, carbon was produced, and the living things on a planet would scratch their heads and ask, "Why is our vacuum energy so low?" If it hadn't been, they—or, more accurately, we—wouldn't be there to ask.

This "anthropic" idea—that the presence of humans tells us something about the properties of the universe in which we live—is quite controversial. Some people regard it as unscientific. They say that our job is to figure out why the universe is the way it is, and that invoking this vague notion is giving up too easily on the quest for understanding. Others think it trivial: Of course we're here, they say, but that doesn't help much in discovering how the world works. Still others are convinced that we don't have any better explanation. For them, a multiverse with many chances for unlikely events to happen, combined with the anthropic principle that selects our unlikely universe, seems the best way to make sense of this very confusing topic.

Section 9: Further Studies of Dark Energy

The problems posed by dark energy are fundamental and quite serious. While the observational programs to improve our measurements are clearly worth pursuing, we can't be sure that they will lead to a deeper understanding. We may need a better idea about gravity even more than precise determinations of cosmic expansion's history, and it seems likely that a truly new idea will seem outrageous at first. Unfortunately, the fact that an idea is outrageous does not necessarily mean that it is a good one. Separating the wild speculations from the useful new ideas is a tricky task, but better observations will help us to weed out some of the impossible ideas and let us see which of the remaining ones best fit the data.

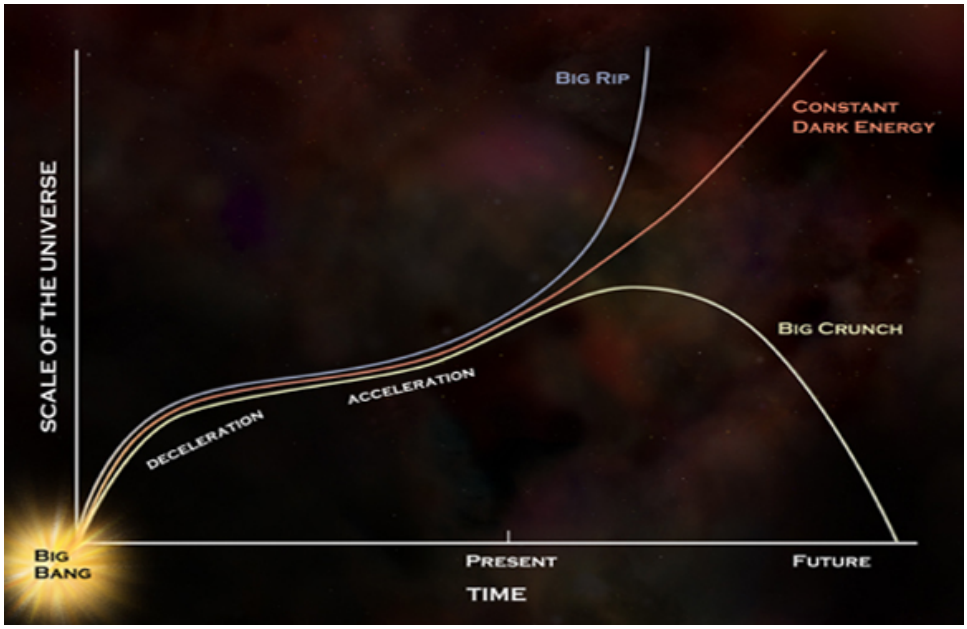


Figure 21: The future scale of the universe depends on the nature of dark energy.
Source: © NASA, CXC, M. Weiss.

One way to approach dark energy is to try to pin down its equation of state. So far, our measurements of the cosmic equation of state are completely consistent with the cosmological constant, but perhaps some variation will show up when we do more precise measurements. Any deviation from a constant energy density would show that the cosmological constant idea was not right, and that we require something more complicated to match the facts for our universe.

Another way to think about cosmic acceleration is to ask whether it is telling us something new about general relativity. After all, though Einstein's theory is a beautiful one and has passed every experimental test to which it has been subjected, it dates back to 1917. Possibly there is some aspect of gravity that he did not get exactly right or that would show up only on cosmic length scales. So far, we have no evidence that Einstein's geometrical theory needs amendment, but physicists are always alert for possible cracks in the foundation of our understanding. They want to use astrophysical measurements to test whether general relativity needs to be modified.

