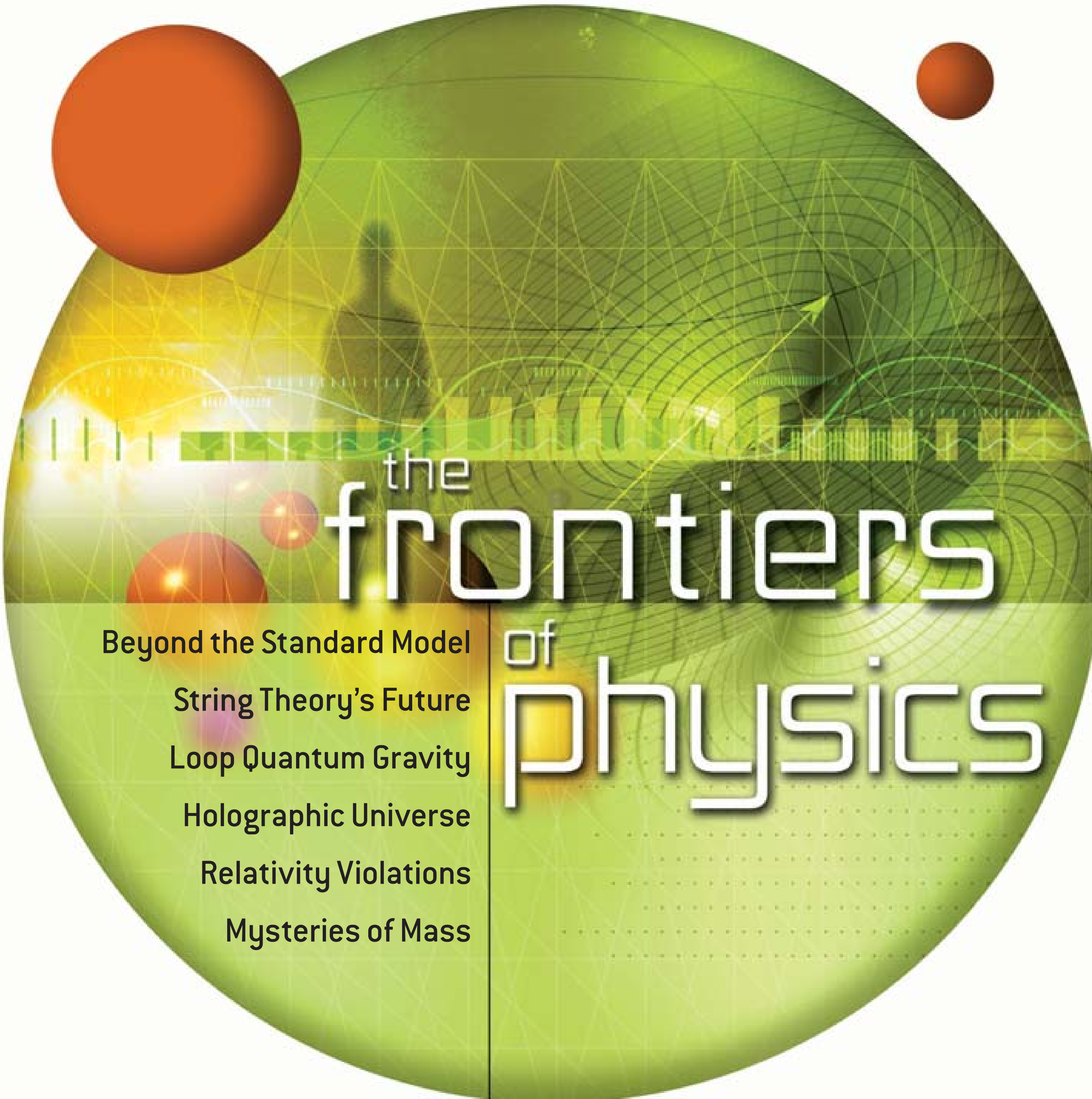
SPECIAL EDITION

SCIENTIFIC AMERICAN

WWW.SCIAM.COM

Display until February 20, 2006



the frontiers of physics

Beyond the Standard Model

String Theory's Future

Loop Quantum Gravity

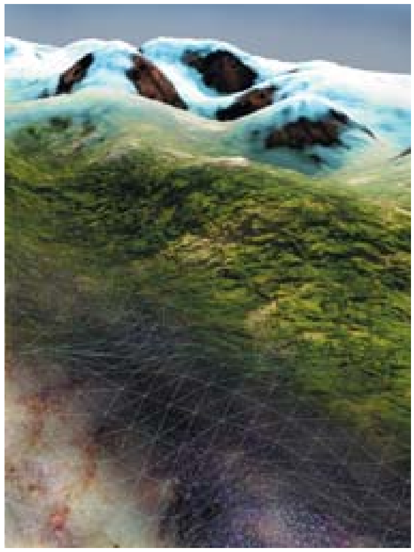
Holographic Universe

Relativity Violations

Mysteries of Mass

letter from the editor

Strange Places



THEORETICAL PLAIN of innumerable universes.

THINGS GET WEIRD—spectacularly so—at the borderlands of physics. The rarefied realms described mathematically and sometimes glimpsed in experiments are all the more extraordinary for *not* being the mere products of someone's hyperactive imagination.

For instance, string theory's equations imply that the universe contains six extra dimensions, which are too tiny to have yet been detected. Some physicists also see innumerable theoretical universes in their equations. In their article "The String Theory Landscape," starting on page 40, Raphael Bousso and Joseph Polchinski provide a view of a theoretical terrain populated with an array of such possible worlds.

We perceive space and time as being continuous, but quantum principles imply that, in fact, at the very smallest scales they actually come in pieces. The effects of this discrete structure could be revealed in experiments in the near future. For a closer look at these ideas and the theory of loop quantum gravity, turn to page 56 for Lee Smolin's "Atoms of Space and Time."

The holographic principle—as Jacob D. Bekenstein explains in "Information in the Holographic Universe," starting on page 74—states that the universe is like a hologram. Just as a trick of light allows a fully three-dimensional image to be recorded on a flat piece of film, our seemingly 3-D universe could be completely equivalent to alternative quantum fields and physical laws "painted" on a distant, vast surface. Our innate perception that the world is three-dimensional could be an extraordinary illusion.

Intellectual enrichment aside, it might be tempting to think that none of what scientists are learning by probing the frontiers of physics truly matters in our everyday lives. Not so. As just one example, consider general relativity, which explains how gravity results from bends in the fabric of spacetime itself. To be accurate, commonplace GPS receivers—which calculate location using a constellation of orbiting satellites—must take the effects of general relativity into account.

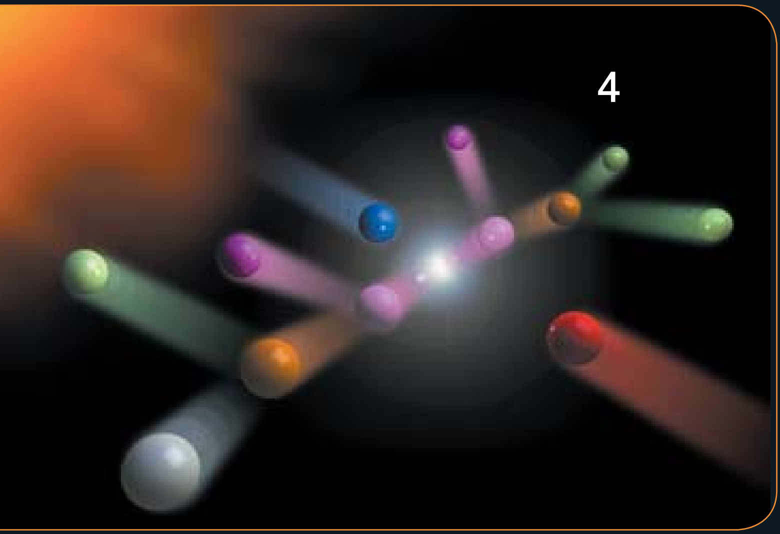
In the pages of this special edition, we invite you to take an armchair journey through our curious universe, with our scientist authors as tour guides. You're in for a mind-boggling treat.

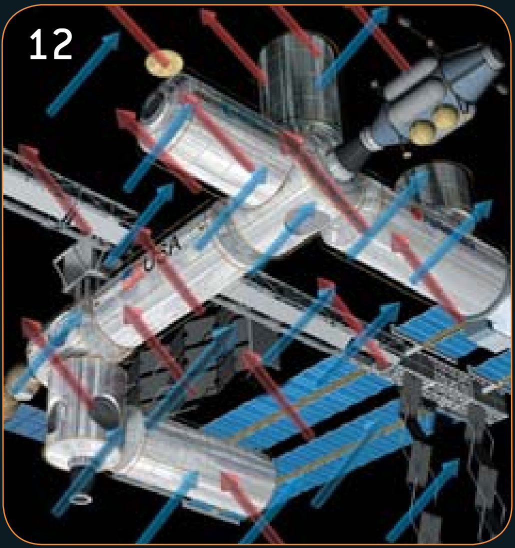
Mariette DiChristina
Executive Editor
Scientific American
editors@sciam.com

Contents 2005

SCIENTIFIC AMERICAN Volume 15, Number 3

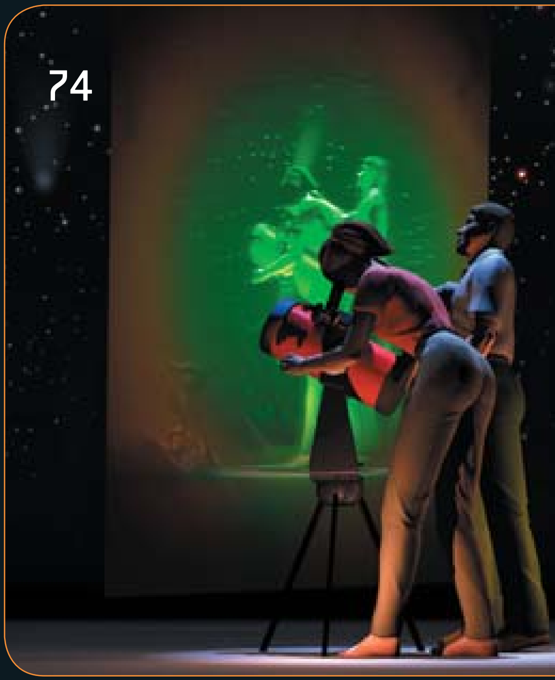
The Frontiers of Physics

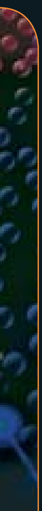




Cover montage by Tom Draper Design

The articles in this special edition have appeared in previous issues of *Scientific American.*

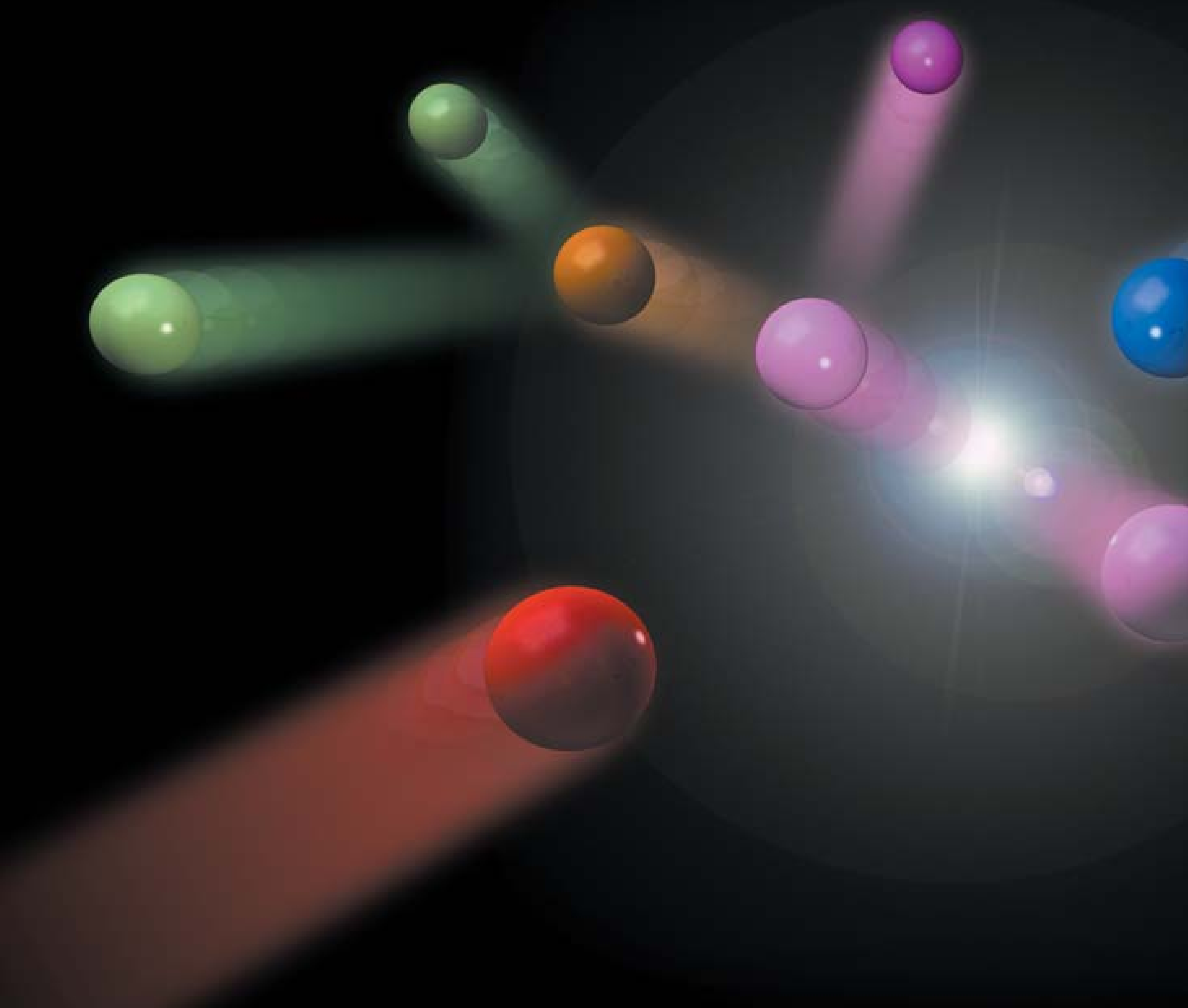




Scientific American Special (ISSN 1048-0943), Volume 15, Number 3, 2005, published by Scientific American, Inc., 415 Madison Avenue, New York, NY 10017-1111. Canadian BN No. 127387652RT; QST No. Q1015332537. To purchase additional quantities: U.S., $10.95 each; elsewhere, $13.95 each. Send payment to Scientific American, Dept. PH, 415 Madison Avenue, New York, NY 10017-1111. Inquiries: fax 212-355-0408 or telephone 212-451-8890. Printed in U.S.A.

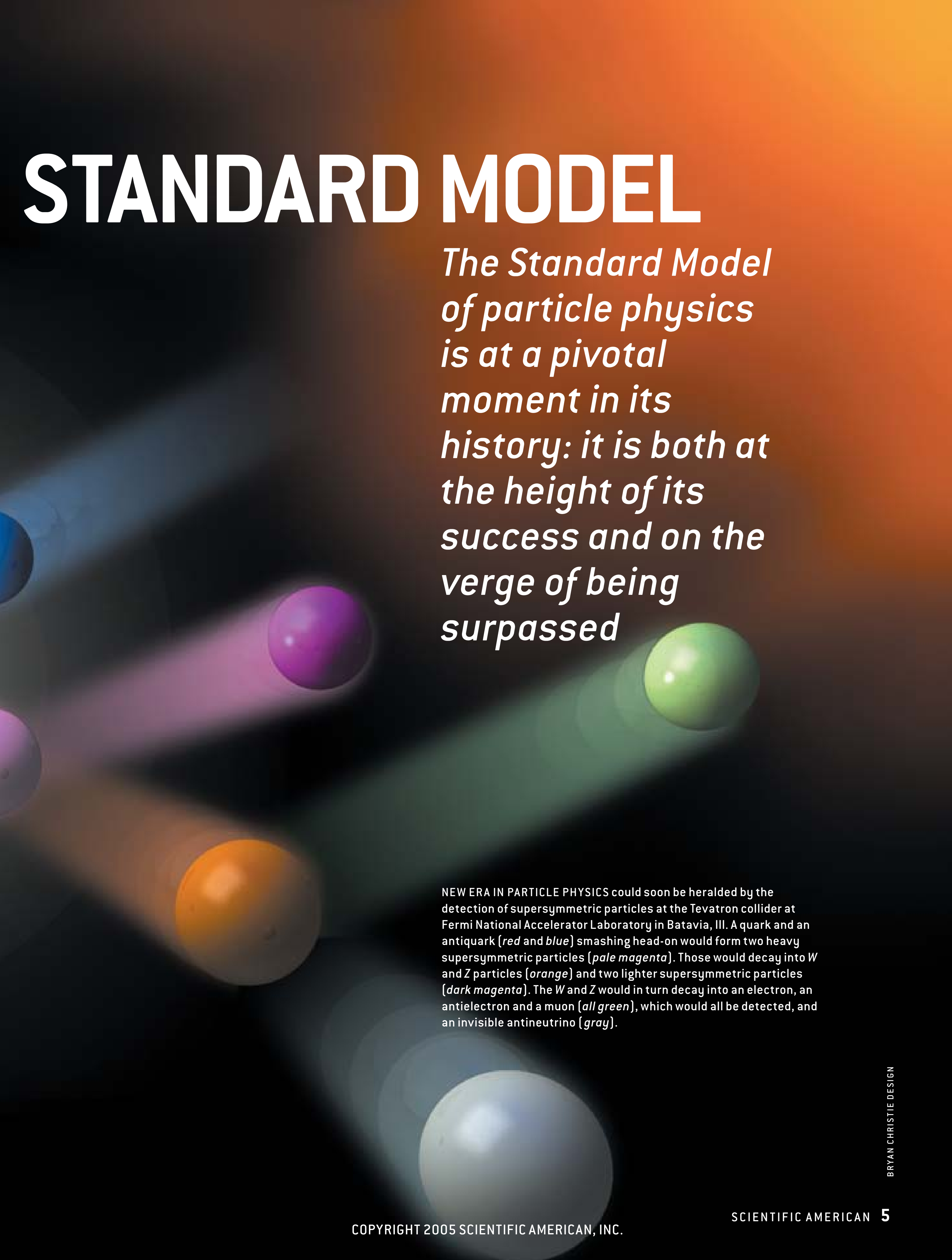
The Dawn of PHYSICS BEYOND THE

By Gordon Kane



STANDARD MODEL

The Standard Model of particle physics is at a pivotal moment in its history: it is both at the height of its success and on the verge of being surpassed



NEW ERA IN PARTICLE PHYSICS could soon be heralded by the detection of supersymmetric particles at the Tevatron collider at Fermi National Accelerator Laboratory in Batavia, Ill. A quark and an antiquark (*red* and *blue*) smashing head-on would form two heavy supersymmetric particles (*pale magenta*). Those would decay into *W* and *Z* particles (*orange*) and two lighter supersymmetric particles (*dark magenta*). The *W* and *Z* would in turn decay into an electron, an antielectron and a muon (*all green*), which would all be detected, and an invisible antineutrino (*gray*).

BRYAN CHRISTIE DESIGN

Centuries after the search began for the fundamental constituents that make up all the complexity and beauty of the everyday world, we have an astonishingly simple answer—it takes just six particles: the electron, the up and the down quarks, the gluon, the photon and the Higgs boson. Eleven additional particles suffice to describe all the esoteric phenomena studied by particle physicists [*see box at right*]. This is not speculation akin to the ancient Greeks' four elements of earth, air, water and fire. Rather it is a conclusion embodied in the most sophisticated mathematical theory of nature in history, the Standard Model of particle physics. Despite the word "model" in its name, the Standard Model is a comprehensive theory that identifies the basic particles and specifies how they interact. Everything that happens in our world (except for the effects of gravity) results from Standard Model particles interacting according to its rules and equations.

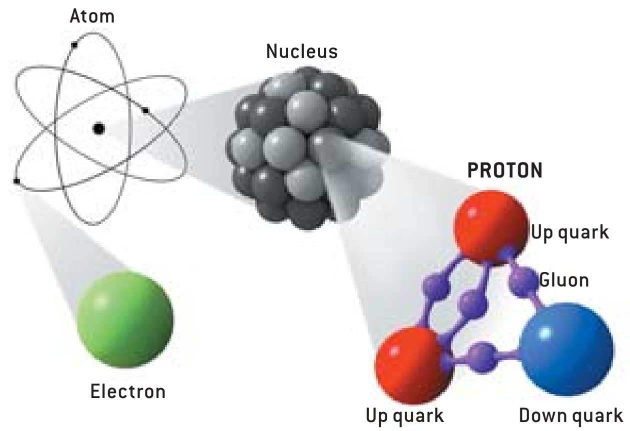
The Standard Model was formulated in the 1970s and tentatively established by experiments in the early 1980s. Nearly three decades of exacting experiments have tested and verified the theory in meticulous detail, confirming all of its predictions. In one respect, this success is rewarding because it confirms that we really understand, at a deeper level than ever before, how nature works. Paradoxically, the success has also been frustrating. Before the advent of the Standard Model, physicists had become used to experiments producing unexpected new particles or other signposts to a new theory almost before the chalk dust had settled on the old one. They have been waiting 30 years for that to happen with the Standard Model.

Their wait should soon be over. Experiments that achieve collisions that are higher in energy than ever before or that study certain key phenomena with greater precision are on the verge

Overview/*A New Era*

- The Standard Model of particle physics is the most successful theory of nature in history, but increasingly there are signs that it must be extended by adding new particles that play roles in high-energy reactions.
- Major experiments are on the verge of providing direct evidence of these new particles. After 30 years of consolidation, particle physics is entering a new era of discovery. Many profound mysteries could be resolved by post–Standard Model physics.
- One element of the Standard Model—a particle called the Higgs boson—also remains to be observed. The Tevatron collider at Fermilab could detect Higgs bosons within the next few years.

THE STANDARD MODEL



The Particles

ALTHOUGH THE STANDARD MODEL needs to be extended, its particles suffice to describe the everyday world (except for gravity) and almost all data collected by particle physicists.

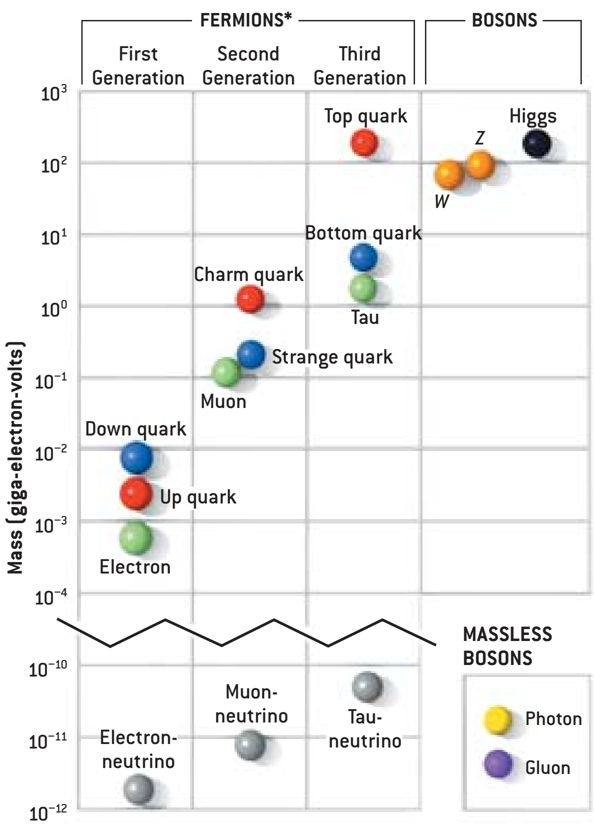
MATTER PARTICLES (FERMIONS) In the Standard Model, the fundamental particles of ordinary matter are the electron, the up quark (u) and the down quark (d). Triplets of quarks bind together to form protons (uud) and neutrons (udd), which in turn make up atomic nuclei (*above*). The electron and the up and the down quarks, together with the electron-neutrino, form the first of three groups of particles called generations. Each generation is identical in every respect except for the remarkably different masses of the fundamental particles (*grid at right*). The values of the masses of the individual neutrinos and the Higgs boson in the grid are speculative but chosen to be consistent with observations and theoretical calculations.

FORCE CARRIERS (BOSONS) The Standard Model describes three of the four known forces: electromagnetism, the weak force (which is involved in the formation of the chemical elements) and the strong force (which holds protons, neutrons and nuclei together). The forces are mediated by force particles: photons for electromagnetism, the *W* and *Z* bosons for the weak force, and gluons for the strong force. For gravity, gravitons are postulated, but the Standard Model does not include gravity. The Standard Model partially unifies the electromagnetic and weak forces—they are facets of one "electroweak" force at high energies or, equivalently, at distances smaller than the diameter of protons.

One of the greatest successes of the Standard Model is that the forms of the forces—the detailed structure of the equations describing them—are largely determined by general principles embodied in the theory rather than being chosen in an ad hoc fashion to match a collection of empirical data. For electromagnetism, for example, the validity of relativistic quantum field theory (on which the Standard Model is based) and the existence of the electron imply that the photon must also exist and interact in the way that it does—we finally understand light. Similar arguments predicted the existence and properties, later confirmed, of gluons and the *W* and *Z* particles.

THE SOURCE OF MASS In addition to the particles described above, the Standard Model predicts the existence of the Higgs boson, which has not yet been directly detected by experiment. The Higgs interacts with the other Standard Model particles in a special manner that gives them mass.

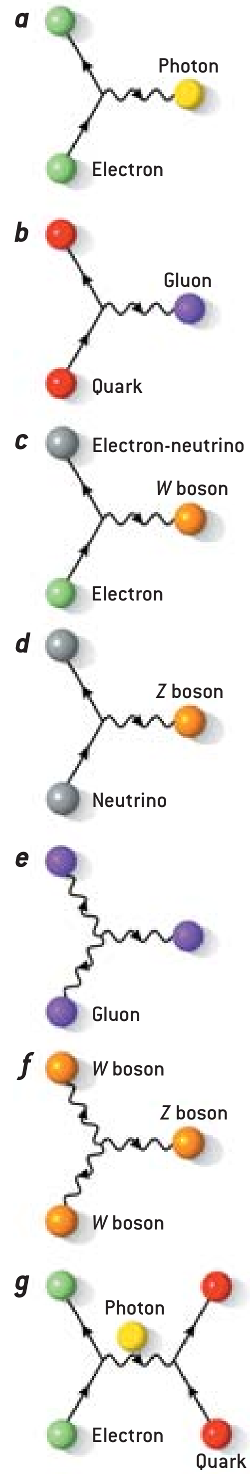
DEEPER LEVELS? Might the Standard Model be superseded by a theory in which quarks and electrons are made up of more fundamental particles? Almost certainly not. Experiments have probed much more deeply than ever before without finding a hint of additional structure. More important, the Standard Model is a consistent theory that makes sense if electrons and quarks are fundamental. There are no loose ends hinting at a deeper underlying structure. Further, all the forces become similar at high energies, particularly if supersymmetry is true [*see box on next page*]. If electrons and quarks are composite, this unification fails: the forces do not become equal. Relativistic quantum field theory views electrons and quarks as being pointlike—they are structureless. In the future, they might be thought of as tiny strings or membranes (as in string theory), but they will still be electrons and quarks, with all the known Standard Model properties of these objects at low energies.



The Rules of the Game

THE STANDARD MODEL describes the fundamental particles and how they interact. For a full understanding of nature, we also need to know what rules to use to calculate the results of the interactions. An example that helps to elucidate this point is Newton's law, $F = ma$. F is any force, m is the mass of any particle, and a is the acceleration of the particle induced by the force. Even if you know the particles and the forces acting on them, you cannot calculate how the particles behave unless you also know the rule $F = ma$. The modern version of the rules is relativistic quantum field theory, which was invented in the first half of the 20th century. In the second half of the 20th century the development of the Standard Model taught researchers about the nature of the particles and forces that were playing by the rules of quantum field theory. The classical concept of a force is also extended by the Standard Model: in addition to pushing and pulling on one another, when particles interact they can change their identity and be created or destroyed.

Feynman diagrams (*a–g, at right*), first devised by physicist Richard P. Feynman, serve as useful shorthand to describe interactions in quantum field theory. The straight lines represent the trajectories of matter particles; the wavy lines represent those of force particles. Electromagnetism is produced by the emission or absorption of photons by any charged particle, such as an electron or a quark. In *a*, the incoming electron emits a photon and travels off in a new direction. The strong force involves gluons emitted (*b*) or absorbed by quarks. The weak force involves *W* and *Z* particles (*c, d*), which are emitted or absorbed by both quarks and leptons (electrons, muons, taus and neutrinos). Notice how the *W* causes the electron to change identity. Gluons (*e*) and *W*s and *Z*s (*f*) also self-interact, but photons do not.

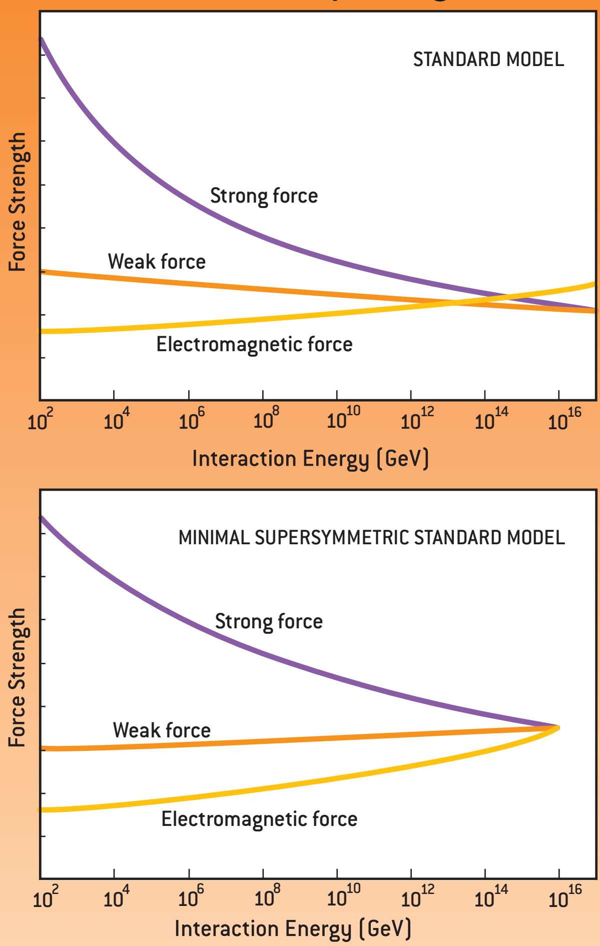


Diagrams *a* through *f* are called interaction vertices. Forces are produced by combining two or more vertices. For example, the electromagnetic force between an electron and a quark is largely generated by the transfer of a photon (*g*). *Everything* that happens in our world, except for gravity, is the result of combinations of these vertices. —*G.K.*

BRYAN CHRISTIE DESIGN

**The fermions are subdivided into quarks and leptons, with leptons including electrons, muons, taus and three forms of neutrino.*

Evidence for Supersymmetry



THE MOST WIDELY FAVORED THEORY to supersede the Standard Model is the Minimal Supersymmetric Standard Model. In this model, every known particle species has a superpartner particle that is related to it by supersymmetry. Particles come in two broad classes: bosons (such as the force particles), which can gather en masse in a single state, and fermions (such as quarks and leptons), which avoid having identical states. The superpartner of a fermion is always a boson and vice versa.

Indirect evidence for supersymmetry comes from the extrapolation of interactions to high energies. In the Standard Model, the three forces become similar but not equal in strength (*top*). The existence of superpartners changes the extrapolation so that the forces all coincide at one energy (*bottom*)—a clue that they become unified if supersymmetry is true.

THE AUTHOR

GORDON KANE, a particle theorist, is Victor Weisskopf Collegiate Professor of Physics at the University of Michigan at Ann Arbor. His work explores ways to test and extend the Standard Model of particle physics. In particular, he studies Higgs physics and the Standard Model's supersymmetric extension, with a focus on relating theory and experiment and on the implications of supersymmetry for particle physics and cosmology. His hobbies include playing squash, exploring the history of ideas, and seeking to understand why science flourishes in some cultures but not in others.

of going beyond the Standard Model. These results will not overturn the Standard Model. Instead they will extend it by uncovering particles and forces not described by it. The most important experiment is occurring at the upgraded Tevatron collider at Fermi National Accelerator Laboratory in Batavia, Ill., which began taking data in 2001. It might produce directly the elusive particles that complete the Standard Model (Higgs bosons) and those predicted by the most compelling extensions of the theory (the "superpartners" of the known particles).

Significant information is also beginning to come from "B factories," particle colliders running in California and Japan configured to create billions of b quarks (one of the 11 additional particles) and their antimatter equivalents to study a phenomenon called CP violation. CP (charge-parity) is the symmetry relating matter to antimatter, and CP violation means that antimatter does not exactly mirror matter in its behavior. The amount of CP violation observed so far in particle decays can be accommodated by the Standard Model, but we have reasons to expect much more CP violation than it can produce. Physics that goes beyond the Standard Model can generate additional CP violation.

Physicists are also studying the precise electric and magnetic properties of particles. The Standard Model predicts that electrons and quarks behave as microscopic magnets with a specific strength and that their behavior in an electric field is determined purely by their electric charge. Most extensions of the Standard Model predict a slightly different magnetic strength and electrical behavior. Experiments are beginning to collect data with enough sensitivity to see the tiny effects predicted.

Looking beyond the earth, scientists studying solar neutrinos and cosmic-ray neutrinos, ghostly particles that barely interact, recently established that neutrinos have masses, a result long expected by theorists studying extensions of the Standard Model [see "Solving the Solar Neutrino Problem," by Arthur B. McDonald, Joshua R. Klein and David L. Wark, on page 22]. The next round of experiments will clarify the form of theory needed to explain the observed neutrino masses.

In addition, experiments are under way to detect mysterious particles that form the cold dark matter of the universe and to examine protons at higher levels of sensitivity to learn whether they decay. Success in either project would be a landmark of post–Standard Model physics.

This research is ushering in a data-rich era in particle physics. Joining the fray by about 2007 will be the Large Hadron Collider (LHC), a machine 27 kilometers in circumference now under construction at CERN, the European laboratory for particle physics near Geneva [see "The Large Hadron Collider," by Chris Llewellyn Smith; SCIENTIFIC AMERICAN, July 2000]. A 30-kilometer-long linear electron-positron collider that will complement the LHC's results is in the design stages.

As the first hints of post–Standard Model physics are glimpsed, news reports often make it sound as if the Standard Model has been found to be wrong, as if it were broken and ready to be discarded, but that is not the right way to think

NINA FINKEL

about it. Take the example of Maxwell's equations, written down in the late 19th century to describe the electromagnetic force. In the early 20th century we learned that at atomic sizes a quantum version of Maxwell's equations is needed. Later the Standard Model included these quantum Maxwell's equations as a subset of its equations. In neither case do we say Maxwell's equations are wrong. They are extended. (And they are still used to design innumerable electronic technologies.)

A Permanent Edifice

SIMILARLY, THE STANDARD MODEL is here to stay. It is a full mathematical theory—a multiply connected and highly stable edifice. It will turn out to be one piece of a larger such edifice, but it cannot be "wrong." No part of the theory can fail without a collapse of the entire structure. If the theory were wrong, many successful tests would be accidents. It will continue to describe strong, weak and electromagnetic interactions at energies in its domain.

The Standard Model is very well tested. It predicted the existence of the *W* and *Z* bosons, the gluon and two of the heavier quarks (the charm and the top quark). All these particles were subsequently found, with precisely the predicted properties.

A second major test involves the electroweak mixing angle, a parameter that plays a role in describing the weak and electromagnetic interactions. That mixing angle must have the same value for every electroweak process. If the Standard Model were wrong, the mixing angle could have one value for one process, a different value for another and so on. It is observed to have the same value everywhere, to an accuracy of about 1 percent.

Third, the Large Electron-Positron (LEP) collider at CERN looked at about 20 million *Z* bosons. Essentially every one of them decayed in the manner expected by the Standard Model, which predicted the number of instances of each kind of decay as well as details of the energies and directions of the outgoing particles. These tests are but a few of the many that have solidly confirmed the Standard Model.

In its full glory, the Standard Model has 17 particles and about as many free parameters—quantities such as particle masses and strengths of interactions [*see box on pages 6 and 7*]. These quantities can in principle take any value, and we learn the correct values only by making measurements. Critics sometimes compare the Standard Model's many parameters with the epicycles on epicycles that medieval theorists used to describe planetary orbits. They imagine that the Standard Model has limited predictive power, or that its content is arbitrary, or that it can explain anything by adjusting some parameter.

The opposite is actually true: once the masses and interaction strengths are measured in any process, they are fixed for the whole theory and for any other experiment, leaving no freedom at all. Moreover, the detailed forms of all the Standard Model's equations are determined by the theory. Every parameter but the Higgs boson mass has been measured. Until we go beyond the Standard Model, the only thing that can change with new results is the precision of our knowledge of the parameters, and as that improves it becomes harder, not easier, for all the experimental data to remain consistent, because measured quantities must agree to higher levels of precision.

Adding further particles and interactions to extend the Standard Model might seem to introduce a lot more freedom, but this is not necessarily the case. The most widely favored extension is the Minimal Supersymmetric Standard Model (MSSM). Supersymmetry assigns a superpartner particle to every particle species. We know little about the masses of those superpartners, but their interactions are constrained by the supersymmetry. Once the masses are measured, the predictions of the MSSM will be even more tightly constrained than the Standard Model because of the mathematical relations of supersymmetry.

Ten Mysteries

IF THE STANDARD MODEL works so well, why must it be extended? A big hint arises when we pursue the long-standing goal of unifying the forces of nature. In the Standard Model, we can extrapolate the forces and ask how they would behave at much higher energies. For example, what were the forces like in the extremely high temperatures extant soon after the big bang? At low energies the strong force is about 30 times as powerful as the weak force and more than 100 times as powerful as electromagnetism. When we extrapolate, we find that the strengths of these three forces become very similar but are never all exactly the same. If we extend the Standard Model to the MSSM, the forces become essentially identical at a specific high energy [*see box on opposite page*]. Even better, the gravitational force approaches the same strength at a slightly higher energy, suggesting a link between the Standard Model forces and gravity. These results seem like strong clues in favor of the MSSM.

Other reasons for extending the Standard Model arise from phenomena it cannot explain or cannot even accommodate:

1. All our theories today seem to imply that the universe should contain a tremendous concentration of energy, even in the emptiest regions of space. The gravitational effects of this so-called vacuum energy would have either quickly curled up the universe long ago or expanded it to much greater size. The Standard Model cannot help us understand this puzzle, called the cosmological constant problem.

2. The expansion of the universe was long believed to be slowing because of the mutual gravitational attraction of all the matter in the universe. We now know that the expansion is accelerating and that whatever causes the acceleration (dubbed "dark energy") cannot be Standard Model physics.

3. There is very good evidence that in the first fraction of a second of the big bang the universe went through a stage of extremely rapid expansion called inflation. The fields responsible for inflation cannot be Standard Model ones.

4. If the universe began in the big bang as a huge burst of energy, it should have evolved into equal parts matter and antimatter (CP symmetry). But instead the stars and nebulae are made of protons, neutrons and electrons and not their antiparticles (their antimatter equivalents). This matter asymmetry cannot be explained by the Standard Model.

5. About a quarter of the universe is invisible cold dark matter that cannot be particles of the Standard Model.

6. In the Standard Model, interactions with the Higgs field (which is associated with the Higgs boson) cause particles to have mass. The Standard Model cannot explain the very special forms that the Higgs interactions must take.

7. Quantum corrections apparently make the calculated Higgs boson mass huge, which in turn would make all particle masses huge. That result cannot be avoided in the Standard Model and thus causes a serious conceptual problem.

8. The Standard Model cannot include gravity, because it does not have the same structure as the other three forces.

9. The values of the masses of the quarks and leptons (such as the electron and neutrinos) cannot be explained by the Standard Model.

10. The Standard Model has three "generations" of particles. The everyday world is made up entirely of first-generation particles, and that generation appears to form a consistent theory on its own. The Standard Model describes all three generations, but it cannot explain why more than one exists.

In expressing these mysteries, when I say the Standard Model *cannot* explain a given phenomenon, I do not mean that the theory has not yet explained it but might do so one day. The Standard Model is a highly constrained theory, and it cannot *ever* explain the phenomena listed above. Possible explanations do exist. One reason the supersymmetric extension is attractive to many physicists is that it can address all but the second and the last three of these mysteries. String theory (in which particles are represented by tiny, one-dimensional entities instead of point objects) addresses the last three [see "The Theory Formerly Known as Strings," by Michael J. Duff; SCIENTIFIC AMERICAN, February 1998]. The phenomena that the Standard Model cannot explain are clues to how it will be extended.

It is no surprise that there are questions the Standard Model cannot answer—every successful theory in science has increased the number of answered questions but has left some unanswered. And even though improved understanding has led to new questions that could not be formulated earlier, the number of unanswered fundamental questions has continued to decrease.

Some of these 10 mysteries demonstrate another reason why particle physics today is entering a new era. It has become clear that many of the deepest problems in cosmology have their solutions in particle physics, so the fields have merged into "particle cosmology." Only from cosmological studies could we learn that the universe is matter (and not antimatter) or that the universe is about a quarter cold dark matter. Any theoretical understanding of these phenomena must explain how they arise as part of the evolution of the universe after the big bang. But cosmology alone cannot tell us what particles make up cold dark matter, or how the matter asymmetry is actually generated, or how inflation originates. Understanding of the largest and the smallest phenomena must come together.

The Higgs

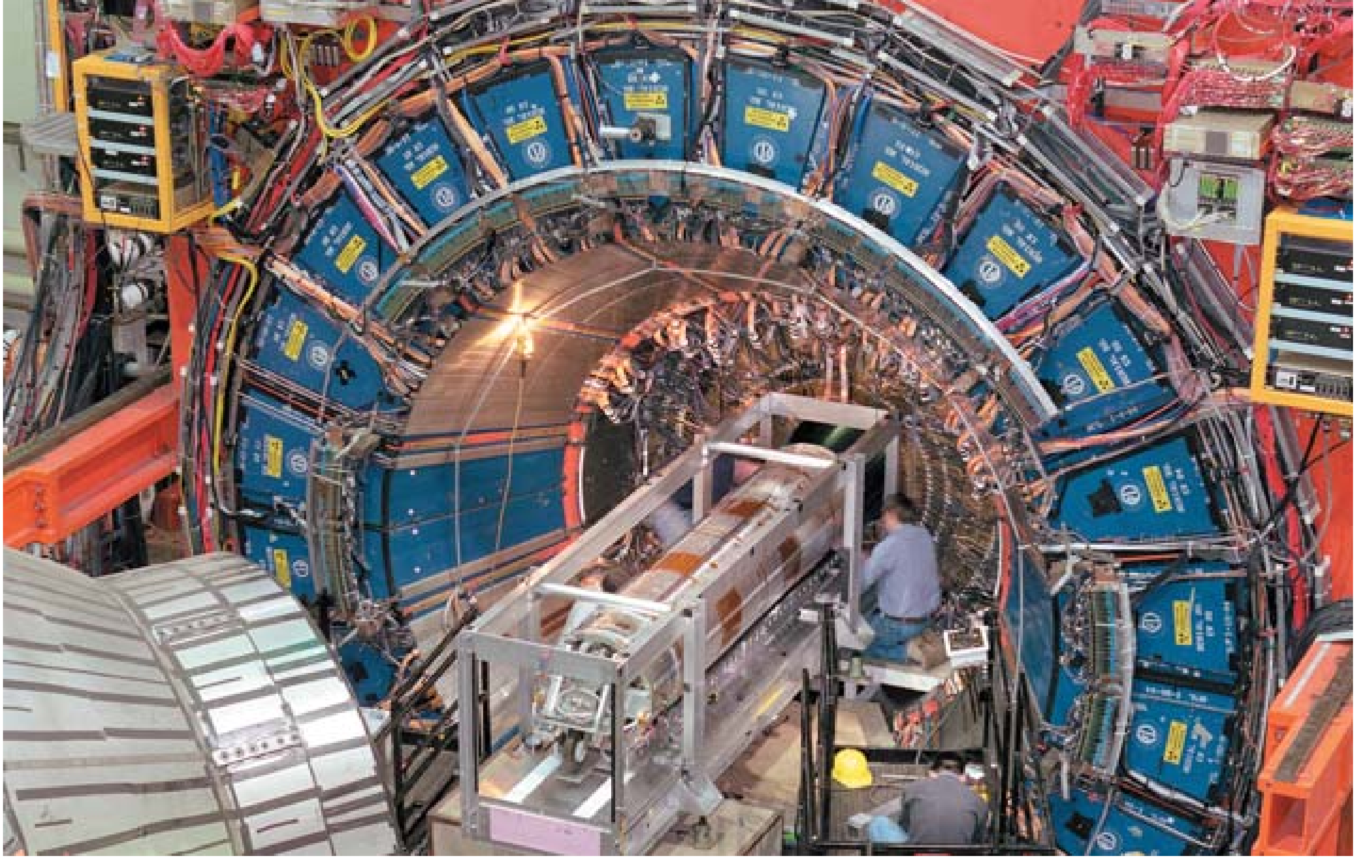
PHYSICISTS ARE TACKLING all these post–Standard Model mysteries, but one essential aspect of the Standard Model also remains to be completed. To give mass to leptons, quarks, and *W* and *Z* bosons, the theory relies on the Higgs field, which has not yet been directly detected.

