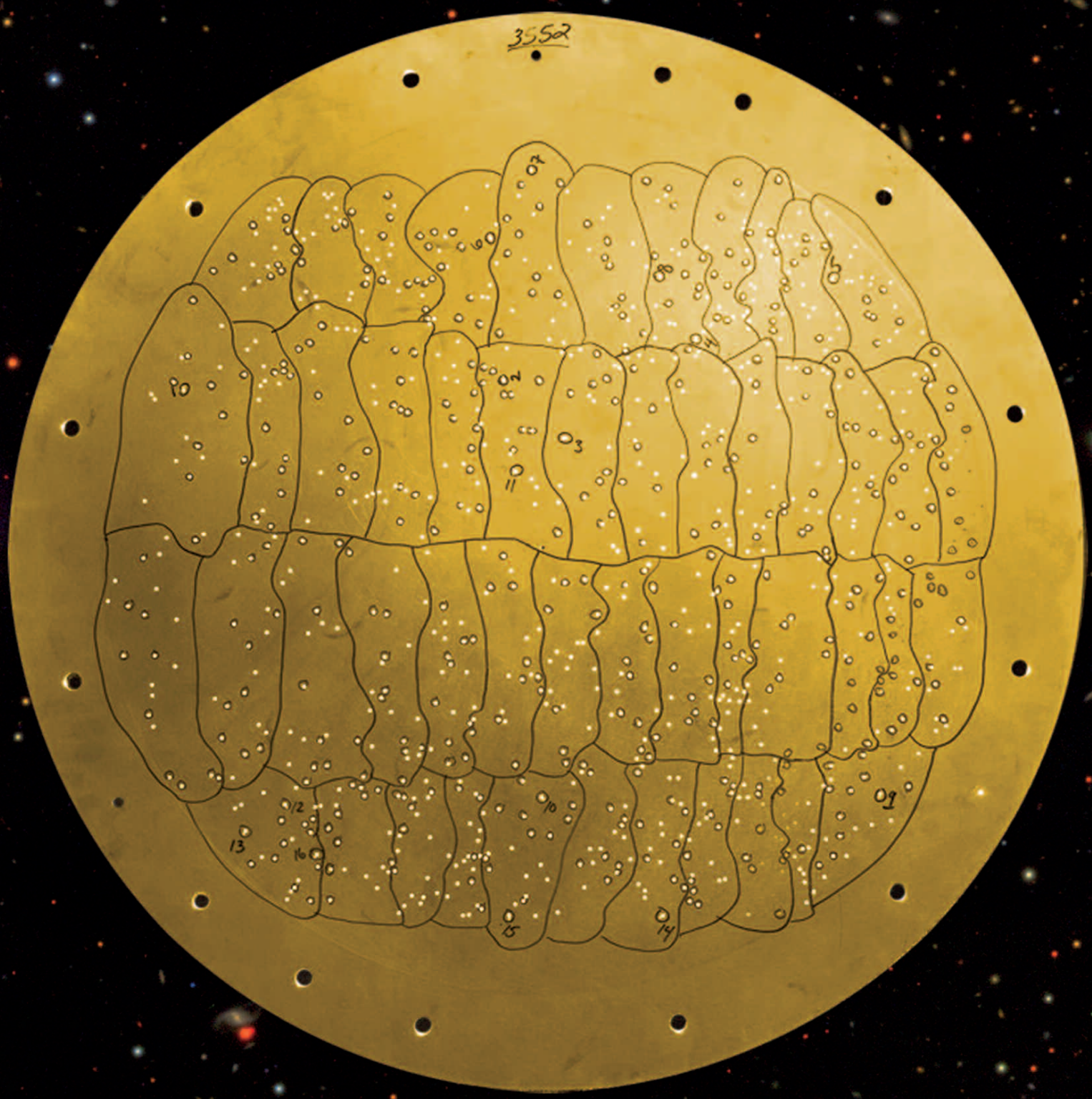
IAS The Institute Letter

Special Edition

Institute for Advanced Study



Entanglement and Spacetime • Detecting Gravitons • Beyond the Higgs
Black Holes and the Birth of Galaxies • Mapping the Early Universe • Measuring the Cosmos • Life on Other Planets • Univalent Foundations of Mathematics
Finding Structure in Big Data • Randomness and Pseudorandomness • The Shape of Data

Entanglement and the Geometry of Spacetime

Can the weird quantum mechanical property of entanglement give rise to wormholes connecting far away regions in space?

BY JUAN MALDACENA

In 1935, Albert Einstein and collaborators wrote two papers at the Institute for Advanced Study. One was on quantum mechanics[1] and the other was on black holes.[2] The paper on quantum mechanics is very famous and influential. It pointed out a feature of quantum mechanics that deeply troubled Einstein. The paper on black holes pointed out an interesting aspect of a black hole solution with no matter, where the solution looks like a wormhole connecting regions of spacetime that are far away. Though these papers seemed to be on two completely disconnected subjects, recent research has suggested that they are closely connected.

Einstein's theory of general relativity tells us that spacetime is dynamical. Spacetime is similar to a rubber sheet that can be deformed by the presence of matter. A very drastic deformation of spacetime is the formation of a black hole. When there is a large amount of matter concentrated in a small enough region of space, this can collapse in an irreversible fashion. For example, if we filled a sphere the size of the solar system with air, it would collapse into a black hole. When a black hole forms, we can define an imaginary surface called "the horizon"; it separates the region of spacetime that can send signals to the exterior from the region that cannot. If an astronaut crosses the horizon, she can never come back out. She does not feel anything special as she crosses the horizon. However, once she crosses, she will be inevitably crushed by the force of gravity into a region called "the singularity" (Figure 1a).

Outside of the distribution of collapsing matter, black holes are described by a spacetime solution found by Karl Schwarzschild in 1916. This solution turned out to be very confusing, and a full understanding of its classical aspects had to wait until the 1960s. The original Schwarzschild solution contains no matter (Figure 1b). It is just vacuum everywhere, but it has both future and past singularities. In 1935, Einstein and Rosen found a curious aspect of this solution: it contains two regions that look like the outside of a black hole. Namely, one starts with a spacetime that is flat at a far distance. As we approach the central region, spacetime is deformed with the same deformation that is generated outside a massive object. At a fixed time, the geometry of space is such that as we move in toward the center, instead of finding a massive object, we find a second asymptotic region (Figure 1c). The geometry of space looks like a wormhole connecting two asymptotically flat regions. This is sometimes called the Einstein–Rosen bridge. They realized this before the full geometry was properly understood. Their motivation was to find a model for elementary particles where particles were represented by smooth geometries. We now think that their original motivation was misguided. This geometry can also be interpreted as a kind of wormhole that connects two distant regions in the same spacetime. John Wheeler and Robert Fuller showed that these wormholes are not traversable, meaning it is not possible to physically travel from one side of the wormhole to the other.[3] We can think of this configuration as a pair of distant black holes. Each black hole has its own horizon. But it is a very particular pair since they are connected through the horizon. The distance from one horizon to the other through the wormhole is zero at one instant of time. Let us consider two observers, Alice and Bob, outside each of the black holes. For a brief moment in time, the horizons of the two black holes touch, then they move away from each other. Alice cannot send a signal to Bob if she stays outside the horizon of her black hole. However, Alice and Bob could both jump into their respective black holes and meet inside. It would be a fatal meeting since they would then die at the singularity. This is a fatal attraction.

Wormholes usually appear in science fiction books or movies as devices that allow us to travel faster than light between very distant points. These are different than the wormhole discussed above. In fact, these science-fiction wormholes would require a type of matter with negative energy, which does not appear to be possible in consistent physical theories.

In black holes that form from collapse, only a part of the Schwarzschild geometry is present, since the presence of matter changes the solution. This case is fairly well understood and there is no wormhole. However, one can still ask about the physical interpretation of the solution with the two asymptotic regions. It is, after all, the general spherically symmetric vacuum solution of general relativity. Surprisingly, the interpretation of this solution involves the paper by Einstein, Podolsky, and Rosen (EPR) written in 1935.[1] By the way, the EPR paper shows that Einstein really did very influential work after he came to the IAS.