Identifying galaxy clusters

To date, the predictions that best test whether general relativity is the right theory of gravity to account for an accelerating universe are those that predict how matter will clump and structure will grow in an expanding universe. By comparing clusters of galaxies, which are the largest aggregations of matter we know of, at large and small distances, astronomers have been able to put some useful limits on this aspect of gravity. Again, everything so far agrees with general relativity: Only more precise measurements could detect a small deviation from that picture.

For more precise measurements, we need better samples of galaxy clusters. Finding such clusters in optical images is tricky because some of the galaxies are in the foreground and others are in the background, making the more distant clusters more and more difficult to distinguish. Since the whole measurement depends on comparing the numbers of distant and nearby clusters, this is a dangerous sort of bias.

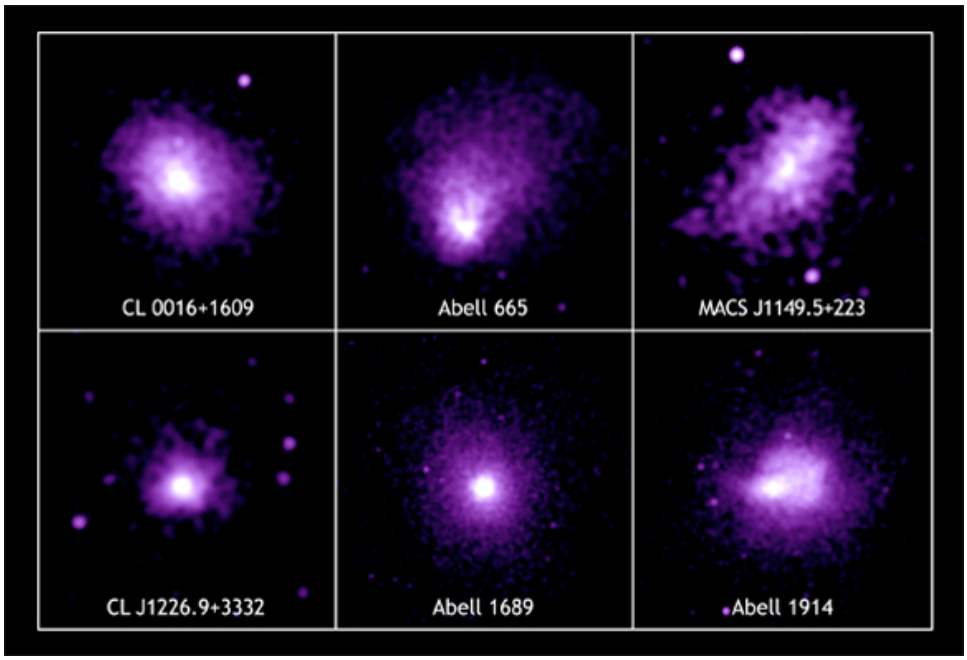


Figure 22: X-ray images of galaxy clusters observed with the Chandra X-Ray Observatory.
Source: © NASA, CXC, MSFC, M. Bonamente et al.

A better way to find galaxy clusters relies on their emission of x-rays. The gravitational well formed by 10^{14} solar masses of (mostly dark) matter in a galaxy cluster means that the gas (mostly hydrogen) that falls into the cluster or that is exhaled from the galaxies can gain a lot of energy in the process. It must be very hot to have enough pressure to keep from collapsing to the center of the cluster. The temperature for the gas in clusters is about 10^7 K, and the emission from such a hot ionized gas occurs principally in the x-ray part of the spectrum.

X-ray telescopes can search the sky for large sources of x-ray emission, and in some cases we can identify the sources with galaxy clusters. However, the inverse square law applies to x-rays just as it does to optical light. So, the more distant clusters are fainter, and the sample becomes more incomplete and harder to interpret as you observe fainter clusters. Although our best data so far comes from the x-ray selected samples, astronomers have a better technique.

Seeking patterns in the CMB

The signature of hot gas in a cluster of galaxies includes more than emitted x-rays. The fast-moving electrons in the gas can sometimes collide with photons from the cosmic microwave background. These

collisions kick up the low-energy radio photons to higher energies. The interactions show up as distinct patterns in the CMB in the neighborhood of a cluster.