The Higgs is fundamentally unlike any other field. To understand how it is different, consider the electromagnetic field. Electric charges give rise to electromagnetic fields such as those all around us (just turn on a radio to sense them). Electromagnetic fields carry energy. A region of space has its lowest possible energy when the electromagnetic field vanishes throughout it. Zero field is the natural state in the absence of charged particles. Surprisingly, the Standard Model requires that the lowest energy occur when the Higgs field has a specific nonzero value. Consequently, a nonzero Higgs field permeates the universe, and particles always interact with this field, traveling through it like people wading through water. The interaction gives them their mass, their inertia.

Associated with the Higgs field is the Higgs boson. In the Standard Model, we cannot predict any particle masses from first principles, including the mass of the Higgs boson itself. One can, however, use other measured quantities to calculate some masses, such as those of the *W* and *Z* bosons and the top quark. Those predictions are confirmed, increasing confidence in the underlying Higgs physics.

Physicists do already know something about the Higgs mass. Experimenters at the LEP collider measured about 20 quantities that are related to one another by the Standard Model. All the parameters needed to calculate predictions for those quantities are already measured—except for the Higgs boson mass. So one can work backward from the data and ask which Higgs mass gives the best fit to the 20 quantities. The answer is that the Higgs mass is less than about 200 giga-electron-volts (GeV). (The proton mass is about 0.9 GeV; the top quark 174 GeV.) That there is an answer at all is strong evidence that the Higgs exists. If the Higgs did not exist and the Standard Model were wrong, it would take a remarkable coincidence for the 20 quantities to be related in the right way to be consistent with a specific Higgs mass. Our confidence in this procedure is bolstered because a similar approach accurately predicted the top quark mass before any top quarks had been detected directly.

LEP also conducted a direct search for Higgs particles, but it could search only up to a mass of about 115 GeV. At that very upper limit of LEP's reach, a small number of events in-



FERMILAB PHOTO

UPGRADING of the Tevatron's huge particle detectors, which was carried out by physicists at Fermilab from 1996 to 2000, has primed the facility for observing Higgs bosons and supersymmetry.

volved particles that behaved as Higgs bosons should. But there were not enough data to be sure a Higgs boson was actually discovered. Together the results suggest the Higgs mass lies between 115 and 200 GeV.

LEP is now dismantled to make way for the construction of the LHC, which is scheduled to begin taking data in 2007. In the meantime, the search for the Higgs continues at the Tevatron at Fermilab [*see illustration above*]. If the Tevatron operates at its design intensity and energy and does not lose running time because of technical or funding difficulties, it could confirm the 115-GeV Higgs boson in about two to three years. If the Higgs is heavier, it will take longer for a clear signal to emerge from the background. The Tevatron will produce more than 10,000 Higgs bosons altogether if it runs as planned, and it could test whether the Higgs boson behaves as predicted. The LHC will be a "factory" for Higgs bosons, producing millions of them and allowing extensive studies.

There are also good arguments that some of the lighter superpartner particles predicted by the MSSM have masses small enough so that they could be produced at the Tevatron as well. Direct confirmation of supersymmetry could come in the next few years. The lightest superpartner is a prime candidate to make up the cold dark matter of the universe—it could be directly observed for the first time by the Tevatron. The LHC will produce large numbers of superpartners if they exist, definitively testing whether supersymmetry is part of nature.

Effective Theories

TO FULLY GRASP the relation of the Standard Model to the rest of physics, and its strengths and limitations, it is useful to think in terms of effective theories. An effective theory is a description of an aspect of nature that has inputs that are, in principle at least, calculable using a deeper theory. For example, in nuclear physics one takes the mass, charge and spin of the proton as inputs. In the Standard Model, one can calculate those quantities, using properties of quarks and gluons as inputs. Nuclear physics is an effective theory of nuclei, whereas the Standard Model is the effective theory of quarks and gluons.

From this point of view, every effective theory is open-ended and equally fundamental—that is, not truly fundamental at all. Will the ladder of effective theories continue? The MSSM solves a number of problems the Standard Model does not solve, but it is also an effective theory because it has inputs as well. Its inputs might be calculable in string theory.

Even from the perspective of effective theories, particle physics may have special status. Particle physics might increase our understanding of nature to the point where the theory can be formulated with no inputs. String theory or one of its cousins might allow the calculation of all inputs—not only the electron mass and such quantities but also the existence of spacetime and the rules of quantum theory. But we are still an effective theory or two away from achieving that goal. SA

MORE TO EXPLORE

The Particle Garden. Gordon Kane. Perseus Publishing, 1996.

The Rise of the Standard Model: A History of Particle Physics from 1964 to 1979. Edited by Lillian Hoddeson, Laurie M. Brown, Michael Riordan and Max Dresden. Cambridge University Press, 1997.

The Little Book of the Big Bang: A Cosmic Primer. Craig J. Hogan. Copernicus Books, 1998.

Supersymmetry: Unveiling the Ultimate Laws of Nature. Gordon Kane. Perseus Publishing, 2001.

An excellent collection of particle physics Web sites is listed at **particleadventure.org/particleadventure/other/othersites.html**

VIOLATIONS OF RELATIVITY could be manifest in the ticking rates of mirror-image, antimatter clocks and the stretching of matter along specific directions.

THE SEARCH FOR RELATIVITY VIOLATIONS

To uncover evidence for an ultimate theory, scientists are looking for infractions of Einstein's once sacrosanct physical principle

By Alan Kostelecký

Relativity lies at the heart of the most fundamental theories of physics. Formulated by Albert Einstein in 1905, relativity is built on the key idea that physical laws take the same form for any inertial observer—that is, for an observer oriented in any direction and moving at any constant speed. The theory predicts an assortment of well-known effects: among them, constancy of the speed of light for all observers, slowing of moving clocks, length contraction of moving objects, and equivalence of mass and energy ($E = mc^2$). These effects have been confirmed in highly sensitive experiments, and relativity is now a basic, everyday tool of experimental physics: particle colliders take advantage of the increase in mass and lifetime of fast particles; experiments with radioactive isotopes depend on the conversion of mass into energy. Even consumer electronics is affected—the Global Positioning System must allow for time dilation, which alters the rates of clocks on its orbiting satellites.

In recent years, however, motivated by attempts to combine all the known forces and particles into one ultimate unified theory, some physicists have been investigating the possibility that relativity's postulates provide only an approximation of nature's workings. The hope is that small relativity violations might offer the first experimental signals of the long-sought ultimate theory.

The unchanging quality, or invariance, of physical laws for different observers represents a symmetry of space and time (spacetime), called Lorentz symmetry after Dutch theoretical physicist Hendrik Antoon Lorentz, who studied it beginning in the 1890s. A perfect sphere illustrates an ordinary symmetry, what is known as symmetry under rotations: no matter how you turn it, the sphere looks the same. Lorentz symmetry is not based on objects looking the same but expresses instead the sameness of the laws of physics under rota-

OVERVIEW

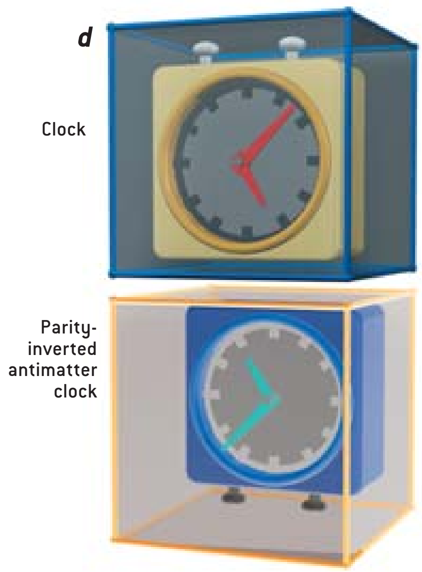
- Although special relativity is among the most fundamental and well verified of all physical theories, tiny violations of it could be predicted by theories that unify quantum mechanics, gravity and the other forces of nature.
- Numerous experiments are under way to uncover such effects, but so far none has proved sensitive enough to succeed.

DAVID EMMITE PHOTOGRAPHY

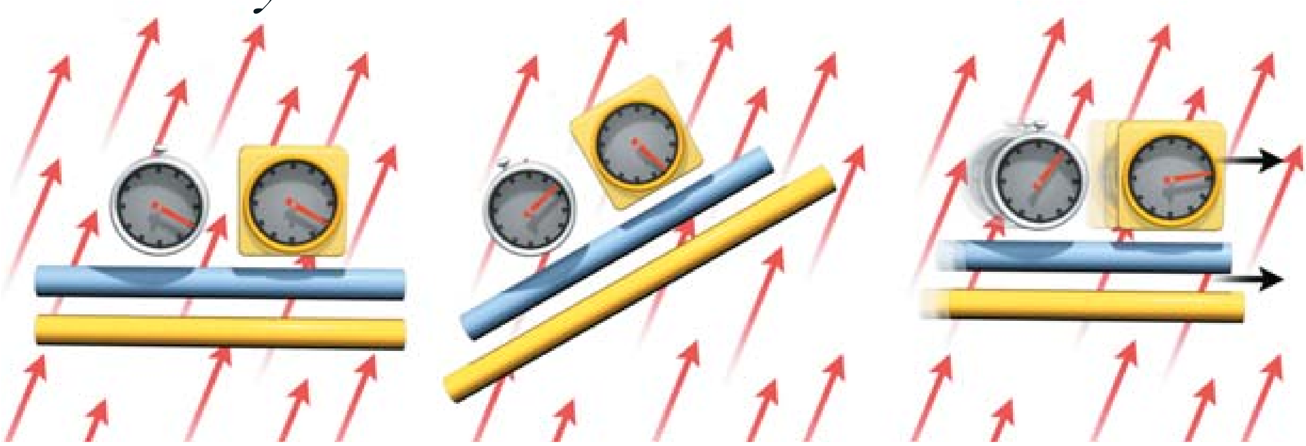
SPACETIME SYMMETRY

Relativity Obeyed

Lorentz symmetry is a fundamental property of the natural world that is of supreme importance for physics. It has two components: rotational symmetry and boost symmetry. Imagine that we have two rods made of dissimilar materials but having identical lengths when placed side by side and two clocks operating by different mechanisms that keep identical time (*a*). Rotational symmetry states that if one rod and one clock are rotated relative to the others, the rods nonetheless retain identical lengths and the clocks remain in sync (*b*). Boost symmetry considers what happens when one rod and one clock are "boosted" so that they move at a constant velocity relative to the other two, which here remain at rest. Boost symmetry predicts that the moving rod will be shorter and that the moving clock will run slower by amounts that depend in a precise way on the relative velocity (*c*). When space and time are combined to form spacetime, boost symmetry actually has almost identical mathematical form to rotational symmetry. A closely related symmetry is CPT symmetry, where the letters stand for charge conjugation, parity inversion and time reversal. This predicts that if a clock is replaced by its antimatter equivalent (charge reversal), which is also inverted (parity) and running backward in time, the two will keep identical time (*d*). A mathematical theorem demonstrates that for a quantum field theory, CPT symmetry must hold whenever Lorentz symmetry is obeyed.



Relativity Violated



Broken Lorentz symmetry can be represented by a field of vectors throughout spacetime. Particles and forces have interactions with this vector field (*arrows*) similar to the interaction of charged particles with an electric field (which is also a vector field). As a result, unlike the Lorentz symmetric case, all directions and all velocities are no longer equivalent. Two dissimilar rods that have equal lengths at one orientation relative to the vector field (*left*) may shrink or expand at another orientation (*center*). Similarly, two dissimilar clocks that are synchronized at the first orientation may run slow or fast at the second orientation. In addition, dissimilar rods and clocks that are boosted (*right*) may undergo different length contractions and time dilations depending on their materials and the direction and magnitude of the boost.

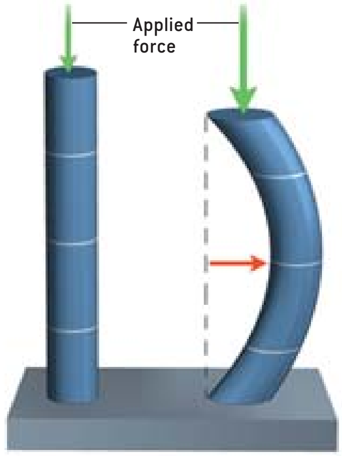
DON FOLEY

tions and also under boosts, which are changes of velocity. An observer sees the same laws of physics at play, no matter what her orientation (rotation) and no matter what her velocity (boost). When Lorentz symmetry holds, spacetime is isotropic in the sense that all directions and all uniform motions are equivalent, so none is singled out as being special.

The Lorentz symmetry of spacetime forms the core of relativity. The details of how boosts work produce all the well-known relativistic effects. Before Einstein's 1905 paper, equations relating to these effects had been developed by several other researchers, including Lorentz, but they typically interpreted the equations as describing physical changes in objects—for example, bond lengths between atoms becoming shorter to generate length contraction. Einstein's great contributions included combining all the pieces and realizing that the lengths and clock rates are intimately linked. The notions of space and time merge into a single concept: spacetime.

Lorentz symmetry is a key element in the very foundations of our best description of the fundamental particles and forces. When combined with the principles of quantum mechanics, Lorentz symmetry produces a framework called relativistic quantum field theory. In this framework, every particle or force is described by a field that permeates spacetime and has the appropriate Lorentz symmetry. Particles such as electrons or photons exist as localized excitations, or quanta, in the relevant field. The Standard Model of particle physics, which describes all known particles and all known nongravitational forces (the electromagnetic, weak and strong forces), is a relativistic quantum field theory. The requirements of Lorentz symmetry strongly constrain how the fields in this theory can behave and interact. Many interactions that one could write down as plausible-looking terms to be added to the theory's equations are excluded because they violate Lorentz symmetry.

The Standard Model does not include the gravitational interaction. Our best description of gravity, Einstein's general relativity, is also founded on Lorentz symmetry. (The term "general" means that gravity is included, whereas "special" relativity excludes it.) In general relativity, the laws of physics at any given location are the same for all observer orientations and velocities, as before, but the effects of gravity make comparisons between experiments at different locations more complicated. General relativity is a classical theory (that is, nonquantum), and no one knows how to combine it with the basic Standard Model in a completely satisfactory way. The two can be partially combined, however, into a theory called "the Standard Model with gravity," which describes all particles and all four forces.



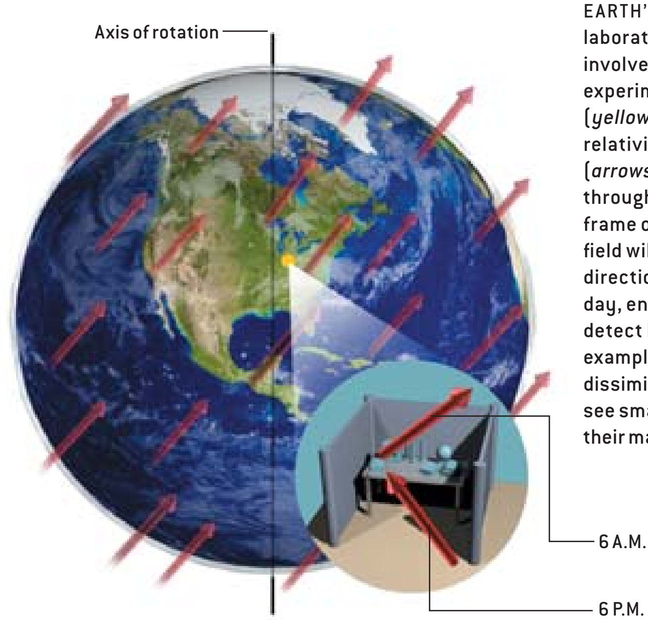
DON FOLEY

SPONTANEOUS SYMMETRY BREAKING occurs when a completely symmetric set of conditions or underlying equations gives rise to an asymmetric result. For example, consider a cylindrical stick with a force applied vertically (*left*). The system is completely symmetrical with respect to rotations around the axis of the stick. If a large enough force is applied, however, the system becomes unstable and the stick will bend in some direction (*right*). The symmetry breaking can be represented by a vector, or an arrow (*red*), that indicates the direction and magnitude of the bending. Lorentz violation involves the emergence of such vector quantities throughout spacetime.

Unification and the Planck Scale

TOGETHER THIS MELDING of the Standard Model and general relativity is astonishingly successful in describing nature. It describes all established fundamental phenomena and experimental results, and no confirmed experimental evidence for physics beyond it exists [see "The Dawn of Physics beyond the Standard Model," by Gordon Kane, on page 4]. Nevertheless, many physicists deem the combination unsatisfactory. One source of difficulty is that although quantum physics and gravity each have an elegant formulation, they seem mathematically incompatible in their present form. In situations where both gravity and quantum physics are important, such as the classic experiment in which cold neutrons rise against the earth's gravitational field, the gravity is incorporated into the quantum description as an externally applied force. That characterization models the experiment extremely well, but it is unsatisfactory as a fundamental and consistent description. It is like describing how a person can lift a heavy object, with the bones' mechanical strength and other properties accurately modeled and explained down to the molecular level, but with the muscles depicted as black-box machines that can supply a specified range of forces.

For these reasons and others, many theoretical physicists believe that it must be possible to formulate an ultimate theory—a complete and unified description of nature that consistently combines quantum physics and gravity. One of the first physicists to work on the idea of a unified theory was Einstein himself, who tackled this problem during the last part of his life. His goal was to find a theory that would describe not only gravity but also electromagnetism. Alas, he had tackled the problem too early. We now believe that electromagnetism is closely related to the strong and weak forces. (The strong force acts between quarks, which make up particles such as protons and neutrons, whereas the weak force is responsible for some kinds of radioactivity and the decay of the neutron.) It was only with experimental facts uncovered after Einstein's death that the strong and weak forces became characterized well enough for them to



EARTH'S ROTATION will turn a laboratory, such as this one involved in a hypothetical experiment at Indiana University (*yellow dot*), relative to any relativity-violating vector field (*arrows*) that is present throughout spacetime. In the lab frame of reference, the vector field will seem to change direction over the course of a day, enabling the experiment to detect Lorentz violations. For example, a comparison of two dissimilar masses in the lab may see small periodic variations in their masses.

DON FOLEY

be understood separately, let alone in combination with electromagnetism and gravity.

One promising and comprehensive approach to this ultimate theory is string theory, which is based on the idea that all particles and forces can be described in terms of one-dimensional objects ("strings"), along with membranes of two dimensions and higher that are called branes [see "The String Theory Landscape," by Raphael Bousso and Joseph Polchinski, on page 40]. Another approach, known as loop quantum gravity, seeks a consistent quantum interpretation of general relativity and predicts that space is a patchwork of discrete pieces (quanta) of volume and area [see "Atoms of Space and Time," by Lee Smolin, on page 56].

Whatever the eventual form of the ultimate theory, quantum physics and gravity are expected to become inextricably intertwined at a fundamental length scale of about 10^{-35} meter, which is called the Planck length, after 19th-century German physicist Max Planck. The Planck length is far too small to be within the direct reach of either conventional microscopes or less conventional ones such as high-energy particle colliders (which probe "merely" down to about 10^{-19} meter). So not only is it very challenging to construct a convincing ultimate theory, but it is also impractical to observe directly the new physics it must surely predict.

Despite these obstacles, a route may exist for obtaining experimental information about the unified theory at the Planck scale. Minuscule indirect effects reflecting the new physics in the unified theory may be detectable in experiments of sufficient sensitivity. An analogy is the image on a television or computer screen, which is composed of many small, bright pixels. The pixels are small compared with the distance at which the screen is viewed, so the image appears smooth to the eye. But in special situations, the pixels become evident—for example, when a newscaster is wearing a tie with narrow stripes that trigger a Moiré pattern on the screen. One class of such "Moiré patterns" from the Planck scale is relativity violations. At macroscopic distances, spacetime appears Lorentz-invariant, but this symmetry may be broken at sufficiently small distances as a consequence of features of the unification of quantum physics and gravity.

The observable effects of Planck-scale relativity violations are likely to lie in the range of 10^{-34} to 10^{-17}. To gain some feeling for these numbers, consider that the thickness of a human hair is about 10^{-30} of the distance across the observable universe. Even 10^{-17} is roughly the ratio of a hair's thickness to the diameter of Neptune's orbit. The detection of relativity violations therefore requires some of the most sensitive experiments ever performed.

Another fundamental spacetime symmetry that could be violated is so-called CPT symmetry. This symmetry holds when the laws of physics are unaffected when three transfomations are all applied at once: interchange of particles and antiparticles (charge conjugation, C), reflection in a mirror (parity inversion, P) and reversal of time (T). The Standard Model obeys CPT symmetry, but theories with relativity violations may break it.

Spontaneous Violations

HOW MIGHT RELATIVITY violations emerge in the ultimate theory? One natural and elegant mechanism is called spontaneous Lorentz violation. It has similarities to the spontaneous breaking of other kinds of symmetry, which occurs whenever the underlying physical laws are symmetrical but the actual system is not. To illustrate the general idea of spontaneous symmetry breaking, consider a slender cylindrical stick, placed vertically with one end on the floor [*see illustration on preceding page*]. Imagine applying a force vertically downward on the stick. This situation is completely symmetrical under rotations around the axis of the stick: the stick is cylindrical, and the force is vertical. So the basic physical equations for this situation are symmetrical under rotation. But if sufficient force is applied, the stick will bend in some particular direction, which spontaneously breaks the rotational symmetry.

THE AUTHOR

ALAN KOSTELECKÝ is professor of theoretical physics at Indiana University. His publications span a broad range of topics in particle physics, gravitation, string theory, mathematical physics and atomic physics. His research on Lorentz and CPT symmetry triggered the recent flood of interest in relativity violations and has led to many new experimental tests.

In the case of relativity violations, the equations describing the stick and the applied force are replaced by the equations of the ultimate theory. In place of the stick are the quantum fields of matter and forces. The natural background strength of such fields is usually zero. In certain situations, however, the background fields acquire a nonzero strength. Imagine that this happened for the electric field. Because the electric field has a direction (technically, it is a vector), every location in space will have a special direction singled out by the direction of the electric field. A charged particle will accelerate in that direction. Rotational symmetry is broken (and so is boost symmetry). The same reasoning applies for any nonzero "tensor" field; a vector is a special case of a tensor.

Such spontaneous nonzero tensor fields do not arise in the Standard Model, but some fundamental theories, including string theory, contain features that are favorable for spontaneous Lorentz breaking. The idea that spontaneous Lorentz breaking and observable relativity violations could occur in string theory and field theories with gravity was originally proposed in 1989 by Stuart Samuel of the City College of New York and me. It was extended in 1991 to include spontaneous CPT violation in string theory by Robertus Potting of Algarve University in Portugal and me. Since then, various additional mechanisms have been proposed for relativity violations arising in string theory and in other approaches to quantum gravity. If spontaneous Lorentz breaking or any other mechanisms do turn out to be part of the ultimate fundamental theory, the concomitant relativity violations could provide the first experimental evidence for the theory.

Standard Model Extension

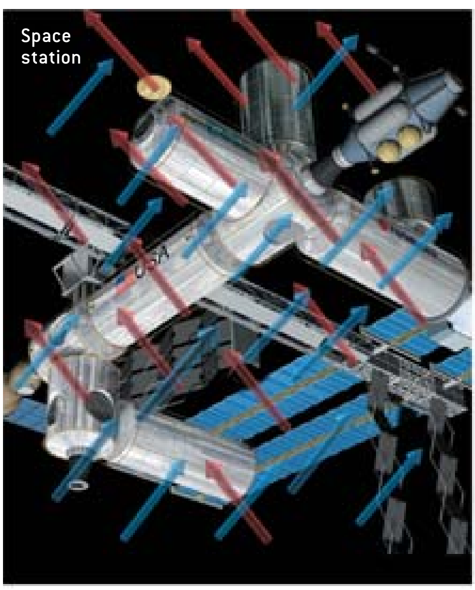
SUPPOSE THAT the fundamental theory of nature does contain Lorentz violation, perhaps with CPT violation, through some mechanism. How would this manifest itself in experiments, and how can it be related to known physics? To answer these questions, we would like to have a general theoretical framework that encompasses all possible effects and that can be applied to analyze any experiment. With such a framework, specific experimental parameters can be calculated, different experiments can be compared, and predictions can be made for the kind of effects to be expected.

Certain criteria guide our construction of this framework. First, all physical phenomena should be independent of the particular coordinate system used to map out space and time. Second, the experimental successes of the Standard Model and general relativity mean that any Lorentz and CPT violations must be small. By following these criteria and using only the known forces and particles, we are led to a set of possible terms—possible interactions—that could be added to the equations of the theory. Each term corresponds to a particular tensor field acquiring a nonzero background value. The coefficients that specify the magnitudes of these terms are unknown, and indeed many might be zero when the ultimate theory is known.

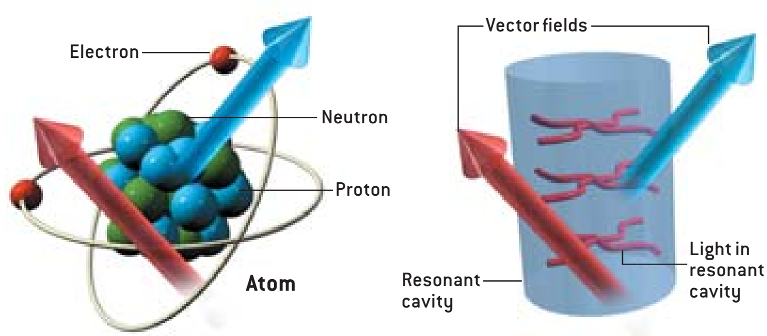
The end result is a theory called the Standard Model Extension, or SME. The beauty of this formulation is its generality: whatever your philosophical or physical preferences for the origin of relativity violations, the resulting effects in nature

ORBITING LABORATORIES

Studying Space in Space



On satellites such as the space station will be experiments that seek evidence of Lorentz violations in comparisons of clocks. The illustration shows the case of two relativity-violating vector fields (*red* and *blue arrows*) with different interactions with particles. Depicted below is a comparison between an atomic clock (*represented by an atom*) and a clock based on light or microwaves (*wavy lines*) in a resonant cavity. The light and electrons (*red*) interact with the red vectors, whereas protons (*blue*) interact with the blue vectors. As the space station rotates, these changing interactions cause the clocks to go in and out of sync, revealing the Lorentz violation. The 92-minute rotation of the space station provides for much faster and more sensitive data taking than a stationary Earth-based experiment.



DON FOLEY

ICONOCLASTS

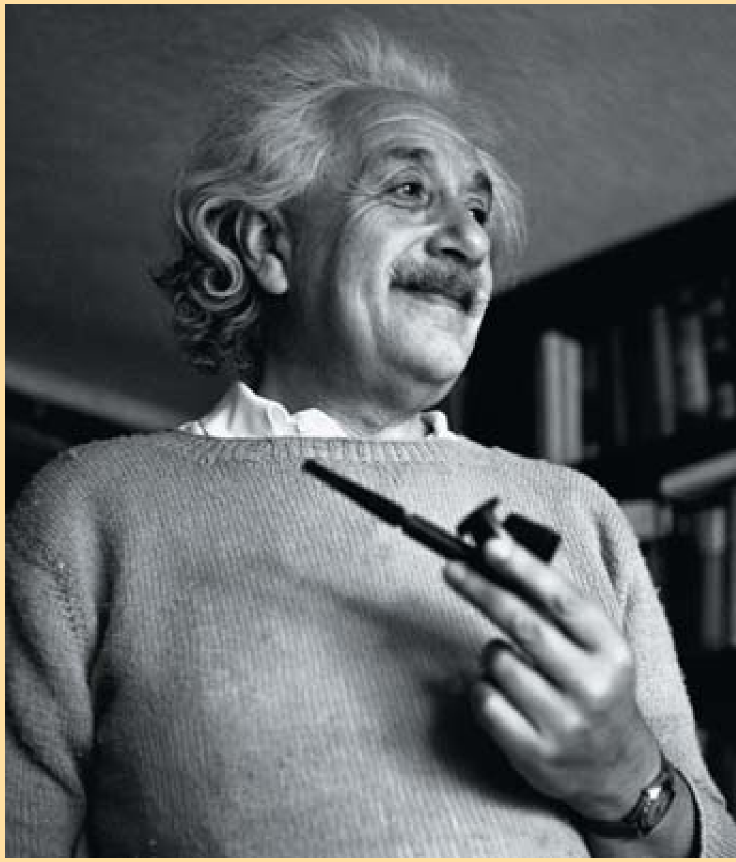
Topping the Giant

Everyone wants to get a piece of Einstein. Two of the three most common crackpot missives received by scientists and science magazines involve Einstein: claims to have a unified theory (succeeding where Einstein failed) and claims to have proved his theories false. (The third big class of craziness: perpetual-motion machines and infinite-energy sources.) Like cannibals seeking the strength and life spirit of their victims, these misguided amateurs seem to think that by outdoing or disproving Einstein they will acquire all his prestige and acclaim. Of course, all that they disprove is their own competence with basic relativity.

But the crazies are not the only iconoclasts. Many serious and well-qualified researchers also seek to go beyond Einstein, in the way that he went beyond Galileo and Newton. The accompanying article by Alan Kostelecký describes the experimental search for departures from Einsteinian relativity. The analysis he discusses is based on a general "Standard Model Extension" in which all plausible relativity-violating terms are added to the equations of particle physics. This all-encompassing model covers every possible deviation that could trickle down to everyday physics from the high-energy pinnacle of the (as yet undiscovered) ultimate unified theory.

Yet certain putative breaches of relativity have attracted specific attention. One class of theories goes by the name "doubly special relativity," which has been studied by Giovanni Amelino-Camelia of the University of Rome since 2000 and later by Lee Smolin of the Perimeter Institute for Theoretical Physics in Waterloo, Ontario, João Magueijo of Imperial College London and others. Magueijo, incidentally, fits the label "iconoclast" to a T—as is apparent from his argumentative book *Faster Than the Speed of Light.*

Doubly special relativity is inspired by quantum gravity theories such as loop quantum gravity [see "Atoms of Space and Time," by Lee Smolin, on page 56], and it imposes a second kind of "speed limit" that works in conjunction with the conventional barrier of the speed of light in a vacuum, also known as *c.* The idea is that at very short distances the smooth continuity of spacetime should break down into something more granular—like grains of



TOWERING FIGURE of Albert Einstein provides a tempting target for physicists of all stripes. He would perhaps look with approval on these efforts to go beyond his theories.

LUCIEN AIGNER *Corbis*

must be described by the SME, because it contains all viable modifications and generalizations of relativity that are compatible with the Standard Model and the known behavior of gravity.

To visualize the effects of Lorentz violation, it is useful to think of spacetime as having an intrinsic orientation. In the case of a vector field causing a particular term in the SME equations, the orientation coincides with the direction of the vector field. The more general case of a tensor field is similar but more complicated. By virtue of couplings to these background fields, the motions and interactions of particles acquire a directional dependence, like charged particles moving in an electric or a magnetic field. A similar visualization works for CPT violation, but in this case the effects occur because particles and antiparticles have different couplings to the background fields.

The SME predicts that the behavior of a particle can be affected by relativity violations in several ways. The particle's properties and interactions can depend on the direction it is moving (rotation violations) and on how fast it is going (boost violations). The particle may have spin, an intrinsic quantity of angular momentum, in which case the relativity-violating behavior can depend on the size and orientation of the spin. The particle can also fail to mirror its antiparticle (CPT violations). Each behavior can vary depending on the species of particle; for instance, protons might be affected more than neutrons, and electrons not at all. These effects combine to produce a plethora of interesting signals that can be sought in experiments. A number of such experiments have begun, but so far none has provided conclusive evidence for relativity violations.

Ancient Light

ONE WAY TO OBTAIN exceptional sensitivity to relativity violations is by studying the properties of polarized light that has traveled billions of light-years across the cosmos. Certain relativity-violating interactions in the SME will change the polarization of light as sand or the network of a spider's web. In quantum physics, short distances and short times correspond to high momenta and high energies. Thus, at sufficiently high energy—the so-called Planck energy—a particle should "see" the graininess of spacetime. That violates relativity, which depends on spacetime being smooth down to the tiniest size scales. Reflecting this, in a doubly special theory, just as a particle cannot be accelerated beyond *c*, it cannot be boosted beyond the Planck energy.

Some of these models predict that extremely high frequency light should travel faster than lower-frequency light. Experimenters are looking for that effect in light from distant explosions called gamma-ray bursts.

But skeptics are unconvinced that these theories are well founded. Some researchers argue, for example, that the equations are physically equivalent to ordinary relativity, just dressed up in enough complexities for that to be unobvious. The proof of the pudding will have to come from a rigorous derivation of such a theory from something more fundamental, such as string theory or loop quantum gravity. Not to mention experimental evidence.

Another infraction that some have contemplated is that *c* itself has varied over the history of the universe. John W. Moffat of the University of Toronto studied models of this type in the early 1990s, and Magueijo has been a more recent champion of them. If *c* had been much greater in the very early moments of the big bang, certain effects could have propagated at an extremely fast rate, which would solve some cosmological puzzles.

If *c* varies, so, too, does the fine structure constant, alpha, which is a dimensionless number that specifies the strength of the electromagnetic interaction. Alpha can be expressed in terms of *c*, Planck's constant and the charge of the electron. Alpha can therefore also change with *c* remaining constant, which might not infringe on relativity but would be equally seismic. Such variation in alpha could occur in string theory, where the magnitude of alpha depends on the precise structure of extra tiny dimensions that are appended to the four dimensions of space and time that we know and love [see "The String Theory Landscape," by Raphael Bousso and Joseph Polchinski, on page 40].

The possibility that alpha might change was considered as long ago as 1955, by the great Russian physicist Lev Landau. Today physicists and astronomers are looking at ancient light from distant quasars for evidence that alpha was slightly different eons ago. Changing alpha would subtly alter the frequency of light emitted or absorbed by atoms and ions. Most searches for such shifts have turned up empty thus far. One exception is the results of a group led by John K. Webb of the University of New South Wales in Australia. Those researchers have used a novel method of analyzing the data to achieve finer precision and have reported evidence (albeit statistically somewhat weak) of shifts: between eight billion and 11 billion years ago, alpha appears to have been about six parts in a million feebler than it is today. Such a minute variation is hard to reconcile with the string theory explanation, which predicts long-term stability of constants such as alpha, punctuated by occasional catastrophic changes of great magnitude.

Some researchers, however, assert that the precision claimed by the new method is not correct and that the "shifts" are just statistical fluctuations. In March 2004 a team of astronomers led by Patrick Petitjean of the Institute of Astrophysics of Paris and the Observatory of Paris and Raghunathan Srianand of the Inter-University Center for Astronomy and Astrophysics in Pune, India, reported using the traditional methods pushed to the limit. They concluded that as far back as 10 billion years, alpha has changed by less than 0.6 part in a million, contradicting the claims of Webb and company.

So far then, Einstein has withstood all challengers. The iconoclasts will have to keep looking for the first chink in his armor.

—Graham P. Collins, staff writer

it travels through otherwise empty space. The change grows as the light travels greater distances.

In the SME, the dominant relativity violations involving light include both ones that break CPT and ones that preserve it. Those that break CPT are expected for technical theoretical reasons to be absent or negligible, and studies of cosmological data have confirmed this down to a sensitivity of 10^{-42}. About half the CPT-preserving relativity violations for light would be observable by measuring cosmological polarization: the change in polarization as the light travels would depend on the color of the light. At Indiana University, Matthew Mewes and I have searched for this effect in polarization data of infrared, visible and ultraviolet light from distant galaxies, obtaining a sensitivity of 10^{-32} on the coefficients controlling these violations.

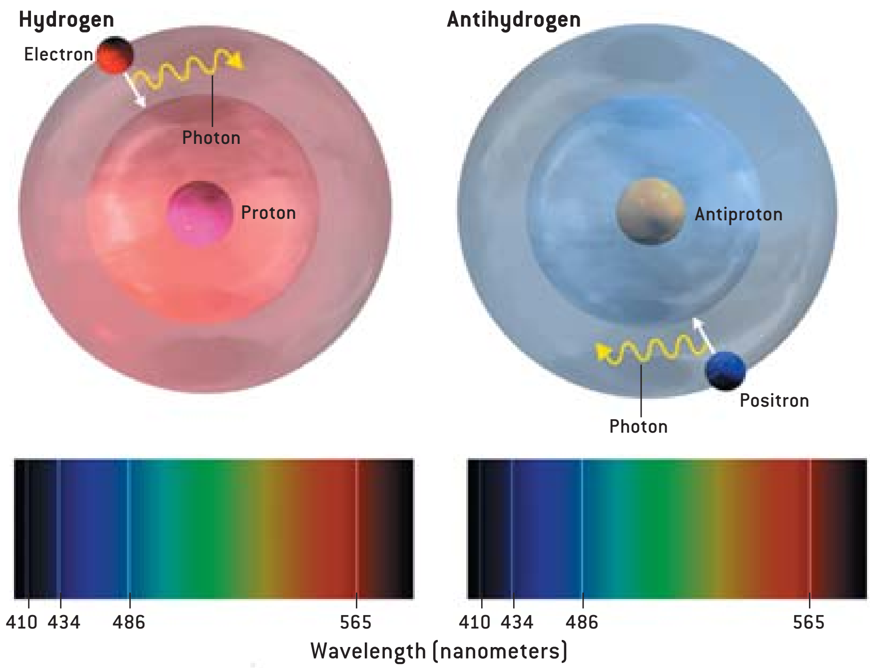
The remaining relativity violations for light can be measured in the laboratory using modern versions of experiments similar to the classic Michelson-Morley test of relativity (named after physicist Albert Michelson and chemist Edward Morley). The original Michelson-Morley experiment sent two beams of light at right angles and verified that their relative speed is independent of direction. The most sensitive experiments nowadays use resonant cavities; for example, rotating one on a turntable and searching for changes in the resonant frequency as it rotates. John A. Lipa's group at Stanford University uses superconducting cavities to study the properties of microwave resonances. Achim Peters of Humboldt University in Berlin, Stephan Schiller of Düsseldorf University in Germany and their collaborators use laser light in sapphire crystal resonators. These experiments and similar ones by other groups have already achieved sensitivities of 10^{-15} to 10^{-11}.

Clock-Comparison Experiments

EXCEPTIONAL SENSITIVITY to relativity violations has also been achieved

CPT SYMMETRY

Antimatter Experiments



Antimatter should behave in identical fashion to matter if a form of spacetime symmetry called CPT invariance holds. Two experiments at CERN near Geneva are testing this hypothesis using antihydrogen atoms. A hydrogen atom emits light with a characteristic color or wavelength when its electron drops from a higher energy level to a lower one (*top left*). The same process in antihydrogen (*top right*) should emit the same color light (photons are their own antiparticles, so it is still a photon that is emitted). Thus, if CPT invariance holds, antihydrogen and hydrogen should have identical emission spectra (*bottom*). The CERN experiments will actually use absorption of ultraviolet laser light (the inverse of the emission process shown here) and transitions involving microwaves, all of which should also be identical for hydrogen and antihydrogen. Any discrepancy would be a signal of CPT violation, which in turn implies Lorentz violation.

DON FOLEY

in clock-comparison experiments, which search for changes in the ticking rate of a clock depending on its orientation. The typical basic "clock" is an atom in a magnetic field, and the ticking rate is the frequency of a transition between two of the atom's energy levels that depends on the strength of the magnetic field. The orientation of the clock is defined by the direction of the applied magnetic field, which is usually fixed in the laboratory and so rotates as the earth rotates. A second clock monitors the ticking rate of the first one. The second clock is often taken to be a different type of atom undergoing the same kind of transition. The ticking rates (the transition frequencies) have to be affected by different amounts for the violation to become apparent.

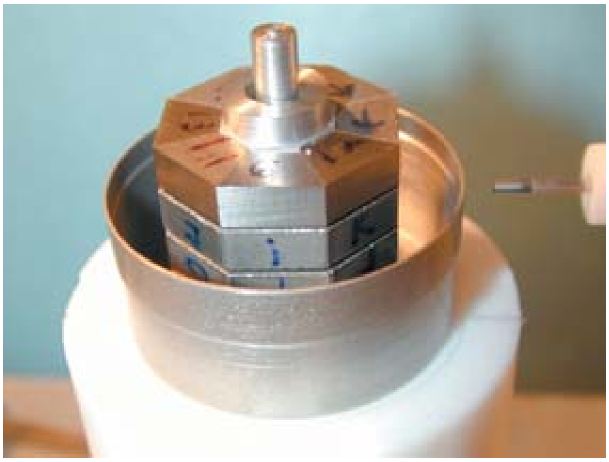
To date, the most sensitive experiments of this type have been performed in Ronald Walsworth's laboratory at the Harvard-Smithsonian Center for Astrophysics. These experiments have attained the remarkable sensitivity of 10^{-31} to a specific combination of SME coefficients for neutrons. Walsworth's group mixes helium and neon in a single glass bulb and turns both gases into masers (microwave lasers), a difficult technical feat. The frequencies of the two masers are compared.

Various clock-comparison experiments with atoms as clocks have been performed at other institutions, achieving sensitivities of 10^{-27} to 10^{-23} for different types of relativity violations involving protons, neutrons and electrons. Other experiments have used (instead of atoms) individual electrons, positrons (antielectrons), negatively charged hydrogen ions and antiprotons in electromagnetic traps, and muonium (an "atom" made of an electron orbiting a positive muon particle).

Researchers have plans for several clock-comparison experiments on the International Space Station (ISS) and other satellites. These experiments would have a number of potential advantages, including easier access to all spatial directions. Typical ground-based clock-comparison experiments use the earth's rotation, but the fixed rotational axis limits sensitivity to some types of rotation violation. Because the ISS's orbital plane is inclined and precesses, all spatial directions could be sampled. Another advantage is that the ISS's orbital period of 92 minutes would allow data to be taken about 16 times as fast as a fixed earth-based experiment. (The ISS is often configured to keep the same side facing the earth, and thus it rotates every 92 minutes as well as orbiting in that time.)

Antimatter

DIRECT TESTS FOR CPT violation can be performed by comparing properties of particles and antiparticles. One of the classic CPT tests involves a type of fundamental particle called the kaon. It turns out that the weak interaction causes a kaon gradually to convert into its antiparticle, the antikaon, and then back again. These kaon oscillations are so finely balanced that even a minuscule CPT violation would change them noticeably. Several large experimental collaborations have studied the oscillations of kaons to search for CPT violation. At present, the most sensitive constraint on Lorentz and CPT violation in kaons has been achieved by the KTeV Collaboration. This experiment used the giant Tevatron accelerator at Fermilab to create vast numbers of kaons. The results yielded two independent measurements



SPIN-COUPLED FORCES are investigated by a University of Washington experiment involving a torsion pendulum experiment (in which the hanging pendulum bob twists back and forth on its wire). The bob (*photograph above*) consists of rings of magnets made of two different materials (*red* and *blue at right*). The field of each magnet type has the same strength but is generated by a different quantity of electron spins (*arrows*). The magnetic field forms a closed loop with very little field outside the bob, reducing spurious signals caused by magnetic interactions. The electron spins, however, are unbalanced. If there is a sufficiently large relativity-violating vector field that interacts with electron spin, it will show up in perturbations of the bob's oscillations.

TED COOK *University of Washington (photograph)*; DON FOLEY *(illustration)*

of SME coefficients at the level of 10^{-21}.

Two experiments, ATHENA and ATRAP, both taking place at CERN (the European laboratory for particle physics near Geneva), are under way to trap antihydrogen and compare its spectroscopic properties with those of hydrogen, which should be identical if CPT is preserved [*see box on opposite page*]. Any difference uncovered would represent a CPT violation and consequently a Lorentz violation.

High-sensitivity tests of relativity have also used objects made of materials in which the spins of many electrons combine to yield a net overall spin. (Think of each electron's "spin" as being a tiny compass needle. Opposite pointing needles cancel, but parallel ones add to give a larger total spin.) Such materials are common—for example, an overall spin produces the magnetic field of a bar magnet. In searching for Lorentz violation, however, the presence of a strong magnetic field is a hindrance. To circumvent this, Eric Adelberger, Blayne Heckel and their colleagues at the University of Washington have designed and built a spin-polarized ring of material that has a net electron spin but no external magnetic field [*see illustration above*]. The ring is used as the bob in a torsion pendulum, which twists back and forth while suspended from a mounting on a rotating platform. A spin-dependent Lorentz violation would show up as a perturbation of the pendulum's oscillations that depends on the pendulum's orientation. This apparatus has been used to set the best current bounds on relativity violations involving electrons, at 10^{-29}.