The EPR paper pointed out that quantum mechanics had a very funny property later called "quantum entanglement,"or, in short, "entanglement." Entanglement is a kind of correlation between two distant physical systems. Of course, correlations between distant systems can exist in classical systems. For example, if I have one glove in my jacket and one in my house, then if one is a left glove, the other will be a right glove. However, entanglement involves correlations between quantum variables. Quantum variables are properties that cannot be known at the same time; they are subject to the Heisenberg uncertainty principle. For example, we cannot know both the position and the velocity of a particle with great precision. If we measure the position very precisely, then the velocity becomes uncertain. Now, the idea in the EPR paper is that we have two distant systems; in each distant system, we can measure two variables that are subject to the uncertainty principle. However, the total state could be such that the results of distant measurements are always perfectly correlated, when they both measure the same variable. The EPR example was the following (Figure 2). Consider a pair of equal-mass particles with a well-defined center of mass, say $x = 0$, and also with a well-defined relative velocity, say $v_{rel} = v_A - v_B$. First, a small clarification. The Heisenberg uncertainty principle says that the position and the velocity cannot be known at the same time. When we have two independent dynamical variables—two independent positions and two independent velocities—then it is possible to know the position of one and the velocity of the other. Since the center of mass and relative position are independent variables, then it is indeed possible to start with the state that EPR postulated. Now for the more surprising part: let us say that two distant observers, call them Alice and Bob, both measure the positions of the respective particles. They find that if Alice measures some value x_A, then Bob should measure $x_B = -x_A$. On the other hand, if Alice measures the velocity v_A, then we know that Bob should measure the definite velocity $v_B = v_A - v_{rel}$. Of course, Alice and Bob should each make a choice of whether they want to measure the velocity or the position. If Alice measures the position and Bob the velocity, they find uncorrelated results. Note that when Alice decides to measure the position, Bob's particle, which could be very distant, seems to "decide" to have a well-defined position

(Continued on page 3)

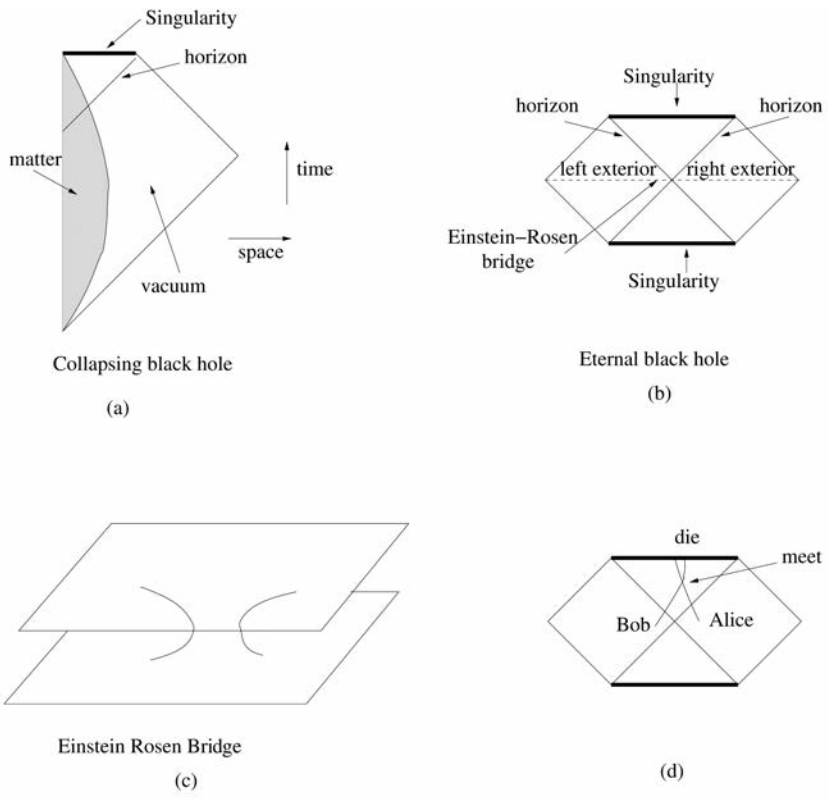


Fig. 1: In (a) we see a spacetime diagram representing a collapsing black hole. We only represent the time and radial directions. At each point in the diagrams, there is also a sphere. Forty-five degree lines represent the direction of propagation of light signals. The shaded region contains the collapsing matter. The thick line is the "singularity." Black holes that are actually formed in our universe are represented by such spacetimes. In (b) we see the diagram for the full eternal black hole. It has two asymptotically flat regions, one to the left and one to the right. Each of these regions looks like the outside of a black hole. The Einstein-Rosen bridge is the spatial section of the geometry that connects the two sides at a special moment in time. In (c) we represent the spatial geometry of the Einstein-Rosen bridge. In (d) we have Bob and Alice jump into their respective black holes and meet in the interior.

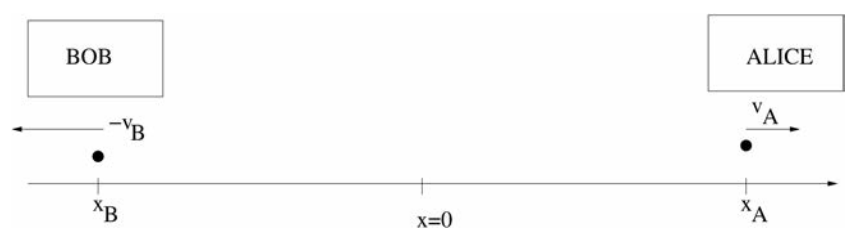


Fig. 2: The set up for the EPR experiment. We generate two particles at x = 0 moving with relative velocity v_{rel}. Alice and Bob are two distant observers who can choose to measure either the position or velocity of their respective particles. They do this many times, always starting from the same state. Alice finds a very random set of possible results for the measurement of the position of her particle and a similarly random set when she decides to measure the velocity. Similar results hold for Bob. However, when both Alice and Bob measure the positions, they are perfectly correlated. The same holds when they both measure the velocity.

also. On the other hand, when Alice measures the velocity, Bob's particle "decides" to have a well-defined velocity. At first sight, this would seem to allow instantaneous communication between Alice and Bob. It would seem that Alice can encode a message of zeros and ones by deciding to measure either her particle's position or velocity and then all that Bob has to do is to see whether his particle has well-defined position or velocity. However, it is possible to show that Bob cannot "read" such a message. These correlations do not allow us to send signals faster than light.

Entanglement appears to be a very esoteric property of quantum mechanical systems. But in the past twenty years, people have found many practical uses for these correlations. Among them is the possibility of Alice and Bob communicating secretly while making sure that the NSA (National Security Agency) is not eavesdropping on the communication.

Let us now return to black holes. There is an important feature of black holes that arises when one considers them as quantum mechanical objects. In 1974, Stephen Hawking argued that quantum mechanics implies that black holes have a temperature, with smaller black holes having a higher temperature. A small enough black hole can be red-hot. In fact, one can even have a white black hole! This is a theoretical prediction that has not yet been verified experimentally because the black holes that are naturally produced by the collapse of stars are too cold for this radiation to be measurable. This thermal property of black holes has an important consequence. As we have known since the nineteenth century, temperature is due to the motion of a large number of microscopic constituents of the system. Thus, black holes should have microscopic constituents that can be in a large number of possible quantum mechanical configurations or "microstates." In fact, we think that black holes, as seen from the outside, behave as ordinary quantum mechanical systems.