The map of the CMB usually shows only slight temperature variations from point to point, but the lack of low-energy photons reveals itself as a large cold spot in the map. If you tune your radio receiver to a slightly higher energy, you'll see a bright area that contains the extra photons kicked up to higher energy by the collision with the fast electrons in the cluster. In 1969, Physicist Yakov Zel'dovich and his student Rashid Sunyaev, now a distinguished astrophysicist, worked out the theory of this pattern. It is only recently that the Sunyaev-Zel'dovich effect has become a practical way to find clusters of galaxies.

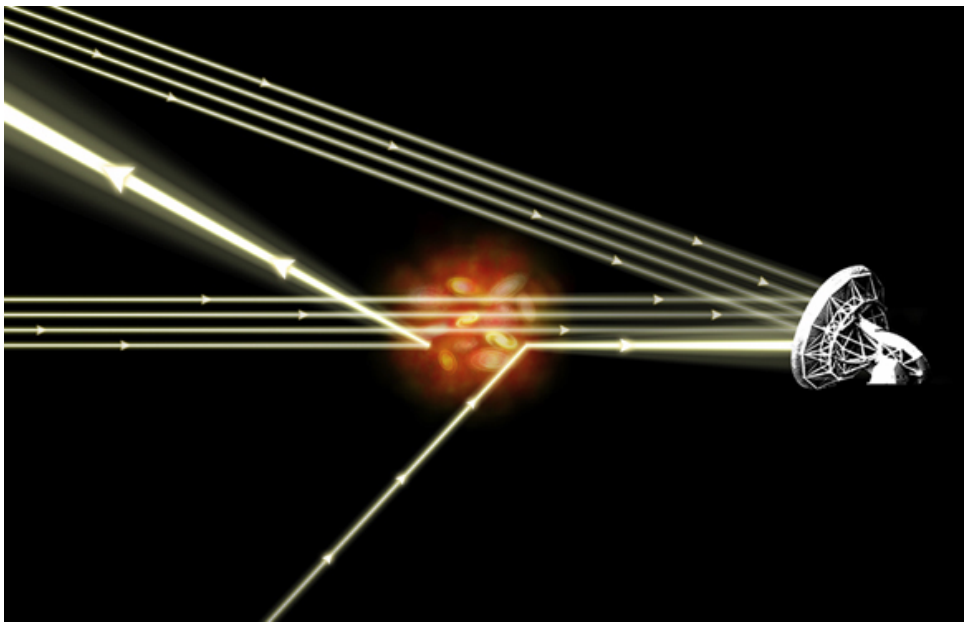


Figure 23: The Sunyaev-Zel'dovich effect allows astronomers to find the signature of galaxy clusters in the CMB.
Source: © NASA, CXC, M. Weiss.

Using the South Pole Telescope, which observes the CMB from Antarctica, astronomers have started to prepare a new, large, and uniform catalog of clusters. One surprising feature of the Sunyaev-Zel'dovich measurements is that the distinctive signature of a cluster does not depend on its distance. So, this method should work just as well at finding distant clusters as at finding nearby ones. This seems likely to be the best way to measure the growth of structure in the universe in the years ahead, and to test whether Einstein's theory of gravity is enough to account for all the facts in our accelerating universe.

Future searches for dark energy

Future measurements will include better samples of supernovae to measure the expansion history of the universe, but astronomers are also developing new ways of monitoring the properties of the universe. These include looking directly at the clustering of galaxies and, less directly, at the gravitational lensing that clumped matter produces.

As shown in Units 3 and 10, mass in the universe curves space. This means that the mass associated with galaxies can act like lenses that distort and magnify the scene behind them. We see this effect in galaxy clusters, which form long, thin arcs of light by bending the light from galaxies behind them. Weaker lenses can distort the shape of background galaxies. By carefully measuring how images of galaxies are warped at low and high redshifts, we can construct another picture of the way in which mass has grown more concentrated over time. This will give us a further clue to whether the growth of structure in the universe we live in matches or contradicts the predictions of general relativity.