It is possible that relativity violations have already been detected but have not been recognized as such. In recent years, ghostly fundamental particles called neutrinos have been shown to oscillate, which requires a modification of the minimal form of the Standard Model [see "Solving the Solar Neutrino Problem," by Arthur B. McDonald, Joshua R. Klein and David L. Wark, on page 22]. The oscillations are usually ascribed to small, previously unknown masses of neutrinos. But unusual oscillation properties for neutrinos are also predicted in the SME. Theorists have shown that the description of neutrino behavior in terms of relativity violations and the SME may be simpler than the conventional description in terms of masses. Future analyses of neutrino data could confirm this idea.

The experiments I have discussed have demonstrated that Planck-scale sensitivities are attainable with existing techniques. Although no compelling evidence for relativity violations has emerged to date, comparatively few types of relativity violations have been studied so far. The next few years will see major improvements both in the scope of relativity tests (more coefficients measured) and in their depth (improved sensitivities). If relativity violations are finally discovered, they will signal a profound change in our understanding of the universe at its most fundamental level. SA

MORE TO EXPLORE

Testing Times in Space. Steve K. Lamoreaux in *Nature,* Vol. 416, pages 803–804; April 25, 2002.

Back to the Future. Philip Ball in *Nature,* Vol. 427, pages 482–484; February 5, 2004.

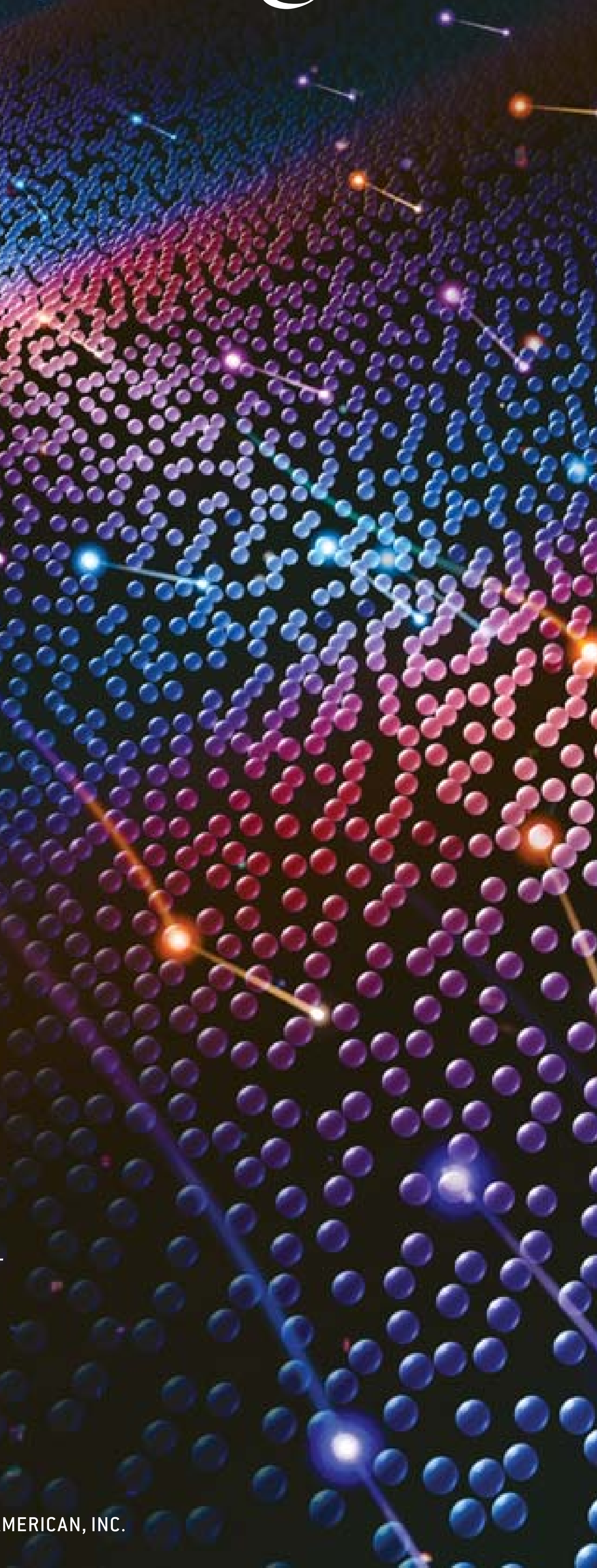
Breaking Lorentz Symmetry. Robert Bluhm in *Physics World,* Vol. 17, No. 3, pages 41–46; March 2004. Available at **physicsweb.org/article/world/17/3/7**

Lorentz Invariance on Trial. Maxim Pospelov and Michael Romalis in *Physics Today,* Vol. 57, No. 7, pages 40–46; July 2004.

Special Relativity Reconsidered. Adrian Cho in *Science,* Vol. 307, pages 866–868; February 11, 2005.

Alan Kostelecký's Web site on Lorentz and CPT violation is at **www.physics.indiana.edu/~kostelec/faq.html**

Solving the



The Sudbury Neutrino Observatory has solved a 30-year-old mystery by showing that neutrinos from the sun change species en route to the earth

By Arthur B. McDonald, Joshua R. Klein and David L. Wark

Building a detector the size of a 10-story building two kilometers underground is a strange way to study solar phenomena. Yet that has turned out to be the key to unlocking a decades-old puzzle about the physical processes occurring inside the sun. English physicist Arthur Eddington suggested as early as 1920 that nuclear fusion powered the sun, but efforts to confirm critical details of this idea in the 1960s ran into a stumbling block: experiments designed to detect a distinctive by-product of solar nuclear fusion reactions—ghostly particles called neutrinos—observed only a fraction of the expected number of them. It was not until 2002, with the results from the underground

Solar Neutrino Problem

Sudbury Neutrino Observatory (SNO) in Ontario, that physicists resolved this conundrum and thereby fully confirmed Eddington's proposal.

Like all underground experiments designed to study the sun, SNO's primary goal is to detect neutrinos, which are produced in great numbers in the solar core. But unlike most of the other experiments built over the previous three decades, SNO detects solar neutrinos using heavy water, in which a neutron has been added to each of the water molecules' hydrogen atoms (making deuterium). The additional neutrons allow SNO to observe solar neutrinos in a way never done before, by counting all three types, or "flavors," of neutrino equally. Using this ability, SNO has demonstrated that the deficit of solar neutrinos seen by earlier experiments resulted not from poor measurements or a misunderstanding of the sun but from a newly discovered property of the neutrinos themselves.

Ironically, the confirmation of our best theory of the sun exposes the first flaw in the Standard Model of particle physics—our best theory of how the most fundamental constituents of matter behave. We now understand the workings of the sun better than we do the workings of the microscopic universe.

The Problem

THE FIRST SOLAR NEUTRINO EXPERIMENT, conducted in the mid-1960s by Raymond Davis, Jr., now at the University of Pennsylvania, was intended to be a triumphant confirmation of the fusion theory of solar power generation and the start of a new field in which neutrinos could be used to learn more about the sun. Davis's experiment, located in the Homestake gold mine near Lead, S.D., detected neutrinos by a radiochemical technique. The detector contained 615 metric tons of liquid tetrachloroethylene, or dry-cleaning fluid, and neutrinos transformed atoms of chlorine in this fluid into atoms of argon. But rather than seeing one atom of argon created each day, as theory predicted, Davis observed just one every 2.5 days. (In 2002 Davis shared the Nobel Prize with Masatoshi Koshiba of the University of Tokyo for pioneering work in neutrino physics.) Thirty years of experiments following Davis's all found similar results despite using a variety of techniques. The number of neutrinos arriving from the sun was always significantly less than the predicted total, in some cases as low as one third, in others as high as three fifths, depending on the energies of the neutrinos studied. With no understanding of why the predictions and the measurements were so different, physicists had to put on hold the original goal of studying the solar core by observing neutrinos.

While experimenters continued to run their neutrino experiments, the late John Bahcall of the Institute for Advanced Study in Princeton, N.J., and other theorists improved the models used to predict the rate of solar neutrino production. Those theoretical models are complex but make only a few assumptions—that the sun is powered by nuclear reactions that change the element abundances, that this power creates an outward pressure balanced by the inward pull of gravity, and that energy is transported by photons and convection. The solar models continued to predict neutrino fluxes that exceeded measurements, but other projections they made, such as the spectrum of helioseismologic vibrations seen on the solar surface, agreed very well with observations.

The mysterious difference between the predictions and the measurements became known as the solar neutrino problem. Although many physicists still believed that inherent difficulties in detecting neutrinos and calculating their production rate in the sun were somehow the cause of the discrepancy, a third alternative became widely favored despite its somewhat revolutionary implications. The Standard Model of particle physics holds that there are three completely distinct, massless flavors of neutrinos: the electron-neutrino, muon-neutrino and tau-neutrino. The fusion reactions in the center of the sun can produce only electron-neutrinos, and experiments like Davis's were designed to look exclusively for this one flavor: at solar neutrino energies, only electron-neutrinos can convert chlorine atoms to argon. But if the Standard Model were incomplete, and the neutrino flavors were not distinct but instead mixed in some way, then an electron-neutrino from the sun might transform into one of the other flavors and thus escape detection.

The most favored mechanism for a change in neutrino flavor is neutrino oscillation [*see box on pages 26 and 27*], which requires that the neutrino flavors (electron-, muon- and tau-neutrinos) are made up of mixtures of neutrino states (denoted as 1, 2 and 3) that have different masses. An electron-neutrino could then be a mixture of states 1 and 2, and a muon-neutrino could be a different mixture of the same two states. Theory predicts that as they travel from the sun to the earth, such mixed neutrinos will oscillate between one flavor and another.

Strong evidence of neutrino oscillation came from the Super-Kamiokande collaboration in 1998, which found that muon-neutrinos formed in the upper atmosphere by cosmic rays were disappearing with a probability that depended on

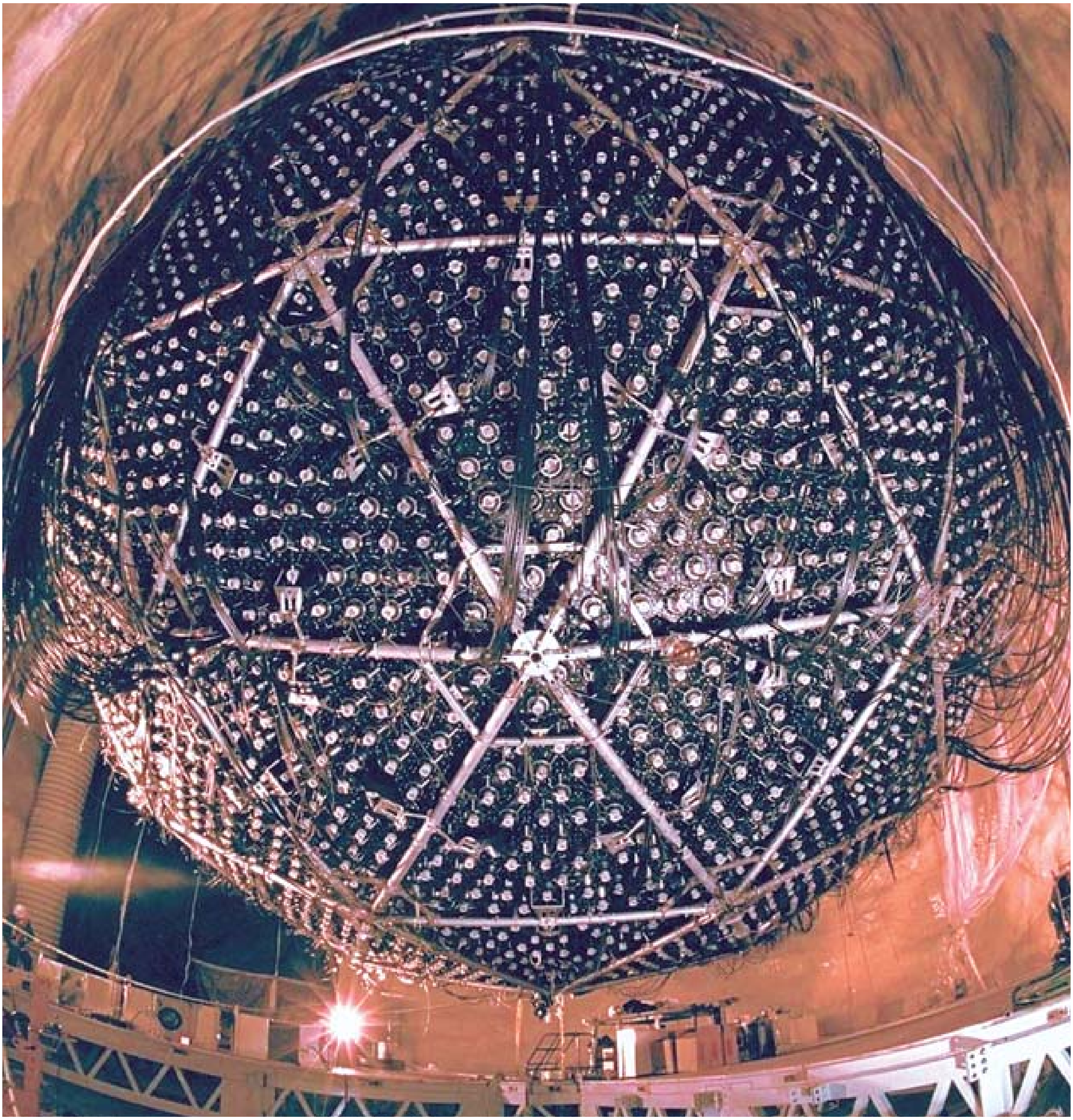
Overview/*Oscillating Neutrinos*

- Since the 1960s, underground experiments have been detecting far fewer electron-neutrinos from the sun than theory predicts. The mystery came to be known as the solar neutrino problem.
- In 2002 the Sudbury Neutrino Observatory (SNO) resolved the solar neutrino problem by determining that many of the electron-neutrinos produced inside the sun change to other flavors of neutrinos before reaching the earth, causing them to go undetected by past experiments.
- SNO's result confirms that we understand how the sun is powered and implies that neutrinos, long thought to be massless, have masses. The Standard Model of particle physics, which is otherwise extraordinarily successful, must be modified to accommodate this change.

SLIM FILMS (*preceding pages*)

their distance traveled. This disappearance is explained very well by neutrino oscillations, in this case muon-neutrinos that are probably turning into tau-neutrinos. The former are easily detected by Super-Kamiokande at cosmic-ray energies and are probably turning into tau-neutrons that mostly evade detection [see "Detecting Massive Neutrinos," by Edward Kearns, Takaaki Kajita and Yoji Totsuka; SCIENTIFIC AMERICAN, August 1999].

A similar process could explain the solar neutrino deficit. In one scenario, the neutrinos would oscillate during their eight-minute journey through the vacuum of space from the sun to the earth. In another model, the oscillation would be



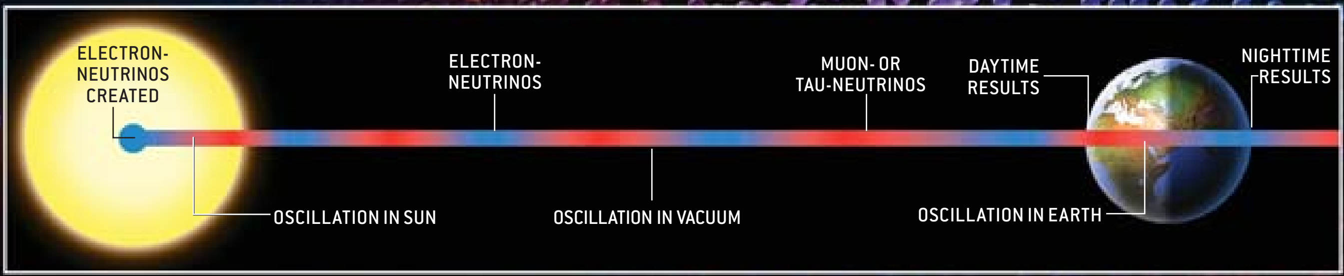
COURTESY OF LAWRENCE BERKELEY NATIONAL LABORATORY (*this page*); SLIM FILMS (*next two pages*)

PHOTOMULTIPLIER TUBES—more than 9,500 of them—on a geodesic sphere 18 meters in diameter act as the eyes of the Sudbury Neutrino Observatory. The tubes surround and monitor a 12-meter-diameter acrylic sphere that contains 1,000 tons of heavy water. Each tube can detect a single photon of light. The entire assembly is suspended in ordinary water. All the materials that make up the detector must be extraordinarily free of natural traces of radioactive elements to avoid overwhelming the tubes with false solar neutrino counts.

DETECTING FICKLE NEUTRINOS

HOW NEUTRINOS OSCILLATE

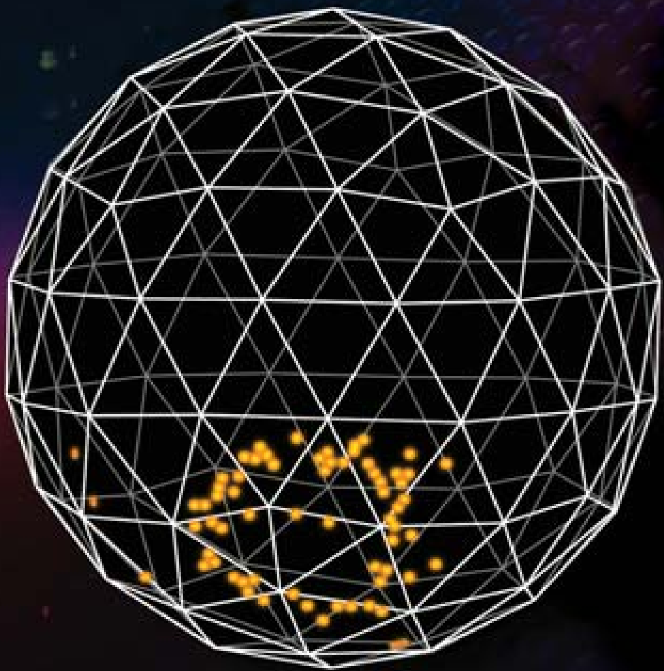
An electron-neutrino (*left*) is actually a superposition of a type 1 and a type 2 neutrino with their quantum waves in phase. Because the type 1 and type 2 waves have different wavelengths, after traveling a distance they go out of phase, making a muon- or a tau-neutrino (*center*). With further travel the neutrino oscillates back to being an electron-neutrino (*right*).



WHERE NEUTRINOS OSCILLATE

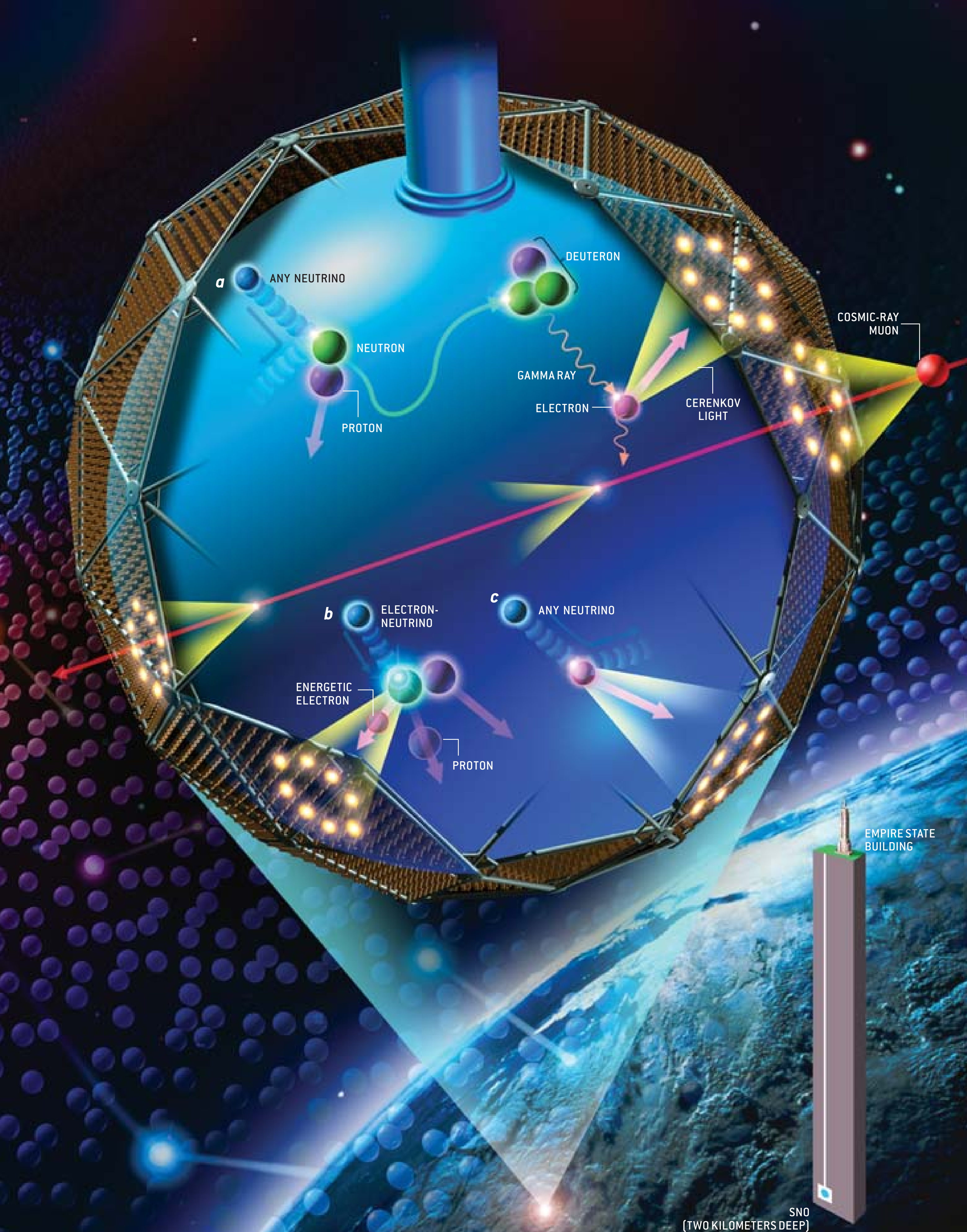
The electron-neutrinos produced at the center of the sun may oscillate while they are still inside the sun or after they emerge on their eight-minute journey to the earth. Which oscillation occurs depends on details such as the mass differences and the intrinsic degree of mixing of type 1 and 2 neutrinos. Extra oscillation may also occur inside the earth, which manifests as a difference between daytime and nighttime results.

ACTUAL DATA OF A CANDIDATE NEUTRINO EVENT



HOW SNO DETECTS NEUTRINOS

The Sudbury Neutrino Observatory, or SNO (*opposite page*), detects a neutrino by seeing a characteristic ring of Cerenkov light emitted by a high-speed electron. The neutrino produces the energetic electron in SNO's heavy water (*large blue sphere*) in one of three ways. In deuteron breakup (*a*), the neutrino (*blue*) splits a deuterium nucleus into its component proton (*purple*) and neutron (*green*). The neutron eventually combines with another deuteron, releasing a gamma ray (*wavy line*), which in turn knocks free an electron (*pink*) whose Cerenkov light (*yellow*) is detected. In neutrino absorption (*b*), a neutron absorbs the neutrino and is thereby turned into a proton and an energetic electron. Only electron-neutrinos can be absorbed in this way. Less often the neutrino may collide directly with an electron (*c*). Cosmic-ray muons (*red*) are distinguished from neutrinos by the amount of Cerenkov light they produce and where they produce it—outside the detector as well as inside. The number of muons is reduced to manageable levels by positioning the detector two kilometers underground.


a
ANY NEUTRINO
NEUTRON
PROTON
DEUTERON
GAMMA RAY
ELECTRON
CERENKOV LIGHT
COSMIC-RAY MUON
b
ELECTRON-NEUTRINO
ENERGETIC ELECTRON
PROTON
c
ANY NEUTRINO
EMPIRE STATE BUILDING
SNO
(TWO KILOMETERS DEEP)

enhanced during the first two seconds of travel through the sun itself, an effect caused by the different ways in which each neutrino flavor interacts with matter. Each scenario requires its own specific range of neutrino parameters—mass differences and the amount of intrinsic mixing of the flavors. Despite the evidence from Super-Kamiokande and other experiments, however, it remained possible that neutrinos were disappearing by some process other than oscillation. Until 2002 scientists had no direct evidence of solar neutrino oscillation, in which the transformed solar neutrinos themselves were detected.

The Observatory

THE SUDBURY NEUTRINO OBSERVATORY was designed to search for this direct evidence, by detecting neutrinos using several different interactions with its 1,000 tons of heavy water. One of these reactions exclusively counts electron-neutrinos; the others count all flavors without distinguishing among them. If the solar neutrinos arriving at the earth consisted only of electron-neutrinos—and therefore no flavor transformation was occurring—then the count of neutrinos of all flavors would be the same as the count of electron-neutrinos alone. On the other hand, if the count of all flavors was far in excess of the count of the electron-neutrinos, that would prove that neutrinos from the sun were changing flavor.

The key to SNO's ability to count both electron-neutrinos alone and all flavors is the heavy water's deuterium nuclei, also called deuterons. The neutron in a deuteron produces two separate neutrino reactions: neutrino absorption, in which an electron-neutrino is absorbed by a neutron and an electron is created, and deuteron breakup, in which a deuterium nucleus is broken apart and the neutron liberated. Only electron-neutrinos can undergo neutrino absorption, but neutrinos of any flavor can break up deuterons. A third reaction detected by SNO, the scattering of electrons by neutrinos, can also be used to count neutrinos other than electron-neutrinos but is much less sensitive to muon- and tau-neutrinos than the deuteron breakup reaction [*see box on preceding two pages*].

SNO was not the first experiment to use heavy water. In the 1960s T. J. Jenkins and F. W. Dix of Case Western Reserve University used heavy water in a very early attempt to observe neutrinos from the sun. They used about 2,000 liters (two tons) of heavy water aboveground, but the signs of solar neutrinos were swamped by the effects of cosmic rays. In 1984 Herb Chen of the University of California, Irvine, proposed bringing 1,000 tons of heavy water from Canada's CANDU nuclear reactor program to the bottom of INCO Ltd.'s Creighton nickel mine in Sudbury—a location deep enough to enable a clear measurement of both neutrino absorption and deuteron breakup for solar neutrinos.

This proposal led to the SNO collaboration—originally headed by Chen and George Ewan of Queen's University in Kingston, Ontario—and ultimately to the creation of the SNO detector. The 1,000 tons of heavy water are held in a 12-meter-diameter transparent acrylic vessel. The heavy water is viewed by more than 9,500 photomultiplier tubes held on an 18-meter-diameter geodesic sphere [*see illustration on page 25*]. Each tube can detect a single photon of light. The entire structure is submerged in ultrapure ordinary water filling a cavity carved out of the rock two kilometers below the surface of the earth.

SNO's Measurement

SOLAR NEUTRINOS CAN BE OBSERVED deep underground because of the extreme weakness of their interaction with matter. During the day, neutrinos easily travel down to SNO through two kilometers of rock, and at night they are almost equally unaffected by the thousands of kilometers that they travel up through the earth. Such feeble coupling makes them interesting from the perspective of solar astrophysics. Most of the energy created in the center of the sun takes tens of thousands of years to reach the solar surface and leave as sunlight. Neutrinos, in contrast, emerge after two seconds, coming to us directly from the point where solar power is created.

With neither the whole sun nor the entire earth able to impede the passage of neutrinos, capturing them with a detector weighing just 1,000 tons poses something of a challenge. But although the vast majority of neutrinos that enter SNO pass through it, on very rare occasions, one will—by chance alone—collide with an electron or an atomic nucleus and deposit enough energy to be observed. With enough neutrinos, even the rarity of these interactions can be overcome. Luckily, the sun's neutrino output is enormous—five million high-energy solar neutrinos pass through every square centimeter of the earth every second—which leads to about 10 observed neutrino events, or interactions, in SNO's 1,000 tons of heavy

EIGHT DECADES OF THE SUN AND NEUTRINOS

IT HAS TAKEN most of a century to verify fully that we understand how the sun generates its power. Along the way, neutrinos have gone from speculative hypothesis to key experimental tool. Their oscillations point to fundamental new physics to be discovered in the decades to come.

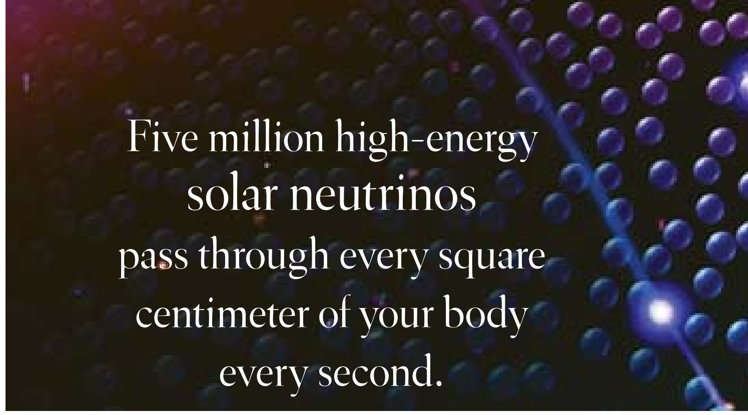
1920 — 1940

1920 **Arthur Eddington** proposes that the sun is powered by nuclear fusion converting hydrogen atoms into helium

1930 **Wolfgang Pauli** rescues conservation of energy by hypothesizing an unseen particle, the neutrino, that carries away energy from some radioactive decays

1938 **Hans Bethe** analyzes the basic nuclear processes that could power the sun and accurately estimates the sun's central temperature

1956 **Frederick Reines** and **Clyde Cowan** first detect the neutrino using the Savannah River nuclear reactor



water every day. The three types of neutrino reaction that occur in SNO all generate energetic electrons, which are detectable through their production of Cerenkov light—a cone of light emitted like a shock wave by the fast-moving particle.

This small number of neutrino events, however, has to be distinguished from flashes of Cerenkov light caused by other particles. In particular, cosmic-ray muons are created continually in the upper atmosphere, and when they enter the detector they can produce enough Cerenkov light to illuminate every photomultiplier tube. The intervening kilometers of rock between the earth's surface and SNO reduce the deluge of cosmic-ray muons to a mere trickle of just three an hour. And although three muons an hour is a far greater rate than the 10 observed neutrino events a day, these muons are easy to distinguish from neutrino events by the Cerenkov light they produce in the ordinary water outside the detector.

A far more sinister source of false neutrino counts is the intrinsic radioactivity in the detector materials themselves. Everything inside the detector—from the heavy water itself to the acrylic vessel that holds it to the glass and steel of the photomultiplier tubes and support structure—has trace amounts of naturally occurring radioactive elements. Similarly, the air in the mine contains radioactive radon gas. Every time a nucleus in these radioactive elements decays inside the SNO detector, it can release an energetic electron or gamma ray and ultimately produce Cerenkov light that mimics the signal of a neutrino. The water and the other materials used in SNO are purified to remove the bulk of the radioactive contaminants (or were chosen to be naturally pure), but even parts per billion are enough to overwhelm the true neutrino signal with false counts.

The task before SNO is therefore very complex—it must count neutrino events, determine how many are caused by each of the three reactions, and estimate how many of the apparent neutrinos are caused by something else, such as radioactive contamination. Errors as small as a few percent in any of the steps of analysis would render meaningless SNO's comparison of the electron-neutrino flux to the total neutrino flux. Over the 306 days of running, from November 1999 to May 2001, SNO recorded nearly half a billion events. By the time the data reduction was complete, only 2,928 of these remained as candidate neutrino events.

SNO cannot uniquely determine whether a given candidate neutrino event was the result of a particular reaction. Typically an event like the one shown on page 26 could equally well be the result of deuteron breakup as neutrino absorption. Fortunately, differences between the reactions show up when we examine many events. For example, deuteron breakup, the splitting of a deuterium nucleus in the heavy water, always leads to a gamma ray of the same energy, whereas the electrons produced by neutrino absorption and electron scattering have a broad spectrum of energies. Similarly, electron scattering produces electrons that travel away from the sun, whereas the Cerenkov light from deuteron breakup can point in any direction. Finally, the locations where the reactions occur differ as well—electron scattering, for instance, occurs as easily in the outer layer of light water as in the heavy water; the other reactions do not. With an understanding of those details, SNO researchers can statistically determine how many of the observed events to assign to each reaction.

Such an understanding is the result of measurements that were complete nuclear physics experiments in their own right: to determine how to measure energy using Cerenkov light, sources of radioactivity with known energies were inserted inside the detector. To measure how the Cerenkov light travels through and reflects off the various media in the detector (the water, the acrylic, the photomultiplier tubes), a variable wavelength laser light source was used. The effects of radioactive contamination were assessed by similar experiments, including radioassays of the water using new techniques designed specifically for SNO.

For the final SNO data set, after statistical analysis, 576 events were assigned to deuteron breakup, 1,967 events to neutrino absorption and 263 to electron scattering. Radioactivity and other backgrounds caused the remaining 122. From these numbers of events, one must calculate how many actual neutrinos must be passing through SNO, based on the tiny prob-

BRYAN CHRISTIE DESIGN

1960 — 1980 — 2000

1964 **John Bahcall** predicts the neutrino flux expected to be observed from the sun

1967 **Raymond Davis, Jr.**, first measures neutrinos from the sun, using 600 tons of dry-cleaning fluid in a mine in Lead, S.D.

1969 **Vladimir Gribov** and **Bruno Pontecorvo** propose that neutrino oscillations explain the anomalously small number of neutrinos detected

1978 and 1985 **Stanislav Mikheyev, Alexei Smirnov** and **Lincoln Wolfenstein** posit that matter can enhance neutrino oscillations

1998 **Super-Kamiokande** assembles evidence of neutrino oscillation in cosmic-ray neutrinos

2002 **SNO** confirms that electron-neutrinos from the sun are oscillating to another flavor, fully resolving the solar neutrino problem

2002 **KamLAND** experiment detects oscillation of antineutrinos emitted from nuclear reactors

Some Other Neutrino Experiments

HOMESTAKE: Solar neutrino detector located in the Homestake gold mine in Lead, S.D. The original chlorine experiment started in 1967, using 600 tons of dry-cleaning fluid.

KAMIOKA: Houses Super-Kamiokande, a 50,000-ton light-water detector studying cosmic-ray and solar neutrinos, as well as muon-neutrinos beamed from the KEK facility 250 kilometers away ("K2K" experiment). Also houses KamLAND, a smaller detector (1,000 tons of liquid scintillator, which emits light when a charged particle passes through) that counts anti-electron-neutrinos emitted by all the nuclear reactors nearby in Japan and South Korea. Originally housed Kamiokande, a light-water detector that observed cosmic-ray and solar neutrinos and was converted to KamLAND.

SAGE (Russian-American Gallium Solar Neutrino Experiment): Located at Baksan in the Caucasus Mountains in Russia. Uses 50 tons of gallium, which is capable of detecting the low-energy neutrinos produced by proton-proton fusion in the sun.

GRAN SASSO: The world's largest underground laboratory, accessed via a highway tunnel, located under the Gran Sasso Mountains about 150 kilometers east of Rome. Solar neutrino experiments include Gallex/GNO, which began in 1991 and uses 30 tons of gallium (as aqueous gallium trichloride), and Borexino, a sphere of 300 tons of scintillator viewed by 2,200 photomultipliers.

MINIBOONE (Booster Neutrino Experiment): Located at Fermilab in Illinois. Beams of muon-neutrinos and anti-muon-neutrinos travel through 500 meters of earth to be detected in an 800-ton tank of mineral oil. Endeavoring to test a controversial result reported by the LSND experiment at Los Alamos National Lab in 1995. Began collecting data in September 2002.

MINOS: Will beam neutrinos from Fermilab to the Soudan detector, 735 kilometers away in Minnesota. Detector is 5,400 tons of iron laced with plastic particle detectors. Began taking data in 2005.

abilities that any particular neutrino will break up a deuteron, be absorbed or scatter an electron. The upshot of all the calculations is that the observed 1,967 neutrino absorption events represent 1.75 million electron-neutrinos passing through each square centimeter of the SNO detector every second. That is only 35 percent of the neutrino flux predicted by solar models. SNO thus first confirms what other solar neutrino experiments have seen—that the number of electron-neutrinos arriving from the sun is far smaller than solar models predict.

The critical question, however, is whether the number of electron-neutrinos arriving from the sun is significantly smaller than the number of neutrinos of all flavors. Indeed, the 576 events assigned to deuteron breakup represent a total neutrino flux of 5.09 million per square centimeter per second—far larger than the 1.75 million electron-neutrinos measured by neutrino absorption. These numbers are determined with high accuracy. The difference between them is more than five times the experimental uncertainty.

THE AUTHORS

ARTHUR B. McDONALD, JOSHUA R. KLEIN and *DAVID L. WARK* are members of the 130-strong Sudbury Neutrino Observatory (SNO) collaboration. McDonald, a native of Sydney, Nova Scotia, has been the director of the SNO Institute since its inception in 1989. He is also professor of physics at Queen's University in Kingston, Ontario. Klein received his Ph.D. from Princeton University in 1994 and began his work on SNO at the University of Pennsylvania. He is now assistant professor of physics at the University of Texas at Austin. Wark has spent the past 13 years in the U.K., at the University of Oxford, the University of Sussex and the Rutherford Appleton Laboratory, trying to explain the infield fly rule to cricket fans. He has worked on a number of neutrino experiments in addition to SNO.

The excess of neutrinos measured by deuteron breakup means that nearly two thirds of the total 5.09 million neutrinos arriving from the sun are either muon- or tau-neutrinos. The sun's fusion reactions can produce only electron-neutrinos, so some of them must be transformed on their way to the earth. SNO has therefore demonstrated directly that neutrinos do not behave according to the simple scheme of three distinct massless flavors described by the Standard Model. In 30 years of trying, only experiments such as Super-Kamiokande and SNO had shown that the fundamental particles have properties not contained in the Standard Model. The observations of neutrino flavor transformation provide direct experimental evidence that there is yet more to be discovered about the microscopic universe.

But what of the solar neutrino problem itself—does the discovery that electron-neutrinos transform into another flavor completely explain the deficit observed for the past 30 years? It does: the deduced 5.09 million neutrinos agrees remarkably well with the predictions of solar models. We can now claim that we really do understand the way the sun generates its power. Having taken a detour lasting three decades, in which we found that the sun could tell us something new about neutrinos, we can finally return to Davis's original goal and begin to use neutrinos to understand the sun. For example, neutrino studies could determine how much of the sun's energy is produced by direct nuclear fusion of hydrogen atoms and how much is catalyzed by carbon atoms.

The implications of SNO's discovery go even further. If neutrinos change flavor through oscillation, then they cannot be massless. After photons, neutrinos are the second most numerous known particles in the universe, so even a tiny mass could have a significant cosmological significance. Neutrino oscillation experiments such as SNO and Super-Kamiokande measure only mass differences, not masses themselves. Showing that mass differences are not zero, however, proves that at least some of the masses are not zero. Combining the oscillation results for mass differences with upper limits for the electron-neutrino mass from other experiments shows that neutrinos make up something between 0.3 and 21 percent of the critical density for a flat universe. (Other cosmological data strongly indicate that the universe is flat.) This amount is not negligible (it is roughly comparable to the 4 percent density that arises from gas, dust and stars), but it is not quite enough to explain all the matter that seems to be present in the universe. Because neutrinos were the last known particles that could have made up the missing dark matter, some particle or particles not currently known to physics must exist—and with a density in excess of everything we know.

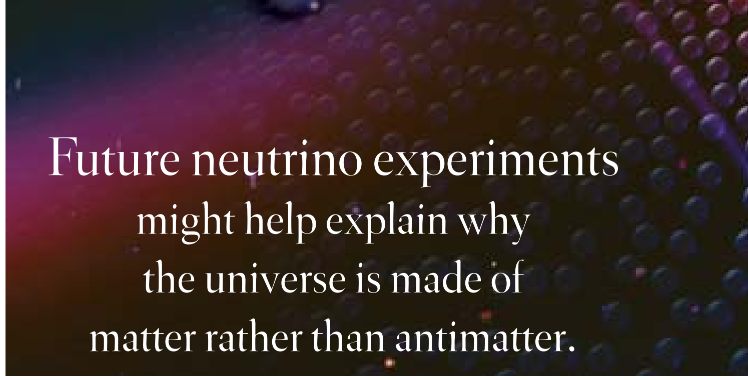
The Future

SNO HAS ALSO BEEN SEARCHING for direct evidence of the effects of matter on neutrino oscillations. As mentioned earlier, travel through the sun can enhance the probability of oscillations. If this occurs, the passage of neutrinos through thousands of kilometers of the earth could lead to a small reversal in the process—the sun might shine more brightly in electron-neutrinos at night than during the day. SNO's data show a small excess of electron-neutrinos arriving at night compared with during the day, but as of now the measurement is not significant enough to decide whether the effect is real.

The SNO results described earlier are just the beginning. For the observations cited here, the neutrons were detected from the critical deuteron breakup events by observing their capture by other deuterium atoms—an inefficient process that produces little light. In May 2001 two tons of highly purified sodium chloride (table salt) were added to the heavy water. Chlorine nuclei capture neutrons with much higher efficiency than deuterium nuclei do, producing events that have more light and can be distinguished from the neutrino absorption reaction and the background.

Thus, SNO has made a separate and more sensitive measurement of the deuteron breakup rate. These measurements, reported in 2003, provided strong confirmation of the previous SNO measurements and determined neutrino properties with increased accuracy. The SNO collaboration also built an array of ultraclean detectors called proportional counters, which were deployed throughout the heavy water in 2003 after the salt was removed, to observe directly the neutrons from the deuteron breakup reaction. Making these detectors was a technical challenge of the first order because they must have a spectacularly low level of intrinsic radioactive background—corresponding to about one count per meter of detector per year. Those devices allow for a number of further detailed measurements of neutrino properties.

SNO has unique capabilities, but it is not the only game in town. In December 2002 the first results from a new Japanese-American experiment called KamLAND were reported.



The KamLAND detector is at the Super-Kamiokande site and studies electron-antineutrinos produced by all the nuclear reactors nearby in Japan and Korea. If matter-enhanced neutrino oscillations explain the flavor change seen by SNO, theory predicts that these antineutrinos should also change flavor over distances of tens or hundreds of kilometers. Indeed, KamLAND has seen too few electron-antineutrinos, implying that they are oscillating en route from the nuclear reactors to the detector. The KamLAND results imply the same neutrino properties as those seen previously by SNO.

Future neutrino experiments might probe one of the biggest mysteries in the cosmos: Why is the universe made of matter rather than antimatter? Russian physicist Andrei Sakharov first pointed out that to get from a big bang of pure energy to the current matter-dominated universe requires the laws of physics to be different for particles and antiparticles. This is called CP (charge-parity) violation, and sensitive measurements of particle decays have verified that the laws of physics violate CP. The problem is that the CP violation seen so far is not enough to explain the amount of matter around us, so phenomena we have not yet observed must be hiding more CP violation. One possible hiding place is neutrino oscillations.

To observe CP-violating neutrino oscillations will be a multistage process. First physicists must see electron-neutrinos appear in intense beams of muon-neutrinos. Second, higher-intensity accelerators must be built to produce beams of neutrinos so intense and pure that their oscillations can be observed in detectors located across continents or on the other side of the earth. Studies of a rare radioactive process called neutrinoless double beta decay will provide further information about neutrino masses and CP violation.

It will probably be more than a decade before these experiments become a reality. A decade may seem a long way off, but the past 30 years, and the sagas of experiments such as SNO, have shown that neutrino physicists are patient and very persistent—one must be to pry out the secrets of these elusive particles. These secrets are intimately tied up with our next level of understanding of particle physics, astrophysics and cosmology, and thus persist we must. SA

MORE TO EXPLORE

The Origin of Neutrino Mass. Hitoshi Murayama in *Physics World,* Vol. 15, No. 5, pages 35–39; May 2002.

The Asymmetry between Matter and Antimatter. Helen R. Quinn in *Physics Today,* Vol. 56, No. 2, pages 30–35; February 2003.

The Neutrino Oscillation Industry Web site, maintained by Argonne National Laboratory, is at **www.neutrinooscillation.org**

The SNO Web site is at **www.sno.phy.queensu.ca**

MALE AFRICAN ELEPHANT (about 6,000 kilograms) and the smallest species of ant (0.01 milligram) differ in mass by more than 11 orders of magnitude—roughly the same span as the top quark and the neutrino. Why the particle masses should differ by such a large amount remains a mystery.

The Mysteries of MASS

Physicists are hunting for an elusive particle that would reveal the presence of a new kind of field that permeates all of reality. Finding that Higgs field will give us a more complete understanding about how the universe works

By Gordon Kane

Most people think they know what mass is, but they understand only part of the story. For instance, an elephant is clearly bulkier and weighs more than an ant. Even in the absence of gravity, the elephant would have greater mass—it would be harder to push and set in motion. Obviously the elephant is more massive because it is made of many more atoms than the ant is, but what determines the masses of the individual atoms? What about the elementary particles that make up the atoms—what determines their masses? Indeed, why do they even have mass?