One can consider, therefore, a pair of black holes where all the microstates are "entangled." Namely, if we observe one of the black holes in one particular microstate, then the other has to be in exactly the same microstate. A pair of black holes in this particular EPR entangled state would develop a wormhole, or Einstein-Rosen bridge, connecting them through the inside. The geometry of this wormhole is given by the fully extended Schwarzschild geometry. It is interesting that both wormholes and entanglement naively appear to lead to a propagation of signals faster than light. But in either case this is not true, for different detailed reasons. The net result is the same: we cannot use either of them to send signals faster than light. This picture was developed through the years starting with work by Werner Israel.[4] Most recently, Leonard Susskind and I emphasized this ER=EPR connection as a way to resolve some apparent paradoxes regarding the black hole interior.[5, 6]

There are several interesting lessons regarding this picture of geometry emerging from entanglement. Perhaps the deepest one is that the peculiar and strange property of quantum mechanical entanglement is behind the beautiful continuity of spacetime. In other words, the solid and reliable structure of spacetime is due to the ghostly features of entanglement. As we entangle two systems with many degrees of freedom, it seems possible to generate a geometric connection between them, even though there is no direct interaction between the two systems. ■

Juan Maldacena, who first came to the Institute as a Member in 1999, has been a Professor in the School of Natural Sciences since 2002. He continues to study a relationship he has proposed between quantum gravity and quantum field theories in an effort to further understand the deep connection between black holes and quantum field theories as well as connections between string theory and cosmology.

Recommended Reading: Dennis Overbye writes about the latest debates involving the quantum mechanical property of entanglement—originating with the EPR paper and arriving at Juan Maldacena's most recent findings with Leonard Susskind—in a recent *New York Times* article, "A Black Hole Mystery Wrapped in a Firewall Paradox"; visit http://ow.ly/nWftw/.

1 "Can Quantum-Mechanical Description of Physical Reality be Considered Complete?" Albert Einstein, Boris Podolsky, Nathan Rosen (Princeton, Institute for Advanced Study), *Physical Review* 47 (1935) 777–80.

2 "The Particle Problem in the General Theory of Relativity," Albert Einstein, Nathan Rosen (Princeton, Institute for Advanced Study), *Physical Review* 48 (1935) 73–77.

3 "Causality and Multiply Connected Space-Time," Robert W. Fuller (Columbia University), John A. Wheeler (Princeton University), *Physical Review* 128 (1962) 919–29.

4 "Thermo Field Dynamics of Black Holes," Werner Israel (Cambridge University, D.A.M.T.P.), *Physics Letters A* 57 (1976) 107–10.

5 "Cool Horizons for Entangled Black Holes," Juan Maldacena (Princeton, Institute for Advanced Study), Leonard Susskind (Stanford University, Institute of Theoretical Physics and Department of Physics), Jun 3, 2013. e-Print: arXiv:1306.0533.

6 "The Black Hole Interior in AdS/CFT and the Information Paradox," Kyriakos Papadodimas, Suvrat Raju. e-Print: arXiv:1310.6334.

From the Spring 2013 Issue

How Incompatible Worldviews Can Coexist

BY FREEMAN DYSON

John Brockman, founder and proprietor of the Edge website, asks a question every New Year and invites the public to answer it. THE EDGE QUESTION 2012 was "What is your favorite deep, elegant, or beautiful explanation?" He got 150 answers that are published in a book, This Explains Everything *(Harper Collins, 2013). Here is my contribution.*

The situation that I am trying to explain is the existence side by side of two apparently incompatible pictures of the universe. One is the classical picture of our world as a collection of things and facts that we can see and feel, dominated by universal gravitation. The other is the quantum picture of atoms and radiation that behave in an unpredictable fashion, dominated by probabilities and uncertainties. Both pictures appear to be true, but the relationship between them is a mystery.

LIGO LIVINGSTON OBSERVATORY

The LIGO Livingston Observatory in Louisiana

The orthodox view among physicists is that we must find a unified theory that includes both pictures as special cases. The unified theory must include a quantum theory of gravitation, so that particles called gravitons must exist, combining the properties of gravitation with quantum uncertainties.

I am looking for a different explanation of the mystery. I ask the question, whether a graviton, if it exists, could conceivably be observed. I do not know the answer to this question, but I have one piece of evidence that the answer may be no. The evidence is the behavior of one piece of apparatus, the gravitational wave detector called LIGO that is now operating in Louisiana and in Washington State. The way LIGO works is to measure very accurately the distance between two mirrors by bouncing light from one to the other. When a gravitational wave comes by, the distance between the two mirrors will change very slightly. Because of ambient and instrumental noise, the actual LIGO detectors can only detect waves far stronger than a single graviton. But even in a totally quiet universe, I can answer the question, whether an ideal LIGO detector could detect a single graviton. The answer is no. In a quiet universe, the limit to the accuracy of measurement of distance is set by the quantum uncertainties in the positions of the mirrors. To make the quantum uncertainties small, the mirrors must be heavy. A simple calculation, based on the known laws of gravitation and quantum mechanics, leads to a striking result. To detect a single graviton with a LIGO apparatus, the mirrors must be exactly so heavy that they will attract each other with irresistible force and collapse into a black hole. In other words, nature herself forbids us to observe a single graviton with this kind of apparatus.

I propose as a hypothesis, based on this single thought-experiment, that single gravitons may be unobservable by any conceivable apparatus.

If this hypothesis were true, it would imply that theories of quantum gravity are untestable and scientifically meaningless. The classical universe and the quantum universe could then live together in peaceful coexistence. No incompatibility between the two pictures could ever be demonstrated. Both pictures of the universe could be true, and the search for a unified theory could turn out to be an illusion. ■

Recommended Reading: Freeman Dyson was awarded the 2012 Henri Poincaré Prize at the International Mathematical Physics Congress. On this occasion, he delivered the lecture "Is a Graviton Detectable?" a PDF of which is available at http://publications.ias.edu/poincare2012/dyson.pdf.

Freeman Dyson, Professor Emeritus in the School of Natural Sciences, first came to the Institute as a Member in 1948 and was appointed a Professor in 1953. His work on quantum electrodynamics marked an epoch in physics. The techniques he used form the foundation for most modern theoretical work in elementary particle physics and the quantum many-body problem. He has made highly original and important contributions to an astonishing range of topics, from number theory to adaptive optics.

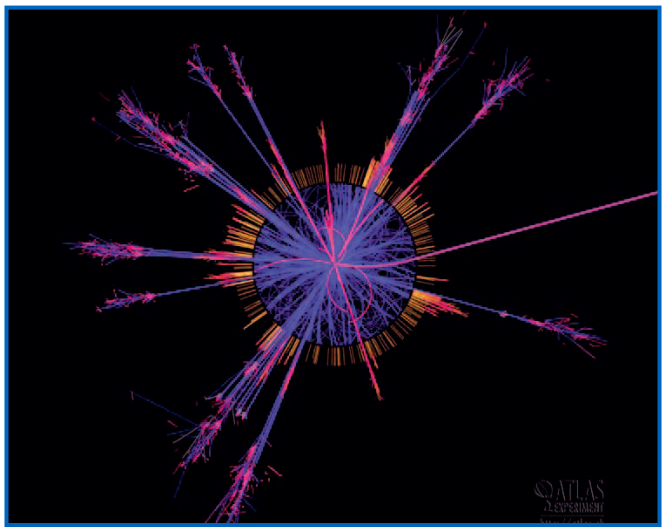
Discovering the Higgs: Inevitability, Rigidity, Fragility, Beauty

Following the discovery in July of a Higgs-like boson—an effort that took more than fifty years of experimental work and more than 10,000 scientists and engineers working on the Large Hadron Collider—Juan Maldacena and Nima Arkani-Hamed, two Professors in the School of Natural Sciences, gave separate public lectures on the symmetry and simplicity of the laws of physics, and why the discovery of the Higgs was inevitable.

Peter Higgs, who predicted the existence of the particle, gave one of his first seminars on the topic at the Institute in 1966, at the invitation of Freeman Dyson. "The discovery attests to the enormous importance of fundamental, deep ideas, the substantial length of time these ideas can take to come to fruition, and the enormous impact they have on the world," said Robbert Dijkgraaf, Director and Leon Levy Professor.

In their lectures "The Symmetry and Simplicity of the Laws of Nature and the Higgs Boson" and "The Inevitability of Physical Laws: Why the Higgs Has to Exist," Maldacena and Arkani-Hamed described the theoretical ideas that were developed in the 1960s and 70s, leading to our current understanding of the Standard Model of particle physics and the recent discovery of the Higgs-like boson. Arkani-Hamed framed the hunt for the Higgs as a detective story with an inevitable ending. Maldacena compared our understanding of nature to the fairytale *Beauty and the Beast*.