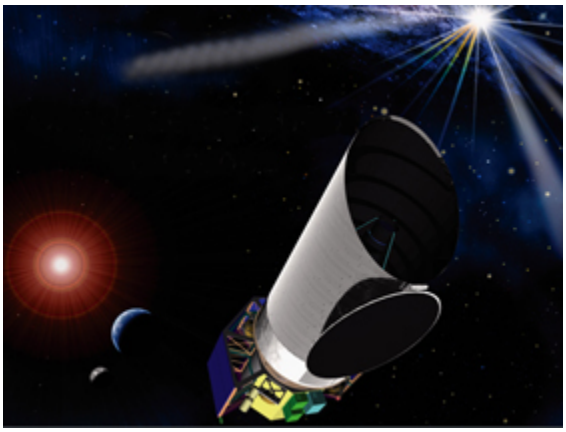


Figure 24: The Joint Dark Energy Mission will make precise measurements of the effects of dark energy from space.
Source: © NASA, GSFC.

Measuring the properties of the universe is the only path we have for learning the properties of dark energy and testing whether Einstein's gravity theory gives us the whole story for assembling galaxies and clusters out of a very subtly rippled past. What's missing is an experimental test on the laboratory scale, or even the solar system scale, that would show us the presence and properties of dark energy. Unlike the case of dark matter, where it seems that we have the technology to detect the phenomenon and are on the brink of doing so, nobody has a clue about how to make a laboratory experiment to detect dark energy. Just as Einstein had to rely on astronomy to test his theory of general relativity, our only "laboratory" for measuring dark energy seems to be the universe itself.

Section 10: Further Reading

- Timothy Clifton and Pedro Ferreira, "Does Dark Energy Really Exist?" *Scientific American*, April 2009, p. 48.
- Robert Kirshner, "The Extravagant Universe," Princeton University Press, 2004.
- Michael Lemonick, "Echoes of the Big Bang," Princeton University Press, 2005.
- Richard Panek, "Going Over the Dark Side," *Sky and Telescope*, Feb. 2009, p.22.

Glossary

Cepheid variable stars: Cepheid variable stars are high-luminosity stars that undergo very regular variations in brightness. A typical Cepheid will dim to a fraction of its maximum brightness and then grow brighter again with a period ranging from a few days to several months. During this cycle, the star is moving between two different states. At maximum brightness, the star is more compact and hotter, and pressure within the star causes it to expand. As the star expands, the pressure is released and the star cools. Eventually, the force of gravity is stronger than the outward pressure on within the star, and it collapses in on itself, heating, becoming brighter, and starting the cycle over again. The absolute luminosity of Cepheids, which are 5 to 20 times more massive than the Sun, is related in a precise way to the period of the brightness oscillation, which allows them to be used as standard candles. See: luminosity, standard candle.

cosmic microwave background: The cosmic microwave background (CMB) radiation is electromagnetic radiation left over from when atoms first formed in the early universe, according to our standard model of cosmology. Prior to that time, photons and the fundamental building blocks of matter formed a hot, dense soup, constantly interacting with one another. As the universe expanded and cooled, protons and neutrons formed atomic nuclei, which then combined with electrons to form neutral atoms. At this point, the photons effectively stopped interacting with them. These photons, which have stretched as the universe expanded, form the CMB. First observed by Penzias and Wilson in 1965, the CMB remains the focus of increasingly precise observations intended to provide insight into the composition and evolution of the universe.

cosmological constant: The cosmological constant is a constant term that Einstein originally included in his formulation of general relativity. It has the physical effect of pushing the universe apart. Einstein's intent was to make his equations describe a static universe. After astronomical evidence clearly indicated that the size of the universe is changing, Einstein abandoned the cosmological constant though other astrophysicists, such as Georges Lemaître and Sir Arthur Stanley Eddington, thought it might be the source of cosmic expansion. The cosmological constant is a simple explanation of dark energy consistent with the observations; however, it is not the only possible explanation, and the value of the cosmological constant consistent with observation is over 60 orders of magnitude different from what theory predicts.

critical density: In cosmology, the critical density is the density of matter and energy that corresponds to a flat geometry in general relativity. The critical density is given by $8\pi H_0/3G$, where G is the gravitational constant. For a Hubble constant of 70 km/sec/Mpc, $\rho_{crit} = 9 \times 10^{-27}$ kg/m^3.

dark energy: Dark energy is the general term for the substance that causes the universe to expand at an accelerated rate. Although dark energy is believed to be 74 percent of the total energy in the universe, we know very few of its properties. One active area of research is to determine whether dark energy behaves like the cosmological constant or changes over time.