We see that the problem of mass has two independent aspects. First, we need to learn how mass arises at all. It turns out mass results from at least three different mechanisms, which I will describe below. A key player in physicists' tentative theories about mass is a new kind of field that permeates all of reality, called the Higgs field. Elementary particle masses are thought to come about from the interaction with the Higgs field. If the Higgs field exists, theory demands that it have an associated particle, the Higgs boson. Using particle accelerators, scientists are now hunting for the Higgs.

The second aspect is that scientists want to know why different species of elementary particles have their specific quantities of mass. Their intrinsic masses span at least 11 orders of magnitude, but we do not yet know why that should be so [*see illustration on page 36*]. For comparison, an elephant and the smallest of ants differ by about 11 orders of magnitude of mass.

What *Is* Mass?

ISAAC NEWTON presented the earliest scientific definition of mass in 1687 in his landmark *Principia:* "The quantity of matter is the measure of the same, arising from its density and bulk conjointly." That very basic definition was good enough for Newton and other scientists for more than 200 years. They understood that science should proceed first by describing how things work and later by understanding why. In recent years, however, the *why* of mass has become a research topic in physics. Understanding the meaning and origins of mass will complete and extend the Standard Model of particle physics, the well-established theory that describes the known elementary particles and their interactions. It will also resolve mysteries such as dark matter, which makes up about 25 percent of the universe.

The foundation of our modern understanding of mass is far more intricate than Newton's definition and is based on the Standard Model. At the heart of the Standard

BRYAN CHRISTIE DESIGN

Model is a mathematical function called a Lagrangian, which represents how the various particles interact. From that function, by following rules known as relativistic quantum theory, physicists can calculate the behavior of the elementary particles, including how they come together to form compound particles, such as protons. For both the elementary particles and the compound ones, we can then calculate how they will respond to forces, and for a force *F,* we can write Newton's equation $F = ma$, which relates the force, the mass and the resulting acceleration. The Lagrangian tells us what to use for m here, and that is what is meant by the mass of the particle.

Why is the Higgs field present throughout the universe? What *is* the Higgs field?

But mass, as we ordinarily understand it, shows up in more than just $F = ma$. For example, Einstein's special relativity theory predicts that massless particles in a vacuum travel at the speed of light and that particles with mass travel more slowly, in a way that can be calculated if we know their mass. The laws of gravity predict that gravity acts on mass and energy as well, in a precise manner. The quantity m deduced from the Lagrangian for each particle behaves correctly in all those ways, just as we expect for a given mass.

Fundamental particles have an intrinsic mass known as their rest mass (those with zero rest mass are called massless). For a compound particle, the constituents' rest mass and also their kinetic energy of motion and potential energy of interactions contribute to the particle's total mass. Energy and mass are related, as described by Einstein's famous equation, $E = mc^2$ (energy equals mass times the speed of light squared).

An example of energy contributing to mass occurs in the most familiar kind of matter in the universe—the protons and neutrons that make up atomic nuclei in stars, planets, people and all that we see. These particles amount to 4 to 5 percent of the mass-energy of the universe [*see box on page 37*]. The Standard Model tells us that protons and neutrons are composed of elementary particles called quarks that are bound together by massless particles called gluons. Although the constituents are whirling around inside each proton, from outside we see a proton as a coherent object with an intrinsic mass, which is given by adding up the masses and energies of its constituents.

The Standard Model lets us calculate that nearly all the mass of protons and neutrons is from the kinetic energy of their constituent quarks and gluons (the remainder is from the quarks' rest mass). Thus, about 4 to 5 percent of the entire universe—almost all the familiar matter around us—comes from the energy of motion of quarks and gluons in protons and neutrons.

Overview/*Higgs Physics*

- Mass is a seemingly everyday property of matter, but it is actually mysterious to scientists in many ways. How do elementary particles acquire mass in the first place, and why do they have the specific masses that they do?
- The answers to those questions will help theorists complete and extend the Standard Model of particle physics, which describes the physics that governs the universe. The extended Standard Model may also help solve the puzzle of the invisible dark matter that accounts for about 25 percent of the cosmos.
- Theories say that elementary particles acquire mass by interacting with a quantum field that permeates all of reality. Experiments at particle accelerators may soon detect direct evidence of this so-called Higgs field.

The Higgs Mechanism

UNLIKE PROTONS and neutrons, truly elementary particles—such as quarks and electrons—are not made up of smaller pieces. The explanation of how they acquire their rest masses gets to the very heart of the problem of the origin of mass. As I noted above, the account proposed by contemporary theoretical physics is that fundamental particle masses arise from interactions with the Higgs field. But why is the Higgs field present throughout the universe? Why isn't its strength essentially zero on cosmic scales, like the electromagnetic field? What *is* the Higgs field?

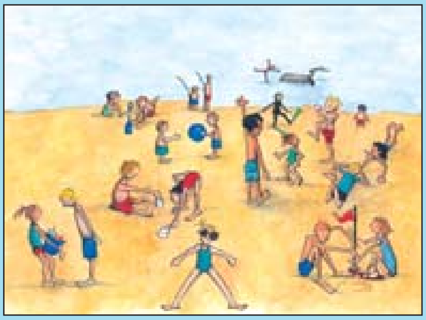
The Higgs field is a quantum field. That may sound mysterious, but the fact is that all elementary particles arise as quanta of a corresponding quantum field. The electromagnetic field is also a quantum field (its corresponding elementary particle is the photon). So in this respect, the Higgs field is no more enigmatic than electrons and light. The Higgs field does, however, differ from all other quantum fields in three crucial ways.

The first difference is somewhat technical. All fields have a property called spin, an intrinsic quantity of angular momentum that is carried by each of their particles. Particles such as electrons have spin ½ and most particles associated with a force, such as the photon, have spin 1. The Higgs boson (the particle of the Higgs field) has spin 0. Having 0 spin enables the Higgs field to appear in the Lagrangian in different ways than the other particles do, which in turn allows—and leads to—its other two distinguishing features.

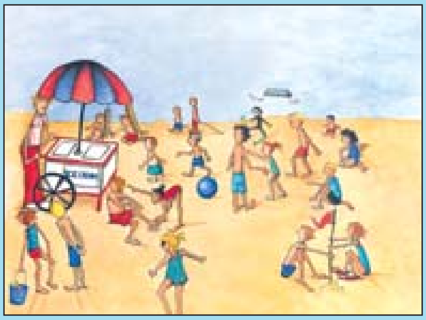
The second unique property of the Higgs field explains how and why it has nonzero strength throughout the universe. Any system, including a universe, will tumble into its lowest energy state, like a ball bouncing down to the bottom of a valley. For the familiar fields, such as the electromagnetic fields that give us

PROPERTIES OF THE ELUSIVE HIGGS

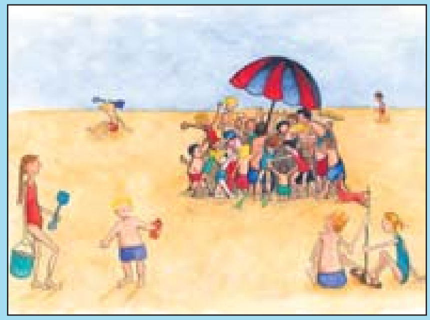
HOW THE HIGGS FIELD GENERATES MASS



"Empty" space, which is filled with the Higgs field, is like a beach full of children.



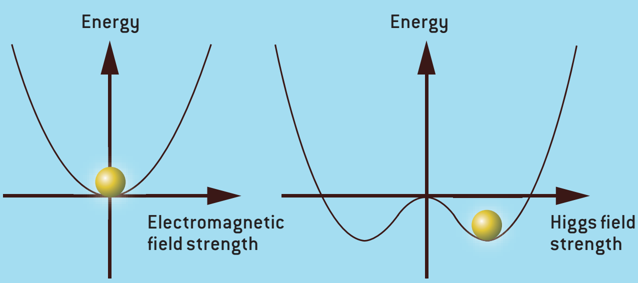
A particle crossing that region of space is like an ice cream vendor arriving . . .



. . . and interacting with kids who slow him down—as if he acquires "mass."

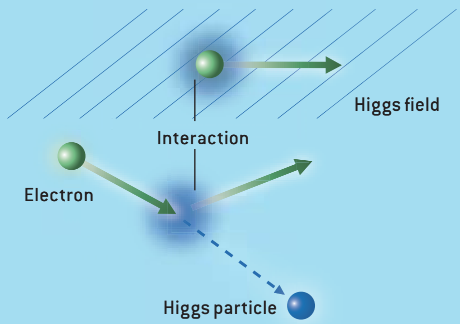
PERMEATING REALITY

A typical field, such as the electromagnetic field, has its lowest energy at zero field strength (*left*). The universe is akin to a ball that rolled around and came to rest at the bottom of the valley—that is, it has settled at a field strength of zero. The Higgs, in contrast, has its minimum energy at a nonzero field strength, and the "ball" comes to rest at a nonzero value (*right*). Thus, the universe, in its natural lowest energy state, is permeated by that nonzero value of the Higgs field.



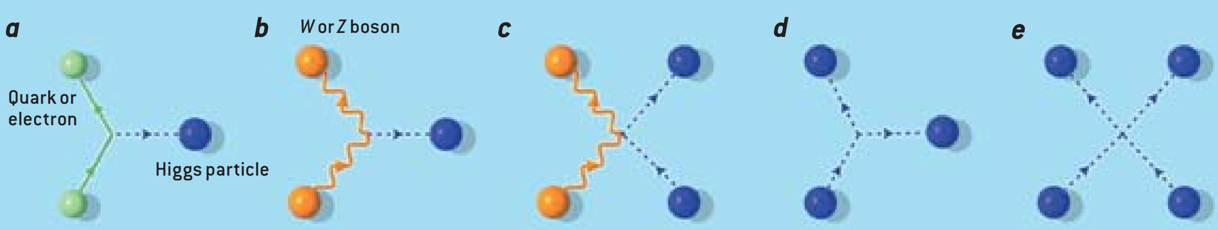
CAUSING TWO PHENOMENA

Two completely different phenomena—the acquisition of mass by a particle (*top*) and the production of a Higgs boson (*bottom*)—are caused by exactly the same interaction. This fact will be of great use in testing the Higgs theory by experiments.

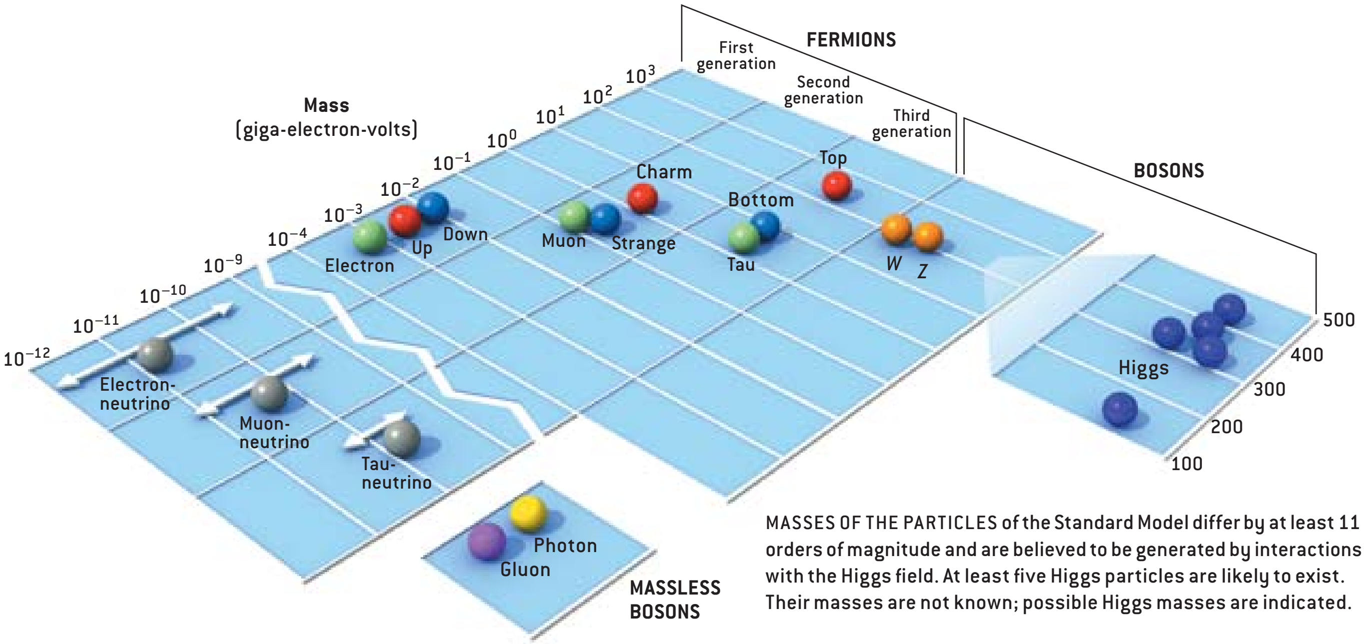


INTERACTING WITH OTHER PARTICLES

Force diagrams called Feynman diagrams represent how the Higgs particle interacts with other particles. Diagram (*a*) represents a particle such as a quark or an electron emitting (*shown*) or absorbing a Higgs particle. Diagram (*b*) shows the corresponding process for a *W* or *Z* boson. The *W* and *Z* can also interact simultaneously with two Higgs, as shown in (*c*), which also represents a *W* or *Z* scattering (roughly speaking, colliding with) a Higgs particle. The interactions represented by diagrams (*a*) through (*c*) are also responsible for generating particles' masses. The Higgs also interacts with itself, as represented by diagrams (*d*) and (*e*). More complicated processes can be built up by joining together copies of these elementary diagrams. Interactions depicted in (*d*) and (*e*) are responsible for the shape of the energy graph (*above left*).



BRYAN CHRISTIE DESIGN; CARTOONS BY EMILY HARRISON, BASED ON CONCEPT BY DAVID MILLER *University College London*



MASSES OF THE PARTICLES of the Standard Model differ by at least 11 orders of magnitude and are believed to be generated by interactions with the Higgs field. At least five Higgs particles are likely to exist. Their masses are not known; possible Higgs masses are indicated.

radio broadcasts, the lowest energy state is the one in which the fields have zero value (that is, the fields vanish)—if any nonzero field is introduced, the energy stored in the fields increases the net energy of the system. But for the Higgs field, the energy of the universe is lower if the field is not zero but instead has a constant nonzero value. In terms of the valley metaphor, for ordinary fields the valley floor is at the location of zero field; for the Higgs, the valley has a hillock at its center (at zero field) and the lowest point of the valley forms a circle around the hillock [*see box on preceding page*]. The universe, like a ball, comes to rest somewhere on this circular trench, which corresponds to a nonzero value of the field. That is, in its natural, lowest energy state, the universe is permeated throughout by a *nonzero* Higgs field.

The final distinguishing characteristic of the Higgs field is the form of its interactions with the other particles. Particles that interact with the Higgs field behave as if they have mass, proportional to the strength of the field times the strength of the interaction. The masses arise from the terms in the Lagrangian that have the particles interacting with the Higgs field.

Our understanding of all this is not yet complete, however, and we are not sure how many kinds of Higgs fields there are. Although the Standard Model requires only one Higgs field to generate all the elementary particle masses, physicists know that the Standard Model must be superseded by a more complete theory. Leading contenders are extensions of the Standard Model known as Supersymmetric Standard Models (SSMs). In these models, each Standard Model particle has a so-called superpartner (as yet undetected) with closely related properties [see "The Dawn of Physics beyond the Standard Model," by Gordon Kane, on page 4]. With the Supersymmetric Standard Model, at least two different kinds of Higgs fields are needed. Interactions with those two fields give mass to the Standard Model particles. They also give some (but not all) mass to the superpartners. The two Higgs fields give rise to five species of Higgs boson: three that are electrically neutral and two that are charged. The masses of particles called neutrinos, which are tiny compared with other particle masses, could arise rather indirectly from these interactions or from yet a third kind of Higgs field.

Theorists have several reasons for expecting the SSM picture of the Higgs interaction to be correct. First, without the Higgs mechanism, the *W* and *Z* bosons that mediate the weak force would be massless, just like the photon (which they are related to), and the weak interaction would be as strong as the electromagnetic one. Theory holds that the Higgs mechanism confers mass to the *W* and *Z* in a very special manner. Predictions of that approach (such as the ratio of the *W* and *Z* masses) have been confirmed experimentally.

Second, essentially all other aspects of the Standard Model have been well tested, and with such a detailed, interlocking theory it is difficult to change one part (such as the Higgs) without affecting the rest. For example, the analysis of precision measurements of *W* and *Z* boson properties led to the accurate prediction of the top quark mass before the top quark had been directly produced. Changing the Higgs mechanism would spoil that and other successful predictions.

Third, the Standard Model Higgs mechanism works very well for giving mass to *all* the Standard Model particles, *W* and *Z* bosons, as well as quarks and leptons; the alternative proposals usually do not. Next, unlike the other theories, the SSM provides a framework to unify our understanding of the forces of nature. Finally, the SSM can explain why the energy "valley" for the universe has the shape needed by the Higgs mechanism. In the basic Standard Model the shape of the valley has to be put in as a postulate, but in the SSM that shape can be derived mathematically.

Testing the Theory

NATURALLY, PHYSICISTS want to carry out direct tests of the idea that mass arises from the interactions with the different Higgs fields. We can test three key features. First, we can look for the signature particles called Higgs bosons. These quanta must exist, or else the explanation is not right. Physicists are currently looking for Higgs bosons at the Tevatron Collider at Fermi National Accelerator Laboratory in Batavia, Ill.

BRYAN CHRISTIE DESIGN

Second, once they are detected we can observe how Higgs bosons interact with other particles. The very same terms in the Lagrangian that determine the masses of the particles also fix the properties of such interactions. So we can conduct experiments to test quantitatively the presence of interaction terms of that type. The strength of the interaction and the amount of particle mass are uniquely connected.

Third, different sets of Higgs fields, as occur in the Standard Model or in the various SSMs, imply different sets of Higgs bosons with various properties, so tests can distinguish these alternatives, too. All that we need to carry out the tests are appropriate particle colliders—ones that have sufficient energy to produce the different Higgs bosons, sufficient intensity to make enough of them and very good detectors to analyze what is produced.

A practical problem with performing such tests is that we do not yet understand the theories well enough to calculate what masses the Higgs bosons themselves should have, which makes searching for them more difficult because one must examine a range of masses. A combination of theoretical reasoning and data from experiments guides us about roughly what masses to expect.

The Large Electron-Positron Collider (LEP) at CERN, the European laboratory for particle physics near Geneva, operated over a mass range that had a significant chance of including a Higgs boson. It did not find one—although there was tantalizing evidence for one just at the limits of the collider's energy and intensity—before it was shut down in 2000 to make room for constructing a newer facility, CERN's Large Hadron Collider (LHC). The Higgs must therefore be heavier than about 120 proton masses. Nevertheless, LEP did produce indirect evidence that a Higgs boson exists: experimenters at LEP made a number of precise measurements, which can be combined with similar measurements from the Tevatron and the collider at the Stanford Linear Accelerator Center. The entire set of data agrees well with theory only

A Cosmic Stocktaking

The theory of the Higgs field explains how elementary particles, the smallest building blocks of the universe, acquire their mass. But the Higgs mechanism is not the only source of mass-energy in the universe ("mass-energy" refers to both mass and energy, which are related by Einstein's $E = mc^2$).

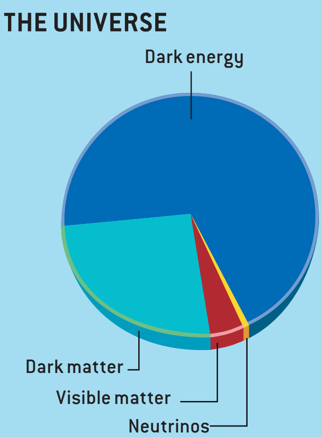
About 70 percent of the mass-energy of the universe is in the form of so-called dark energy, which is not directly associated with particles. The chief sign of the existence of dark energy is that the universe's expansion is accelerating. The precise nature of dark energy is one of the most profound open questions in physics [see "A Cosmic Conundrum," by Lawrence M. Krauss and Michael S. Turner, on page 66].

The remaining 30 percent of the universe's mass-energy comes from matter, particles with mass. The most familiar kinds of matter are protons, neutrons and electrons, which make up stars, planets, people and all that we see. These particles provide about one sixth of the matter of the universe, or 4 to 5 percent of the entire universe. As is explained in the main text, most of this mass arises from the energy of motion of quarks and gluons whirling around inside protons and neutrons.

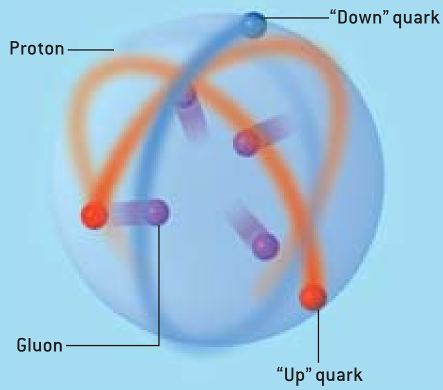
A smaller contribution to the universe's matter comes from particles called neutrinos, which come in three varieties. Neutrinos have mass but surprisingly little. The absolute masses of neutrinos are not yet measured, but the existing data put an upper limit on them—less than half a percent of the universe.

Almost all the rest of the matter—around 25 percent of the universe's total mass-energy—is matter we do not see, called dark matter. We deduce its existence from its gravitational effects on what we do see. We do not yet know what this dark matter actually is, but there are good candidates, and experiments are under way to test different ideas [see "The Search for Dark Matter," by David B. Cline; SCIENTIFIC AMERICAN, March 2003]. The dark matter should be composed of massive particles because it forms galaxy-size clumps under the effects of the gravitational force. A variety of arguments have let us conclude that the dark matter cannot be composed of any of the normal Standard Model particles.

The leading candidate particle for dark matter is the lightest superpartner (LSP), which is discussed in greater detail in the main text. The lightest superpartner occurs in extensions of the Standard Model called Supersymmetric Standard Models. The mass of the LSP is thought to be about 100 proton masses. That the LSP was a good candidate for the dark matter was recognized by theorists before cosmologists knew that a new form of fundamental matter was needed to explain dark matter. *—G.K.*



MASS-ENERGY of the universe mainly comes in four broad types: mysterious dark energy that causes the universe's expansion to accelerate; invisible dark matter that we can detect by its gravitational effects; visible matter; and neutrinos.



MOST VISIBLE MASS is locked up in protons and neutrons. Each of these consists of quarks and gluons flying around. Almost all of the proton's or neutron's mass is from the energy of motion of the quarks and gluons.

BRYAN CHRISTIE DESIGN

if certain interactions of particles with the lightest Higgs boson are included and only if the lightest Higgs boson is not heavier than about 200 proton masses. That provides researchers with an upper limit for the mass of the Higgs boson, which helps to focus the search.

For the next few years, the only collider that could produce direct evidence for Higgs bosons will be the Tevatron. Its energy is sufficient to discover a Higgs boson in the range of masses implied by the indirect LEP evidence, *if* it can consistently achieve the beam intensity it was expected to have, which so far has not been possible. In 2007 the LHC, which is seven times more energetic and is designed to have far more intensity than the Tevatron, is scheduled to begin taking data. It will be a factory for Higgs bosons (meaning it will produce many of the particles a day). Assuming the LHC functions as planned, gathering the relevant data and learning how to interpret it should take one to two years. Carrying out the complete tests that show in detail that the interactions with Higgs fields are providing the mass will require a new electron-positron collider in addition to the LHC (which collides protons) and the Tevatron (which collides protons and antiprotons).

The LEP collider saw tantalizing evidence for the Higgs particle.

Dark Matter

WHAT IS DISCOVERED about Higgs bosons will not only test whether the Higgs mechanism is indeed providing mass, it will also point the way to how the Standard Model can be extended to solve problems such as the origin of dark matter.

With regard to dark matter, a key particle of the SSM is the lightest superpartner (LSP). Among the superpartners of the known Standard Model particles predicted by the SSM, the LSP is the one with the lowest mass. Most superpartners decay promptly to lower-mass superpartners, a chain of decays that ends with the LSP, which is stable because it has no lighter particle that it can decay into. (When a superpartner decays, at least one of the decay products should be another superpartner; it should not decay entirely into Standard Model particles.) Superpartner particles would have been created early in the big bang but then promptly decayed into LSPs. The LSP is the leading candidate particle for dark matter.

The Higgs bosons may also directly affect the amount of dark matter in the universe. We know that the amount of LSPs today should be less than the amount shortly after the big bang, because some would have collided and annihilated into quarks and leptons and photons, and the annihilation rate may be dominated by LSPs interacting with Higgs bosons.

As mentioned earlier, the two basic SSM Higgs fields give mass to the Standard Model particles and *some* mass to the superpartners, such as the LSP. The superpartners acquire more mass via additional interactions, which may be with still further Higgs fields or with fields similar to the Higgs. We have theoretical models of how these processes can happen, but until we have data on the superpartners themselves we will not know how they work in detail. Such data are expected from the LHC or perhaps even from the Tevatron.

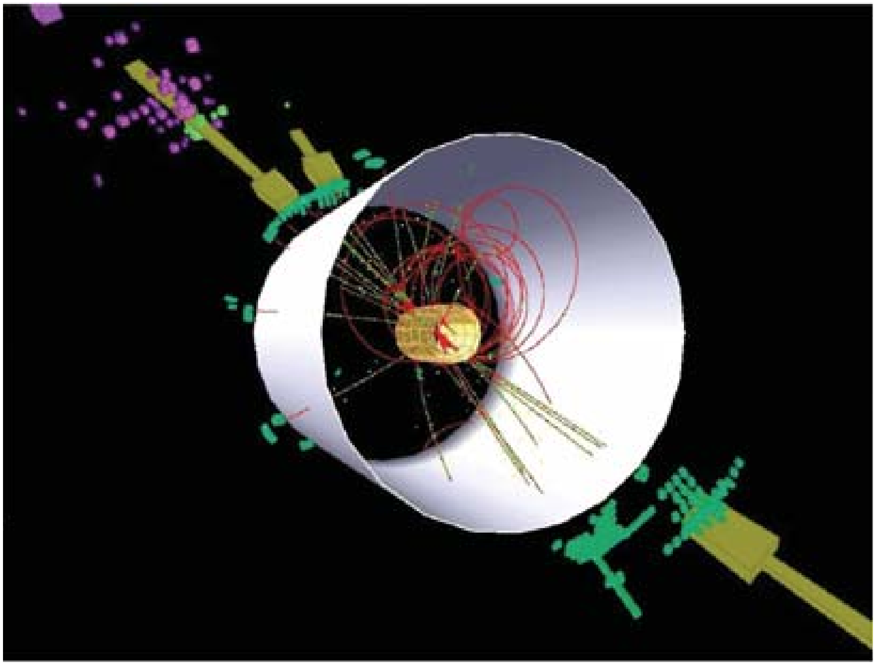
Neutrino masses may also arise from interactions with additional Higgs or Higgs-like fields, in a very interesting way. Neutrinos were originally assumed to be massless, but since 1979 theorists have predicted that they have small masses, and over the past decade or so several impressive experiments have confirmed the predictions [see "Solving the Solar Neutrino Problem," by Arthur B. McDonald, Joshua R. Klein and David L. Wark, on page 22]. The neutrino masses are less than a millionth the size of the next smallest mass, the electron mass. Because neutrinos are electrically neutral, the theoretical description of their masses is more subtle than for charged particles. Several processes contribute to the mass of each neutrino species, and for technical reasons the actual mass value emerges from solving an equation rather than just adding the terms.

Thus, we have understood the three ways that mass arises: The main form of mass we are familiar with—that of protons and neutrons and therefore of atoms—comes from the motion of quarks bound into protons and neutrons. The proton mass would be about what it is even without the Higgs field. The masses of the quarks themselves, however, and also the mass of the electron, are entirely caused by the Higgs field. Those masses would vanish without the Higgs. Last, but certainly not least, most of the amount of superpartner masses, and therefore the mass of the dark matter particle (if it is indeed the lightest superpartner), comes from additional interactions beyond the basic Higgs one.

Finally, we consider an issue known as the family problem. Over the past half a century physicists have shown that the world we see, from people to flowers to stars, is constructed from just six particles: three matter particles (up quarks, down quarks and electrons), two force quanta (photons and gluons), and Higgs bosons—a remarkable and surprisingly

THE AUTHOR

GORDON KANE, a particle theorist, is Victor Weisskopf Collegiate Professor of Physics at the University of Michigan at Ann Arbor. His work explores ways to test and extend the Standard Model of particle physics. In particular, he studies Higgs physics and the Standard Model's supersymmetric extension and cosmology, with a focus on relating theory and experiment. Recently he has emphasized integrating these topics with string theory and studying the implications for collider experiments.



A HIGGS PARTICLE might have been created when a high-energy positron and electron collided in the L3 detector of the Large Electron-Positron Collider at CERN. The lines represent particle tracks. The green and purple blobs and gold histograms depict amounts of energy deposited in layers of the detector by particles flying away from the reaction. Only by combining many such events can physicists conclude whether Higgs particles were present in some of the reactions or if all the data were produced by other reactions that happened to mimic the Higgs signal.

L3 COLLABORATION AND PRINCETON UNIVERSITY

simple description. Yet there are four more quarks, two more particles similar to the electron, and three neutrinos. All are very short-lived or barely interact with the other six particles. They can be classified into three families: up, down, electron-neutrino, electron; charm, strange, muon-neutrino, muon; and top, bottom, tau-neutrino, tau. The particles in each family have interactions identical to those of the particles in other families. They differ only in that those in the second family are heavier than those in the first, and those in the third family are heavier still. Because these masses arise from interactions with the Higgs field, the particles must have different interactions with the Higgs field.

Hence, the family problem has two parts: Why are there three families when it seems only one is needed to describe the world we see? Why do the families differ in mass and have the masses they do? Perhaps it is not obvious why physicists are astonished that nature contains three almost identical families even if one would do. It is because we want to fully understand the laws of nature and the basic particles and forces. We expect that every aspect of the basic laws is a necessary one. The goal is to have a theory in which all the particles and their mass ratios emerge inevitably, without making ad hoc assumptions about the values of the masses and without adjusting parameters. If having three families is essential, then it is a clue whose significance is currently not understood.

Tying It All Together

THE STANDARD MODEL and the SSM can accommodate the observed family structure, but they cannot explain it. This is a strong statement. It is not that the SSM has not *yet* explained the family structure but that it *cannot*. For me, the most exciting aspect of string theory is not only that it may provide us with a quantum theory of all the forces but also that it may tell us what the elementary particles are and why there are three families. String theory seems able to address the question of why the interactions with the Higgs field differ among the families. In string theory, repeated families can occur, and they are not identical. Their differences are described by properties that do not affect the strong, weak, electromagnetic or gravitational forces but that do affect the interactions with Higgs fields, which fits with our having three families with different masses. Although string theorists have not yet fully solved the problem of having three families, the theory seems to have the right structure to provide a solution. String theory allows many different family structures, and so far no one knows why nature picks the one we observe rather than some other [see "The String Theory Landscape," by Raphael Bousso and Joseph Polchinski, on page 40]. Data on the quark and lepton masses and on their superpartner masses may provide major clues to teach us about string theory.

One can now understand why it took so long historically to begin to understand mass. Without the Standard Model of particle physics and the development of quantum field theory to describe particles and their interactions, physicists could not even formulate the right questions. Whereas the origins and values of mass are not yet fully understood, it is likely that the framework needed to understand them is in place. Mass could not have been comprehended before theories such as the Standard Model and its supersymmetric extension and string theory existed. Whether they indeed provide the complete answer is not yet clear, but mass is now a routine research topic in particle physics. SA

MORE TO EXPLORE

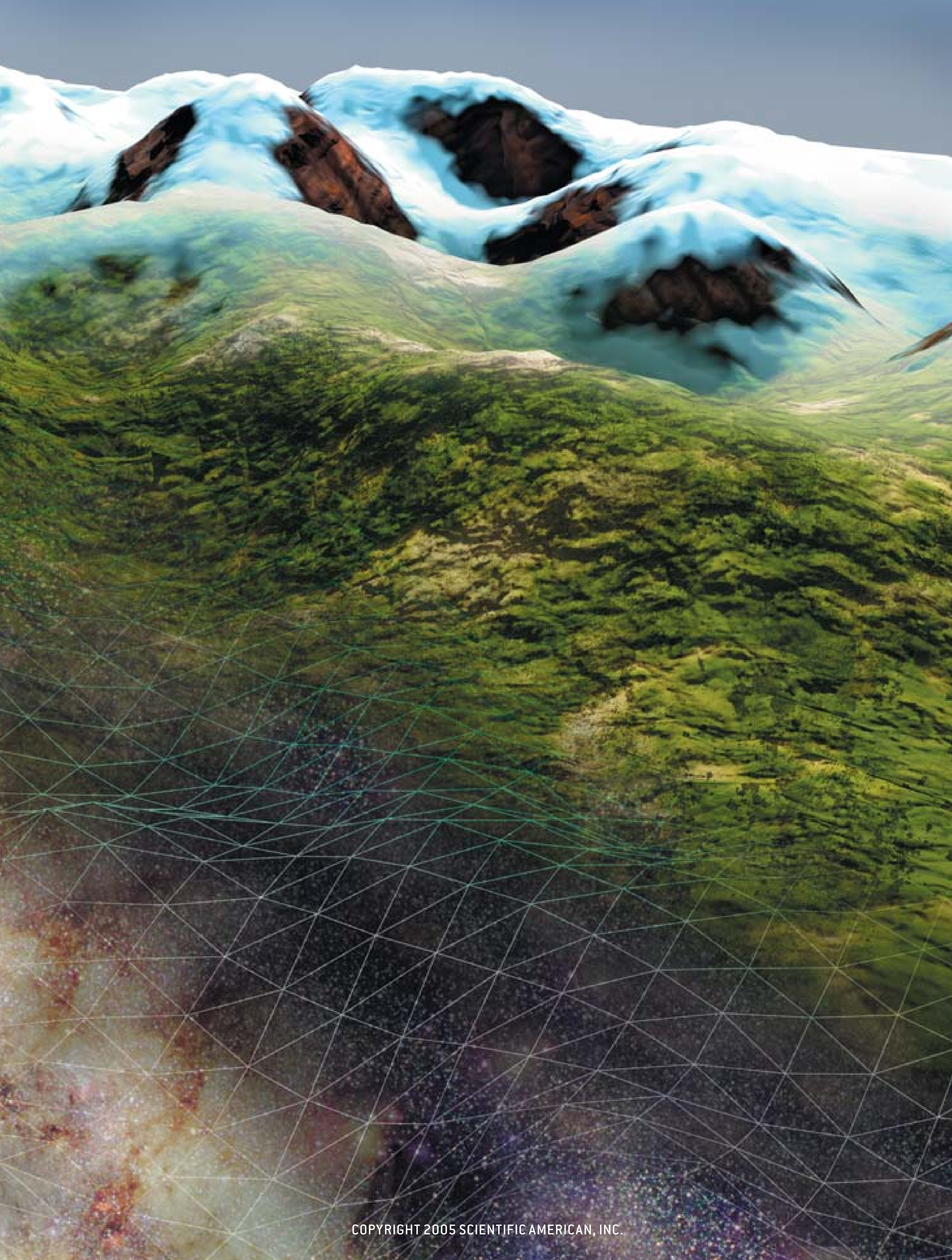
The Particle Garden. Gordon Kane. Perseus Publishing, 1996.

The Little Book of the Big Bang: A Cosmic Primer. Craig J. Hogan. Copernicus Books, 1998.

Mass without Mass II: The Medium Is the Mass-age. Frank Wilczek in *Physics Today,* Vol. 53, No. 1, pages 13–14; January 2000.

Supersymmetry: Unveiling the Ultimate Laws of Nature. Gordon Kane. Perseus Publishing, 2001.

An excellent collection of particle physics Web sites is listed at **particleadventure.org/particleadventure/other/othersites.html**



THE STRING THEORY LANDSCAPE

The theory of strings predicts that the universe might occupy one random "valley" out of a virtually infinite selection of valleys in a vast landscape of possibilities

By Raphael Bousso and Joseph Polchinski

DON FOLEY

THEORETICAL LANDSCAPE populated with an array of innumerable possible universes is predicted by string theory. The landscape has perhaps 10^{500} valleys, each one of which corresponds to a set of laws of physics that may operate in vast bubbles of space. Our visible universe would be one relatively small region within one such bubble.

OVERVIEW

- According to string theory, the laws of physics that we see operating in the world depend on how extra dimensions of space are curled up into a tiny bundle.
- A map of all possible configurations of the extra dimensions produces a "landscape" wherein each valley corresponds to a stable set of laws.
- The entire visible universe exists within a region of space that is associated with a valley of the landscape that happens to produce laws of physics suitable for the evolution of life.

According to Albert Einstein's theory of general relativity, gravity arises from the geometry of space and time, which combine to form spacetime. Any massive body leaves an imprint on the shape of spacetime, governed by an equation Einstein formulated in 1915. The earth's mass, for example, makes time pass slightly more rapidly for an apple near the top of a tree than for a physicist working in its shade. When the apple falls, it is actually responding to this warping of time. The curvature of spacetime keeps the earth in its orbit around the sun and drives distant galaxies ever farther apart. This surprising and beautiful idea has been confirmed by many precision experiments.

Given the success of replacing the gravitational force with the dynamics of space and time, why not seek a geometric explanation for the other forces of nature and even for the spectrum of elementary particles? Indeed, this quest occupied Einstein for much of his life. He was particularly attracted to work by German Theodor Kaluza and Swede Oskar Klein, which proposed that whereas gravity reflects the shape of the four familiar spacetime dimensions, electromagnetism arises from the geometry of an additional fifth dimension that is too small to see directly (at least so far). Einstein's search for a unified theory is often remembered as a failure. In fact, it was premature: physicists first had to understand the nuclear forces and the crucial role of quantum field theory in describing physics—an understanding that was only achieved in the 1970s.

The search for a unified theory is a central activity in theoretical physics today, and just as Einstein foresaw, geometric concepts play a key role. The Kaluza-Klein idea has been resurrected and extended as a feature of string theory, a promising framework for the unification of quantum mechanics, general relativity and particle physics. In both the Kaluza-Klein conjecture and string theory, the laws of physics that we see are controlled by the shape and size of additional microscopic dimensions. What determines this shape? Recent experimental and theoretical developments suggest a striking and controversial answer that greatly alters our picture of the universe.

> String theory's equations imply that six extra dimensions exist that are too small to have yet been detected.

THE AUTHORS

RAPHAEL BOUSSO and *JOSEPH POLCHINSKI*'s work together began at a workshop on string duality in Santa Barbara. It grew out of the synergy between Bousso's background in quantum gravity and inflationary cosmology and Polchinski's background in string theory. Bousso is assistant professor of physics at the University of California, Berkeley. His research includes a general formulation of the holographic principle, which relates spacetime geometry to its information content. Polchinski is professor at the Kavli Institute for Theoretical Physics at the University of California, Santa Barbara. His contributions to string theory include the seminal idea that branes constitute a significant feature of the theory.

Kaluza-Klein Theory and Strings

KALUZA AND KLEIN put forth their concept of a fifth dimension in the early part of the 20th century, when scientists knew of two forces—electromagnetism and gravity. Both fall off inversely proportional to the square of the distance from their source, so it was tempting to speculate that they were connected in some way. Kaluza and Klein noticed that Einstein's geometric theory of gravity might provide this connection if an additional spatial dimension existed, making spacetime five-dimensional.

This idea is not as wild as it seems. If the extra spatial dimension is curled up into a small enough circle, it will have eluded our best microscopes—that is, the most powerful particle accelerators [*see box on opposite page*]. Moreover, we already know from general relativity that space is flexible. The three dimensions that we see are expanding and were once much smaller, so it is not such a stretch to imagine that there is another dimension that remains small today.

Although we cannot detect it directly, a small extra dimension would have important indirect effects that could be observed. General relativity would then describe the geometry of a five-dimensional spacetime. We can split this geometry into three elements: the shape of the four large spacetime dimensions, the angle between the small dimension and the others, and the circumference of the small dimension. The large spacetime behaves according to ordinary four-dimensional general relativity. At every location within it, the angle and circumference have some value, just like two fields permeating spacetime and taking on certain values at each location. Amazingly, the angle field turns out to mimic an electromagnetic field living in the four-dimensional world. That is, the equations governing its behavior are identical to those of electromagnetism. The circumference determines the relative strengths of the electromagnetic and gravitational forces. Thus, from a theory of gravity alone in five dimensions, we obtain a theory of both gravity and electromagnetism in four dimensions.

The possibility of extra dimensions has also come to play a vital role in unifying general relativity and quantum mechanics. In string theory, a leading approach to that unification, particles are in actuality one-dimensional objects, small vibrating loops or strands. The typical size of a string is near the Planck length, or 10^{-33} centimeter (less than a billionth of a billionth of the size of an atomic nucleus). Consequently, a string looks like a point under anything less than Planckian magnification.

For the theory's equations to be mathematically consistent, a string has to vibrate in 10 spacetime dimensions, which implies that six extra dimensions exist that are too small to have yet been detected. Along with the strings, sheets known as "branes" (derived from "membranes") of various dimensions can be immersed in spacetime. In the original Kaluza-Klein idea, the quantum wave functions of ordinary particles would fill the extra dimension—in effect, the particles themselves would be smeared across the extra dimension. Strings, in contrast, can be confined to lie on a brane. String theory also contains fluxes, or forces that can be represented by field lines, much as forces are represented in classical (nonquantum) electromagnetism.

Altogether the string picture looks more complicated than Kaluza-Klein theory, but the underlying mathematical structure is actually more unified and complete. The central theme of Kaluza-Klein theory remains: the physical laws that we see depend on the geometry of hidden extra dimensions.

Too Many Solutions?

THE KEY QUESTION IS, What determines this geometry? The answer from general relativity is that spacetime must satisfy Einstein's equations—in the words of John A. Wheeler of Princeton University, matter tells spacetime how to curve, and spacetime tells matter how to move. But the solution to the equations is not unique, so many different geometries are allowed. The case of five-dimensional Kaluza-Klein geometry provides a simple example of this nonuniqueness. The circumference of the small dimension can take any size at all: in the absence of matter, four large flat dimensions, plus a circle of any size, solve Einstein's equations. (Similar multiple solutions also exist when matter is present.)

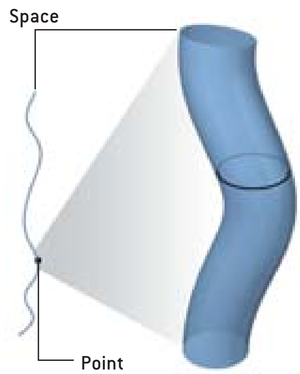
In string theory we have several extra dimensions, which results in many more adjustable parameters. One extra dimension can be wrapped up only in a circle. When more than one extra dimension exists, the bundle of extra dimensions can have many different shapes (technically, "topologies"), such as a sphere, a doughnut, two doughnuts joined together and so on. Each doughnut loop (a "handle") has a length and a circumference, resulting in a huge assortment of possible geometries for the small dimensions. In addition to the handles, further parameters correspond to the locations of branes and the different amounts of flux wound around each loop [*see box on page 45*].

Yet the vast collection of solutions are not all equal: each configuration has a potential energy, contributed by fluxes, branes and the curvature itself of the curled-up dimensions. This energy is called the vacuum energy, because it is the energy of the spacetime when the large four dimensions are completely devoid of matter or fields. The geometry of the small dimensions will try to adjust to minimize this energy, just as a ball

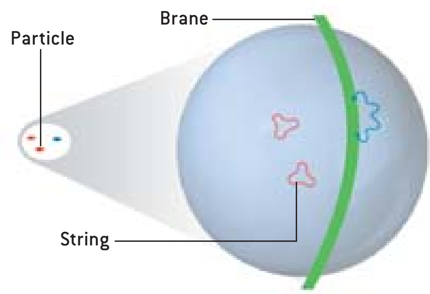
EXTRA DIMENSIONS

Strings and Tubes

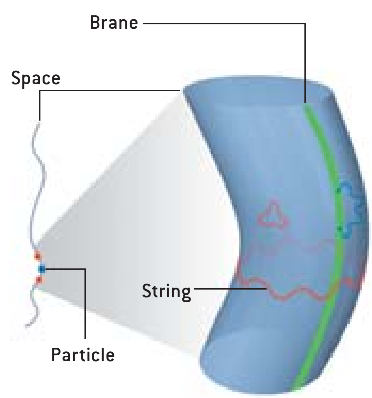
Extra spatial dimensions beyond the three we perceive are postulated by Kaluza-Klein theory and string theory. To imagine those dimensions, which are tiny, consider a space that consists of a long, very thin tube. Viewed from a distance, the tube looks like a one-dimensional line, but under high magnification, its cylindrical shape becomes apparent. Each zero-dimensional point on the line is revealed to be a one-dimensional circle of the tube. In the original Kaluza-Klein theory, every point in our familiar three-dimensional space is actually a tiny circle.