"What we know already is incredibly rigid. The laws are very rigid within the structure we have, and they are very fragile to monkeying with the structure," said Arkani-Hamed. "Often in physics and mathematics, people will talk about beauty. Things that are beautiful, ideas that are beautiful, theoretical structures that are beautiful, have this feeling of inevitability, and this flip side of rigidity and fragility about them."



Gauge symmetries determine the interactions and production of particles (as depicted here). Juan Maldacena used a monetary analogy to describe the gauge symmetries of the electromagnetic and weak force.

The recent discovery of the Higgs-like boson is "a triumph for experiment but also a triumph for theory," said Arkani-Hamed. "We were led to saying, 'This thing has got to be there. We've never seen one before, but by these arguments, by our little detective story, it's gotta be there.' And by God, it is. It's allowed to be there. It can be there. It is there."

In Maldacena's comparison, beauty is the fundamental forces of nature—gravity, electromagnetism, the strong force, and the weak force—and the beast is the Higgs mechanism. "We really need both to understand nature," said Maldacena. "We are, in some sense, the children of this marriage."

Current knowledge of the fundamental forces of physics is based on two well established theories: the Standard Model of particle physics, a set of equations that gives an impressively accurate description of elementary particles and their interactions, but omits gravity and only accounts for about one-sixth of the matter in the universe; and Einstein's theory of general relativity, which describes the observed gravitational behavior of large objects in the universe, such as galaxies and clusters of galaxies, but has yet to be reconciled with quantum principles.

Ordinary matter—the material we see and are familiar with, such as the planets, the stars, human bodies, and everyday objects—is acted on by gravity, electromagnetism, the strong force, and the weak force. These interactions apply over an enormous range of distances—from the size of the observable universe (around 10^{28} centimeters) down to the weak scale (around 10^{-17} centimeters).

In the Standard Model of particle physics, nature is built out of elementary building blocks, such as electrons and quarks. Forces between particles are transmitted by other particles, such as photons, the carrier of electromagnetic forces, and W and Z particles, the basic particles that transmit the weak interactions. The Higgs isn't the first particle that the Standard Model has predicted and that has been later discovered experimentally. The model also has led to the prediction and discovery of the W and Z particles, the top quark, and the tau neutrino.

The Higgs boson explains how most fundamental particles acquire mass as they interact with a Higgs field that exists everywhere in the universe. It is the final element of the Standard Model that needed to be confirmed experimentally and its discovery promises to provide further understanding of the origin of mass and help clarify some long-standing mysteries.

The weak scale is the distance that is being probed at the Large Hadron Collider, where the Higgs-like boson was discovered. With all ordinary matter and interactions, the force between two electrons (the size of the quantum mechanical fluctuations) gets weaker as you go to longer distances (lower energies) and stronger at shorter distances (higher energies), a basic consequence of the Heisenberg uncertainty principle.

"We've learned that the essential unity and simplicity of the laws of nature become manifest at short distances," explained Arkani-Hamed. "They're hidden at large distances by a variety of accidents, but when we go to short distances we finally see them. We see for the first time all these different interactions described in a common way."

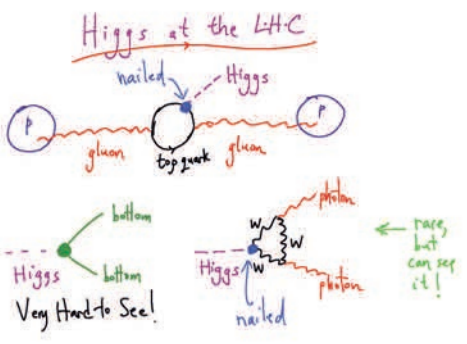
In the Standard Model, all particles intrinsically have some spin and an angular momentum that is associated with that spin. Known particles have angular momenta, measured in H-bar (Planck's constant) units, in multiples of 1/2. According to the model, the only allowable spins are 0, 1/2, 1, 3/2, and 2, but we have seen only a subset of that: 1/2, 1, and 2. The electron has spin 1/2. The photon has spin 1. The graviton, which interacts the same with everything, is the only particle that has spin 2.

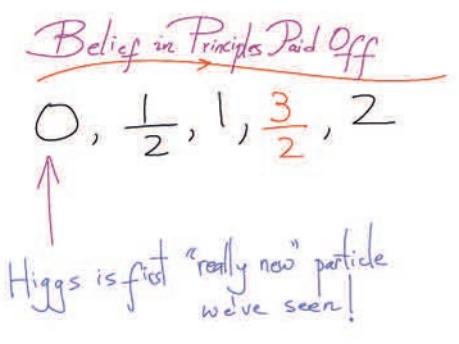
The story of the Higgs starts by trying to understand why some particles have mass. According to the Standard Model, the W and Z particles that carry the electroweak force should have zero mass to allow for the unification of the electromagnetic and weak nuclear forces in a single electroweak force. Between theory and experiment, it was determined that the Higgs particle had to enter the picture under 200 GeV (a unit to measure mass), that it had to interact with W, Z, and top quark particles, and that it had to have 0 spin. While the Standard Model did not predict the exact mass of a Higgs particle, from precise measurements, it was known that it had to be somewhere between 80 to around 200 times the mass of a proton. The Higgs-like boson, which was discovered last summer in the mass region of around 126 GeV, allows once-massless particles to have mass without destroying the principles of the Standard Model.

"People sometimes ask, what is this [the discovery of the Higgs] useful for?" said Maldacena. "I have to be honest, I don't know of any technological application. There is the apocryphal quote of [Michael] Faraday. When asked what the possible technological application of electricity was, he said to the prime minister, 'Someday we will be able to tax it.' I think, maybe, we could say the same thing about the Higgs boson. Something we do know is that it is helping us understand what happens in nature at very short distances."

Gauge symmetries determine the interactions and production of particles, and Maldacena used a monetary analogy to describe the gauge symmetries of the electromagnetic and weak force. In his analogy, the magnetic field is a gauge symmetry where each country is identical except they can choose their own currency. All money must be changed to the new currency when moving from country to country.

In physics, the currency is the rotations within a circle at each point in spacetime, and the exchange rate is the electromagnetic potential, or the displacement that results from traveling from one small spacetime region (country) to the next. Following a quantum mechanic understanding of the probabilistic laws of nature, "these exchange rates are random with a probabilistic distribution that depends on the opportunity to speculate," said Maldacena. "Nature doesn't like speculation, and will not offer you these opportunities very easily, but it will offer them to you, if you can find them."





Slide images from Nima Arkani-Hamed's lecture "The Inevitability of Physical Laws: Why the Higgs Has to Exist."

The gauge symmetry of weak interactions involves symmetries of spheres rather than circles at each point in spacetime. Maldacena described the Higgs mechanism as an object sitting at each point on these weak spheres. When a rotation is made—even in a vacuum and empty space—this mechanism causes a transformation or change.

Continuing the monetary analogy, Maldacena introduced the notion of being able to buy something, in this case gold, in each country. The gold can be taken from one country to the next, its price is set by each of the countries, and money can be earned by going back and forth between the countries. In this analogy, the price of gold in each country is the Higgs field. Once the price or gauge is set to a constant value everywhere in space, this leads to a preferential value for the exchange rates, and leads to the masses for the W and Z weak bosons. In Maldacena's analogy, the Higgs boson arises when there are two objects, such as gold and silver, to purchase. The relative price of gold and silver is the Higgs boson; the ratio behaves as a massive particle. According to Maldacena, it is necessary to have at least two objects to buy so that when the distances between points in spacetime becomes very small we can still retain interesting interactions at long distances.

The Higgs-like boson was produced at the LHC in an indirect way but according to similar gauge symmetries derived from the Standard Model. When protons collide, they produce many particles. Very rarely, they produce Higgs bosons. These Higgs bosons decay very quickly into particles, such as two photons. Since the Higgs bosons decay too quickly to discern, theorists predicted that experimentalists could detect the Higgs by

(Continued on page 5)

HIGGS *(Continued from page 4)*

looking at events that have two photons and finding a bump in the data where two photons would amount to the mass of the Higgs boson.