Hubble's Law: Hubble's Law states that the redshift, or apparent recessional velocity, of a distant galaxy is equal to a constant called "Hubble's constant" times the distance to the galaxy. See: Hubble's constant, Hubble diagram, megaparsec, parsec.

Hubble constant: The Hubble constant, defined as "the ratio of the present rate of expansion to the current size of the universe," is a measure of the expansion of the universe. Measurements of Cepheid variable stars made using the Hubble Space Telescope give the present value of the Hubble constant as 72 ± 3 kilometers per second per megaparsec. See: Hubble diagram, megaparsec, parsec.

Hubble diagram: The Hubble diagram is a graph that compares the brightness (or distance) of objects observed in the universe to their redshift (or apparent recessional velocity). Edwin Hubble, for whom the diagram is named, plotted his observations of galaxies outside the Milky Way in this format. In doing so, Hubble showed that the universe is expanding because the recessional velocities of the galaxies are proportional to their distances. Modern Hubble diagrams are based on observations of Type Ia supernovae, and suggest that the expansion rate of the universe is increasing.

light curve: The light curve of an astronomical object is a graph of the object's brightness as a function of time. The light curve of a Cepheid variable star rises and falls in a characteristic pattern that looks somewhat like the teeth of a saw, while the light curve of a supernova rises and falls sharply over the course of a few weeks, followed by a long, slow decline.

light-year: A light-year is the distance that light, which moves at a constant speed, travels in one year. One light-year is equivalent to 9.46×10^{15} meters, or 5,878 billion miles.

luminosity: The luminosity of an object is defined as the total energy per unit time emitted by the object. Another term for luminosity is power. For example, the Sun's luminosity, or power output, is 3.8×10^{26} watts.

redshift: The term redshift is used for a number of different physical effects that lengthen the wavelength of photons, shifting them toward the red end of the spectrum. The Doppler shift of an object moving away from an observer is a redshift, as are the gravitational redshift (Unit 3), and the cosmological redshift due to the expansion of the universe (Unit 11).

standard candle: In astronomy, a standard candle is a class of objects whose distances can be computed by comparing their observed brightness with their known luminosity. Cepheid variable stars are useful as standard candles because their pulsation period is related to their luminosity in a known way. To use a Cepheid variable star to make a distance measurement, an astronomer would measure the apparent brightness of the star and its pulsation period, then calculate the luminosity of the star from its period, and finally compute the distance by comparing the apparent brightness to the calculated luminosity.

Sunyaev-Zel'dovich Effect: The Sunyaev-Zel'dovich, or S-Z, effect creates a distinctive pattern of temperature variation in the cosmic microwave background as CMB photons pass through galaxy clusters. Between the galaxies in a cluster, there is a gas containing energetic electrons that scatter about 1 percent of the CMB photons. The scattered photons gain energy from the electrons, leaving a cold spot in the CMB when observed at the original photon energy. The more energetic scattered photons appear as a hot spot when the CMB is observed with a receiver tuned to a slightly higher radio frequency. The S-Z effect allows astronomers to identify galaxy clusters across the entire sky in a manner that is independent of redshift. See: cosmic microwave background, galaxy cluster.

supernova: A supernova is an exploding star that can reach a luminosity of well over 100 million times that of the Sun. A supernova's brightness rises and falls rapidly over the course of about a month, then fades slowly over months and years. There are two broad classes of supernovae: those that get their energy from a sudden burst of fusion energy and those whose energy comes from gravitational collapse. In practice, these are distinguished on the basis of their different light curves and spectral characteristics. The type Ia supernovae used as standard candles in measurements of the expansion rate of the universe are thought to arise from the explosion of white dwarf stars in a binary system. As the white dwarf draws matter from its companion star, its carbon core reaches the temperature and density at which it can ignite

and fuse explosively in a nuclear flame to iron. This violent explosion destroys the star, and creates about half a solar mass of radioactive isotopes that power the bright peak of the light curve.