String theory predicts that what appear to be pointlike particles are actually tiny strings. In addition, it predicts the existence of membranelike objects called branes (*green*), which can come in a variety of dimensionalities. Strings that have end points (*blue*) always have their ends on a brane. Those that are closed loops (*red*) are free from that restriction.



String theory also incorporates Kaluza-Klein theory, which we again represent by showing a line of space that is actually a tube. This tube has a one-dimensional brane running through it and is populated by strings, some of which loop around the circumference of the tube one or more times. At lower magnification, the strings look like point particles, and the extra dimension, including its brane, is not apparent.



DON FOLEY

placed on a slope will start to roll downhill to a lower position.

To understand what consequences follow from this minimization, focus first on a single parameter: the overall size of the hidden space. We can plot a curve showing how the vacuum energy changes as this parameter varies. An example is shown in the top illustration on page 47. At very small sizes, the energy is high, so the curve starts out high at the left. Then, from left to right, it dips down into three valleys, each one lower than the previous one. Finally, at the right, after climbing out of the last valley, the curve trails off down a shallow slope to a constant value. The bottom of the leftmost valley is above zero energy; the middle one is at exactly zero; and the right-hand one is below zero.

How the hidden space behaves depends on the initial conditions—where the "ball" that represents it starts on the curve. If the configuration starts out to the right of the last peak, the ball will roll off to infinity, and the size of the hidden space will increase without bound (it will cease to be hidden). Otherwise it will settle down at the bottom of one of the troughs—the size of the hidden space adjusts to minimize the energy. These three local minima differ by virtue of whether the resulting vacuum energy is positive, negative or zero. In our universe the size of the hidden dimensions is not changing with time: if it were, we would see the constants of nature changing. Thus, we must be sitting at a minimum. In particular, we seem to be sitting at a minimum with a slightly positive vacuum energy.

Because there is more than one parameter, we should actually think of this vacuum energy curve as one slice through a complex, multidimensional mountain range, which Leonard Susskind of Stanford University has described as the landscape of string theory [*see middle illustration on page* 47]. The minima of this multidimensional landscape—the bottoms of depressions where a ball could come to rest—correspond to the stable configurations of spacetime (including branes and fluxes), which are called stable vacua.

A real landscape allows only two independent directions (north-south and east-west), and this is all we can draw. But the landscape of string theory is much more complicated, with hundreds of independent directions. The landscape dimensions should not be confused with the actual spatial dimensions of the world; each axis measures not some position in physical space but some aspect of the geometry, such as the size of a handle or the position of a brane.

The landscape of string theory is far from being fully mapped out. Calculating the energy of a vacuum state is a difficult problem and usually depends on finding suitable approximations. Researchers have made steady progress recently, most notably in 2003, when Shamit Kachru, Renata Kallosh and Andrei Linde, all at Stanford, and Sandip Trivedi of the Tata Institute of Fundamental Research in Mumbai, India, found strong evidence that the landscape does have minima where a universe can get stuck.

We cannot be sure how many stable vacua there are—that is, how many points where a ball could rest. But the number could very well be enormous. Some research suggests that there are solutions with up to about 500 handles, but not many more. We can wrap different numbers of flux lines around each handle, but not too many, because they would make the space unstable, like the right part of the curve in the figure. If we suppose that each handle can have from zero to nine flux lines (10 possible values), then there would be 10^{500} possible configurations. Even if each handle could have only zero or one flux unit, there are 2^{500}, or about 10^{150}, possibilities.

As well as affecting the vacuum energy, each of the many solutions will conjure up different phenomena in the four-dimensional macroscopic world by defining which kinds of particles and forces are present and what masses and interaction strengths they have. String theory may provide us with a unique set of fundamental laws, but the laws of physics that we see in the macroscopic world will depend on the geometry of the extra dimensions.

Each solution will conjure up different phenomena in the macroscopic world by defining which kinds of particles and forces are present.

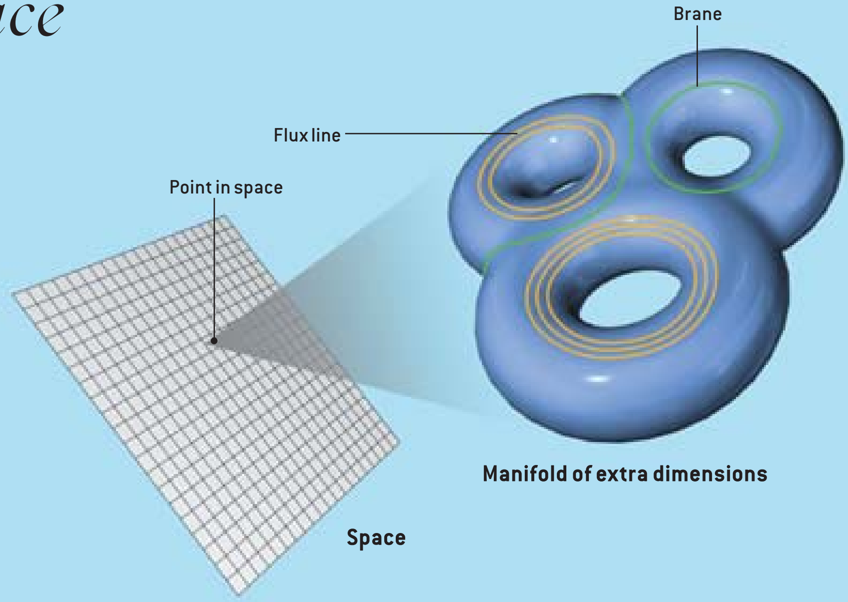
Many physicists hope that physics will ultimately explain why the universe has the specific laws that it does. But if that hope is to come true, many profound questions about the string theory landscape must be answered. Which stable vacuum describes the physical world we experience? Why has nature chosen this particular vacuum and not any other? Have all other solutions been demoted to mere mathematical possibilities, never to come true? String theory, if correct, would be the ultimate failure in democracy: richly populated with possible worlds but granting the privilege of reality to only one of its many citizens.

Instead of reducing the landscape to a single chosen vacuum, in 2000 we proposed a very different picture based on two important ideas. The first is that the world need not be stuck with one configuration of the small dimensions for good, because a rare quantum process allows the small dimensions to jump from one configuration to another. The second is that Einstein's general relativity theory, which is a part of string theory, implies that the universe can grow so rapidly that different configurations will coexist side by side in different subuniverses, each large enough to be unaware of the others.

VACUUM STATE

The Hidden Space

Any given solution to the equations of string theory represents a specific configuration of space and time. In particular, it specifies the arrangement of the small dimensions, along with their associated branes (*green*) and lines of force known as flux lines (*orange*). Our world has six extra dimensions, so every point of our familiar three-dimensional space hides an associated tiny six-dimensional space, or manifold—a six-dimensional analogue of the circle in the top illustration on page 43. The physics that is observed in the three large dimensions depends on the size and the structure of the manifold: how many doughnutlike "handles" it has, the length and circumference of each handle, the number and locations of its branes, and the number of flux lines wrapped around each doughnut.



Thus, the mystery of why our particular vacuum should be the only one to exist is eliminated. Moreover, we proposed that our idea resolves one of the greatest puzzles in nature.

A Trail through the Landscape

AS OUTLINED BEFORE, each stable vacuum is characterized by its numbers of handles, branes and flux quanta. But now we take into account that each of these elements can be created and destroyed, so that after periods of stability, the world can snap into a different configuration. In the landscape picture, the disappearance of a flux line or other change of topology is a quantum jump over a mountain ridge into a lower valley.

Consequently, as time goes on, different vacua can come into existence. Suppose that each of the 500 handles in our earlier example starts out with nine units of flux. One by one, the 4,500 flux units will decay in some sequence governed by the probabilistic predictions of quantum theory until all the energy stored in fluxes is used up. We start in a high mountain valley and leap randomly over the adjoining ridges, visiting 4,500 successively lower valleys. We are led through some varied scenery, but we pass by only a minuscule fraction of the 10^{500} possible solutions. It would seem that most vacua never get their 15 minutes of fame.

Yet we are overlooking a key part of the story: the effect of the vacuum energy on how the universe evolves. Ordinary objects such as stars and galaxies tend to slow down an expanding universe and can even cause it to recollapse. Positive vacuum energy, however, acts like antigravity: according to Einstein's equation, it causes the three dimensions that we see to grow more and more rapidly. This rapid expansion has an important and surprising effect when the hidden dimensions tunnel to a new configuration.

Remember that at every point in our three-dimensional space there sits a small six-dimensional space, which lives at some point on the landscape. When this small space jumps to a new configuration, the jump does not happen at the same instant everywhere. The tunneling first happens at one place in the three-dimensional universe, and then a bubble of the new low-energy configuration expands rapidly [*see box on page 48*]. If the three large dimensions were not expanding, this growing bubble would eventually overrun every point in the universe. But the old region is also expanding, and this expansion can easily be faster than that of the new bubble.

Everybody wins: both the old and the new regions increase in size. The new never completely obliterates the old. What makes this outcome possible is Einstein's dynamical geometry. General relativity is not a zero-sum game—the stretching of the spatial fabric allows new volume to be created for both the old and the new vacua. This trick will work as the new vacuum ages as well. When its turn comes to decay, it will not disappear altogether; rather it will sprout a growing bubble, occupied by a vacuum with yet lower energy.

Because the original configuration keeps growing, eventually it will decay again at another location, to another nearby minimum in the landscape. The process will continue infinitely many times, decays happening in all possible ways, with far separated regions losing fluxes from different handles. In this manner, every bubble will be host to many new solutions. Instead of a single sequence of flux decay, the universe thus experiences all possible sequences, resulting in a hierarchy of nested

DON FOLEY

bubbles, or subuniverses. The result is very similar to the eternal inflation scenario proposed by Alan Guth of the Massachusetts Institute of Technology, Alexander Vilenkin of Tufts University, and Linde [see "The Self-Reproducing Inflationary Universe," by Andrei Linde; SCIENTIFIC AMERICAN, November 1994].

Our scenario is analogous to an infinite number of explorers embarking on all possible paths through every minimum in the landscape. Each explorer represents some location in the universe far away from all the others. The path taken by that explorer is the sequence of vacua experienced at his location in the universe. As long as the explorers' starting point in the landscape is high up in the glaciers, practically all the minima will be visited. In fact, each one will be reached infinitely many times by every possible path downhill from the higher minima. The cascade comes to a halt only where it drops below sea level—into negative energy. The characteristic geometry associated with negative vacuum energy does not allow the game of perpetual expansion and bubble formation to continue. Instead a localized "big crunch" occurs, much like in the interior of a black hole.

In each bubble, an observer conducting experiments at low energies (like we do) will see a specific four-dimensional universe with its own characteristic laws of physics. Information from outside our bubble cannot reach us, because the intermediate space is expanding too rapidly for light to outrun it. We see only one set of laws, those corresponding to our local vacuum, simply because we do not see very far. In our scenario, what we think of as the big bang that began our universe was no more than the most recent jump to a new string configuration in this location, which has now spread across many billions of light-years. One day (probably too far off to worry about) this part of the world may experience another such transition.

Think of the landscape of string theory as a complex, **multidimensional mountain range**, with hundreds of independent directions.

The Vacuum Energy Crisis

THE PICTURE WE HAVE DESCRIBED explains how all the different stable vacua of the string landscape come into existence at various locations in the universe, thus forming innumerable subuniverses. This result may solve one of the most important and long-standing problems in theoretical physics—one related to the vacuum energy. To Einstein, what we now think of as vacuum energy was an arbitrary mathematical term—a "cosmological constant"—that could be added to his equation of general relativity to make it consistent with his conviction that the universe was static [see "A Cosmic Conundrum," by Lawrence M. Krauss and Michael S. Turner, on page 66]. To obtain a static universe, he proposed that this constant takes a positive value, but he abandoned the idea after observations proved the universe to be expanding.

With the advent of quantum field theory, empty space—the vacuum—became a busy place, full of virtual particles and fields popping in and out of existence, and each particle and field carries some positive or negative energy. According to the simplest computations based on this theory, these energies should add up to a tremendous density of about 10^{94} grams per cubic centimeter, or one Planck mass per cubic Planck length. We denote that value by Λ_P. This result has been called the most famous wrong prediction in physics because experiments have long shown that the vacuum energy is definitely no greater than $10^{-120}\Lambda_P$. Theoretical physics thus stumbled into a major crisis.

Understanding the origin of this great discrepancy has been one of the central goals of theoretical physics for more than three decades, but none of the numerous proposals for a resolution has gained wide acceptance. It was frequently assumed that the vacuum energy is exactly zero—a reasonable guess for a number that is known to have at least 120 zeros after the decimal point. So the apparent task was to explain how physics could produce the value zero. Many attempts centered on the idea that the vacuum energy can adjust itself to zero, but there were no convincing explanations of how this adjustment would take place or why the end result should be anywhere near zero.

In our 2000 paper, we combined the wealth of string theory solutions and their cosmological dynamics with a 1987 insight of Steven Weinberg of the University of Texas at Austin to provide both a how and a why.

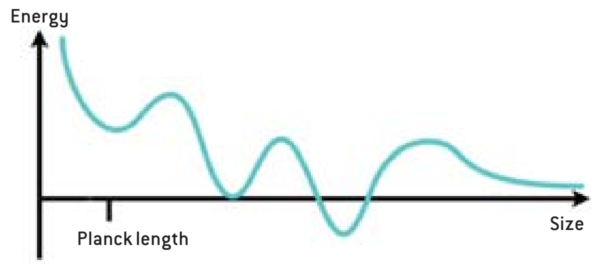
First consider the wealth of solutions. The vacuum energy is just the vertical elevation of a point in the landscape. This elevation ranges from around $+\Lambda_P$ at the glacial peaks to $-\Lambda_P$ at the bottom of the ocean. Supposing that there are 10^{500} minima, their elevations will lie randomly between these two values. If we plot all these elevations on the vertical axis, the average spacing between them will be $10^{-500}\Lambda_P$. Many, albeit a very small fraction of the total, will therefore have values between zero and $10^{-120}\Lambda_P$. This result explains how such small values come about.

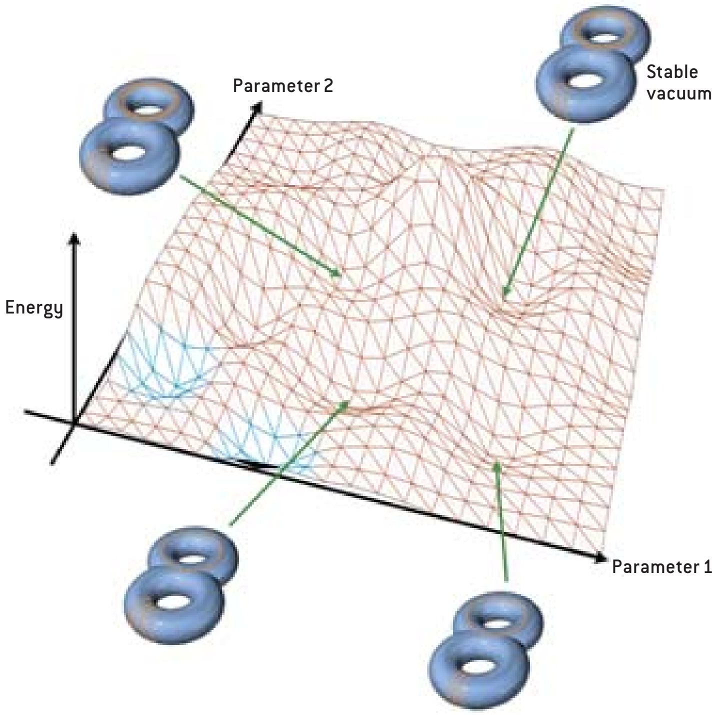
The general idea is not new. Andrei Sakharov, the Soviet physicist and dissident, suggested as early as 1984 that the complicated geometries of hidden dimensions might produce a spectrum for vacuum energy that includes values in the experimental window. Other researchers have made alternative proposals that do not seem to be realized in string theory.

STRING LANDSCAPE

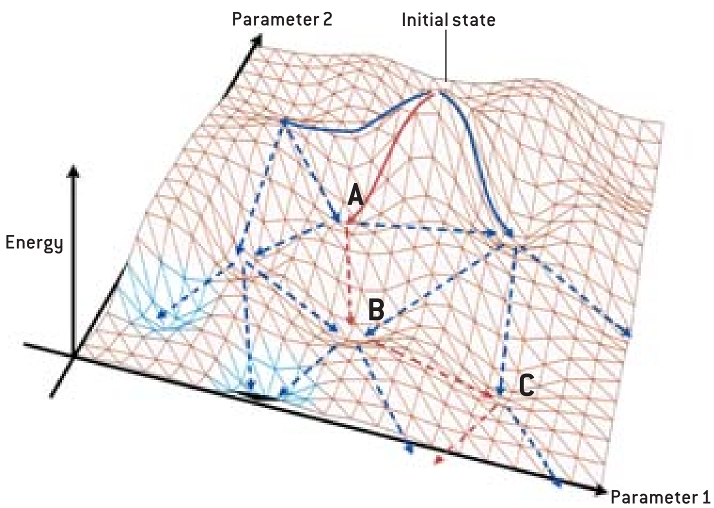
Topography of Energy

A landscape emerges when the energy of each possible string solution is plotted as a function of the parameters that define the six-dimensional manifold associated with that solution. If only one parameter is varied—say, the overall size of that manifold—the landscape forms a simple line graph. Here three particular sizes (all close to the Planck scale) have energies in the troughs, or minima, of the curve. The manifold will naturally tend to adjust its size to end up at one of the three minima, like a ball rolling around on the slope (it might also "roll off" to infinity at the right-hand end of the graph in this example).





The true string theory landscape reflects all parameters and thus would form a topography with a vast number of dimensions. We represent it by a landscape showing the variation of the energy contained in empty space when only two features change. The manifold of extra dimensions tends to end up at the bottom of a valley, which is a stable string solution, or a stable vacuum—that is, a manifold in a valley tends to stay in that state for a long while. Blue regions are below zero energy.



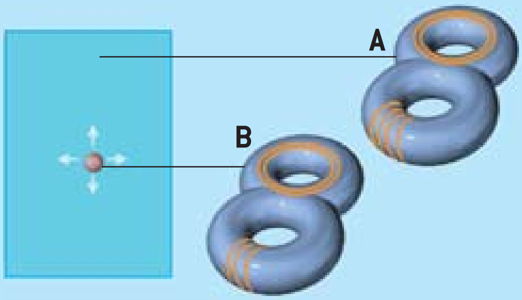
Quantum effects, however, allow a manifold to change state abruptly at some point—to tunnel through the intervening ridge to a nearby lower valley. The red arrows show how one region of the universe might evolve: starting out at a high mountaintop, rolling down into a nearby valley (*vacuum A*), eventually tunneling through to another, lower valley (*vacuum B*), and so on. Different regions of the universe will randomly follow different paths. The effect is like an infinite number of explorers traversing the landscape, passing through all possible valleys (*blue arrows*).

DON FOLEY

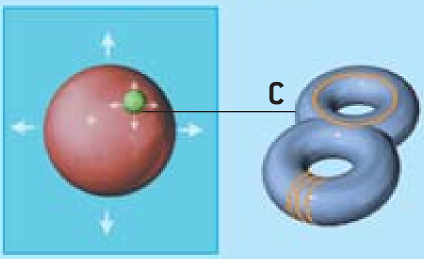
THE MULTIVERSE

Bubbles of Reality

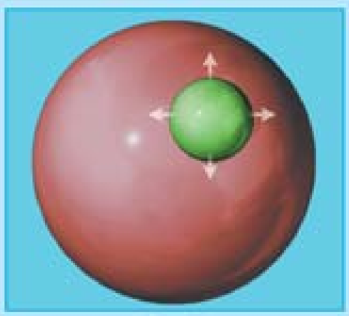
The possibility of decay from one stable vacuum to another suggests a radical new picture of our universe at the largest scales.



Tunneling from one stable vacuum to another would not occur everywhere in the universe at once. Instead it would occur at one random location, producing an expanding bubble of space (*arrows*) having the new vacuum. In this example, the blue region of space has vacuum A, whose manifold of small extra dimensions consists of a two-handled doughnut with groups of two and four flux lines wrapped around the handles. The red region, which has vacuum B, emerges when one of the four flux lines decays. Corresponding to their different manifolds, the two regions will have different kinds of particles and forces and thus different laws of physics.



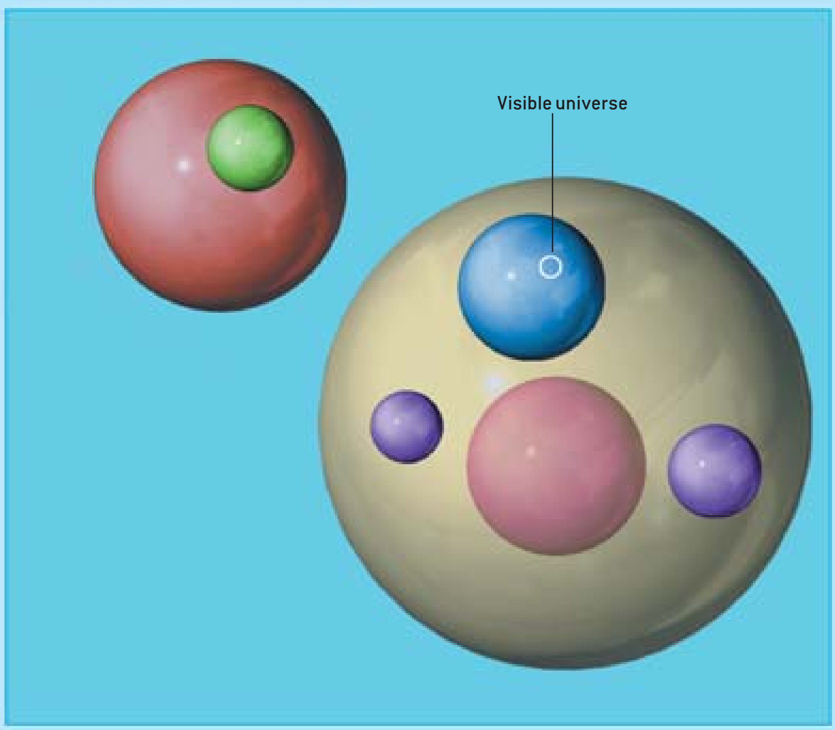
The red region grows rapidly, potentially becoming billions of light-years in diameter. Eventually another transition occurs within the red region, this time a decay of one of the two flux lines. This decay generates the green region, which has vacuum C and still another set of particles and forces.



The green region also grows rapidly, but it never catches up with the red region. Similarly, the red region never completely replaces the original blue vacuum.

Because the quantum tunneling is a random process, widely separated locations in the universe will decay through different sequences of vacua. In this way, the entire landscape is explored; every stable vacuum occurs in many different places in the universe.

The whole universe is therefore a foam of expanding bubbles within bubbles, each with its own laws of physics. Extremely few of the bubbles are suitable for the formation of complex structures such as galaxies and life. Our entire visible universe (more than 20 billion light-years in diameter) is a relatively small region within one of these bubbles.



DON FOLEY

We have explained how cosmology populates most of the minima, resulting in a complicated universe that contains bubbles with every imaginable value of the vacuum energy. In which of these bubbles will we find ourselves? Why should our vacuum energy be so close to zero? Here Weinberg's insight comes into play. Certainly an element of chance is involved. But many places are so inhospitable, it is no wonder we do not live there. This logic is familiar on smaller scale—you were not born in Antarctica, at the bottom of the Marianas Trench or on the airless wastes of the moon. Rather you find yourself in the tiny fraction of the solar system that is hospitable to life. Similarly, only a small fraction of the stable vacua are hospitable to life. Regions of the universe with large positive vacuum energy experience expansions so virulent that a supernova explosion would seem peaceful in comparison. Regions with large negative vacuum energy rapidly disappear in a cosmic crunch. If the vacuum energy in our bubble had been greater than $+10^{-118}\Lambda_P$ or less than $-10^{-120}\Lambda_P$, we could not have lived here, just as we do not find ourselves roasting on Venus or crushed on Jupiter. This type of reasoning is called anthropic.

Plenty of minima will be in the sweet spot, a hair's breadth above or below the water line. We live where we can, so we should not be surprised that the vacuum energy in our bubble is tiny. But neither should we expect it to be exactly zero! About 10^{380} vacua lie in the sweet spot, but at most only a tiny fraction of them will be exactly zero. If the vacua are distributed completely randomly, 90 percent of them will be somewhere in the range of 0.1 to $1.0 \times 10^{-118}\Lambda_P$. So if the landscape picture is right, a nonzero vacuum energy should be observed, most likely not much smaller than $10^{-118}\Lambda_P$.

In one of the most stunning developments in the history of experimental physics, recent observations of distant supernovae have shown that the visible universe's expansion is accelerating—the telltale sign of positive vacuum energy [see "Surveying Space-time with Supernovae," by Craig J. Hogan, Robert P. Kirshner and Nicholas B. Suntzeff; SCIENTIFIC AMERICAN, January 1999]. From the rate of acceleration, the value of the energy was determined to be about $10^{-120}\Lambda_P$, just small enough to have evaded detection in other experiments and large enough for the anthropic explanation to be plausible.

The landscape picture seems to resolve the vacuum energy crisis, but with some unsettling consequences. Einstein asked whether God had a choice in how the universe was made or whether its laws are completely fixed by some fundamental principle. As physicists, we might hope for the latter. The underlying laws of string theory, although they are still not completely known, appear to be completely fixed and inevitable: the mathematics does not allow any choices. But the laws that we see most directly are not the underlying laws. Rather our laws depend on the shape of the hidden dimensions, and for this the choices are many. The details of what we see in nature are not inevitable but are a consequence of the particular bubble that we find ourselves in.

Does the string landscape picture make other predictions, beyond the small but nonzero value of the vacuum energy? Answering this question will require a much greater understanding of the spectrum of vacua and is the subject of active research on several fronts. In particular, we have not yet located a specific stable vacuum that reproduces the known laws of physics in our four-dimensional spacetime. The string landscape is largely uncharted territory. Experiments could help. We might someday see the higher-dimensional physical laws directly, via strings, black holes or Kaluza-Klein particles using accelerators. Or we might even make direct astronomical observations of strings of cosmic size, which could have been produced in the big bang and then expanded along with the rest of the universe.

In each bubble, an observer will see
a specific four-dimensional universe
with its own characteristic laws of physics.

The picture that we have presented is far from certain. We still do not know the precise formulation of string theory—unlike general relativity, where we have a precise equation based on a well-understood underlying physical principle, the exact equations of string theory are unclear, and important physical concepts probably remain to be discovered. These may completely change or do away with the landscape of string vacua or with the cascade of bubbles that populate the landscape. On the experimental side, the existence of nonzero vacuum energy now seems an almost inevitable conclusion from observations, but cosmological data are notoriously fickle and surprises are still possible.

It is far too early to stop seeking competing explanations for the existence of vacuum energy and its very small size. But it would be equally foolish to dismiss the possibility that we have emerged in one of the gentler corners of a universe more varied than all the landscapes of planet Earth. SA

MORE TO EXPLORE

The Elegant Universe. Brian Greene. W. W. Norton, 1999.

The Cosmological Constant Problem. Thomas Banks in *Physics Today,* Vol. 57, No. 3, pages 46–51; March 2004.

A First Course in String Theory. Barton Zwiebach. Cambridge University Press, 2004.

The official string theory Web site is at **www.superstringtheory.com/**

The Future of String Theory

A Conversation with Brian Greene

String theory used to get everyone all tied up in knots. Even its practitioners fretted about how complicated it was, while other physicists mocked its lack of experimental predictions. The rest of the world was largely oblivious. Scientists could scarcely communicate just why string theory was so exciting—why it could fulfill Albert Einstein's dream of the ultimate unified theory, how it could give insight into such deep questions as why the universe exists at all. But in the mid-1990s the theory started to click together conceptually. Researchers came up with ways it might be tested experimentally. The outside world began to pay attention. Woody Allen satirized the theory in a *New Yorker* column in July 2003—probably the first time anyone has used Calabi-Yau spaces to tell a story about interoffice romance.

Few people can take more credit for demystifying string theory than Brian Greene, a Columbia University physics professor and a major contributor to the theory. His 1999 book *The Elegant Universe* reached number four on the *New York Times* best-seller list and was a finalist for the Pulitzer Prize. In 2003 Greene hosted a three-part Nova series on PBS based on that book and in 2004 published *The Fabric of the Cosmos,* a best-seller on the nature of space and time. Scientific American staff editor George Musser spoke with him over a plate of stringy spaghetti. Here is an abridged, edited version of that conversation.

RANDY HARRIS

SCIENTIFIC AMERICAN: Sometimes when our readers hear the words "string theory" or "cosmology," they throw up their hands and say, "I'll never understand it."

BRIAN GREENE: I've definitely encountered a certain amount of intimidation at the outset when it comes to ideas like string theory or cosmology. But what I have found is that the basic interest is so widespread and so deep in most people that I've spoken with, that there is a willingness to go a little bit further than you might with other subjects that are more easily taken in.

SA: I noticed that at several points in *The Elegant Universe,* you first gave a rough idea of the physics concepts and then the detailed version.

BG: I found that to be a useful way of going about it, especially in the harder parts. It gives the reader permission: If the rough idea is the level at which you want to take it in, that's great; feel free to skip this next stuff. If not, go for it. I like to say things more than one way. I just think that when it comes to abstract ideas, you need many roads into them. From the scientific point of view,

The difference between making a breakthrough and not can often be just a small element of perception.

if you stick with one road, I think you really compromise your ability to make breakthroughs. I think that's really what breakthroughs are about. Everybody's looking at a problem one way, and you come at it from the back. That different way of getting there somehow reveals things that the other approach didn't.

SA: What are some examples of that back-door approach?

BG: Well, probably the biggest ones are Ed Witten's breakthroughs. Ed [of the Institute for Advanced Study in Princeton, N.J.] just walked up the mountain and looked down and saw the connections that nobody else saw and in that way united the five string theories that previously were thought to be completely distinct. It was all out there; he just took a different perspective, and bang, it all came together. And that's genius.

To me that suggests what a fundamental discovery is. The universe in a sense guides us toward truths, because those truths are the things that govern what we see. If we're all being governed by what we see, we're all being steered in the same direction. Therefore, the difference between making a breakthrough and not can often be just a small element of perception, either true perception or mathematical perception, that puts things together in a different way.

SA: Do you think that these discoveries would have been made without the intervention of genius?

BG: Well, it's tough to say. In the case of string theory, I think so, because the pieces of the puzzle were really becoming clearer and clearer. It may have been five or

10 years later, but I suspect it would have happened. But with general relativity, I don't know. General relativity is such a leap, such a monumental rethinking of space, time and gravity, that it's not obvious to me how and when that would have happened without Einstein.

SA: Are there examples in string theory that you think are analogous to that huge leap?

BG: I think we're still waiting for a leap of that magnitude. String theory has been built up out of a lot of smaller ideas that a lot of people have contributed and been slowly stitching together into an ever more impressive theoretical edifice. But what idea sits at the top of that edifice, we still don't really know. When we do have that idea, I believe that it will be like a beacon shining down; it will illuminate the edifice, and it will also, I believe, give answers to critical questions that remain unresolved.

SA: In the case of relativity, you had the equivalence principle and general covariance in that beacon role.

NOVA

In the Standard Model, it's gauge invariance. In *The Elegant Universe,* you suggested the holographic principle could be that principle for string theory [see also "Information in the Holographic Universe," by Jacob D. Bekenstein, on page 74]. What's your thinking on that now?

BG: Well, the past few years have only seen the holographic principle rise to a yet greater prominence and believability. Back in the mid-1990s, shortly after the holographic ideas were suggested, the supporting ideas were rather abstract and vague, all based on features of black holes: Black hole entropy resides on the surface; therefore, maybe the degrees of freedom reside on the surface; therefore, maybe that's true of all regions that have a horizon; maybe it's true of cosmological horizons; maybe we're living within a cosmological region that has its true degrees of freedom far away. Wonderfully strange ideas, but the supporting evidence was meager.

But that changed with the work of Juan Maldacena [of the Institute for Advanced Study], in which he found an explicit example within string theory, where physics in the bulk—that is, in the arena that we consider to be real—would be exactly mirrored by physics taking place on a bounding surface. There'd be no difference in terms of the ability of either description to truly describe what's going on, yet in detail the descriptions would be vastly different. One would be in five dimensions, the other in four. So even the number of dimensions seems not to be something that you can count on, because there can be alternative descriptions that would accurately reflect the physics you're observing.

So to my mind, that makes the abstract ideas now concrete; it makes you believe the abstract ideas. And even if the details of string theory change, I think, as many others do—not everyone, though—that the holographic idea will persist and will guide us. Whether it truly is *the* idea, I don't know. I don't think so. But I think that it could well be one of the key steppingstones toward finding the essential ideas of the theory. It steps outside the details of the theory and just says, Here's a very general feature of a world that has quantum mechanics and gravity.

SA: Let's talk a bit about loop quantum gravity and some of the other approaches. You've always described string theory as the only game in town when it comes to quantum gravity. Do you still feel that way?

BG: I think it's the most fun game in town! But to be fair, the loop quantum gravity community has made tremendous progress. There are still many very basic questions that I don't feel have been answered, not to my satisfaction. But it's a viable approach, and it's great there are such large numbers of extremely talented people working on it. My hope—and it has been one that Lee Smolin has championed—is that ultimately

Relativity is a monumental rethinking of space and time. We're still waiting for another leap of that magnitude.

we're developing the same theory from different angles [see "Atoms of Space and Time," by Lee Smolin, on page 56]. It's far from impossible that we're going down our route to quantum gravity, they're going down their route to quantum gravity, and we're going to meet someplace. Because it turns out that many of their strengths are our weaknesses. Many of our strengths are their weaknesses.

One weakness of string theory is that it's so-called background dependent. We need to assume an existing spacetime within which the strings move. You'd hope, though, that a true quantum theory of gravity would have spacetime emerge from its fundamental equations. They [the loop quantum gravity researchers], however, do have a background-independent formulation in their approach, where spacetime does emerge more fundamentally from the theory itself. On the other hand, we are able to make very direct contact with Einstein's general relativity on large scales. We see it in our equations. They have some difficulty making contact with ordinary gravity. So naturally, you'd think maybe one could put together the strengths of each.

SA: Has that effort been made?

BG: Slowly. There are very few people who are really well versed in both theories. These are both two huge subjects, and you can spend your whole life, every moment of your working day, just in your own subject, and you still won't know everything that's going on. But many people are heading down that path and starting to think along those lines, and there have been some joint meetings.

SA: If you have this background dependence, what hope is there to really understand, in a deep sense, what space and time are?

BG: Well, you can chip away at the problem. For instance, even with background dependence, we've learned things like mirror symmetry—there can be two spacetimes, one physics. We've learned topology change—that space can evolve in ways that we wouldn't have thought possible before. We've learned that the microworld might be governed by noncommutative geometry, where the coordinates, unlike real numbers, depend on the order in which you multiply them. So you can get hints. You can get isolated glimpses of what's truly going on down there. But I think without the background-independent formalism, it's going to be hard to put the pieces together on their own.

SA: The mirror symmetry is incredibly profound, because it divorces spacetime geometry from physics. The connection between the two was always the Einsteinian program.

BG: That's right. Now, it doesn't divorce them completely. It simply says that you're missing half of the story. Geometry is tightly tied to physics, but it's a two-to-one map. It's not physics and geometry. It's physics and geometry-geometry, and which geometry you want to pick is up to you. Sometimes using one geometry gives you more insight than the other. Again, different ways of looking at one and the same physical system: two different geometries and one physics. And people have found there are mathematical questions about certain physical and geometrical systems that

The theory seems to be able to give rise to many different universes, of which ours seems to be only one.

people couldn't answer using the one geometry. Bring in the mirror geometry that had previously gone unrealized, and, all of a sudden, profoundly difficult questions, when translated, were mind-bogglingly simple.

SA: Can you describe noncommutative geometry?

BG: Since the time of Descartes, we've found it very powerful to label points by their coordinates, either on Earth by their latitude and longitude or in three-space by the three Cartesian coordinates, *x*, *y* and *z*, that you learn in high school. And we've always imagined that those numbers are like ordinary numbers, which have the property that, when you multiply them together—which is often an operation you need to do in physics—the answer doesn't depend on the order of operation: 3 times 5 is 5 times 3. What we seem to be finding is that when you coordinatize space on very small scales, the numbers involved are not like 3's and 5's, which don't depend on the order in which they're multiplied. There's a new class of numbers that *do* depend on the order of multiplication.

They're actually not that new, because for a long time we have known of an entity called the matrix. Sure as shooting, matrix multiplication depends on the order of multiplication. *A* times *B* does not equal *B* times *A* if *A* and *B* are matrices. String theory seems to indicate that points described by single numbers are replaced by geometrical objects described by matrices. On big scales, it turns out that these matrices become more and more diagonal, and diagonal matrices do have the property that they commute when you multiply. It doesn't matter how you multiply *A* times *B* if they're diagonal matrices. But then if you venture into the microworld, the off-diagonal entries in the matrices get bigger and bigger and bigger until way down in the depths, they are playing a significant part.

NOVA

Noncommutative geometry is a whole new field of geometry that some people have been developing for years without necessarily an application of physics in mind. French mathematician Alain Connes has this big thick book called *Noncommutative Geometry*. Euclid and Gauss and Riemann and all those wonderful geometers were working in the context of commutative geometry, and now Connes and others are taking off and developing the newer structure of noncommutative geometry.

SA: It is baffling to me—maybe it *should* be baffling—that you would have to label points with a matrix or some nonpure number. What does that mean?

BG: The way to think about it is: There is no notion of a point. A point is an approximation. If there is a point, you should label it by a number. But the claim is that, on sufficiently small scales, that language of points becomes such a poor approximation that it just isn't relevant. When we talk about points in geometry, we really talk about how something can move through points. It's the motion of objects that ultimately is what's relevant. Their motion, it turns out, can be more complicated than just sliding back and forth. All those motions are captured by a matrix. So rather than labeling an object by what point it's passing through, you need to label its motion by this matrix of degrees of freedom.


NOVA

IF YOU WERE A STRING, spacetime might look something like this: six extra dimensions curled into a so-called Calabi-Yau shape.

SA: What is your current thinking on anthropic and multiverse-type ideas? You talked about it in *The Elegant Universe* in the context of whether there is some limit to the explanatory power of string theory.

BG: I and many others have never been too happy with any of these anthropic ideas, largely because it seems to me that at any point in the history of science, you can say, "Okay, we're done, we can't go any further, and the final answer to every currently unsolved question is: 'Things are the way they are because had they not been this way, we wouldn't have been here to ask the question.'" So it sort of feels like a cop-out. Maybe that's the wrong word. Not necessarily like a cop-out; it feels a little dangerous to me, because maybe you just needed five more years of hard work and you would have answered those unresolved questions, rather than just chalking them up to, "That's just how it is." So that's my concern: that one doesn't stop looking by virtue of having this fallback position.

But you know, it's definitely the case that the anthropic ideas have become more developed. They're now real proposals whereby you would have many universes, and those many universes could all have different properties, and it very well could be that we're simply in this one because the properties are right for us to be here, and we're not in those others because we couldn't survive there. It's less of just a mental exercise.

SA: String theory, and modern physics generally, seems to be approaching a single logical structure that *had* to be the way it is; the theory is the way it is because there's no other way it could be. On the one hand, that would argue against an anthropic direction. But on the other hand, there's a flexibility in the theory that leads you to an anthropic direction.

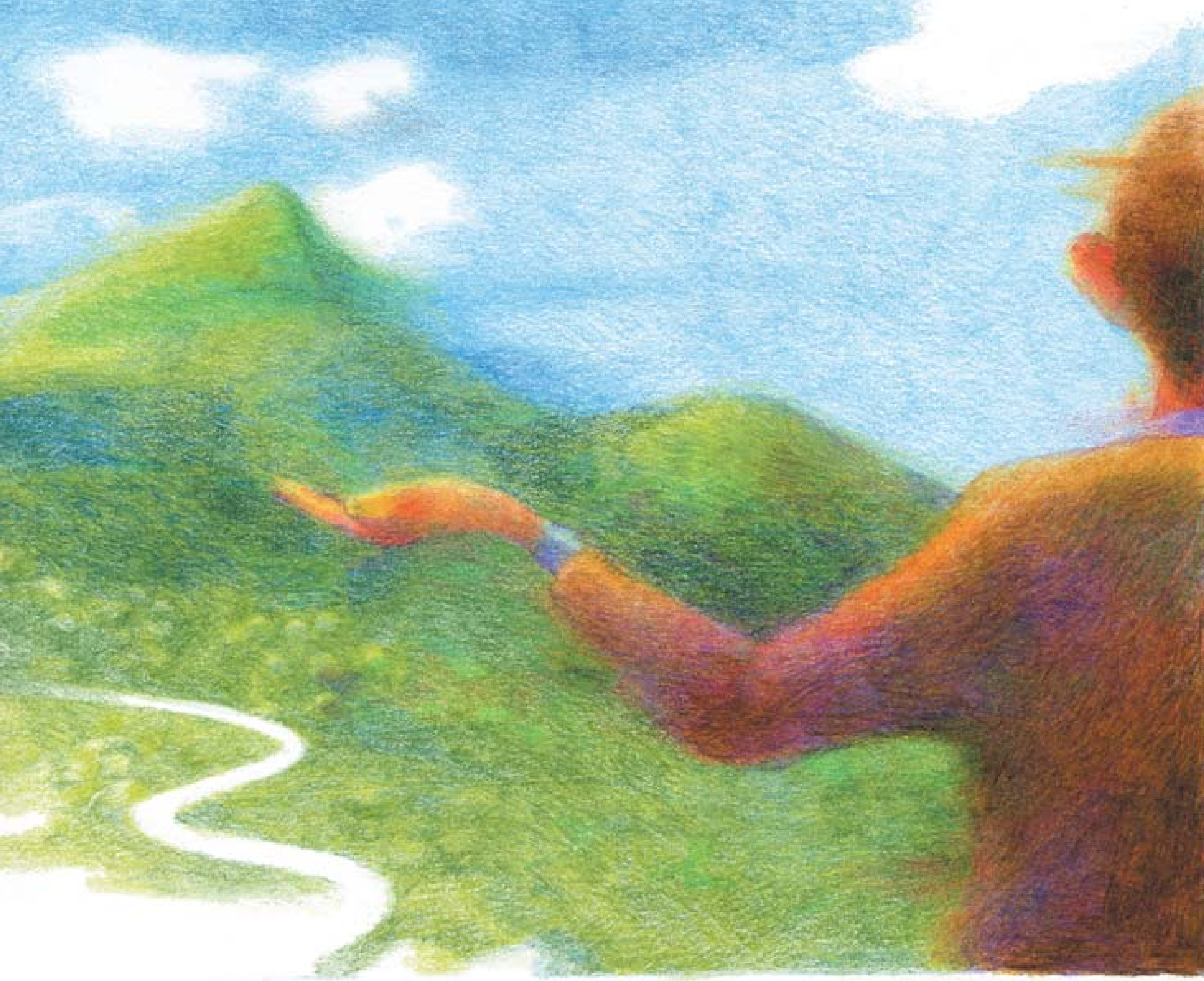
BG: The flexibility may or may not truly be there. That really could be an artifact of our lack of full understanding. But were I to go by what we understand today, the theory seems to be able to give rise to many different worlds, of which ours seems to be potentially one, but not even necessarily a very special one. So yes, there is a tension with the goal of absolute, rigid inflexibility.

SA: If you had other grad students waiting in the wings, what would you steer them to?

BG: Well, the big questions are, I think, the ones that we've discussed. Can we understand where space and time come from? Can we figure out the fundamental ideas of string theory? Can we show that this fundamental idea yields a unique theory with a unique solution, which happens to be the world as we know it? Is it possible to test these ideas through astronomical observations or through accelerator-based experiment?

Can we even take a step further back and understand why quantum mechanics had to be part and parcel of the world as we know it? How many of the things that we rely on at a very deep level in any physical theory that has a chance of being right—such as space, time, quantum mechanics—are truly essential, and how many of them can be relaxed and potentially still yield the world that appears close to ours?