The Higgs boson is the first particle with spin 0. This leaves only spin 3/2 unrealized in nature. But there is a strong candidate. Supersymmetry is associated with 3/2, and it is possible that the LHC will confirm the existence of supersymmetry, which extends the Standard Model and unites matter particles and force particles by pairing them in a single framework. It suggests that the strong force, the weak force, and the electromagnetic force become one at very short distances.

Supersymmetry also naturally leads to a new dark matter particle that does not emit or absorb light, and can only be detected from its gravitational effects. Ordinary matter that is explained by the Standard Model makes up about 4 percent of the universe; dark matter comprises about 22 percent.

"We know from astrophysical observations that there is more matter than what we see," said Maldacena. "If we look at the sky, we see some galaxies sitting there in the sky, surrounded by what looks like the blackness of empty space. What we don't know is whether this dark matter particle will or will not be produced at the LHC."

In the last decade, astronomical observations of several kinds, particularly of distant supernovae and the cosmic microwave background, also indicate the existence of what is known as dark energy, a uniform background field that makes up about 74 percent of the universe and is credited with accelerating the expansion of the universe. The presence of dark energy suggests a fundamental gap in our current understanding of the basic forces of nature.

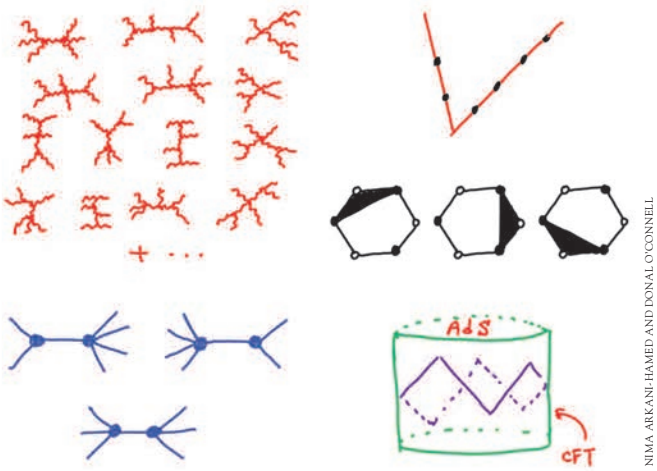
"Space, time, and quantum mechanics framed the central dramas of the twentieth century, and really have taken us shockingly far. The story of the Higgs is the last example of how far they took us. But in a sense, the story of the Higgs is one of the last embers of the set of ideas that we dealt with and understood in the twentieth century," said Arkani-Hamed.

"Relativity and quantum mechanics—the picture of spacetime that Einstein gave us and quantum mechanics—are incredibly rigid and powerful. The next set of questions is: Where do these things come from? That's the one thing I didn't question. I just took spacetime and quantum mechanics and the rest of it followed. What is the deeper origin of spacetime and quantum mechanics? This is what you should ask your friendly neighborhood string theorist."—*Kelly Devine Thomas*

From the Spring 2009 Issue

Feynman Diagrams and the Evolution of Particle Physics

For sixty years, Feynman diagrams have been an essential calculational and conceptual tool for theoretical physicists striving to deepen our understanding of the fundamental forces and particles of nature. Members of the Institute have played leading roles in the development of their use, from Freeman Dyson in the late 1940s and early 1950s to the current generation of theoretical physicists in the School of Natural Sciences. Most recently, clues provided by Feynman diagrams have led to powerful new methods that are revolutionizing our ability to understand the fundamental particle collisions that will occur at the Large Hadron Collider (LHC). At the same time, these clues have motivated Institute theorists to pursue a radical transcription of our ordinary physics formulated in space and time in terms of a theory without explicit reference to spacetime. The story of these developments connects one of the most pressing practical issues in theoretical particle physics with perhaps the most profound insight in string theory in the last decade—and at the same time provides a window into the history of physics at the Institute.



NIMA ARKANI-HAMED AND DONAL O'CONNELL

These figures, drawn by Nima Arkani-Hamed, show how calculating scattering amplitudes has evolved from using Feynman and BCFW diagrams to a "holographically dual" AdS/CFT formulation.

Surprising as it now seems, when Richard Feynman first introduced his diagrams at a meeting at a hotel in the Pocono Mountains in the spring of 1948, they were not immediately embraced by the physicists present, who included J. Robert Oppenheimer, then Director of the Institute and organizer of the meeting, and a number of then Members of the Institute, including Niels Bohr and Paul Dirac. The main event of the meeting, whose topic was how to calculate observable quantities in quantum electrodynamics, was an eight-hour talk by Julian Schwinger of Harvard, whose well-received analysis used calculations founded in an orthodox understanding of quantum mechanics. On the other hand, Feynman struggled to explain the rules and the origins of his diagrams, which used simple pictures instead of complicated equations to describe particle interactions, also known as scattering amplitudes.

Traveling on a Greyhound bus from San Francisco to Princeton at the end of the summer of 1948 to take up his appointment as a Member of the Institute, twenty-four-year-old Dyson had an epiphany that would turn Feynman diagrams into the working language of particle physicists all over the world. Earlier, in June, Dyson had embarked on a four-day road trip to Albuquerque with Feynman, whom he had met at Cornell the previous year. Then he spent five weeks at a summer school at the University of Michigan in Ann Arbor where Schwinger presented detailed lectures about his theory. Dyson had taken these opportunities to speak at length with both Feynman and Schwinger and, as the bus was making its way across Nebraska, Dyson began to fit Feynman's pictures and Schwinger's equations together. "Feynman and Schwinger were just looking at the same set of ideas from two different sides," Dyson explains in his autobiographical book, *Disturbing the Universe*. "Putting their methods together, you would have a theory of quantum electrodynamics that combined the mathematical precision of Schwinger with the practical flexibility of Feynman." Dyson combined these ideas with those of a Japanese physicist, Shinichiro Tomonaga, whose paper Hans Bethe had passed on to him at Cornell, to map out the seminal paper, "The Radiation Theories of Tomonaga, Schwinger and Feynman," as the bus sped on through the Midwest. Published in the *Physical Review* in 1949, this work marked an epoch in physics.

While Feynman, Schwinger, and Tomonaga were awarded the Nobel Prize in Physics in 1965 for their contributions to developing an improved theory of quantum electrodynamics, it was Dyson who derived the rules and provided instructions about how the Feynman diagrams should be drawn and how they should be translated into their associated mathematical expressions. Moreover, he trained his peers to use the diagrams during the late 1940s and 1950s, turning the Institute into a hotbed of activity in this area. According to David Kaiser of the Massachusetts Institute of Technology, author of *Drawing Theories Apart: The Dispersion of Feynman Diagrams in Postwar Physics*, "Feynman diagrams spread throughout the U.S. by means of a postdoc cascade emanating from the Institute for Advanced Study."

Feynman diagrams are powerful tools because they provide a transparent picture for particle interactions in spacetime and a set of rules for calculating the scattering amplitudes describing these interactions that are completely consistent with the laws of quantum mechanics and special relativity. These rules allow any process involving particle scattering to be converted into a collection of diagrams representing all the ways the collision can take place. Each of these diagrams corresponds to a particular mathematical expression that can be evaluated. The exact description of the scattering process involves summing an infinite number of diagrams. But in quantum electrodynamics, a simplification occurs: because the electric charge is a small number, the more interactions a diagram involves the smaller the contribution it makes to the sum. Thus, to describe a process to a given accuracy, one only has to sum up a finite number of diagrams.

"Freeman was the person who realized that once you force the quantum mechanical answer to look like it is consistent with the laws of special relativity, then it is very natural to do the calculations in terms of Feynman diagrams," says Nima Arkani-Hamed, Professor in the School of Natural Sciences. "Almost nobody thinks about Feynman diagrams the way Feynman originally arrived at them. You open up any textbook and the derivation of these things uses this very beautiful, profound set of insights that Freeman came up with."