Could physics have taken a different path that would have been experimentally as successful but completely different? I don't know. But I think it's a real interesting question to ask. How much of what we believe is truly fundamentally driven in a unique way by data and mathematical consistency, and how much of it could have gone one way or another, and we just happened to go down one path because that's what we happened to discover? Could beings on another planet have completely different sets of laws that somehow work just as well as ours? SA



We perceive space and time to be continuous, but if the amazing theory of loop quantum gravity is correct, they actually come in discrete pieces

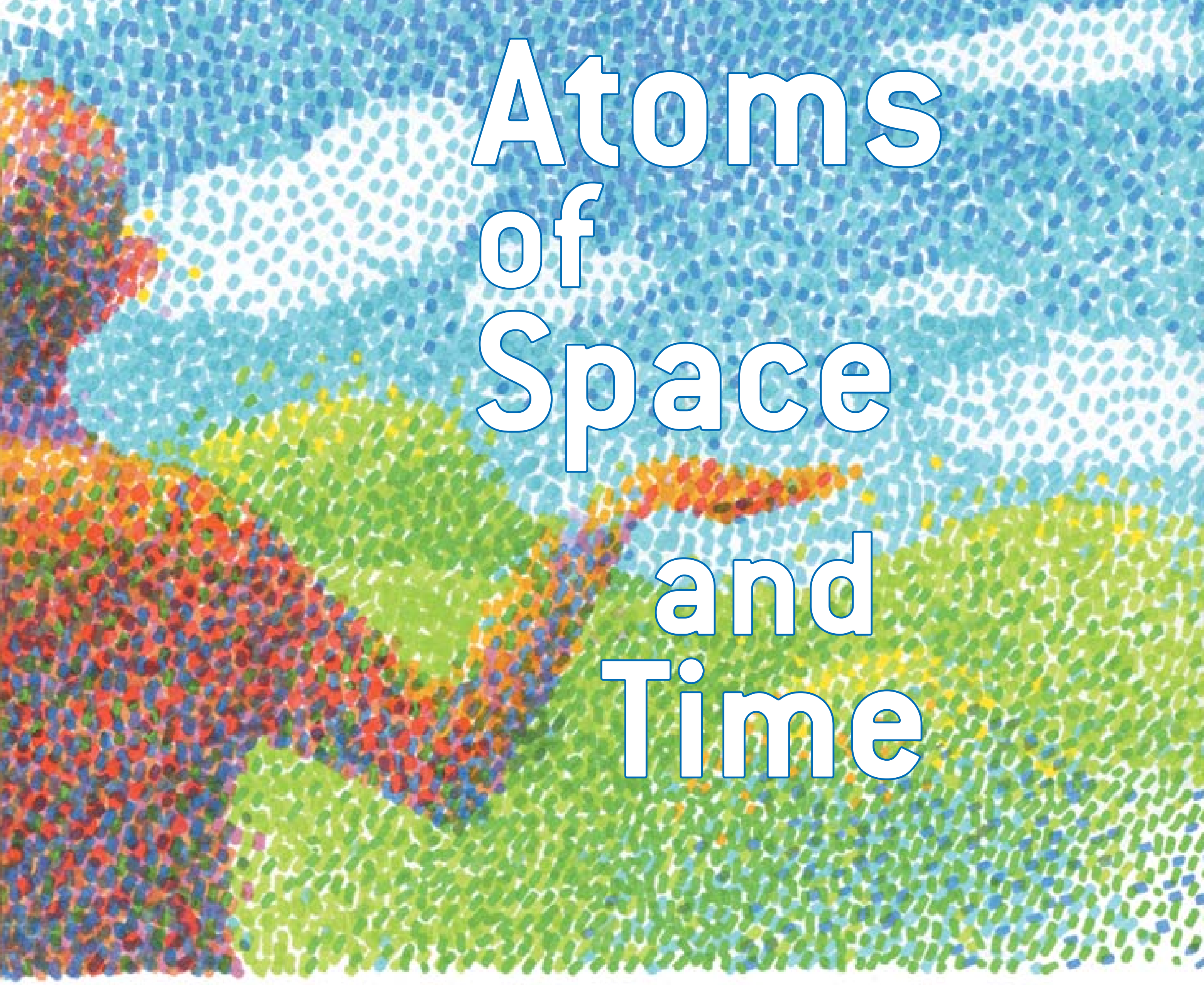
By Lee Smolin

A little more than 100 years ago most people—and most scientists—thought of matter as continuous. Although since ancient times some philosophers and scientists had speculated that if matter were broken up into small enough bits, it might turn out to be made up of very tiny atoms, few thought the existence of atoms could ever be proved. Today we have imaged individual atoms and have studied the particles that compose them. The granularity of matter is old news.

In recent decades, physicists and mathematicians have asked if space is also made of discrete pieces. Is it continuous, as we learn in school, or is it more like a piece of cloth, woven out of individual fibers? If we could probe to size scales that were small enough, would we see "atoms" of space, irreducible pieces of volume that cannot be broken into anything smaller? And what about time: Does nature change continuously, or does the world evolve in series of very tiny steps, acting more like a digital computer?

DUSAN PETRICIC

Atoms of Space and Time



The past two decades have seen great progress on these questions. A theory with the strange name of "loop quantum gravity" predicts that space and time are indeed made of discrete pieces. The picture revealed by calculations carried out within the framework of this theory is both simple and beautiful. The theory has deepened our understanding of puzzling phenomena having to do with black holes and the big bang. Best of all, it is testable; it makes predictions for experiments that can be done in the near future that will enable us to detect the atoms of space, if they are really there.

Quanta

MY COLLEAGUES AND I developed the theory of loop quantum gravity while struggling with a long-standing problem in physics: Is it possible to develop a quantum theory of gravity? To explain why this is an important question—and what it has to do with the granularity of space and time—I must first say a bit about quantum theory and the theory of gravity.

The theory of quantum mechanics was formulated in the first quarter of the 20th century, a development that was closely connected with the confirmation that matter is made of atoms. The equations of quantum mechanics require that certain quantities, such as the energy of an atom, can come only in specific, discrete units. Quantum theory successfully predicts the properties and behavior of atoms and the elementary particles and forces that compose them. No theory in the history of science has been more successful than quantum theory. It underlies our understanding of chemistry, atomic and subatomic physics, electronics and even biology.

In the same decades that quantum mechanics was being formulated, Albert Einstein constructed his general theory of relativity, which is a theory of gravity. In his theory, the gravitational force arises as a consequence of space and time (which together form "spacetime") being curved by the presence of matter. A loose analogy is that of a bowling ball placed on a rubber sheet along with a marble that is rolling around nearby. The balls could represent the sun and the earth, and the sheet is space. The bowling ball creates a deep indentation in the rubber sheet, and the slope of this indentation causes the marble to be deflected toward the larger ball, as if some force—gravity—were pulling it in that direction. Similarly, any piece of matter or concentration of energy distorts the geometry of spacetime, causing other particles and light rays to be deflected toward it, a phenomenon we call gravity.

Quantum theory and Einstein's theory of general relativity separately have each been fantastically well confirmed by experiment—but no experiment has explored the regime where both theories predict significant effects. The problem is that quantum effects are most prominent at small size scales, whereas general relativistic effects require large masses, so it takes extraordinary circumstances to combine both conditions.

Allied with this hole in the experimental data is a huge conceptual problem: Einstein's theory of general relativity is thoroughly classical, or nonquantum. For physics as a whole to be logically consistent, there has to be a theory that somehow unites quantum mechanics and general relativity. This long-sought-after theory is called quantum gravity. Because general relativity deals in the geometry of spacetime, a quantum theory of gravity will in addition be a quantum theory of spacetime.

SPACE IS WOVEN out of distinct threads.

Physicists have developed a considerable collection of mathematical procedures for turning a classical theory into a quantum one. Many theoretical physicists and mathematicians have worked on applying those standard techniques to general relativity. Early results were discouraging. Calculations carried out in the 1960s and 1970s seemed to show that quantum theory and general relativity could not be successfully combined. Consequently, something fundamentally new seemed to be required, such as additional postulates or principles not included in quantum theory and general relativity, or new particles or fields, or new entities of some kind. Perhaps with the right additions or a new mathematical structure, a quantumlike theory could be developed that would successfully approximate general relativity in the nonquantum regime. To avoid spoiling the successful predictions of quantum theory and general relativity, the exotica contained in the full theory would remain hidden from experiment except in the extraordinary circumstances where both quantum theory and general relativity are expected to have large effects. Many different approaches along these lines have been tried, with names such as twistor theory, noncommutative geometry and supergravity.

An approach that is very popular with physicists is string theory, which postulates that space has six or seven dimensions—all so far completely unobserved—in addition to the three that we are familiar with. String theory also predicts the existence of a great many new elementary particles and forces, for which there is so far no observable evidence. Some researchers believe that string theory is subsumed in a theory called M-theory [see "The Theory Formerly Known as Strings," by Michael J. Duff; SCIENTIFIC AMERICAN, February 1998], but unfortunately no precise definition of this conjectured theory has ever been given. Thus, many physicists and mathematicians are convinced that alternatives must be studied. Our loop quantum gravity theory is the best-developed alternative.

Overview/*Quantum Spacetime*

- To understand the structure of space on the very smallest size scale, we must turn to a quantum theory of gravity. Gravity is involved because Einstein's general theory of relativity reveals that gravity is caused by the warping of space and time.
- By carefully combining the fundamental principles of quantum mechanics and general relativity, physicists are led to the theory of "loop quantum gravity." In this theory, the allowed quantum states of space turn out to be related to diagrams of lines and nodes called spin networks. Quantum spacetime corresponds to similar diagrams called spin foams.
- Loop quantum gravity predicts that space comes in discrete lumps, the smallest of which is about a cubic Planck length, or 10^{-99} cubic centimeter. Time proceeds in discrete ticks of about a Planck time, or 10^{-43} second. The effects of this discrete structure might be seen in experiments in the near future.

A Big Loophole

IN THE MID-1980S a few of us—including Abhay Ashtekar, now at Pennsylvania State University, Ted Jacobson of the University of Maryland and Carlo Rovelli, now at the University of the Mediterranean in Marseille—decided to reexamine the question of whether quantum mechanics could be combined consistently with general relativity using the standard techniques. We knew that the negative results from the 1970s had an important loophole. Those calculations assumed that the geometry of space is continuous and smooth, no matter how minutely we examine it, just as people had expected matter to be before the discovery of atoms. Some of our

DUSAN PETRICIC

teachers and mentors had pointed out that if this assumption was wrong, the old calculations would not be reliable.

So we began searching for a way to do calculations without assuming that space is smooth and continuous. We insisted on not making any assumptions beyond the experimentally well tested principles of general relativity and quantum theory. In particular, we kept two key principles of general relativity at the heart of our calculations.

The first is known as background independence. This principle says that the geometry of spacetime is not fixed. Instead the geometry is an evolving, dynamical quantity. To find the geometry, one has to solve certain equations that include all the effects of matter and energy. Incidentally, string theory, as currently formulated, is not background independent; the equations describing the strings are set up in a predetermined classical (that is, nonquantum) spacetime.

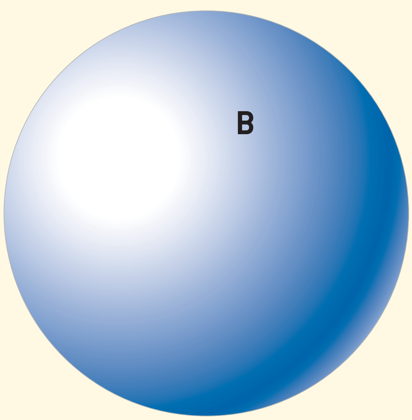
The second principle, known by the imposing name diffeomorphism invariance, is closely related to background independence. This principle implies that, unlike theories prior to general relativity, one is free to choose any set of coordinates to map spacetime and express the equations. A point in spacetime is defined only by what physically happens at it, not by its location according to some special set of coordinates (no coordinates are special). Diffeomorphism invariance is very powerful and is of fundamental importance in general relativity.

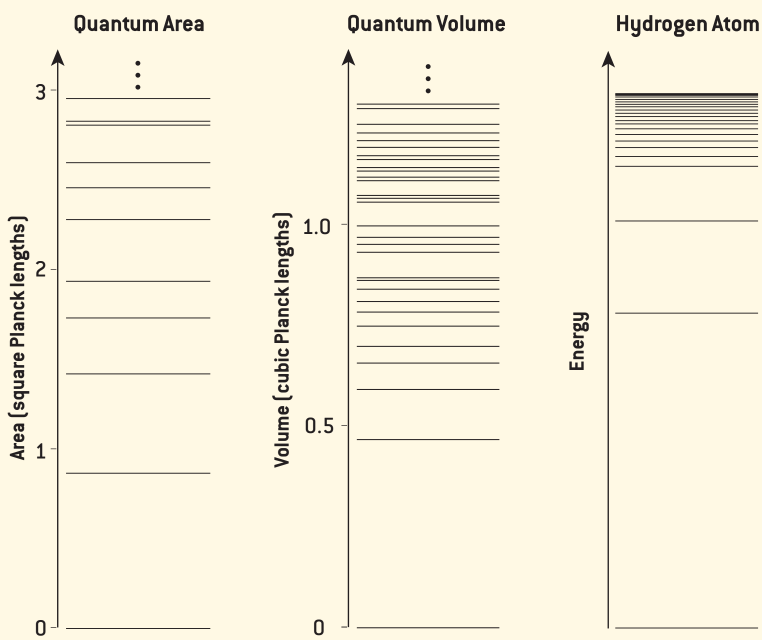
By carefully combining these two principles with the standard techniques of quantum mechanics, we developed a mathematical language that allowed us to do a computation to determine whether space is continuous or discrete. That calculation revealed, to our delight, that space is quantized. We had laid the foundations of our theory of loop quantum gravity. The term "loop," by the way, arises from how some computations in the theory involve small loops marked out in spacetime.

The calculations have been redone by a number of physicists and mathematicians using a range of methods. Over the years since, the study of loop quantum gravity has grown into a healthy field of research, with many contributors around the world; our combined efforts give us confidence in the picture of spacetime I will describe.

Ours is a quantum theory of the structure of spacetime at the smallest size scales, so to explain how the theory works we need to consider what it predicts for a small region or volume. In dealing with quantum physics, it is essential to specify precisely what physical quantities are to be measured. To do so, we consider a region somewhere that is marked out by a boundary, B [*see box below*]. The boundary may be defined by some matter, such as a cast-iron shell, or it may be defined by the geometry of spacetime itself, as in the event horizon of a black hole (a surface from within

QUANTUM STATES OF VOLUME AND AREA





A central prediction of the loop quantum gravity theory relates to volumes and areas. Consider a spherical shell that defines the boundary, B, of a region of space having some volume (*above*). According to classical (nonquantum) physics, the volume could be any positive real number. The loop quantum gravity theory says, however, that there is a nonzero absolute minimum volume (about one cubic Planck length, or 10^{-99} cubic centimeter), and it restricts the set of larger volumes to a discrete series of numbers. Similarly, there is a nonzero minimum area (about one square Planck length, or 10^{-66} square centimeter) and a discrete series of larger allowed areas. The discrete spectrum of allowed quantum areas (*left*) and volumes (*center*) is broadly similar to the discrete quantum energy levels of a hydrogen atom (*right*).

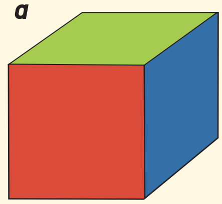
NADIA STRASSER; SOURCE: AREA AND VOLUME DATA FROM ROBERTO DE PIETRI AND CARLO ROVELLI (http://arxiv.org/abs/gr-qc/9602023)

VISUALIZING QUANTUM STATES OF VOLUME

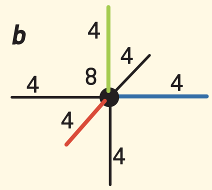
Diagrams called spin networks are used by physicists who study loop quantum gravity to represent quantum states of space at a minuscule scale. Some such diagrams correspond to polyhedra-shaped volumes. For example, a cube (*a*) consists of a volume enclosed within six square faces. The corresponding spin network (*b*) has a dot, or node, representing the volume and six lines that represent the six faces. The complete spin network has a number at the node to indicate the cube's volume and a number on each line to indicate the area of the corresponding face. Here the volume is eight cubic Planck lengths, and the faces are each four square Planck lengths. (The rules of loop quantum gravity restrict the allowed volumes and areas to specific quantities: only certain combinations of numbers are allowed on the lines and nodes.)

If a pyramid sat on the cube's top face (*c*), the line representing that face in the spin network would connect the cube's node to the pyramid's node (*d*). The lines corresponding to the four exposed faces of the pyramid and the five exposed faces of the cube would stick out from their respective nodes. (The numbers have been omitted for simplicity.)

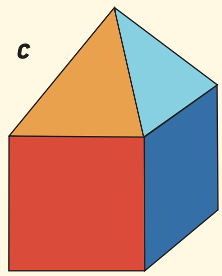
a



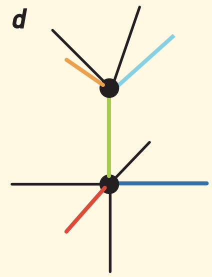
b



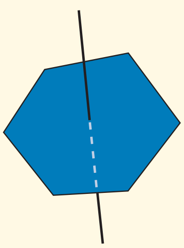
c



d

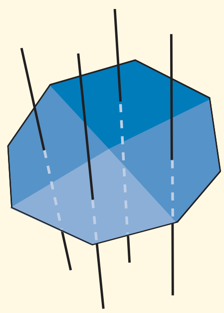


e



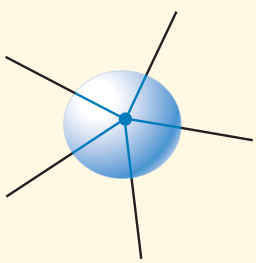
One quantum of area

f



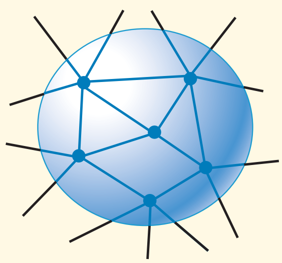
Larger area

g



One quantum of volume

h



Larger volume

In general, in a spin network, one quantum of area is represented by a single line (*e*), whereas an area composed of many quanta is represented by many lines (*f*). Similarly, a quantum of volume is represented by one node (*g*), whereas a larger volume takes many nodes (*h*). If we have a region of space defined by a spherical shell, the volume inside the shell is given by the sum of all the enclosed nodes and its surface area is given by the sum of all the lines that pierce it.

The spin networks are more fundamental than the polyhedra: any arrangement of polyhedra can be represented by a spin network in this fashion, but some valid spin networks represent combinations of volumes and areas that cannot be drawn as polyhedra. Such spin networks would occur when space is curved by a strong gravitational field or in the course of quantum fluctuations of the geometry of space at the Planck scale.

NADIA STRASSER

which even light cannot escape the black hole's gravitational clutches).

What happens if we measure the volume of the region? What are the possible outcomes allowed by both quantum theory and diffeomorphism invariance? If the geometry of space is continuous, the region could be of any size and the measurement result could be any positive real number; in particular, it could be as close as one wants to zero volume. But if the geometry is granular, then the measurement result can come from just a discrete set of numbers and it cannot be smaller than a certain minimum possible volume. The question is similar to asking how much energy electrons orbiting an atomic nucleus have. Classical mechanics predicts that an electron can possess any amount of energy, but quantum mechanics allows only specific energies (amounts in between those values do not occur). The difference is like that between the measure of something that flows continuously, like the 19th-century conception of water, and something that can be counted, like the atoms in that water.

The theory of loop quantum gravity predicts that space is like atoms: there is a discrete set of numbers that the volume-measuring experiment can return. Volume comes in distinct pieces. Another quantity we can measure is the area of the boundary B. Again, calculations using the theory return an unambiguous result: the area of the surface is discrete as well. In other words, space is not continuous. It comes only in specific quantum units of area and volume.

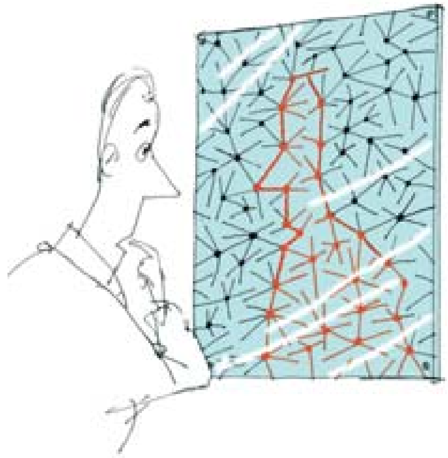
The possible values of volume and area are measured in units of a quantity called the Planck length. This length is related to the strength of gravity, the size of quanta and the speed of light. It measures the scale at which the geometry of space is no longer continuous. The Planck length is very small: 10^{-33} centimeter. The smallest possible nonzero area is about a square Planck length, or 10^{-66} cm^2. The smallest nonzero volume is approximately a cubic Planck length, 10^{-99} cm^3. Thus, the theory predicts that there are about 10^{99} atoms of volume in every cubic centimeter of space. The quantum of volume is so tiny that there are more such quanta in a cubic centimeter than there are cubic centimeters in the visible universe (10^{85}).

Spin Networks

WHAT ELSE DOES our theory tell us about spacetime? To start with, what do these quantum states of volume and area look like? Is space made up of a lot of little cubes or spheres? The answer is no—it's not that simple. Nevertheless, we can draw diagrams that represent the quantum states of volume and area. To those of us working in this field, these diagrams are beautiful because of their connection to an elegant branch of mathematics.

To see how these diagrams work, imagine that we have a lump of space shaped like a cube, as shown in the box on the opposite page. In our diagrams, we would depict this cube as a dot, which represents the volume, with six lines sticking out, each of which represents one of the cube's faces. We have to write a number next to the dot to specify the quantity of volume, and on each line we write a number to specify the area of the face that the line represents.

Next, suppose we put a pyramid on top of the cube. These two polyhedra, which share a common face, would be depicted as two dots (two volumes) connected by one of the lines (the face that joins the two volumes). The cube has five



MATTER EXISTS at the nodes of the spin network.

DUSAN PETRICIC

other faces (five lines sticking out), and the pyramid has four (four lines sticking out). It is clear how more complicated arrangements involving polyhedra other than cubes and pyramids could be depicted with these dot-and-line diagrams: each polyhedron of volume becomes a dot, or node, and each flat face of a polyhedron becomes a line, and the lines join the nodes in the way that the faces join the polyhedra together. Mathematicians call these line diagrams graphs.

Now in our theory, we throw away the drawings of polyhedra and just keep the graphs. The mathematics that describes the quantum states of volume and area gives us a set of rules for how the nodes and lines can be connected and what numbers can go where in a diagram. Every quantum state corresponds to one of these graphs, and every graph that obeys the rules corresponds to a quantum state. The graphs are a convenient shorthand for all the possible quantum states of space. (The mathematics and other details of the quantum states are too complicated to discuss here; the best we can do is show some of the related diagrams.)

The graphs are a better representation of the quantum states than the polyhedra are. In particular, some graphs connect in strange ways that cannot be converted into a tidy picture of polyhedra. For example, whenever space is curved, the polyhedra will not fit together properly in any drawing we could do, yet we can still easily draw a graph. Indeed, we can take a graph and from it calculate how much space is distorted. Because the distortion of space is what produces gravity, this is how the diagrams form a quantum theory of gravity.

For simplicity, we often draw the graphs in two dimensions, but it is better to imagine them filling three-dimensional space, because that is what they represent. Yet there is a conceptual trap here: the lines and nodes of a graph do not live at specific locations in space. Each graph is defined only by the way its pieces connect together and how they relate to well-defined boundaries such as boundary B. The continuous, three-dimensional space that you are imagining the graphs occupy *does not exist* as a separate entity. All that exist are the lines and nodes; they *are* space, and the way they connect defines the geometry of space.

These graphs are called spin networks because the numbers on them are related to quantities called spins. Roger Penrose of the University of Oxford first proposed in the early 1970s that spin networks might play a role in theories of quantum gravity. We were very pleased when we found, in 1994, that precise calculations confirmed his intuition. Readers familiar with Feynman diagrams should note that our spin networks are not Feynman diagrams, despite the superficial resemblance. Feynman diagrams represent quantum interactions between particles, which proceed from one quantum state to another. Our diagrams represent fixed quantum states of spatial volumes and areas.

The individual nodes and edges of the diagrams represent extremely small regions of space: a node is typically a

THE AUTHOR

LEE SMOLIN is a researcher at the Perimeter Institute for Theoretical Physics in Waterloo, Ontario, and adjunct professor of physics at the University of Waterloo. He has a B.A. from Hampshire College and a Ph.D. from Harvard University and has been on the faculty of Yale, Syracuse and Pennsylvania State universities. In addition to his work on quantum gravity, he is interested in elementary particle physics, cosmology and the foundations of quantum theory. His 1997 book, *The Life of the Cosmos* (Oxford University Press), explored the philosophical implications of developments in contemporary physics and cosmology.

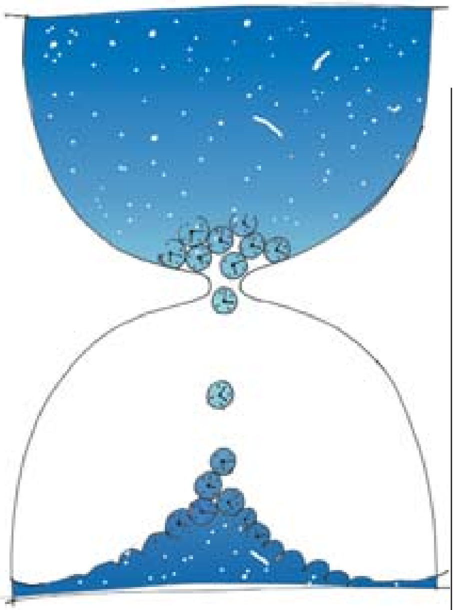
volume of about one cubic Planck length, and a line is typically an area of about one square Planck length. But in principle, nothing limits how big and complicated a spin network can be. If we could draw a detailed picture of the quantum state of our universe—the geometry of its space, as curved and warped by the gravitation of galaxies and black holes and everything else—it would be a gargantuan spin network of unimaginable complexity, with approximately 10^{184} nodes.

These spin networks describe the geometry of space. But what about all the matter and energy contained in that space? How do we represent particles and fields occupying positions and regions of space? Particles, such as electrons, correspond to certain types of nodes, which are represented by adding more labels on nodes. Fields, such as the electromagnetic field, are represented by additional labels on the lines of the graph. We represent particles and fields moving through space by these labels moving in discrete steps on the graphs.

Moves and Foams

PARTICLES AND FIELDS are not the only things that move around. According to general relativity, the geometry of space changes in time. The bends and curves of space change as matter and energy move, and waves can pass through it like ripples on a lake [see "Ripples in Space and Time," by W. Wayt Gibbs; SCIENTIFIC AMERICAN, April 2002]. In loop quantum gravity, these processes are represented by changes in the graphs. They evolve in time by a succession of certain "moves" in which the connectivity of the graphs changes [*see box on opposite page*].

When physicists describe phenomena quantum-mechanically, they compute probabilities for different processes. We do the same when we apply loop quantum gravity theory to describe phenomena, whether it be particles and fields moving on the spin networks or the geometry of space itself evolving in time. In particular, Thomas Thiemann of the Perimeter Institute for Theoretical Physics in Waterloo, Ontario, has de-



TIME ADVANCES by the discrete ticks of innumerable clocks.

rived precise quantum probabilities for the spin network moves. With these the theory is completely specified: we have a well-defined procedure for computing the probability of any process that can occur in a world that obeys the rules of our theory. It remains only to do the computations and work out predictions for what could be observed in experiments of one kind or another.

Einstein's theories of special and general relativity join space and time together into the single, merged entity known as spacetime. The spin networks that represent space in loop quantum gravity theory accommodate the concept of spacetime by becoming what we call spin "foams." With the addition of another dimension—time—the lines of the spin networks grow to become two-dimensional surfaces, and the nodes grow to become lines. Transitions where the spin networks change (the moves discussed earlier) are now represented by nodes where the lines meet in the foam. The spin foam picture of spacetime was proposed by several people, including Carlo Rovelli, Mike Reisenberger (now at the University of Montevideo), John Barrett of the University of Nottingham, Louis Crane of Kansas State University, John Baez of the University of California, Riverside, and Fotini Markopoulou of the Perimeter Institute for Theoretical Physics.

In the spacetime way of looking at things, a snapshot at a specific time is like a slice cutting across the spacetime. Taking such a slice through a spin foam produces a spin network. But it would be wrong to think of such a slice as moving continuously, like a smooth flow of time. Instead, just as space is defined by a spin network's discrete geometry, time is defined by the sequence of distinct moves that rearrange the network, as shown in the box on the opposite page. In this way, time also becomes discrete. Time flows not like a river but like the ticking of a clock, with "ticks" that are about as long as the Planck time: 10^{-43} second. Or, more precisely, time in our universe flows by the ticking of innumerable clocks—in a sense, at every location in the spin foam where a quantum "move" takes place, a clock at that location has ticked once.

Predictions and Tests

I HAVE OUTLINED what loop quantum gravity has to say about space and time at the Planck scale, but we cannot verify the theory directly by examining spacetime on that scale. It is too small. So how can we test the theory? An important test is whether one can derive classical general relativity as an approximation to loop quantum gravity. In other words, if the spin networks are like the threads woven into a piece of cloth, this is analogous to asking whether we can compute the right elastic properties for a sheet of the material by averaging over thousands of threads. Similarly, when averaged over many Planck lengths, do spin networks describe the geometry of space and its evolution in a way that agrees roughly with the "smooth cloth" of Einstein's classical theory? This is a difficult problem, but recently researchers have made progress for some cases—for certain configurations of the material, so to speak. For example, long-wavelength gravitational waves propagating on otherwise flat (uncurved) space can be described as excitations of specific quantum states described by the loop quantum gravity theory.

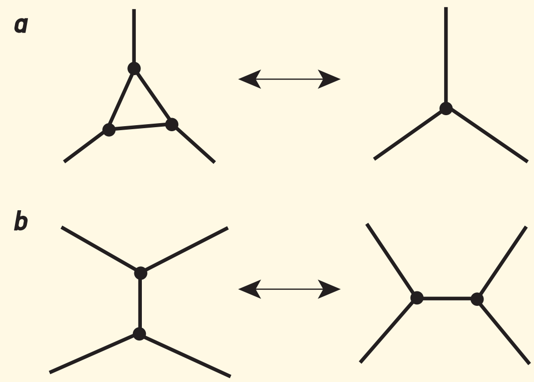
Another fruitful test is to see what loop quantum gravity has to say about one of the long-standing mysteries of

DUSAN PETRICIC

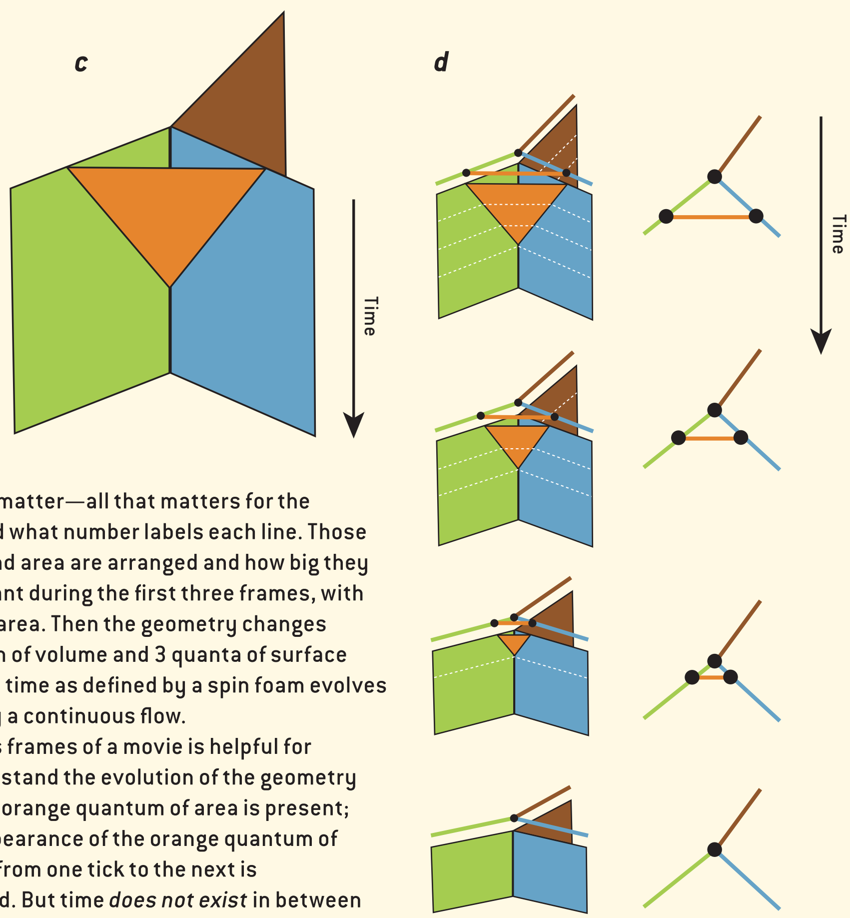
gravitational physics and quantum theory: the thermodynamics of black holes, in particular their entropy, which is related to disorder. Physicists have computed predictions regarding black hole thermodynamics using a hybrid, approximate theory in which matter is treated quantum-mechanically but spacetime is not. A full quantum theory of gravity, such as loop quantum gravity, should be able to reproduce these predictions. Specifically, in the 1970s Jacob D. Bekenstein, now at the Hebrew University of Jerusalem, inferred that black holes must be ascribed an entropy proportional to their surface area [see "Information in the Holographic Universe," by Jacob D. Bekenstein, on page 74]. Shortly after, Stephen W. Hawking of the University of Cambridge deduced that black holes, particularly small ones, must emit radiation. These predictions are among the greatest results of theoretical physics in the past 30 years.

EVOLUTION OF GEOMETRY IN TIME

Changes in the shape of space—such as those occurring when matter and energy move around within it and when gravitational waves flow by—are represented by discrete rearrangements, or moves, of the spin network. In *a,* a connected group of three volume quanta merge to become a single volume quantum; the reverse process can also occur. In *b,* two volumes divide up space and connect to adjoining volumes in a different way. Represented as polyhedra, the two polyhedra would merge on their common face and then split like a crystal cleaving on a different plane. These spin-network moves take place not only when large-scale changes in the geometry of space occur but also incessantly as quantum fluctuations at the Planck scale.



Another way to represent moves is to add the time dimension to a spin network—the result is called a spin foam (*c*). The lines of the spin network become planes, and the nodes become lines. Taking a slice through a spin foam at a particular time yields a spin network; taking a series of slices at different times produces frames of a movie showing the spin network evolving in time (*d*). But notice that the evolution, which at first glance appears to be smooth and continuous, is in fact discontinuous. All the spin networks that include the orange line (*first three frames shown*) represent exactly the same geometry of space. The length of the orange line doesn't matter—all that matters for the geometry is how the lines are connected and what number labels each line. Those are what define how the quanta of volume and area are arranged and how big they are. Thus, in *d,* the geometry remains constant during the first three frames, with 3 quanta of volume and 6 quanta of surface area. Then the geometry changes discontinuously, becoming a single quantum of volume and 3 quanta of surface area, as shown in the last frame. In this way, time as defined by a spin foam evolves by a series of abrupt, discrete moves, not by a continuous flow.

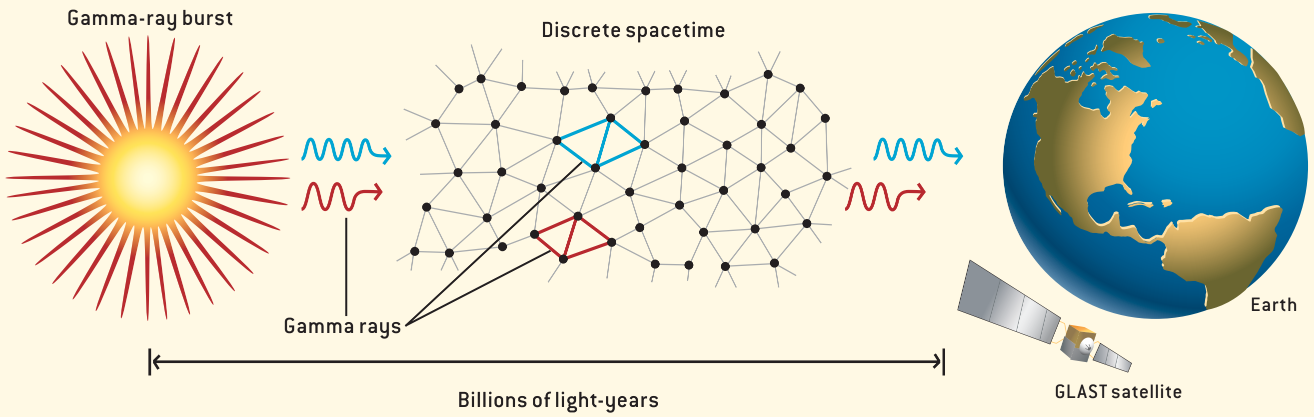


Although speaking of such sequences as frames of a movie is helpful for visualization, the more correct way to understand the evolution of the geometry is as discrete ticks of a clock. At one tick the orange quantum of area is present; at the next tick it is gone—in fact, the disappearance of the orange quantum of area *defines* the tick. The difference in time from one tick to the next is approximately the Planck time, 10^{-43} second. But time *does not exist* in between the ticks; there is no "in between," in the same way that there is no water in between two adjacent molecules of water.

NADIA STRASSER; SOURCE: *A* AND *B* ADAPTED FROM FOTINI MARKOPOULOU (http://arxiv.org/abs/gr-qc/9704013/); *C* ADAPTED FROM CARLO ROVELLI (http://arxiv.org/abs/gr-qc/9806121/)

AN EXPERIMENTAL TEST

Radiation from distant cosmic explosions called gamma-ray bursts might provide a way to test whether the theory of loop quantum gravity is correct. Gamma-ray bursts occur billions of light-years away and emit a huge amount of gamma rays within a short span. According to loop quantum gravity, each photon occupies a region of lines at each instant as it moves through the spin network that is space (in reality a very large number of lines, not just the five depicted here). The discrete nature of space causes higher-energy gamma rays to travel slightly faster than lower-energy ones. The difference is tiny, but its effect steadily accumulates during the rays' billion-year voyage. If a burst's gamma rays arrive at Earth at slightly different times according to their energy, that would be evidence for loop quantum gravity. The GLAST satellite, which is scheduled to be launched in 2007, will have the required sensitivity for this experiment.



NADIA STRASSER

To do the calculation in loop quantum gravity, we pick the boundary B to be the event horizon of a black hole. When we analyze the entropy of the relevant quantum states, we get *precisely* the prediction of Bekenstein. Similarly, the theory reproduces Hawking's prediction of black hole radiation. In fact, it makes further predictions for the fine structure of Hawking radiation. If a microscopic black hole were ever observed, this prediction could be tested by studying the spectrum of radiation it emits. That may be far off in time, however, because we have no technology to make black holes, small or otherwise.

Indeed, any experimental test of loop quantum gravity would appear at first to be an immense technological challenge. The problem is that the characteristic effects described by the theory become significant only at the Planck scale, the very tiny size of the quanta of area and volume. The Planck scale is 16 orders of magnitude below the scale probed in the highest-energy particle accelerators currently planned (higher energy is needed to probe shorter-distance scales). Because we cannot reach the Planck scale with an accelerator, many people have held out little hope for the confirmation of quantum gravity theories.

In the past several years, however, a few imaginative young researchers have thought up new ways to test the predictions of loop quantum gravity that can be done now. These methods depend on the propagation of light across the universe. When light moves through a medium, its wavelength suffers some distortions, leading to effects such as bending in water and the separation of different wavelengths, or colors. These effects also occur for light and particles moving through the discrete space described by a spin network.

Unfortunately, the magnitude of the effects is proportional to the ratio of the Planck length to the wavelength. For visible light, this ratio is smaller than 10^{-28}; even for the most powerful cosmic rays ever observed, it is about one billionth. For any radiation we can observe, the effects of the granular structure of space are very small. What the young researchers spotted is that these effects accumulate when light travels a long distance. And we detect light and particles that come from billions of light years away, from events such as gamma-ray bursts [see "The Brightest Explosions in the Universe," by Neil Gehrels, Luigi Piro and Peter J. T. Leonard; SCIENTIFIC AMERICAN, December 2002].

A gamma-ray burst spews out photons in a range of energies in a very brief explosion. Calculations in loop quantum gravity, by Rodolfo Gambini of the University of the Republic in Uruguay, Jorge Pullin of Louisiana State University and others, predict that photons of different energies should travel at slightly different speeds and therefore arrive at slightly different times [*see box above*]. We can look for this effect in data from satellite observations of gamma-ray bursts. So far the precision is about a factor of 1,000 below what is needed, but a new satellite observatory called GLAST, planned for 2007, will have the precision required.

The reader may ask if this result would mean that Einstein's theory of special relativity is wrong when it predicts a universal speed of light. Several people, including Giovanni Amelino-Camelia of the University of Rome "La Sapienza" and João Magueijo of Imperial College London, as well as myself, have developed modified versions of

Einstein's theory that will accommodate high-energy photons traveling at different speeds. Our theories propose that the universal speed is the speed of very low energy photons or, equivalently, long-wavelength light.

Another possible effect of discrete spacetime involves very high energy cosmic rays. More than 30 years ago researchers predicted that cosmic-ray protons with an energy greater than 3×10^{19} electron volts would scatter off the cosmic microwave background that fills space and should therefore never reach the earth. Puzzlingly, a Japanese experiment called AGASA has detected more than 10 cosmic rays with an energy over this limit. But it turns out that the discrete structure of space can raise the energy required for the scattering reaction, allowing higher-energy cosmic-ray protons to reach the earth. If the AGASA observations hold up, and if no other explanation is found, then it may turn out that we have already detected the discreteness of space.

The Cosmos

IN ADDITION to making predictions about specific phenomena such as high-energy cosmic rays, loop quantum gravity has opened up a new window through which we can study deep cosmological questions such as those relating to the origins of our universe. We can use the theory to study the earliest moments of time just after the big bang. General relativity predicts that there was a first moment of time, but this conclusion ignores quantum physics (because general relativity is not a quantum theory). Recent loop quantum gravity calculations by Martin Bojowald of the Max Planck Institute for Gravitational Physics in Golm, Germany, indicate that the big bang is actually a big bounce; before the bounce the universe was rapidly contracting. Theorists are now hard at work developing predictions for the early universe that may be testable in future cosmological observations. It is not impossible that in our lifetime we could see evidence of the time before the big bang.

A question of similar profundity concerns the cosmological constant—a positive or negative energy density that could permeate "empty" space. Recent observations of distant supernovae and the cosmic microwave background strongly indicate that this energy does exist and is positive, which accelerates the universe's expansion [see "The Quintessential Universe," by Jeremiah P. Ostriker and Paul J. Steinhardt; SCIENTIFIC AMERICAN, January 2001]. Loop quantum gravity has no trouble incorporating the positive energy density. This fact was demonstrated in 1990, when Hideo Kodama of Kyoto University wrote down equations describing an exact quantum state of a universe having a positive cosmological constant.

HOW CLASSICAL REALITY arises from quantum spacetime is still being worked out.

DUSAN PETRICIC

Many open questions remain to be answered in loop quantum gravity. Some are technical matters that need to be clarified. We would also like to understand how, if at all, special relativity must be modified at extremely high energies. So far our speculations on this topic are not solidly linked to loop quantum gravity calculations. In addition, we would like to know that classical general relativity is a good approximate description of the theory for distances much larger than the Planck length, in all circumstances. (At present we know only that the approximation is good for certain states that describe rather weak gravitational waves propagating on an otherwise flat spacetime.) Finally, we would like to understand whether or not loop quantum gravity has anything to say about unification: Are the different forces, including gravity, all aspects of a single, fundamental force? String theory is based on a particular idea about unification, but we also have ideas for achieving unification with loop quantum gravity.

Loop quantum gravity occupies a very important place in the development of physics. It is arguably *the* quantum theory of general relativity, because it makes no extra assumptions beyond the basic principles of quantum theory and relativity theory. The remarkable departure that it makes—proposing a discontinuous spacetime described by spin networks and spin foams—emerges from the mathematics of the theory itself, rather than being inserted as an ad hoc postulate.

Still, everything I have discussed is theoretical. It could be that in spite of all I have described here, space really is continuous, no matter how small the scale we probe. Then physicists would have to turn to more radical postulates, such as those of string theory. Because this is science, in the end experiment will decide. The good news is that the decision may come soon. SA

MORE TO EXPLORE

Three Roads to Quantum Gravity. Lee Smolin. Basic Books, 2001.

The Quantum of Area? John Baez in *Nature,* Vol. 421, pages 702–703; February 2003.

How Far Are We from the Quantum Theory of Gravity? Lee Smolin. March 2003. Preprint available at **http://arxiv.org/hep-th/0303185**

Welcome to Quantum Gravity. Special section. *Physics World,* Vol. 16, No. 11, pages 27–50; November 2003.

Loop Quantum Gravity. Lee Smolin. Online at **www.edge.org/3rd_culture/smolin03/smolin03_index.html**

A COSMIC CONUNDRUM

A new incarnation of Einstein's cosmological constant may point the way beyond general relativity

By Lawrence M. Krauss and Michael S. Turner

In 1917 Albert Einstein faced a confusing problem as he tried to reconcile his new theory of gravity, the general theory of relativity, with the limited understanding of the universe at the time. Like most of his contemporaries, Einstein was convinced that the universe must be static—neither expanding nor contracting—but this desired state was not consistent with his equations of gravity. In desperation, Einstein added an extra, ad hoc cosmological term to his equations to counterbalance gravity and allow for a static solution.