By the 1980s and 1990s, Feynman diagrams were being used to calculate increasingly complicated processes. These included not only collisions of the familiar electrons and photons, governed by quantum electrodynamics, but also the interaction of particles such as quarks and gluons, which are the basic constituents of protons and neutrons, and are governed by the theory known as quantum chromodynamics. These calculations are essential to understanding and interpreting the physics probed at modern high-energy particle accelerators. However, theorists found that using Feynman diagrams in this wider context led to an explosion in their number and complexity.

Increasingly clever tricks were developed in the late 1980s to calculate these more complicated processes without actually calculating the Feynman diagrams directly. This led to a surprising realization. While each step in the calculation was very complicated, involving a huge number of terms, cancellations between them led to a final answer that was often stunningly simple. "The answer seems to have all sorts of incredible properties in it that we are learning more and more about, which are not obvious when you draw Feynman diagrams. In fact, keeping spacetime manifest is forcing the introduction of so much redundancy in how we talk about things that computing processes involving only a few gluons can require thousands of diagrams, while the final answer turns out to be given by a few simple terms," says Arkani-Hamed. "A big, overarching goal is to figure out some way of getting to that answer directly without going through this intermediary spacetime description."

In 2003, Edward Witten, Charles Simonyi Professor in the School of Natural Sciences, came up with a proposal along these lines. He found a remarkable rewriting of the leading approximation to the interactions between gluons that led directly to the simple form of their scattering amplitudes, without using Feynman diagrams. This work immediately led to a major innovation: a new diagrammatic representation of amplitudes, called

(Continued on page 6)

"CSW diagrams" (after Freddy Cachazo, then a Member, Witten's student Peter Svrcek, and Witten). This led to a number of new insights into amplitudes that, via a circuitous path, led to a second, apparently unrelated representation of the amplitudes known as "BCFW diagrams" (after former Members Ruth Britto and Bo Feng, as well as Cachazo and Witten). These powerful new diagrams highlight and exploit properties of the physics that are invisible in Feynman diagrams, and they provide a much more efficient route to the final answer.

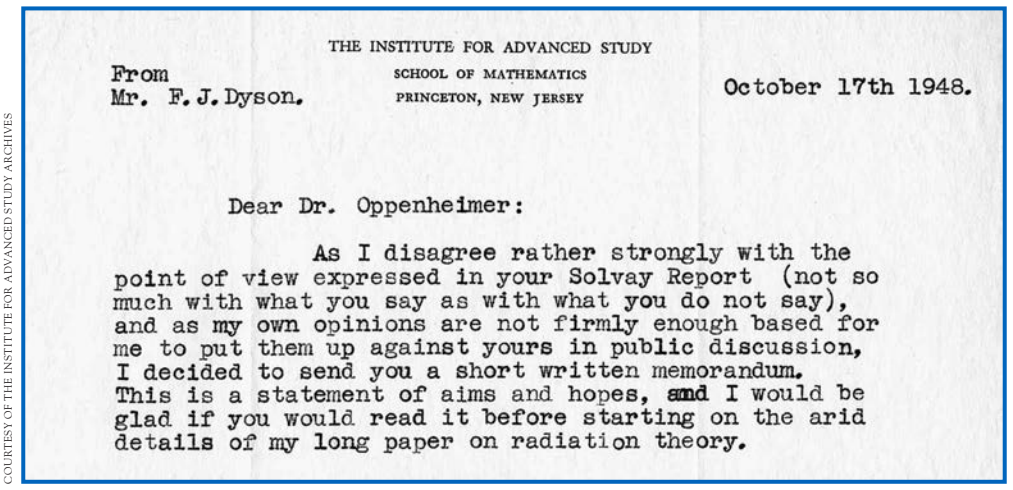
COURTESY OF THE INSTITUTE FOR ADVANCED STUDY ARCHIVES

COURTESY OF MELANIE JACKSON AGENCY AND CALTECH ARCHIVES

Freeman Dyson (left) and Richard Feynman, circa 1950

These new methods have triggered a breakthrough in a critical area relevant to describing physics at the LHC. This enormous machine will experimentally probe our understanding of nature at extremely short distances, and could reveal major new physical principles, such as an extended quantum notion of spacetime known as supersymmetry. In order to establish the discovery of new particles and forces, it is necessary to accurately understand the predictions from current theories. But these calculations had been hampered by the complexity of the relevant Feynman graphs. Many processes experimental physicists were interested in were considered to be impossible to calculate theoretically in practice. Now, this is no longer the case, and already computer code exploiting the BCFW technique is being developed for application to the data the LHC will produce.

In addition to their practical value, these new ideas have opened up a number of new frontiers of purely theoretical research, both in exploring further the inner workings of scattering amplitudes and in investigating their relationship with deeper theories of space and time. About a year and a half ago, Arkani-Hamed became intrigued by the BCFW formalism, and with his student Jared Kaplan he found a simple physical argument for why it is applicable to calculating scattering amplitudes for gluons and gravitons, not just in four dimensions of spacetime as originally formulated, but in any number of dimensions. "This idea of BCFW is somehow a powerful and general fact about physics in any number of dimensions," says Arkani-Hamed. Their work also suggested that the amplitudes for the scattering of gravitons might be especially simple. "Even the simplest processes for gravity involve ludicrously complicated Feynman diagrams, and yet not only are the amplitudes just as simple in this new language, there is even some indication that they might be simpler," says Arkani-Hamed. "Perhaps this is because the things that are the most complicated from the Feynman diagram perspective are the simplest from this other perspective that we are searching for."



COURTESY OF THE INSTITUTE FOR ADVANCED STUDY ARCHIVES

THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

From
Mr. F. J. Dyson.

October 17th 1948.

Dear Dr. Oppenheimer:

As I disagree rather strongly with the point of view expressed in your Solvay Report (not so much with what you say as with what you do not say), and as my own opinions are not firmly enough based for me to put them up against yours in public discussion, I decided to send you a short written memorandum. This is a statement of aims and hopes, and I would be glad if you would read it before starting on the arid details of my long paper on radiation theory.

This excerpt is from a letter written to J. Robert Oppenheimer by Freeman Dyson shortly after his arrival at the Institute in 1948. Oppenheimer initially expressed deep resistance to Dyson's work on quantum electrodynamics, which drew on the ideas of Richard Feynman, Julian Schwinger, and Shinichiro Tomonaga. Oppenheimer eventually conceded his position with a small piece of paper left in Dyson's mailbox with the handwritten words "Nolo contendere. R.O."

Expressing amplitudes in terms of variables that encode directly only the physical properties of particles, such as their momentum and spin, is a key step in this program. This was achieved in the 1980s for four-dimensional theories. But if a general rewriting of basic interactions for particles like gluons and gravitons is sought, it must be possible to extend these successes to higher dimensions, especially since extra dimensions appear naturally in string theory. This year, Institute Member Donal O'Connell and one of Arkani-Hamed's students, Clifford Cheung, showed that this could also be achieved in six dimensions. Using their formalism, and BCFW diagrams in six dimensions, O'Connell and Cheung were able to discover very compact expressions for scattering gauge bosons and gravitons in higher dimensions, which also unify a multitude of four-dimensional expressions.

Witten's 2003 proposal used a set of ideas that Roger Penrose first suggested in the 1960s called twistor theory, which posits that instead of keeping track of points in spacetime, physicists should look at the light rays that come out of those points and follow them out to infinity. Witten's method of calculating the scattering amplitudes suggested a new string theory that lives in twistor space rather than in ordinary spacetime; the structure of this string theory is directly related to the CSW diagrammatic construction.

The BCFW diagrams arose from studying general properties of relativistic quantum theory formulated in spacetime. However, in a recent collaboration, Arkani-Hamed, Cachazo, Cheung, and Kaplan found to their surprise that the BCFW formalism is also most naturally expressed in twistor space. Their reasoning leads to a direct mapping of ordinary physics in spacetime to a simpler description in twistor space. "What is extremely cool about this business is that we are trying to come up with an explanation for marvelous patterns found in theories describing our world," says Arkani-Hamed. "We have a lot of clues now, and I think there is a path towards a complete theory that will rewrite physics in a language that won't have spacetime in it but will explain these patterns."