Twelve years later, though, American astronomer Edwin Hubble discovered that the universe was far from static. He found that remote galaxies were swiftly receding from our own at a rate that was proportional to their distance. A cosmological term was not needed to explain an expanding universe, so Einstein abandoned the concept. Russian-American physicist George Gamow declared in his autobiography that "when I was discussing cosmological problems with Einstein, he remarked that the introduction of the cosmological term was the biggest blunder he ever made in his life."

In the past seven years, however, the cosmological term—now called the cosmological constant—has reemerged to play a central role in 21st-century physics. But the motivation for this resurrection is actually very different from Einstein's original thinking; the new version of the term arises from recent observations of an accelerating universe and, ironically, from the principles of quantum mechanics, the branch of physics that Einstein so famously abhorred. Many physicists now expect the cosmological term to provide the key to moving beyond Einstein's theory to a deeper understanding of space, time, and gravity and perhaps to a quantum theory that unifies gravity with the

DON DIXON

LONELY UNIVERSE may be our ultimate fate if the cosmic expansion keeps accelerating—a phenomenon believed to be caused by the cosmological constant. The orange spheres represent the observable universe, which grows at the speed of light; the blue spheres represent an expanding patch of space. As expansion accelerates, fewer galaxy clusters are observable.

OVERVIEW

- Quantum mechanics and relativity, combined with recent evidence of an accelerating universe, have led physicists to resurrect the cosmological term that Einstein introduced and later repudiated. But this term now represents a mysterious form of energy that permeates empty space and drives an accelerated cosmic expansion.
- The efforts to explain the origin of this energy may help scientists move beyond Einstein's theory in ways that are likely to change our fundamental understanding of the universe.

other fundamental forces of nature. It is too soon to say what the ultimate resolution will be, but it is likely to change our picture of the universe.

Birth of a Constant

GENERAL RELATIVITY grew out of a decade-long struggle by Einstein to follow up on his pivotal observation in 1907 that gravity and accelerated motion are equivalent. As expressed in Einstein's well-known thought experiment, the physics inside an elevator sitting at rest in a uniform gravitational field of strength g is exactly the same as the physics inside an elevator that is rocketing through empty space with a uniform acceleration of g.

Einstein was also strongly influenced by the philosophical notions of Austrian physicist Ernst Mach, who rejected the idea of an absolute frame of reference for spacetime. In Newtonian physics, inertia refers to the tendency of an object to move with constant velocity unless acted on by a force. The notion of constant velocity requires an inertial (that is, not accelerating) frame of reference. But not accelerating with respect to what? Newton postulated the existence of absolute space, an immovable frame of reference that defined all local inertial frames. Mach, though, proposed that the distribution of matter in the universe defined inertial frames, and to a large extent Einstein's general theory of relativity embodies this notion.

Einstein's theory was the first concept of gravity that offered a hope of providing a self-consistent picture of the whole universe. It allowed a description not only of how objects move through space and time but of how space and time themselves dynamically evolve. In using his new theory to try to describe the universe, Einstein sought a solution that was finite, static and adhered to Mach's principles (for instance, a finite distribution of matter trailing off into emptiness did not seem to satisfy Mach's notion of matter being necessary to define space). These three prejudices led Einstein to introduce the cosmological term to construct a static solution that was finite and yet had no boundaries—his universe curved back on itself like the surface of a balloon [*see box on page 70*]. Physically, the cosmological term would have been unobservable on the scale of our solar system, but it would produce a cosmic repulsion on larger scales that would counteract the gravitational attraction of distant objects.

> **In its current incarnation,** the cosmological term arises not from relativity but from quantum mechanics.

Einstein's enthusiasm for the cosmological term began to wane quickly, however. In 1917 Dutch cosmologist Willem de Sitter demonstrated that he could produce a spacetime solution with a cosmological term even in the absence of matter—a very non-Machian result. This model was later shown to be nonstatic. In 1922 Russian physicist Alexander Friedmann constructed models of expanding and contracting universes that did not require a cosmological term. And in 1930 British astrophysicist Arthur Eddington showed that Einstein's universe was not really static: because gravity and the cosmological term were so precariously balanced, small perturbations would lead to runaway contraction or expansion. By 1931, with the expansion of the universe firmly established by Hubble, Einstein formally abandoned the cosmological term as "theoretically unsatisfactory anyway."

Hubble's discovery obviated the need for the cosmological term to counteract gravity; in an expanding universe, gravity simply slows the expansion. The question then became, Is gravity strong enough to eventually stop the expansion and cause the universe to collapse, or will the cosmos expand forever? In the Friedmann models, the answer is tied to the average density of matter: a high-density universe will collapse, whereas a low-density universe will expand eternally. The dividing point is the critical-density universe, which expands forever albeit at an ever decreasing rate. Because, according to Einstein's theory, the average curvature of the universe is tied to its average density, geometry and destiny are linked. The high-density universe is positively curved like the surface of a balloon, the low-density universe is negatively curved like the surface of a saddle, and the critical-density universe is spatially flat. Thus, cosmologists came to believe that determining the universe's geometry would reveal its ultimate fate.

The Energy of Nothing

THE COSMOLOGICAL TERM was banished from cosmology for the next six decades (except for a brief reappearance as part of the steady-state universe, a theory propounded in the late 1940s but decisively ruled out in the 1960s). But perhaps the most surprising thing about the term is that even if Einstein had not introduced it in a rush of confusion following his development of general relativity, we now realize that its presence seems to be inevitable. In its current incarnation, the cosmological term arises not from relativity, which governs nature on its largest scales, but from quantum mechanics, the physics of the smallest scales.

This new concept of the cosmological term is quite different from the one Einstein introduced. His original field equation, $G_{\mu\nu} = 8\pi G T_{\mu\nu}$, relates the curvature of space, $G_{\mu\nu}$, to the distribution of matter and energy, $T_{\mu\nu}$, where G is Newton's constant characterizing the strength of gravity. When Einstein added the cosmological term, he placed it on the left-hand side of the equation, suggesting it was a property of space itself [*see box at right*]. But if one moves the cosmological term to the right-hand side, it takes on a radically new meaning, the one it has today. It now represents a bizarre new form of energy density that remains constant even as the universe expands and whose gravity is repulsive rather than attractive.

Lorentz invariance, the fundamental symmetry associated with both the special and general theories of relativity, implies that only empty space can have this kind of energy density. Put in this perspective, the cosmological term seems even more bizarre. If asked what the energy of empty space is, most people would say "nothing." That is, after all, the only intuitively sensible value.

Alas, quantum mechanics is anything but intuitive. On the very small scales where quantum effects become important, even empty space is not really empty. Instead virtual particle-antiparticle pairs pop out of the vacuum, travel for short distances and then disappear again on timescales so fleeting that one cannot observe them directly. Yet their indirect effects are very important and can be measured. For example, the virtual particles affect the spectrum of hydrogen in a calculable way that has been confirmed by measurements.

Once we accept this premise, we should be prepared to contemplate the possibility that these virtual particles might endow empty space with some nonzero energy. Quantum mechanics thus makes the consideration of Einstein's cosmological term obligatory rather than optional. It cannot be dismissed as "theoretically unsatisfactory." The problem, however, is that all calculations and estimates of the magnitude of the empty-space energy lead to absurdly large values—ranging from 55 to 120 orders of magnitude greater than the energy of all the matter and radiation in the observable universe. If the vacuum energy density were really

THE COSMOLOGICAL TERM

A Change of Meaning

The heart of Einstein's general theory of relativity is the field equation, which states that the geometry of spacetime ($G_{\mu\nu}$, Einstein's curvature tensor) is determined by the distribution of matter and energy ($T_{\mu\nu}$, the stress-energy tensor), where G is Newton's constant characterizing the strength of gravity. (A tensor is a geometric or physical quantity that can be represented by an array of numbers.) In other words, matter and energy tell space how to curve.

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}$$

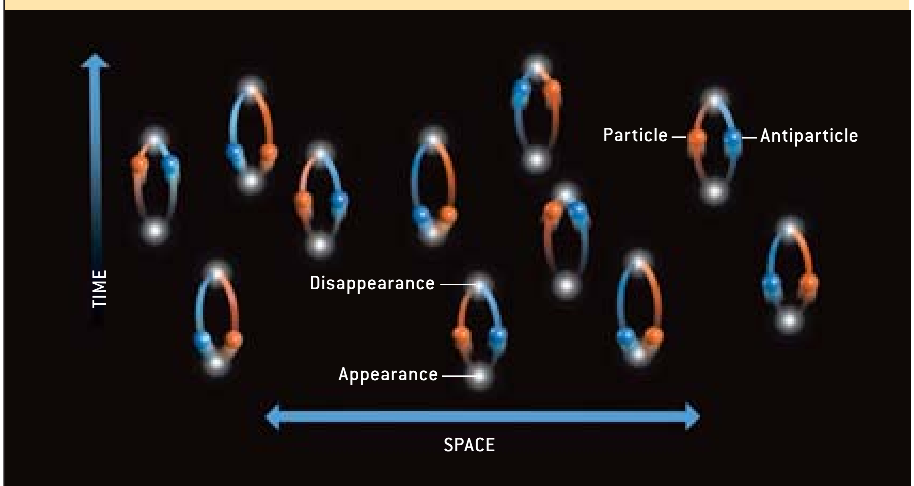
To create a model of a static universe, Einstein introduced the cosmological term Λ to counterbalance gravity's attraction on cosmic scales. He added the term (multiplied by $g_{\mu\nu}$, the spacetime metric tensor, which defines distances) to the left side of the field equation, suggesting that it was a property of space itself. But he abandoned the term once it became clear that the universe was expanding.

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}$$

The new cosmological term now being studied by physicists is necessitated by quantum theory, which holds that empty space may possess a small energy density. This term—ρ_{VAC}, the energy density of the vacuum, multiplied by $g_{\mu\nu}$—must go on the right side of the field equation with the other forms of energy.

$$G_{\mu\nu} = 8\pi G \left(T_{\mu\nu} + \rho_{VAC}\, g_{\mu\nu}\right)$$

Although Einstein's cosmological term and the quantum vacuum energy are mathematically equivalent, conceptually they could not be more different: the former is a property of space, the latter a form of energy that arises from virtual particle-antiparticle pairs. Quantum theory holds that these virtual particles constantly pop out of the vacuum, exist for a very brief time and then disappear (*below*).



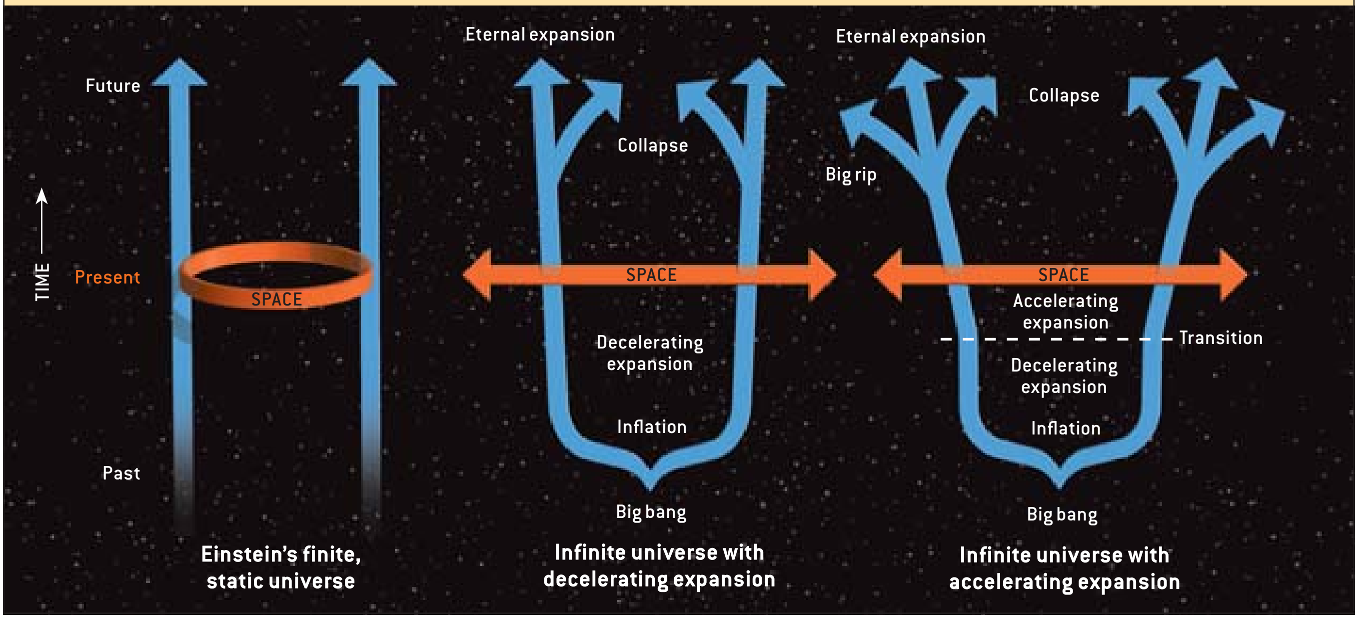
SLIM FILMS

THE AUTHORS

LAWRENCE M. KRAUSS and *MICHAEL S. TURNER* were among the first cosmologists to argue that the universe is dominated by a cosmological term that is radically different from the one introduced and then repudiated by Einstein. Their 1995 prediction of cosmic acceleration was confirmed by astronomical observations three years later. Krauss, director of the Center for Education and Research in Cosmology and Astrophysics at Case Western Reserve University, has also written seven popular books, including *Hiding in the Mirror: The Mysterious Allure of Extra Dimensions,* published in October 2005. Turner, who is Rauner Distinguished Service Professor at the University of Chicago, is now serving as assistant director for mathematical and physical sciences at the National Science Foundation.

Models of the Cosmos: Then and Now

Einstein's cosmological model (*left*) was a universe finite in space but infinite in time, remaining the same fixed size for eternity. This universe has no spatial boundaries; it curves back on itself like a circle. After the discovery of cosmic expansion, cosmologists constructed a model of an infinite universe in which the rate of expansion continuously slowed because of gravity (*center*), possibly leading to collapse. In the 1980s theorists added an early phase of rapid growth called inflation, for which there is now good evidence. In the past six years observations have shown that the cosmic expansion began to accelerate about five billion years ago (*right*). The ultimate fate of the universe—continued expansion, collapse or a hyper speedup called the big rip—depends on the nature of the mysterious dark energy driving the accelerated expansion.



that high, all matter in the universe would instantly fly apart.

This problem has been a thorn in the side of theorists for at least 30 years. In principle, it should have been recognized as early as the 1930s, when calculations of the effects of virtual particles were first performed. But in all areas of physics other than those related to gravity, the absolute energy of a system is irrelevant; what matters are the energy differences between states (for example, the energy differences between an atom's ground state and its excited states). If a constant is added to all the energy values, it drops out of such calculations, making it easy to ignore. Moreover, at that time few physicists took cosmology seriously enough to worry about applying quantum theory to it.

But general relativity implies that all forms of energy, even the energy of nothing, act as a source of gravity. Russian physicist Yakov Borisovich Zel'dovich realized the significance of this problem in the late 1960s, when he made the first estimates of the energy density of the vacuum. Since that time, theorists have been trying to figure out why their calculations yield such absurdly high values. Some undiscovered mechanism, they reasoned, must cancel the great bulk of the vacuum energy, if not all of it. Indeed, they assumed that the most plausible value for the energy density is zero—even quantum nothingness should weigh nothing.

As long as theorists believed in the back of their minds that such a canceling mechanism might exist, they could place the cosmological term problem on the back burner. Although it was fascinating, it could be ignored. Nature, however, has intervened.

Back with a Vengeance

THE FIRST DEFINITIVE evidence that something was amiss came from measurements of the slowing of the expansion rate of the universe. Recall that Hubble found that the relative velocities of remote galaxies were proportional to their distance from our own galaxy. From the point of view of general relativity, this relation arises from the expansion of space itself, which should slow down over time because of gravitational attraction. And because very distant galaxies are seen as they were billions of years ago, the slowing of the expansion should lead to a curvature of the otherwise linear Hubble relation—the most distant galaxies should be receding faster than Hubble's law would predict. The trick, though, is accurately determining the distances and velocities of very remote galaxies.

Such measurements rely on finding standard candles—objects of known intrinsic luminosity that are bright enough to be seen across the universe. A breakthrough came in the 1990s with

SLIM FILMS

the calibration of type Ia supernovae, which are believed to be the thermonuclear explosions of white dwarf stars about 1.4 times the mass of the sun. Two teams—the Supernova Cosmology Project, led by Saul Perlmutter of Lawrence Berkeley National Laboratory, and the High-z Supernova Search Team, led by Brian Schmidt of Mount Stromlo and Siding Spring Observatories—set out to measure the slowing of the expansion of the universe using this type of supernova. In early 1998 both groups made the same startling discovery: over the past five billion years, the expansion has been speeding up, not slowing down [see "Cosmological Antigravity," by Lawrence M. Krauss; SCIENTIFIC AMERICAN, January 1999]. Since then, the evidence for a cosmic speedup has gotten much stronger and has revealed not only a current accelerating phase but an earlier epoch of deceleration [see "From Slowdown to Speedup," by Adam G. Riess and Michael S. Turner; SCIENTIFIC AMERICAN, February 2004].

The supernova data, however, are not the only evidence pointing to the existence of some new form of energy driving the cosmic expansion. Our best picture of the early universe comes from observations of the cosmic microwave background (CMB), residual radiation from the big bang that reveals features of the universe at an age of about 400,000 years. In 2000, measurements of the angular size of variations of the CMB across the sky were good enough for researchers to determine that the geometry of the universe is flat. This finding was confirmed by a CMB-observing spacecraft called the Wilkinson Microwave Anisotropy Probe and other experiments.

A spatially flat geometry requires that the universe's average density must equal the critical density. But many different measurements of all forms of matter—including cold dark matter, a putative sea of slowly moving particles that do not emit light but do exert attractive gravity—showed that matter contributes only about 30 percent of the critical density. A flat universe therefore requires some other form of smoothly distributed energy that would have no observable influence on local clustering and yet could account for 70 percent of the critical density. Vacuum energy, or something very much like it, would produce precisely the desired effect.

In addition, a third line of reasoning suggested that cosmic acceleration was the missing piece of the cosmological puzzle. For two decades, the paradigm of inflation plus cold dark matter has been the leading explanation for the structure of the universe. The theory of inflation holds that in its very first moments the universe underwent a tremendous burst of expansion that smoothed and flattened its geometry and blew up quantum fluctuations in energy density from subatomic to cosmic size. This event produced the slightly inhomogeneous distribution of matter that led to the variations seen in the CMB and to the observed structures in the universe today. The gravity of cold dark matter, which far outweighs ordinary matter, governed the formation of these structures.

By the mid-1990s, however, this paradigm was seriously challenged by observational data. The predicted level of matter clustering differed from what was being measured. Worse, the predicted age of the universe appeared to be younger than the ages of the oldest stars. In 1995 the two of us pointed out that these contradictions would disappear if vacuum energy accounted for about two thirds of the critical density. (This model was very different from Einstein's closed universe, in which the density of the cosmological term was half that of matter.) Given the checkered history of vacuum energy, our proposal was, at the very least, provocative.

A decade later, though, everything fits together. In addition to explaining the current cosmic acceleration and the earlier period of deceleration, a resurrected cosmological term pushes the age of the universe to almost 14 billion years (comfortably above the ages of the oldest stars) and adds exactly enough energy to bring the universe to the critical

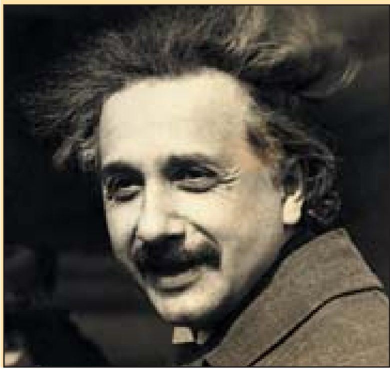
Cosmological observations may illuminate the relation between gravity and quantum mechanics at a fundamental level.

density. But physicists still do not know whether this energy actually comes from the quantum vacuum. The importance of discovering the cause of cosmic acceleration has brought a whole new urgency to the efforts to quantify vacuum energy. The problem of determining the weight of nothing can no longer be put aside for future generations. And the puzzle now seems even more confounding than it did when physicists were trying to devise a theory that would cancel vacuum energy. Now theorists must explain why vacuum energy might not be zero but so small that its effects on the cosmos became relevant only a few billion years ago.

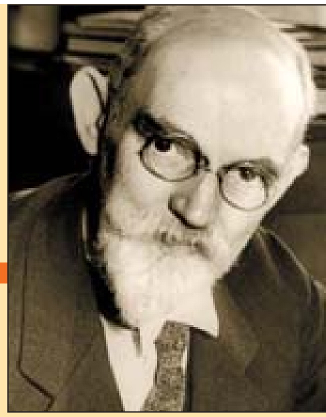
Of course, nothing could be more exciting to scientists than a puzzle of this magnitude, richness and importance. Just as Einstein was led to general relativity by considering the incompatibility of special relativity and Newton's theory of gravity, physicists today believe that Einstein's theory is incomplete because it cannot consistently incorporate the laws of quantum mechanics. But cosmological observations may illuminate the relation between gravity and quantum mechanics at a fundamental level. It was the equivalence of accel-

A Checkered History

Since Albert Einstein conceived the cosmological term almost 90 years ago, it has been repudiated, refashioned and resurrected. Here are some highlights.

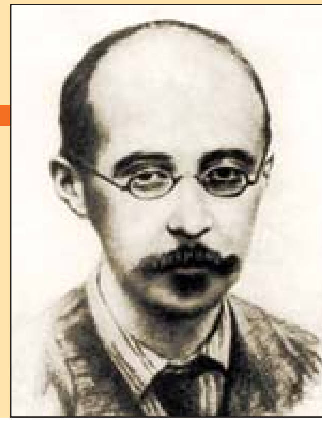


FEB. 1917: Einstein introduces the cosmological term to counteract gravity, allowing him to build a theoretical model of a static, finite universe



MARCH 1917: Dutch cosmologist Willem de Sitter produces an alternative model with a cosmological term. This model is later shown to have accelerating expansion

1922: Russian physicist Alexander Friedmann constructs models of expanding and contracting universes without a cosmological term



erated frames and gravity that pointed the way for Einstein; perhaps another kind of acceleration, the cosmic speed-up, will point the way today. And theorists have already outlined some ideas about how to proceed.

The Superworld

STRING THEORY, which is now often called M-theory, is viewed by many physicists as a promising approach to marrying quantum mechanics with gravity. One of the basic ideas underlying this theory is called supersymmetry, or SUSY. SUSY is a symmetry between particles of half-integer spin (fermions such as quarks and leptons) and those of whole-integer spin (bosons such as photons, gluons and other force carriers). In a world in which SUSY were fully manifest, a particle and its superpartner would have the same mass; for example, the supersymmetric electron (called the selectron) would be as light as the electron, and so on. In this superworld, moreover, it could be proved that quantum nothingness would weigh nothing and that the vacuum would have zero energy.

In the real world, however, we know that no selectron as light as the electron could exist because physicists would have already detected it in particle accelerators. (Theorists speculate that superpartner particles are millions of times heavier than electrons and thus cannot be found without the help of more powerful accelerators.) SUSY must therefore be a broken symmetry, which suggests that quantum nothingness might weigh something.

Physicists have produced models of broken supersymmetry yielding a vacuum energy density that is many orders of magnitude smaller than the absurdly high estimates made previously. But even this theorized density is far larger than that indicated by cosmological observations. Recently, however, researchers have recognized that M-theory appears to allow for an almost infinite number of different solutions. Although almost all these possible solutions would indeed result in a vacuum energy that is far too high, some might produce a vacuum energy as low as the value that cosmologists have observed [see "The String Theory Landscape," by Raphael Bousso and Joseph Polchinski, on page 40].

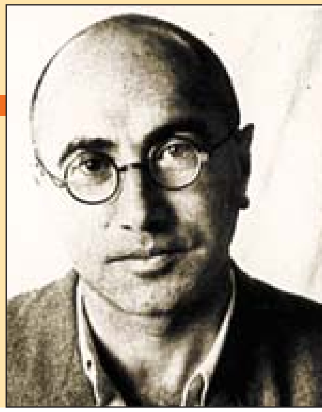
Another hallmark of string theory is the positing of additional dimensions. Current theory adds six or seven spatial dimensions, all hidden from view, to the usual three. This construct offers another approach to explaining cosmic acceleration. Georgi Dvali of New York University and his collaborators have suggested that the effect of extra dimensions may show up as an additional term in Einstein's field equation that leads to an accelerated expansion of the universe [see "Out of the Darkness," by Georgi Dvali; SCIENTIFIC AMERICAN, February 2004]. This approach runs counter to long-held expectations: for decades, it had been assumed that the place to look for differences between general relativity and its successor theory would be at short distances, not cosmic ones. Dvali's plan flies in the face of this wisdom—if he is correct, the first harbinger of a new

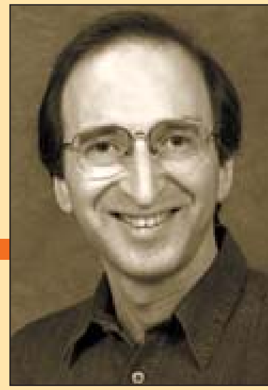
> The discovery of cosmic acceleration has forever altered our thinking about the future. Destiny is no longer tied to geometry.

UNDERWOOD & UNDERWOOD/CORBIS (*Einstein*); YERKES OBSERVATORY (*de Sitter*); *SOVIET PHYSICS-USPEKHI*, COURTESY OF AIP EMILIO SEGRÈ VISUAL ARCHIVES (*Friedmann*)

1929: American astronomer Edwin Hubble discovers that the universe is expanding. Two years later Einstein abandons the cosmological term, calling it "theoretically unsatisfactory anyway"

1967: Russian physicist Yakov Borisovich Zel'dovich estimates the energy density of the quantum vacuum and finds that it would make an immense cosmological term





1998: Two teams of supernova hunters led by Saul Perlmutter (*left*) and Brian Schmidt (*right*) report that the cosmic expansion is accelerating. A refashioned cosmological term would produce this effect. Since 1998 the evidence for cosmic acceleration has strengthened

cosmic understanding will be at the largest distances, not the smallest.

It is possible that the explanation of cosmic acceleration will have nothing to do with resolving the mystery of why the cosmological term is so small or how Einstein's theory can be extended to include quantum mechanics. General relativity stipulates that an object's gravity is proportional to its energy density plus three times its internal pressure. Any energy form with a large, negative pressure—which pulls inward like a rubber sheet instead of pushing outward like a ball of gas—will therefore have repulsive gravity. So cosmic acceleration may simply have revealed the existence of an unusual energy form, dubbed dark energy, that is not predicted by either quantum mechanics or string theory.

Geometry vs. Destiny

IN ANY CASE, the discovery of cosmic acceleration has forever altered our thinking about the future. Destiny is no longer tied to geometry. Once we allow for the existence of vacuum energy or something similar, anything is possible. A flat universe dominated by positive vacuum energy will expand forever at an ever increasing rate [*see illustration on page 66*], whereas one dominated by negative vacuum energy will collapse. And if the dark energy is not vacuum energy at all, then its future impact on cosmic expansion is uncertain. It is possible that, unlike a cosmological constant, the density of dark energy may rise or fall over time. If the density rises, the cosmic acceleration will increase, tearing apart galaxies, solar systems, planets and atoms, in that order, after a finite amount of time. But if the density falls, the acceleration could stop. And if the density becomes negative, the universe could collapse. The two of us have demonstrated that without knowing the detailed origin of the energy currently driving the expansion, no set of cosmological observations can pin down the ultimate fate of the universe.

To resolve this puzzle, we may need a fundamental theory that allows us to predict and categorize the gravitational impact of every single possible contribution to the energy of empty space. In other words, the physics of nothingness will determine the fate of our universe! Finding the solution may require new measurements of the cosmic expansion and of the structures that form within it to provide direction for theorists. Fortunately, many experiments are being planned, including a space telescope dedicated to observing distant supernovae and new telescopes on the ground and in space to probe dark energy through its effect on the development of large-scale structures.

Our knowledge of the physical world usually develops in an atmosphere of creative confusion. The fog of the unknown led Einstein to consider a cosmological term as a desperate solution to constructing a static, Machian universe. Today our confusion about cosmic acceleration is driving physicists to explore every avenue possible to understand the nature of the energy that is driving the speedup. The good news is that although many roads may lead to dead ends, the resolution of this profound and perplexing mystery may eventually help us unify gravity with the other forces in nature, which was Einstein's fondest hope. SA

MORE TO EXPLORE

Subtle Is the Lord: The Science and Life of Albert Einstein. Abraham Pais. Oxford University Press, 1982.

The Cosmological Constant Problem. Steven Weinberg in *Reviews of Modern Physics,* Vol. 61, No. 1, pages 1–23; 1989.

The Observational Case for a Low Density Universe with a Non-Zero Cosmological Constant. J. P. Ostriker and P. J. Steinhardt in *Nature,* Vol. 377, pages 600–602; October 19, 1995.

The Cosmological Constant Is Back. Lawrence M. Krauss and Michael S. Turner in *General Relativity and Gravitation,* Vol. 27, No. 11, page 1135; 1995.

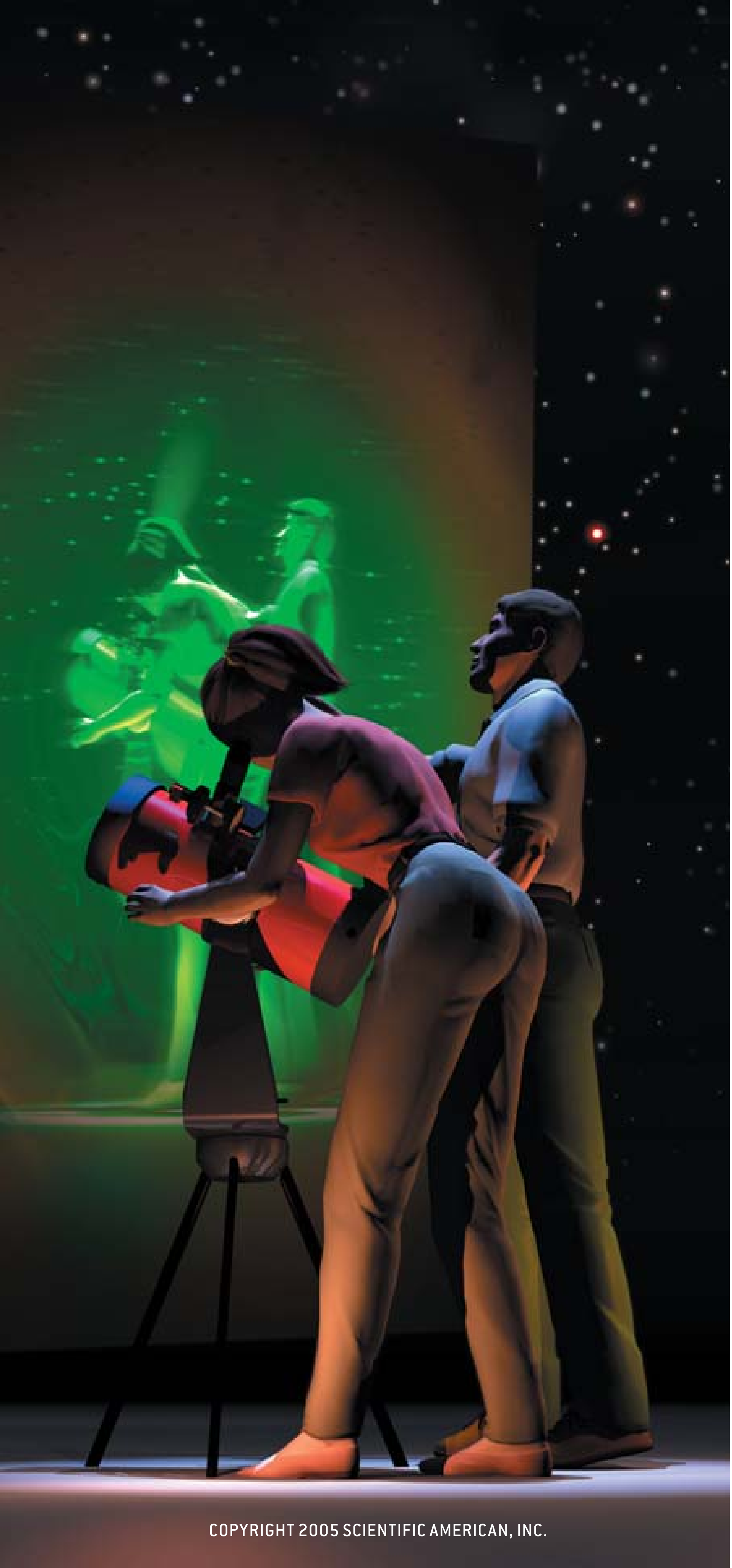
Geometry and Destiny. Lawrence M. Krauss and Michael S. Turner in *General Relativity and Gravitation,* Vol. 31, No. 10, pages 1453–1459; October 1999.

Information in the HOLOGRAPHIC UNIVERSE

Theoretical results about black holes suggest that the universe could be like a gigantic hologram

By Jacob D. Bekenstein

Illustrations by Alfred T. Kamajian



Ask anybody what the physical world is made of, and you are likely to be told "matter and energy."

Yet if we have learned anything from engineering, biology and physics, information is just as crucial an ingredient. The robot at the automobile factory is supplied with metal and plastic but can make nothing useful without copious instructions telling it which part to weld to what and so on. A ribosome in a cell in your body is supplied with amino acid building blocks and is powered by energy released by the conversion of ATP to ADP, but it can synthesize no proteins without the information brought to it from the DNA in the cell's nucleus. Likewise, a century of developments in physics has taught us that information is a crucial player in physical systems and processes. Indeed, a current trend, initiated by John A. Wheeler of Princeton University, is to regard the physical world as made of information, with energy and matter as incidentals.

This viewpoint invites a new look at venerable questions. The information storage capacity of devices such as hard-disk drives has been increasing by leaps and bounds. When will such progress halt? What is the ultimate information capacity of a device that weighs, say, less than a gram and can fit inside a cubic centimeter (roughly the size of a computer chip)? How much information does it take to describe a whole universe? Could that description fit in a computer's memory? Could we, as William Blake memorably penned, "see a world in a grain of sand," or is that idea no more than poetic license?

Remarkably, recent developments in theoretical physics answer some of these questions, and the answers might be important clues to the ultimate theory of reality. By studying the mysterious properties of black holes, physicists have deduced absolute limits on how much information a region of space or a quantity of matter and energy can hold. Related results suggest that our universe, which we perceive to have three spatial dimensions, might instead be "written" on a two-dimensional surface, like a hologram. Our everyday perceptions of the world as three-dimensional would then be either a profound illusion or merely one of two alternative ways of viewing reality. A grain of sand may not encompass our world, but a flat screen might.

A Tale of Two Entropies

FORMAL INFORMATION theory originated in seminal 1948 papers by American applied mathematician Claude E. Shannon, who introduced today's most widely used measure of information content: entropy. Entropy had long been a central concept of thermodynamics, the branch of physics dealing with heat. Thermodynamic entropy is popularly described as the disorder in a physical system. In 1877 Austrian physicist Ludwig Boltzmann characterized it more precisely in terms of the number of distinct microscopic states that the particles composing a chunk of matter could be in while still looking like the same macroscopic chunk of matter. For example, for the air in the room around you, one would count all the ways that the individual gas molecules could be distributed in the room and all the ways they could be moving.

When Shannon cast about for a way to quantify the information contained in, say, a message, he was led by logic to a formula with the same form as Boltzmann's. The Shannon entropy of a message is the number of binary digits, or bits, needed to encode it. Shannon entropy does not enlighten us about the value of information, which is highly dependent on context. Yet as an objective measure of quantity of information, it has been enormously useful in science and technology. For instance, the design of every modern communications device—from cellular phones to modems to compact-disc players—relies on Shannon entropy.

Thermodynamic entropy and Shannon entropy are conceptually equivalent: the number of arrangements that are counted by Boltzmann entropy reflects the amount of Shannon information one would need to implement any particular arrangement. The two entropies have two salient differences, though. First, the thermodynamic entropy used by a chemist or a refrigeration engineer is expressed in units of energy divided by temperature, whereas the Shannon entropy used by a communications engineer is in bits, essentially dimensionless. That difference is merely a matter of convention.

Even when reduced to common units, however, typical values of the two entropies differ vastly in magnitude. A silicon microchip carrying a gigabyte of data, for instance, has a Shannon entropy of about 10^{10} bits (one byte is eight bits), tremendously smaller than the chip's thermodynamic entropy, which is about 10^{23} bits at room temperature. This discrepancy occurs because the entropies are computed for different degrees of freedom. A degree of freedom is any quantity that can vary, such as a coordinate specifying a particle's location or one component of its velocity. The Shannon entropy of the chip cares only about the overall state of each tiny transistor etched in the silicon crystal—the transistor is on or off; it is a 0 or a 1—a single binary degree of freedom.

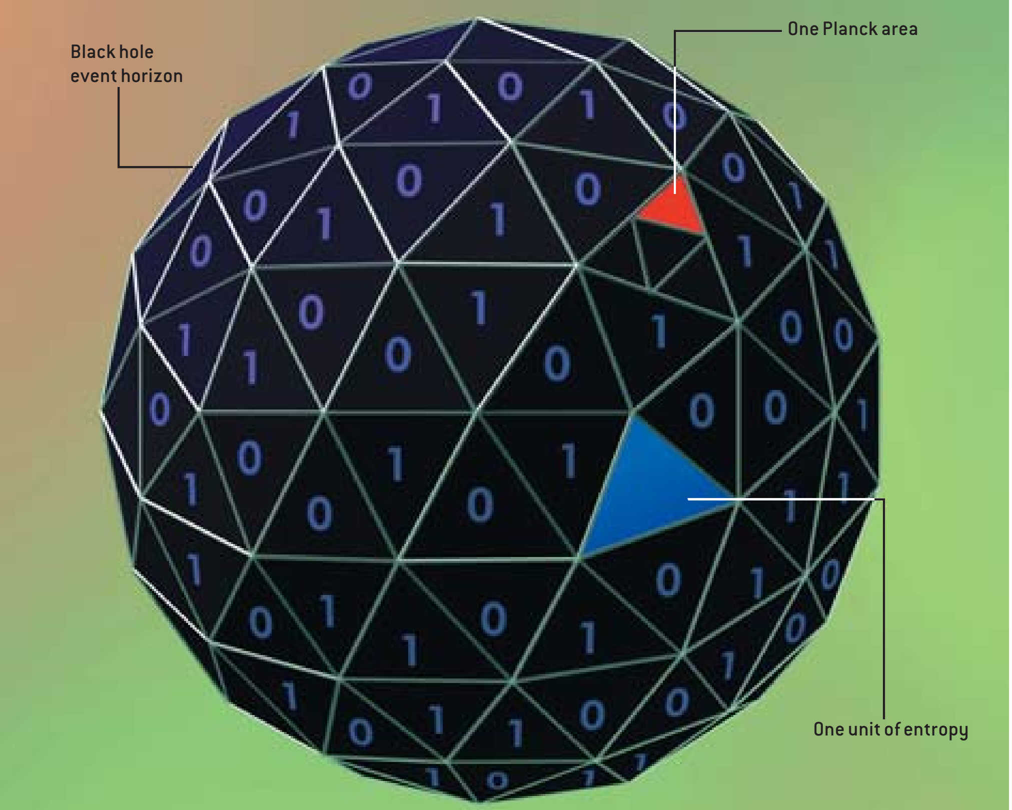
Thermodynamic entropy, in contrast, depends on the states of all the billions of atoms (and their roaming electrons) that make up each transistor. As miniaturization brings closer the day when each atom will store one bit of information for us, the useful Shannon entropy of the state-of-the-art microchip will edge closer in magnitude to its material's thermodynamic entropy. When the two entropies are calculated for the same degrees of freedom, they are equal.

What are the ultimate degrees of freedom? Atoms, after all, are made of electrons and nuclei, nuclei are agglomerations of protons and neutrons, and those in turn are composed of quarks. Many physicists today consider electrons and quarks to be excitations of superstrings, which they hypothesize to be the most fundamental entities. But the vicissitudes of a century of revelations in physics warn us not to be dogmatic. There could be more levels of structure in our universe than are dreamt of in today's physics.

One cannot calculate the ultimate information capacity of a chunk of matter or, equivalently, its true thermodynamic entropy, without knowing the nature of the ultimate constituents of matter or of the deepest level of structure, which I shall refer to as level X. (This ambiguity causes no problems in analyzing practical thermodynamics, such as that of car engines, for example, because the quarks within the atoms can be ignored—they do not change their states under the relatively benign conditions in the engine.) Given the dizzying progress

Overview/*The World as a Hologram*

- An astonishing theory called the holographic principle holds that the universe is like a hologram: just as a trick of light allows a fully three-dimensional image to be recorded on a flat piece of film, our seemingly three-dimensional universe could be completely equivalent to alternative quantum fields and physical laws "painted" on a distant, vast surface.
- The physics of black holes—immensely dense concentrations of mass—provides a hint that the principle might be true. Studies of black holes show that, contrary to common sense, the maximum entropy or information content of any region of space is defined not by its volume but by its surface area.
- Physicists hope this surprising finding is a clue to the ultimate theory of reality.



ENTROPY OF A BLACK HOLE is proportional to the area of its event horizon, the surface from within which even light cannot escape the gravity of the hole. Specifically, a hole with a horizon spanning *A* Planck areas has $A/4$ units of entropy. (The Planck area, approximately 10^{-66} square centimeter, is the fundamental quantum unit of area determined by the strength of gravity, the speed of light and the size of quanta.) Considered as information, it is as if the entropy were written on the event horizon, with each bit (each digital 1 or 0) corresponding to four Planck areas.

in miniaturization, one can playfully contemplate a day when quarks will serve to store information, one bit apiece perhaps. How much information would then fit into our one-centimeter cube? And how much if we harness superstrings or even deeper, yet undreamt of levels? Surprisingly, developments in gravitation physics in the past three decades have supplied some clear answers to what seem to be elusive questions.

Black Hole Thermodynamics

A CENTRAL PLAYER in these developments is the black hole. Black holes are a consequence of general relativity, Albert Einstein's 1915 geometric theory of gravitation. In this theory, gravitation arises from the curvature of spacetime, which makes objects move as if they were pulled by a force. Conversely, the curvature is caused by the presence of matter and energy. According to Einstein's equations, a sufficiently dense concentration of matter or energy will curve spacetime so extremely that it rends, forming a black hole. The laws of relativity forbid anything that went into a black hole from coming out again, at least within the classical (nonquantum) description of the physics. The point of no return, called the event horizon of the black hole, is of crucial importance. In the simplest case, the horizon is a sphere, whose surface area is larger for more massive black holes.

It is impossible to determine what is inside a black hole. No detailed information can emerge across the horizon and escape into the outside world. In disappearing forever into a black hole, however, a piece of matter does leave some traces. Its energy (we count any mass as energy in accordance with Einstein's $E = mc^2$) is permanently reflected in an increment in the black hole's mass. If the matter is captured while circling the hole, its associated angular momentum is added to the black hole's angular momentum. Both the mass and angular momentum of a black hole are measurable from their effects on spacetime around the hole. In this way, the laws of conservation of energy and angular momentum are upheld by black holes. Another fundamental law, the second law of thermodynamics, appears to be violated.

The second law of thermodynamics summarizes the familiar observation that most processes in nature are irreversible: a teacup falls from the table and shatters, but no one has ever seen shards jump up of their own accord and assemble into a teacup. The second law of thermodynamics forbids such inverse processes. It states that the entropy of an isolated physical system can never decrease; at best, entropy remains constant, and usually it increases. This law is central to physical chemistry and engineering; it is arguably the physical law with the greatest impact outside physics.

As first emphasized by Wheeler, when matter disappears into a black hole, its entropy is gone for good, and the second law seems to be transcended, made irrelevant. A clue to resolving this puzzle came in 1970, when Demetrious Christodoulou, then a graduate student of Wheeler's at Princeton, and Stephen W. Hawking of the University of Cambridge independently proved that in various processes, such as black hole mergers, the total area of the event horizons never decreases. The analogy with the tendency of entropy to increase led me to propose in 1972 that a black hole has entropy proportional to the area of its horizon [*see illustration above*]. I conjectured that when matter falls into a black hole, the increase in black hole entropy always compensates or overcompensates for the "lost" entropy of the

THE AUTHOR

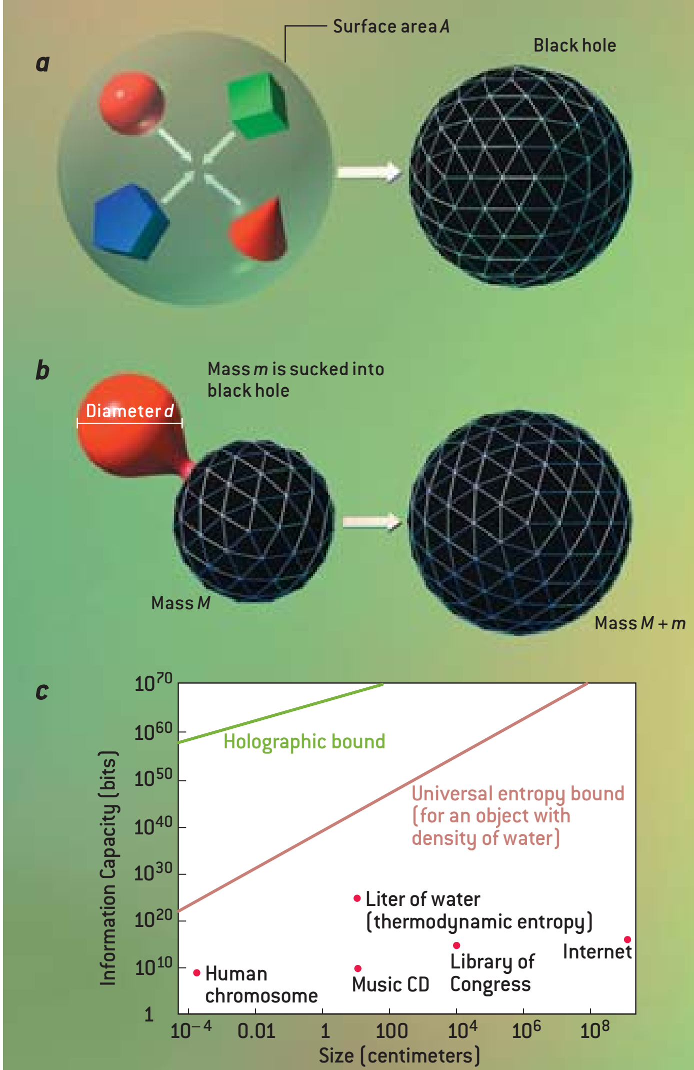
JACOB D. BEKENSTEIN has contributed to the foundation of black hole thermodynamics and to other aspects of the connections between information and gravitation. He is Polak Professor of Theoretical Physics at the Hebrew University of Jerusalem, a member of the Israel Academy of Sciences and Humanities, and a recipient of the Rothschild and the Israel prizes. Bekenstein dedicates this article to John Archibald Wheeler (his Ph.D. supervisor 30 years ago). Wheeler belongs to the third generation of Ludwig Boltzmann's students: Wheeler's Ph.D. adviser, Karl Herzfeld, was a student of Boltzmann's student Friedrich Hasenöhrl.

matter. More generally, the sum of black hole entropies and the ordinary entropy outside the black holes cannot decrease. This is the generalized second law—GSL for short.

The GSL has passed a large number of stringent, if purely theoretical, tests. When a star collapses to form a black hole, the black hole entropy greatly exceeds the star's entropy. In 1974 Hawking demonstrated that a black hole spontaneously emits thermal radiation, now known as Hawking radiation, by a quantum process [see "The Quantum Mechanics of Black Holes," by Stephen W. Hawking; Scientific American, January 1977]. The Christodoulou-Hawking theorem fails in the face of this phenomenon (the mass of the black hole, and therefore its horizon area, decreases), but the GSL copes with it: the entropy of the emergent radiation more than compensates for the decrement in black hole entropy, so the GSL is preserved. In 1986 Rafael D. Sorkin of Syracuse University exploited the horizon's role in barring information inside the black hole from influencing affairs outside to show that the GSL (or something very similar to it) must be valid for any conceivable process that black holes undergo. His deep argument makes it clear that the entropy entering the GSL is that calculated down to level X, whatever that level may be.

Hawking's radiation process allowed him to determine the proportionality constant between black hole entropy and horizon area: black hole entropy is precisely one quarter of the event horizon's area measured in Planck areas. (The Planck length, about 10^{-33} centimeter, is the fundamental length scale related to gravity and quantum mechanics. The Planck area is its square.) Even in thermodynamic terms, this is a vast quantity of entropy. The entropy of a black hole one centimeter in diameter would be about 10^{66} bits, roughly equal to the thermodynamic entropy of a cube of water 10 billion kilometers on a side.

LIMITS ON INFORMATION DENSITY



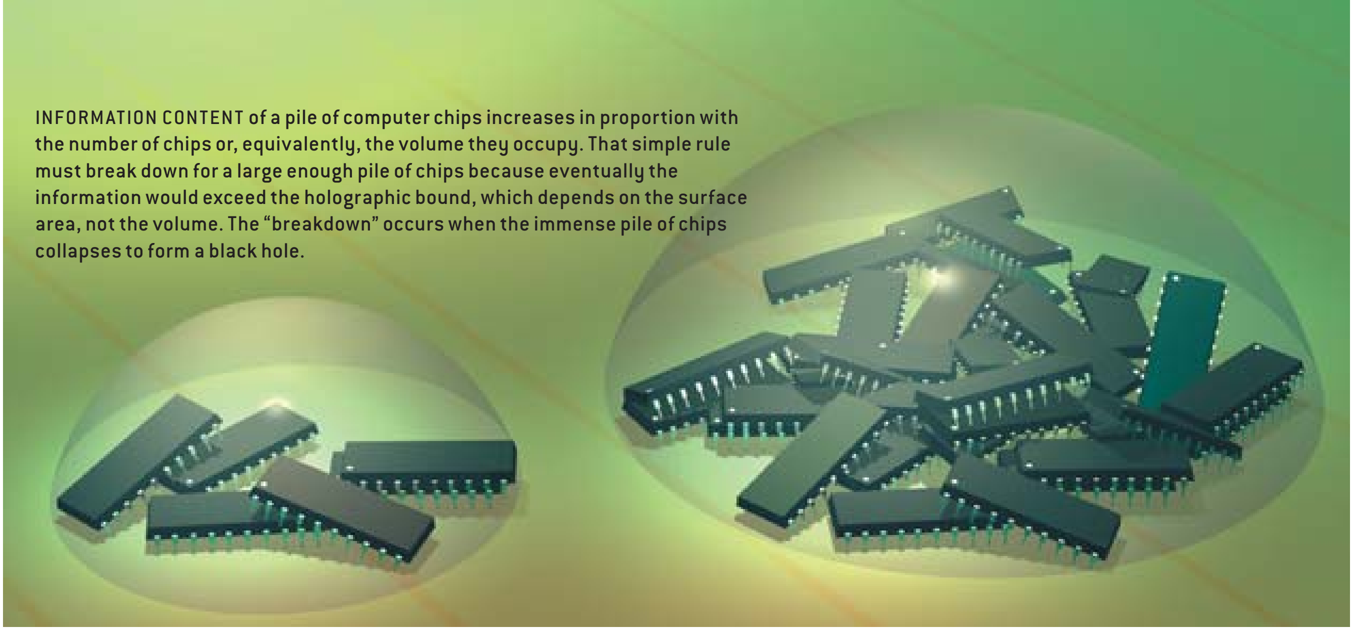
The thermodynamics of black holes allows one to deduce limits on the density of entropy or information in various circumstances.

The holographic bound defines how much information can be contained in a specified region of space. It can be derived by considering a roughly spherical distribution of matter that is contained within a surface of area *A*. The matter is induced to collapse to form a black hole (*a*). The black hole's area must be smaller than *A*, so its entropy must be less than $A/4$ [*see illustration on preceding page*]. Because entropy cannot decrease, one infers that the original distribution of matter also must carry less than $A/4$ units of entropy or information. This result—that the maximum information content of a region of space is fixed by its area—defies the commonsense expectation that the capacity of a region should depend on its volume.

The universal entropy bound defines how much information can be carried by a mass *m* of diameter *d*. It is derived by imagining that a capsule of matter is engulfed by a black hole not much wider than it (*b*). The increase in the black hole's size places a limit on how much entropy the capsule could have contained. This limit is tighter than the holographic bound, except when the capsule is almost as dense as a black hole (in which case the two bounds are equivalent).

The holographic and universal information bounds are far beyond the data storage capacities of any current technology, and they greatly exceed the density of information on chromosomes and the thermodynamic entropy of water (*c*). —*J.D.B.*

LAURIE GRACE (*graph*)



INFORMATION CONTENT of a pile of computer chips increases in proportion with the number of chips or, equivalently, the volume they occupy. That simple rule must break down for a large enough pile of chips because eventually the information would exceed the holographic bound, which depends on the surface area, not the volume. The "breakdown" occurs when the immense pile of chips collapses to form a black hole.

The World as a Hologram

THE GSL ALLOWS US to set bounds on the information capacity of any isolated physical system, limits that refer to the information at all levels of structure down to level X. In 1980 I began studying the first such bound, called the universal entropy bound, which limits how much entropy can be carried by a specified mass of a specified size [*see box on opposite page*]. A related idea, the holographic bound, was devised in 1995 by Leonard Susskind of Stanford University. It limits how much entropy can be contained in matter and energy occupying a specified volume of space.

In his work on the holographic bound, Susskind considered any approximately spherical isolated mass that is not itself a black hole and that fits inside a closed surface of area *A*. If the mass can collapse to a black hole, that hole will end up with a horizon area smaller than *A*. The black hole entropy is therefore smaller than $A/4$. According to the GSL, the entropy of the system cannot decrease, so the mass's original entropy cannot have been bigger than $A/4$. It follows that the entropy of an isolated physical system with boundary area *A* is necessarily less than $A/4$. What if the mass does not spontaneously collapse? In 2000 I showed that a tiny black hole can be used to convert the system to a black hole not much different from the one in Susskind's argument. The bound is therefore independent of the constitution of the system or of the nature of level X. It just depends on the GSL.

We can now answer some of those elusive questions about the ultimate limits of information storage. A device measuring a centimeter across could in principle hold up to 10^{66} bits—a mind-boggling amount. The visible universe contains at least 10^{100} bits of entropy, which could in principle be packed inside a sphere a tenth of a light-year across. Estimating the entropy of the universe is a difficult problem, however, and much larger numbers, requiring a sphere almost as big as the universe itself, are entirely plausible.

But it is another aspect of the holographic bound that is truly astonishing. Namely, that the maximum possible entropy depends on the boundary area instead of the volume. Imagine that we are piling up computer memory chips in a big heap. The number of transistors—the total data storage capacity—increases with the volume of the heap. So, too, does the total thermodynamic entropy of all the chips. Remarkably, though, the theoretical ultimate information capacity of the space occupied by the heap increases only with the surface area. Because volume increases more rapidly than surface area, at some point the entropy of all the chips would exceed the holographic bound. It would seem that either the GSL or our commonsense ideas of entropy and information capacity must fail. In fact, what fails is the pile itself: it would collapse under its own gravity and form a black hole before that impasse was reached. Thereafter each additional memory chip would increase the mass and surface area of the black hole in a way that would continue to preserve the GSL.

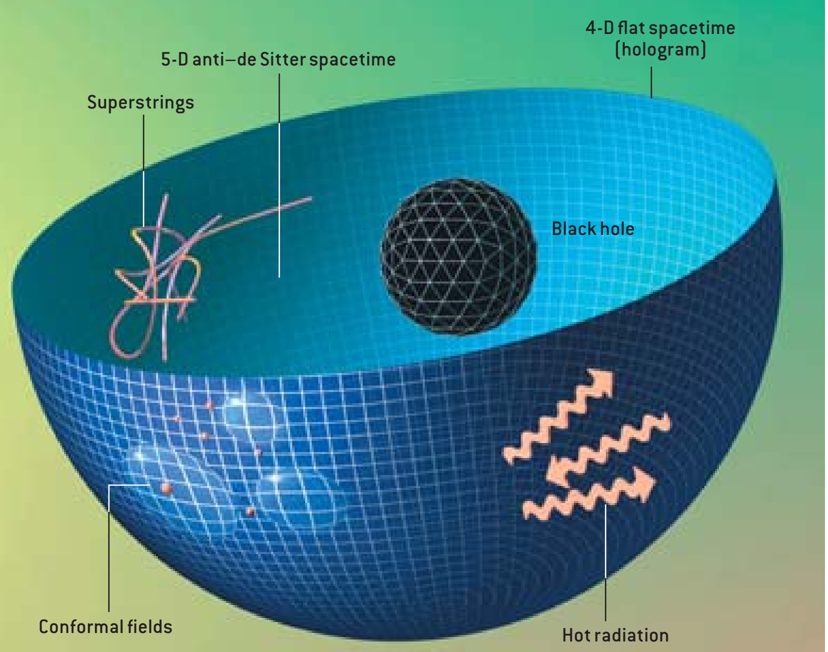
This surprising result—that information capacity depends on surface area—has a natural explanation if the holographic *principle* (proposed in 1993 by Nobelist Gerard 't Hooft of the University of Utrecht in the Netherlands and elaborated by Susskind) is true. In the everyday world, a hologram is a special kind of photograph that generates a full three-dimensional image when it is illuminated in the right manner. All the information describing the 3-D scene is encoded into the pattern of light and dark areas on the two-dimensional piece of film, ready to be regenerated. The holographic principle contends that an analogue of this visual magic applies to the full physical description of any system occupying a 3-D region: it proposes that another physical theory defined only on the 2-D boundary of the region completely describes the 3-D physics. If a 3-D system can be fully described by a physical theory operating solely on its 2-D boundary, one would expect the information content of the system not to exceed that of the description on the boundary.

A Universe Painted on Its Boundary

CAN WE APPLY the holographic principle to the universe at large? The real universe is a 4-D system: it has volume and extends in time. If the physics of our universe is holographic, there would be an alternative set of physical laws, operating on a 3-D boundary of spacetime somewhere, that would be equivalent to

A HOLOGRAPHIC SPACETIME

Two universes of different dimensions and obeying disparate physical laws are rendered completely equivalent by the holographic principle. Theorists have demonstrated this principle mathematically for a specific type of five-dimensional spacetime ("anti–de Sitter") and its four-dimensional boundary. In effect, the 5-D universe is recorded like a hologram on the 4-D surface at its periphery. Superstring theory rules in the 5-D spacetime, but a so-called conformal field theory of point particles operates on the 4-D hologram. A black hole in the 5-D spacetime is equivalent to hot radiation on the hologram—for example, the hole and the radiation have the same entropy even though the physical origin of the entropy is completely different for each case. Although these two descriptions of the universe seem utterly unalike, no experiment could distinguish between them, even in principle. *—J.D.B.*



our known 4-D physics. We do not yet know of any such 3-D theory that works in that way. Indeed, what surface should we use as the boundary of the universe? One step toward realizing these ideas is to study models that are simpler than our real universe.

A class of concrete examples of the holographic principle at work involves so-called anti–de Sitter spacetimes. The original de Sitter spacetime is a model universe first obtained by Dutch astronomer Willem de Sitter in 1917 as a solution of Einstein's equations, including the repulsive force known as the cosmological constant. De Sitter spacetime is empty, expands at an accelerating rate and is very highly symmetrical. In 1997 astronomers studying distant supernova explosions concluded that our universe now expands in an accelerated fashion and will probably become increasingly like a de Sitter spacetime in the future. Now, if the repulsive cosmological constant is replaced by an attractive one, de Sitter's solution turns into anti–de Sitter spacetime, which has equally as much symmetry. More important for the holographic concept, it possesses a boundary, which is located "at infinity" and is a lot like our everyday spacetime.

Using anti–de Sitter spacetime, theorists have devised a concrete example of the holographic principle at work: a universe described by superstring theory functioning in an anti–de Sitter spacetime is completely equivalent to a quantum field theory operating on the boundary of that spacetime [*see box above*]. Thus, the full majesty of superstring theory in an anti–de Sitter universe is painted on the boundary of the universe. Juan Maldacena, then at Harvard University, first conjectured such a relation in 1997 for the 5-D anti–de Sitter case, and it was later confirmed for many situations by Edward Witten of the Institute for Advanced Study in Princeton, N.J., and Steven S. Gubser, Igor R. Klebanov and Alexander M. Polyakov of Princeton University. Examples of this holographic correspondence are now known for spacetimes with a variety of dimensions.

This result means that two ostensibly very different theories—not even acting in spaces of the same dimension—are equivalent. Creatures living in one of these universes would be incapable of determining if they inhabited a 5-D universe described by string theory or a 4-D one described by a quantum field theory of point particles. (Of course, the structures of their brains might give them an overwhelming "commonsense" prejudice in favor of one description or another, in just the way that our brains construct an innate perception that our universe has three spatial dimensions; see the illustration on the opposite page.)

The holographic equivalence can allow a difficult calculation in the 4-D boundary spacetime, such as the behavior of quarks and gluons, to be traded for another, easier calculation in the highly symmetric, 5-D anti–de Sitter spacetime. The correspondence works the other way, too. Witten has shown that a black hole in anti–de Sitter spacetime corresponds to hot radiation in the alternative physics operating on the bounding spacetime. The entropy of the hole—a deeply mysterious concept—equals the radiation's entropy, which is quite mundane.

The Expanding Universe

HIGHLY SYMMETRIC and empty, the 5-D anti–de Sitter universe is hardly like our universe existing in 4-D, filled with matter and radiation and riddled with violent events. Even if we approximate our real universe with one that has matter and radiation spread uniformly throughout, we get not an anti–de Sitter universe but rather a "Friedmann-Robertson-Walker" universe. Most cosmologists today concur that our universe resembles an FRW universe, one that is infinite, has no boundary and will go on expanding ad infinitum.

Does such a universe conform to the holographic principle or the holographic bound? Susskind's argument based on collapse to a black hole is of no help here. Indeed, the holographic bound deduced from black holes must break down in a uniform expanding universe. The entropy of a region uniformly filled with matter and radiation is truly proportional to its volume. A sufficiently large region will therefore violate the holographic bound.

In 1999 Raphael Bousso, then at Stanford, proposed a modified holographic bound, which has since been found to work even in situations where the bounds we discussed earlier cannot be applied. Bousso's formulation starts with any suitable 2-D surface; it may be closed like a sphere or open like a sheet of paper. One then imagines a brief burst of light issuing simultaneously and perpendicularly from all over one side of the surface. The only demand is that the imaginary light rays are converging to start with. Light emitted from the inner surface of a spherical shell, for instance, satisfies that requirement. One then considers the entropy of the matter and radiation that these imaginary rays traverse, up to the points where they start crossing. Bousso conjectured that this entropy cannot exceed the entropy represented by the initial surface—one quarter of its area, measured in Planck areas. This is a different way of tallying up the entropy than that used in the original holographic bound. Bousso's bound refers not to the entropy of a region at one time but rather to the sum of entropies of locales at a variety of times: those that are "illuminated" by the light burst from the surface.

Bousso's bound subsumes other entropy bounds while avoiding their limitations. Both the universal entropy bound and the 't Hooft-Susskind form of the holographic bound can be deduced from Bousso's for any isolated system that is not evolving rapidly and whose gravitational field is not strong. When these conditions are overstepped—as for a collapsing sphere of matter already inside a black hole—these bounds eventually fail, whereas Bousso's bound continues to hold. Bousso has also shown that his strategy can be used to locate the 2-D surfaces on which holograms of the world can be set up.

Augurs of a Revolution

RESEARCHERS HAVE proposed many other entropy bounds. The proliferation of variations on the holographic motif makes it clear that the subject has not yet reached the status of physical law. But although the holographic way of thinking is not yet fully understood, it seems to be here to stay. And with it comes a realization that the fundamental belief, prevalent for 50 years, that field theory is the ultimate language of physics must give way. Fields, such as the electromagnetic field, vary continuously from point to point, and they thereby describe an infinity of degrees of freedom. Superstring theory also embraces an infinite number of degrees of freedom. Holography restricts the number of degrees of freedom that can be present inside a bounding surface to a finite number; field theory with its infinity cannot be the final story. Furthermore, even if the infinity is tamed, the mysterious dependence of information on surface area must be somehow accommodated.

Holography may be a guide to a better theory. What is the fundamental theory like? The chain of reasoning involving holography suggests to some, notably Lee Smolin of the Perimeter Institute for Theoretical Physics in Waterloo, Ontario, that such a final theory must be concerned not with fields, not even with spacetime, but rather with information exchange among physical processes. If so, the vision of information as the stuff the world is made of will have found a worthy embodiment. SA

OUR INNATE PERCEPTION that the world is three-dimensional could be an extraordinary illusion.

MORE TO EXPLORE

Black Hole Thermodynamics. Jacob D. Bekenstein in *Physics Today,* Vol. 33, No. 1, pages 24–31; January 1980.

Black Holes and Time Warps: Einstein's Outrageous Legacy. Kip S. Thorne. W. W. Norton, 1995.

Black Holes and the Information Paradox. Leonard Susskind in *Scientific American,* Vol. 276, No. 4, pages 52–57; April 1997.

The Universe in a Nutshell. Stephen Hawking. Bantam Books, 2001.

Three Roads to Quantum Gravity. Lee Smolin. Basic Books, 2002.

From the fixed past to the tangible present to the undecided future, it feels as though time flows inexorably on. But that is an illusion **By Paul Davies**

THAT MYSTERIOUS FLOW

OVERVIEW

- Our senses tell us that time flows: namely, that the past is fixed, the future undetermined, and reality lives in the present. Yet various physical and philosophical arguments suggest otherwise.
- The passage of time is probably an illusion. Consciousness may involve thermodynamic or quantum processes that lend the impression of living moment by moment.

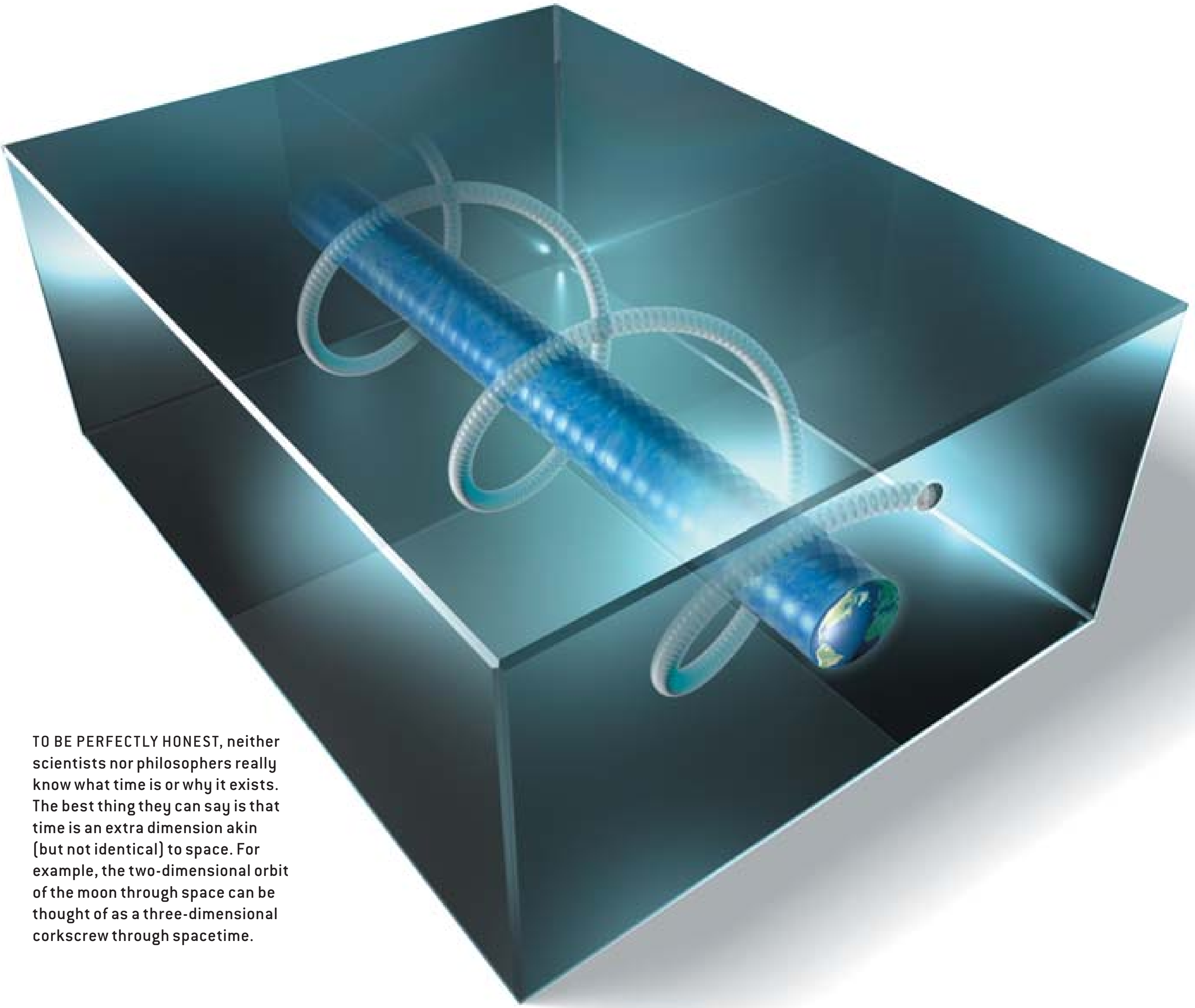
"Gather ye rosebuds while ye may, / Old Time is still a-flying."

So wrote 17th-century English poet Robert Herrick, capturing the universal cliché that time flies. And who could doubt that it does? The passage of time is probably the most basic facet of human perception, for we feel time slipping by in our innermost selves in a manner that is altogether more intimate than our experience of, say, space or mass. The passage of time has been compared to the flight of an arrow and to an ever rolling stream, bearing us inexorably from past to future. Shakespeare wrote of "the whirligig of time," his countryman Andrew Marvell of "Time's winged chariot hurrying near."

Evocative though these images may be, they run afoul of a deep and devastating paradox. Nothing in known physics corresponds to the passage of time. Indeed, physicists insist that time doesn't flow at all; it merely is. Some philosophers argue that the very notion of the passage of time is nonsensical and that talk of the river or flux of time is founded on a misconception. How can something so basic to our experience of the physical world turn out to be a case of mistaken identity? Or is there a key quality of time that science has not yet identified?

Time Isn't of the Essence

IN DAILY LIFE we divide time into three parts: past, present and future. The grammatical structure of language revolves around this fundamental distinction. Reality is associated with the present moment. The past we think of as having slipped out of existence, whereas the future is even more shadowy, its details still unformed. In this simple picture, the "now" of our conscious awareness glides steadily onward,



TO BE PERFECTLY HONEST, neither scientists nor philosophers really know what time is or why it exists. The best thing they can say is that time is an extra dimension akin (but not identical) to space. For example, the two-dimensional orbit of the moon through space can be thought of as a three-dimensional corkscrew through spacetime.

transforming events that were once in the unformed future into the concrete but fleeting reality of the present, and thence relegating them to the fixed past.

Obvious though this commonsense description may seem, it is seriously at odds with modern physics. Albert Einstein famously expressed this point when he wrote to a friend, "The past, present and future are only illusions, even if stubborn ones." Einstein's startling conclusion stems directly from his special theory of relativity, which denies any absolute, universal significance to the present moment. According to the theory, simultaneity is relative. Two events that occur at the same moment if observed from one reference frame may occur at different moments if viewed from another.

An innocuous question such as "What is happening on Mars now?" has no definite answer. The key point is that Earth and Mars are a long way apart—up to about 20 light-minutes. Because information cannot travel faster than light, an Earth-based observer is unable to know the situation on Mars at the same instant. He must infer the answer after the event, when light has had a chance to pass between the planets. The inferred past event will be different depending on the observer's velocity.

For example, during a future manned expedition to Mars, mission controllers back on Earth might say, "I wonder what Commander Jones is doing at Alpha Base now." Looking at their clock and seeing that it was 12:00 p.m. on Mars, their answer might be "Eating lunch." But an astronaut zooming past Earth at near the speed of light at the same moment could, on looking at his clock, say that the time on Mars was earlier or later than 12:00, depending on his direction of motion. That astronaut's answer to the question about Commander Jones's activities would be "Cooking lunch" or "Washing dishes" [*see box on page 86*]. Such mismatches make a mockery of any attempt to confer special status on the present moment, for whose "now" does that moment refer to? If you and I were in relative motion, an event that I might

BRYAN CHRISTIE DESIGN

judge to be in the as yet undecided future might for you already exist in the fixed past.

The most straightforward conclusion is that both past and future are fixed. For this reason, physicists prefer to think of time as laid out in its entirety—a timescape, analogous to a landscape—with all past and future events located there together. It is a notion sometimes referred to as block time. Completely absent from this description of nature is anything that singles out a privileged special moment as the present or any process that would systematically turn future events into present, then past, events. In short, the time of the physicist does not pass or flow.

> Physicists think of time as laid out in its entirety—a timescape, analogous to a landscape.

How Time Doesn't Fly

A NUMBER OF PHILOSOPHERS over the years have arrived at the same conclusion by examining what we normally mean by the passage of time. They argue that the notion is internally inconsistent. The concept of flux, after all, refers to motion. It makes sense to talk about the movement of a physical object, such as an arrow through space, by gauging how its location varies with time. But what meaning can be attached to the movement of time itself? Relative to what does it move? Whereas other types of motion relate one physical process to another, the putative flow of time relates time to itself. Posing the simple question "How fast does time pass?" exposes the absurdity of the very idea. The trivial answer "One second per second" tells us nothing at all.

Although we find it convenient to refer to time's passage in everyday affairs, the notion imparts no new information that cannot be conveyed without it. Consider the following scenario: Alice was hoping for a white Christmas, but when the day came she was disappointed that it only rained; however, she was happy that it snowed the following day. Although this description is replete with tenses and references to time's passage, exactly the same information is conveyed by simply correlating Alice's mental states with dates, in a manner that omits all reference to time passing or the world changing. Thus, the following cumbersome and rather dry catalogue of facts suffices:

December 24: Alice hopes for a white Christmas.
December 25: There is rain. Alice is disappointed.
December 26: There is snow. Alice is happy.

In this description, nothing happens or changes. There are simply states of the world at different dates and associated mental states for Alice.

Similar arguments go back to ancient Greek philosophers such as Parmenides and Zeno. A century ago British philosopher John McTaggart sought to draw a clear distinction between the description of the world in terms of events happening, which he called the A series, and the description in terms of dates correlated with states of the world, the B series. Each seems to be a true description of reality, and yet the two points of view are seemingly in contradiction. For example, the event "Alice is disappointed" was once in the future, then in the present and afterward in the past. But past, present and future are exclusive categories, so how can a single event have the character of belonging to all three? McTaggart used this clash between the A and B series to argue for the unreality of time as such, perhaps a rather drastic conclusion. Most physicists would put it less dramatically: the flow of time is unreal, but time itself is as real as space.

Just in Time

A GREAT SOURCE of confusion in discussions of time's passage stems from its link with the so-called arrow of time. To deny that time flows is not to claim that the designations "past" and "future" are without physical basis. Events in the world undeniably form a unidirectional sequence. For instance, an egg dropped on the floor will smash into pieces, whereas the reverse process—a broken egg spontaneously assembling itself into an intact egg—is never witnessed. This is an example of the second law of thermodynamics, which states that the entropy of a closed system—roughly defined as

NOBODY REALLY KNOWS . . .

What Is Time, Anyway?

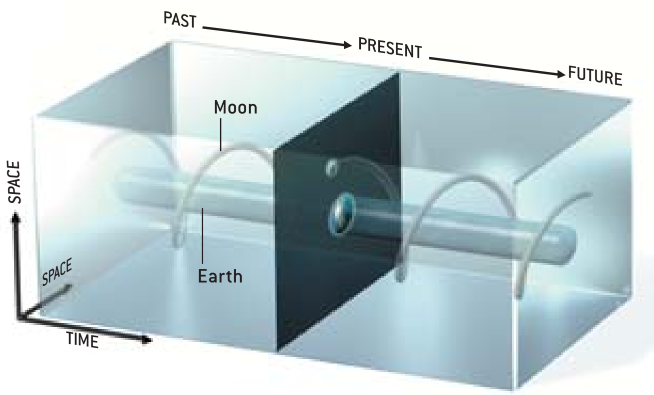
Saint Augustine of Hippo, the famous fifth-century theologian, remarked that he knew well what time is—until somebody asked. Then he was at a loss for words. Because we sense time psychologically, definitions of time based on physics seem dry and inadequate. For the physicist, time is simply what (accurate) clocks measure. Mathematically, it is a one-dimensional space, usually assumed to be continuous, although it might be quantized into discrete "chronons," like frames of a movie.

The fact that time may be treated as a fourth dimension does not mean that it is identical to the three dimensions of space. Time and space enter into daily experience and physical theory in distinct ways. For instance, the formula for calculating spacetime distances is not the same as the one for calculating spatial distances. The distinction between space and time underpins the key notion of causality, stopping cause and effect from being hopelessly jumbled. On the other hand, many physicists believe that on the very smallest scale of size and duration, space and time might lose their separate identities. *—P.D.*

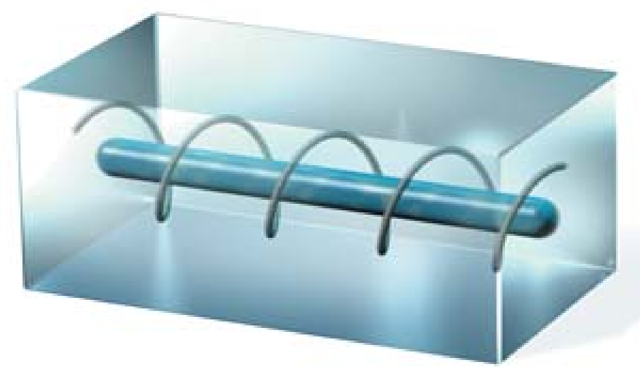
All Time Like the Present

According to conventional wisdom, the present moment has special significance. It is all that is real. As the clock ticks, the moment passes and another comes into existence—a process that we call the flow of time. The moon, for example, is located at only one position in its orbit around Earth. Over time, it ceases to exist at that position and is instead found at a new position.

Researchers who think about such things, however, generally argue that we cannot possibly single out a present moment as special when every moment considers itself to be special. Objectively, past, present and future must be equally real. All of eternity is laid out in a four-dimensional block composed of time and the three spatial dimensions. (This diagram shows only two of these spatial dimensions.) —*P.D.*



CONVENTIONAL VIEW: Only the present is real



BLOCK UNIVERSE: All times are equally real

how disordered it is—will tend to rise with time. An intact egg has lower entropy than a shattered one.

Because nature abounds with irreversible physical processes, the second law of thermodynamics plays a key role in imprinting on the world a conspicuous asymmetry between past and future directions along the time axis. By convention, the arrow of time points toward the future. This does not imply, however, that the arrow is moving toward the future, any more than a compass needle pointing north indicates that the compass is traveling north. Both arrows symbolize an asymmetry, not a movement. The arrow of time denotes an asymmetry of the world in time, not an asymmetry or flux of time. The labels "past" and "future" may legitimately be applied to temporal directions, just as "up" and "down" may be applied to spatial directions, but talk of the past or the future is as meaningless as referring to the up or the down.

The distinction between pastness or futureness and "the" past or "the" future is graphically illustrated by imagining a movie of, say, the egg being dropped on the floor and breaking. If the film were run backward through the projector, everyone would see that the sequence was unreal. Now imagine if the film strip were cut up into frames and the frames shuffled randomly. It would be a straightforward task for someone to rearrange the stack of frames into a correctly ordered sequence, with the broken egg at the top of the stack and the intact egg at the bottom. This vertical stack retains the asymmetry implied by the arrow of time because it forms an ordered sequence in vertical space, proving that time's asymmetry is actually a property of states of the world, not a property of time as such. It is not necessary for the film actually to be run as a movie for the arrow of time to be discerned.

Given that most physical and philosophical analyses of time fail to uncover any sign of a temporal flow, we are left with something of a mystery. To what should we attribute the powerful, universal impression that the world is in a continual state of flux? Some researchers, notably the late Nobel laureate chemist Ilya Prigogine, have contended that the subtle physics of irreversible processes make the flow of time an objective aspect of the world. But I and others argue that it is some sort of illusion.

After all, we do not really observe the passage of time. What we actually observe is that later states of the world differ from earlier states that we still remember. The fact that we remember the past, rather than the future, is an observation not of the passage of time but of the asymmetry of time. Nothing other than a conscious observer registers the flow of time. A clock measures durations between events much as a measuring tape

BRYAN CHRISTIE DESIGN

THE AUTHOR

PAUL DAVIES is a theoretical physicist and professor of natural philosophy at Macquarie University's Australian Center for Astrobiology in Sydney. He is one of the most prolific writers of popular-level books in physics. His scientific research interests include black holes, quantum field theory, the origin of the universe, the nature of consciousness and the origin of life.

SIMULTANEITY

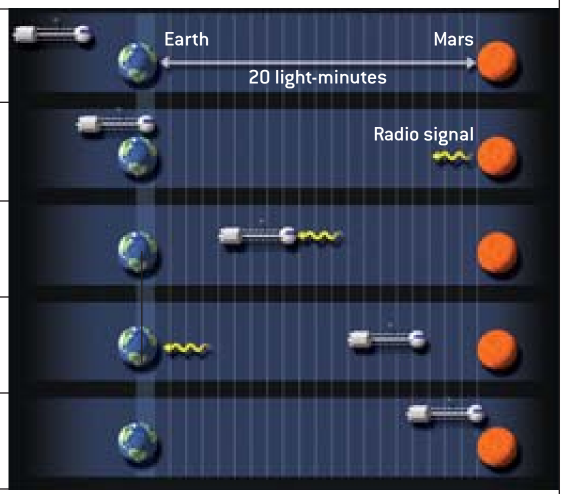
It's All Relative

What is happening on Mars right now? Such a simple question, such a complex answer. The trouble stems from the phrase "right now." Different people, moving at different velocities, have different perceptions of what the present moment is. This strange fact is known as the relativity of simultaneity. In the following scenario, two people—an Earthling sitting in Houston and a rocketman crossing the solar system at 80 percent of the speed of light—attempt to answer the question of what is happening on Mars right now. A resident of Mars has agreed to eat lunch when his clock strikes 12:00 P.M. and to transmit a signal at the same time. *—P.D.*

As Seen from Earth

From the Earthling's perspective, Earth is standing still, Mars is a constant distance (20 light-minutes) away, and the rocket ship is moving at 80 percent of the speed of light. The situation looks exactly the same to the Martian.

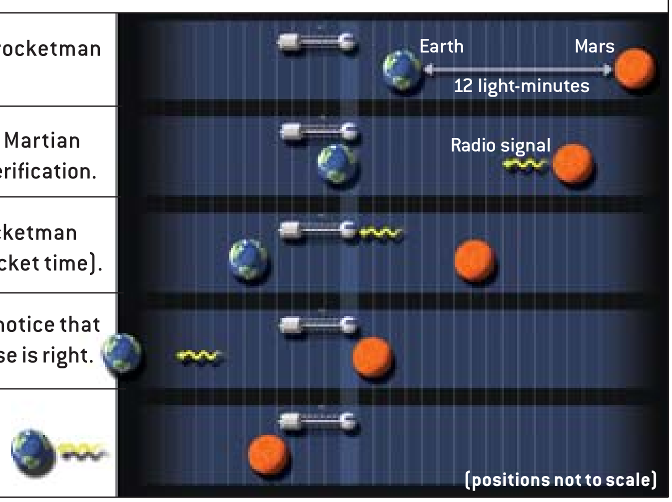
Before noon	By exchanging light signals, the Earthling and Martian measure the distance between them and synchronize their clocks.
12:00 P.M.	The Earthling hypothesizes that the Martian has begun to eat lunch. He prepares to wait 20 minutes for verification.
12:11 P.M.	Knowing the rocket's speed, the Earthling deduces that it encounters the signal while on its way to Mars.
12:20 P.M.	The signal arrives at Earth. The Earthling has confirmed his earlier hypothesis. Noon on Mars is the same as noon on Earth.
12:25 P.M.	The ship arrives at Mars.



As Seen from the Rocket

From the rocketman's perspective, the rocket is standing still. It is the planets that are hurtling through space at 80 percent of the speed of light. His measurements show the two planets to be separated by 12 light-minutes—a different distance than the Earthling inferred. This discrepancy, a well-known effect of Einstein's theory, is called length contraction. A related effect, time dilation, causes clocks on the ship and planets to run at different rates. (The Earthling and Martian think the ship's clock is slow; the rocketman thinks the planets' are.) As the ship passes Earth, it synchronizes its clock to Earth's.

Before noon	By exchanging light signals with his colleagues, the rocketman measures the distance between the planets.
12:00 P.M.	Passing Earth, the rocketman hypothesizes that the Martian has begun to eat. He prepares to wait 12 minutes for verification.
12:07 P.M.	The signal arrives, disproving the hypothesis. The rocketman infers that the Martian ate sometime before noon (rocket time).
12:15 P.M.	Mars arrives at the ship. The rocketman and Martian notice that their two clocks are out of sync but disagree as to whose is right.
12:33 P.M.	The signal arrives at Earth. The clock discrepancies demonstrate that there is no universal present moment.



BRYAN CHRISTIE DESIGN

measures distances between places; it does not measure the "speed" with which one moment succeeds another. Therefore, it appears that the flow of time is subjective, not objective.

Living in the Present

THIS ILLUSION CRIES OUT for explanation, and that explanation is to be sought in psychology, neurophysiology, and maybe linguistics or culture. Modern science has barely begun to consider the question of how we perceive the passage of time; we can only speculate about the answer. It might have something to do with the functioning of the brain. If you spin around several times and stop suddenly, you will feel giddy. Subjectively, it seems as if the world is rotating relative to you, but the evidence of your eyes is clear enough: it is not. The apparent movement of your surroundings is an illusion created by the rotation of fluid in the inner ear. Perhaps temporal flux is similar.

There are two aspects to time asymmetry that might create the false impression that time is flowing. The first is the thermodynamic distinction between past and future. As physicists have realized over the past few decades, the concept of entropy is closely related to the information content of a system. For this reason, the formation of memory is a unidirectional process—new memories add information and raise the entropy of the brain. We might perceive this unidirectionality as the flow of time.

A second possibility is that our perception of the flow of time is linked in some way to quantum mechanics. It was appreciated from the earliest days of the formulation of quantum mechanics that time enters into the theory in a unique manner, quite unlike space. The special role of time is one reason it is proving so difficult to merge quantum mechanics with general relativity. Heisenberg's uncertainty principle, according to which nature is inherently indeterministic, implies an open future (and, for that matter, an open past). This indeterminism manifests itself most conspicuously on an atomic scale of size and dictates that the observable properties that characterize a physical system are generally undecided from one moment to the next.

For example, an electron hitting an atom may bounce off in one of many directions, and it is normally impossible to predict in advance what the outcome in any given case will be. Quantum indeterminism implies that for a particular quantum state there are many (possibly infinite) alternative futures or potential realities. Quantum mechanics supplies the relative probabilities for each observable outcome, although it won't say which potential future is destined for reality.

But when a human observer makes a measurement, one and only one result is obtained; for example, the rebounding electron will be found moving in a certain direction. In the act of measurement, a single, specific reality gets projected out from a vast array of possibilities. Within the observer's mind, the possible makes a transition to the actual, the open future to the fixed past—which is precisely what we mean by the flux of time.

There is no agreement among physicists on how this transition from many potential realities into a single actuality takes place. Many physicists have argued that it has something to do with the consciousness of the observer, on the basis that it is the act of observation that prompts nature to make up its mind. A few researchers, such as Roger Penrose of the University of Oxford, maintain that consciousness—including the impression of temporal flux—could be related to quantum processes in the brain.

Although researchers have failed to find evidence for a single "time organ" in the brain, in the manner of, say, the visual cortex, it may be that future work will pin down those brain processes responsible for our sense of temporal passage. It is possible to imagine drugs that could suspend the subject's impression that time is passing. Indeed, some practitioners of meditation claim to be able to achieve such mental states naturally.

And what if science were able to explain away the flow of time? Perhaps we would no longer fret about the future or grieve for the past. Worries about death might become as irrelevant as worries about birth. Expectation and nostalgia might cease to be part of human vocabulary. Above all, the sense of urgency that attaches to so much of human activity might evaporate. No longer would we be slaves to Henry Wadsworth Longfellow's entreaty to "act, act in the living present," for the past, present and future would literally be things of the past. SA

> **Modern science** has barely begun to consider the question of how we perceive the passage of time.

MORE TO EXPLORE

The Unreality of Time. John Ellis McTaggart in *Mind,* Vol. 17, pages 456–473; 1908.

Can Time Go Backward? Martin Gardner in *Scientific American,* Vol. 216, No. 1, pages 98–108; January 1967.

What Is Time? G. J. Whitrow. Thames & Hudson, 1972.

The Physics of Time Asymmetry. Paul Davies. University of California Press, 1974.

Time and Becoming. J.J.C. Smart in *Time and Cause.* Edited by Peter van Inwagen. Reidel Publishing, 1980.

About Time: Einstein's Unfinished Revolution. Paul Davies. Simon & Schuster, 1995.