This line of thought has connections to the concepts of duality and holography, which grew out of developments in string theory in the 1990s and have dominated much of the activity in the field for the past decade. A "duality" is an exact quantum equivalence between two completely different classical theories. The first examples of this remarkable phenomenon were discovered in four-dimensional supersymmetric theories by Nathan Seiberg, Professor in the School of Natural Sciences, in collaboration with Witten. This led to the realization that all string theories are different manifestations of a single underlying theory. Holography is the most striking example of duality to date, relating a gravitational theory in a curved spacetime to a non-gravitational particle theory in a lower-dimensional flat space. The analogy is to familiar holograms, which encode three-dimensional information on a two-dimensional surface. Juan Maldacena, also a Professor in the School of Natural Sciences, found the first example of a holographic duality, in which everything that happens in the bulk of spacetime can be mapped to processes occurring on its boundary. Maldacena's conjecture is now known as the AdS/CFT correspondence, and provides a dictionary for translating the physics of anti–de Sitter space (AdS)—a negatively curved space with an extra fifth dimension, containing gravity and strings—to a conformal field theory (CFT), a four-dimensional particle theory that lives on the boundary of the spacetime. "Things about gravity are mysterious; things about particle theories are much less mysterious. Incredibly, the AdS/CFT correspondence maps mysterious things about gravity to well-understood things about particle physics, giving us the first working example of what a theory with emergent spacetime looks like," says Arkani-Hamed. "It encourages the thought that, even in a nearly flat spacetime like our own, there is a picture of scattering processes, which takes incoming particles and converts them to outgoing particles with some very simple rules that bypass evolution through the intermediary spacetime."

Exploitation of the AdS/CFT correspondence has led to many remarkable new developments. A key point is that the four-dimensional CFT involved in the correspondence is a close cousin of quantum chromodynamics, which is the theory relevant at the LHC. AdS/CFT thus provides a sort of theoretical laboratory for the exploration of phenomena related to the hadron collider. While the CFT is similar to the theories that describe nature, it is different in that it is far more symmetric. In fact, the theory enjoys so much symmetry that it has a property known as integrability, which has allowed, for the first time, the exact computation of a quantity relevant to scattering amplitudes. Already there is much progress in an area where a Feynman diagram computation is hopeless: when the coupling analogous to the electric charge is large, one would have to sum all possible diagrams. But via AdS/CFT, Maldacena and Member Fernando Alday have shown that in the large coupling regime, the scattering amplitude can be computed by turning it into a tractable calculation in a string theory living in AdS space. This work led to another major surprise: the scattering amplitudes were shown to have unexpected symmetries that were later sought and found in diagrammatic calculations at weak coupling. These symmetries are related to the integrability properties, and give new hope that scattering amplitudes in the CFT can be determined exactly.

Arkani-Hamed suspects that the key to further progress will be finding the analogue for the AdS/CFT correspondence for flat space. An essential problem in gravity is the inability to precisely talk about physical observables that are localized in space and time. "The general rule seems to be that we can only describe gravitational systems when we're sitting at the boundary of spacetime at infinity, because that is where notionally we can repeat experiments infinitely many times with an infinitely big apparatus to make infinitely precise measurements. None of these things are exactly possible with finite separations," says Arkani-Hamed. "The AdS/CFT correspondence already tells us how to formulate physics in this way for negatively curved spacetimes; we are trying to figure out if there is some analogue of that picture for describing scattering amplitudes in flat space. Since a sufficiently small portion of any spacetime is flat, figuring out how to talk about the physics of flat space holographically will likely represent a real step forward in theoretical physics."—*KDT*

Nima Arkani-Hamed: Unraveling Nature's Mysteries

"Everything here is fraught with danger and excitement," says Nima Arkani-Hamed, Professor in the School of Natural Sciences. With a broad sweep of his hand, he motions to the diagram he has drawn on the chalkboard in his office of the range of distance scales for known phenomena—from 10^{-33} cm, which is associated with quantum gravity and string theory, to 10^{+28} cm, which is the size of the universe.

"Why is the universe big, why is gravity so weak? You would think after 2,000 years of thinking about physics we would have good answers to questions like that. We have lousy answers to these questions," says Arkani-Hamed. "Our current laws of nature—the Standard Model of particle physics—are perfectly consistent. No experiments contradict them, but they give such lousy answers to these questions that we think we are missing something very, very big."

With the imminent start-up of the Large Hadron Collider (LHC), a particle accelerator that will collide protons together and allow us to probe the laws of nature down to distances of 10^{-17} cm, a billion times smaller than the atom, and ten times smaller than the tiniest distances we have probed to date, fundamental particle physics is on the threshold of a new era.

RANDALL HAGADORN

Nima Arkani-Hamed

Arkani-Hamed, one of the world's leading phenomenologists who joined the Faculty in January, has taken a lead in building models of the universe that relate to theories that can be tested at the LHC—from supersymmetry to large extra dimensions of space to the idea that our universe exists in a sea of universes, each governed by a different set of principles.

"I try to take ideas that are in the theoretical zeitgeist and see if they might be relevant to solving any of the outstanding mysteries, and then see what experimental consequences can be derived," says Arkani-Hamed. "Phenomenologists are jacks of all trades. We try to propose theories that extend things that we have seen, figure out the direct observational consequences of those theories, and work closely with our experimental colleagues to see how we can actually extract information about nature directly from an experiment."

Among the ideas that will be tested at the LHC is the existence of supersymmetry, which involves the ordinary dimensions of space and time having quantum mechanical partners, and the possibility that there may be extra spatial dimensions aside from the three spatial dimensions familiar to us. Both supersymmetry and extra dimensions are essential components of string theory, the leading candidate for unifying general relativity and quantum mechanics. These are all subjects that Institute physicists have taken a lead in developing.

Just as for every particle there exists an antiparticle, supersymmetry predicts that for every known particle there also exists a superpartner particle. Part of the strong theoretical appeal of supersymmetry is its possible connection to dark energy and the fact that it provides a natural candidate for dark matter—a new weakly interacting massive particle (WIMP) with mass close to the scale that will be probed at the LHC.

"Often people will describe the LHC or accelerators in general as microscopes for probing short distances. But normally, a microscope is looking at something. What is the LHC looking at? It is looking at the vacuum," says Arkani-Hamed. "People like to say the dark energy is very mysterious and we don't know what it is but that is a bit of an exaggeration, because there is an extremely simple thing that it could be. It could be the energy of the vacuum. It's not that we don't know how to accommodate it in our equations. We definitely know how to accommodate it. The problem is that we get it 120 orders of magnitude bigger than it apparently is."

To accommodate dark energy in particle physics requires unnatural fine-tuning, which also arises in another aspect of Arkani-Hamed's research—a paradox of the Standard Model called the "hierarchy problem" that relates to the extreme weakness of gravity in comparison to the other forces of nature—electromagnetism, the strong nuclear force, and the weak nuclear force. Violent short distance quantum fluctuations in the vacuum would naturally lead to the prediction that the strength of gravity is thirty orders of magnitude larger than its observed strength, requiring inordinate fine-tuning for the parameters of the theory.

"Fine-tuning is like walking into a room and seeing a pencil standing on its tip in the middle of a table," says Arkani-Hamed. "If you saw it, you would think that maybe there is a little string hanging from the ceiling that you missed, or maybe there is a little hand holding it up or something. The dark energy problem and the hierarchy problem are conceptually identical puzzles. In both cases, we have to do a tremendous amount of fine-tuning in order to explain some very obvious property of our world because we don't yet see any dynamics or any mechanism for explaining why it is what it is."

Particle physics data point to another mysterious component of empty space, the Higgs field, a force that fills space and gives particles the property of mass and might be related to dark energy. Arkani-Hamed is willing to bet several months' salary that the Higgs particle, the last element predicted by the Standard Model that has not been confirmed experimentally, will be discovered at the LHC.

While supersymmetry is the most popular solution to the hierarchy problem, Arkani-Hamed has proposed other possibilities, including the existence of large extra dimensions of space, which dilute gravity's strength, and a theory called split supersymmetry, in which only half of all particles have superpartners. "One of the confusing things about supersymmetry," says Arkani-Hamed, "is that people have mounted tons of experiments to look for possible signals of these partner particles and so far there has been no hint of it directly or indirectly."

Split supersymmetry finds a common explanation for the cosmological constant and the hierarchy problem. It relates to the theory of the multiverse, in which our entire observable universe might be a tiny part of a much larger multiverse, in which many universes function according to distinct and self-containing physical laws with one common exception: gravity, which can travel freely between them. "This is a very controversial idea, because to invoke universes you can't see to explain properties of our own universe is obviously a tricky proposition," says Arkani-Hamed. "But it is not obviously wrong. It is a subject of lots of continuing activity and thinking right now."

In a multiverse, a near-infinite number of universes exist, but ours is the only one we can observe because it is the only one in which we can live—a concept also known as the anthropic principle. "It is very interesting that the observed value of the cosmological constant, the observed value of the vacuum of dark energy, if you interpret the dark energy as a cosmological constant, is right around the value where if it was a little bigger then the universe would be empty," says Arkani-Hamed.

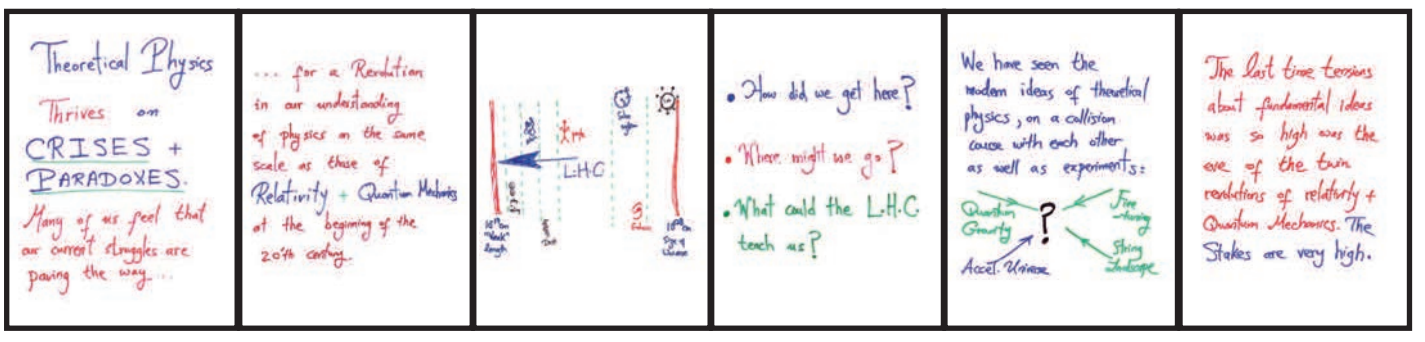
In his recent talk on dark energy at the Space Telescope Science Institute, Edward Witten, Charles Simonyi Professor in the School of Natural Sciences, addressed the theoretical possibility of a multiverse in which the aim is not to explain why the vacuum has a very tiny energy but rather to look for a theory that generates all kinds of vacua with different properties that are realized in different times and places in a multiverse, perhaps as a result of cosmic inflation.

The good news, if we are living in a multiverse in which the only vacuum we can observe is the one that allows our existence, Witten says, is that the Standard Model as we know it may be fairly accurate. "If the universe is really a multiverse, finding the vacuum state we observe should be like searching for a needle in a haystack," says Witten. "But this comes with a hefty dose of bad news: if the vacuum of the real world is really a needle in a haystack, it is hard to see how we are supposed to be able to understand it."

At the Institute, Arkani-Hamed will be looking at LHC data to interpret signals that underscore these and other theoretical possibilities while helping to attract and mentor highly talented postdoctoral fellows, with a diversity of theoretical skills, in anticipation of a golden era of discovery. Speaking of his decision to come to the Institute from his position as Professor of Physics at Harvard, Arkani-Hamed says, "This is a very, very special group. I couldn't possibly ask for better or more stimulating colleagues in high energy theory and string theory, quantum field theory, and astronomy."

"Here I have the ability to really focus very sharply on doing science. Again, I keep bringing this up because it looms so large in most of our minds: an experiment like the LHC is a once-in-a-lifetime opportunity. There really seem to be very fundamental scientific issues at stake. And there really is a chance to unravel them and learn something potentially revolutionary and new about nature. You can't give up on opportunities like that."—*KDT*

Nima Arkani-Hamed was one of the organizers and lecturers of the School of Natural Sciences's 2008 Prospects in Theoretical Physics program, "Strings and Phenomenology." The July 14–25 program was designed for string theorists wanting to learn about issues of compactification relevant to phenomenology and cosmology, and for cosmologists seeking to learn about strings and their applications to phenomenology.



Slides drawn from Arkani-Hamed's lecture "The Future of Fundamental Physics"

Most Detailed Map of Infant Universe Shows Conditions for Structure Formation

The most detailed map of the infant universe to date was publicly released in March, showing relic radiation from the Big Bang, imprinted when the universe was just 380,000 years old. This was the first release of cosmological data from the Planck satellite, a mission of the European Space Agency that was initiated in 1996 and involved hundreds of scientists in over thirteen countries. In a lecture in May, Matias Zaldarriaga, Professor in the School of Natural Sciences, explained how theoretical models allowed the Planck team to determine the composition of the universe, map the seeds for the formation of structure, and confirm our broad understanding of the beginnings and evolution of the universe.

Our current understanding of the history of the universe began to take shape around the 1930s, after Edwin Hubble discovered that the universe was expanding. Since then, there have been great advances in understanding the composition of the universe and how it has evolved through cosmic history. According to the standard cosmology model, in the current phase in the history of the Big Bang, the universe began about fourteen billion years ago. Initially the universe was hot and dense with interacting particles. It has been conjectured that prior to this phase, the universe underwent a brief period of accelerated expansion known as inflation when quantum fluctuations, stretched to cosmologically large scales, became the seeds of the universe's stars and galaxies.

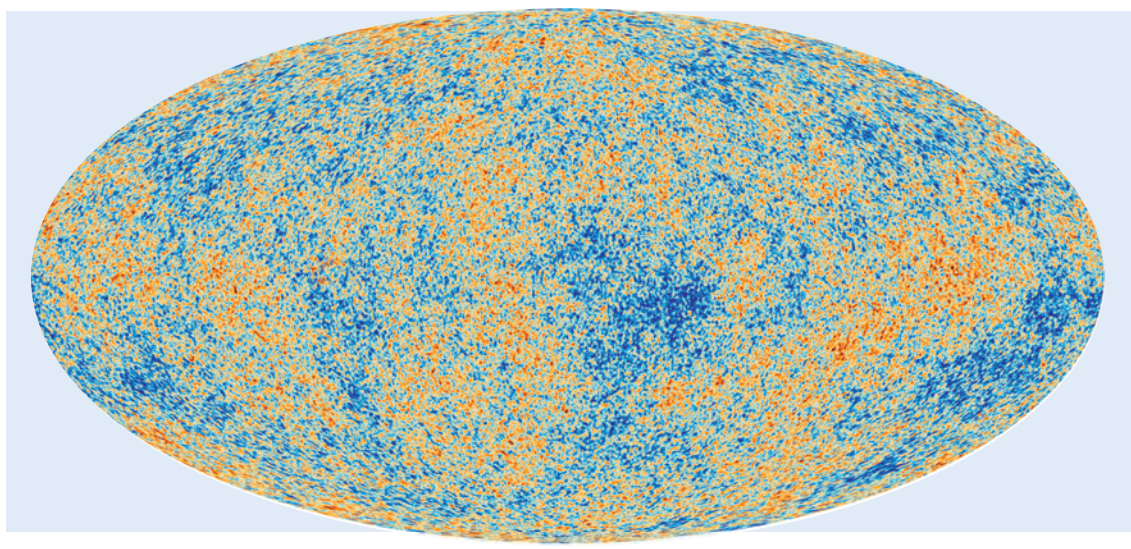
The Planck map—a composite made from nine maps of the sky in nine different frequencies by the Planck satellite—captures the early light from the cosmic microwave background radiation that is remnant from the Big Bang. The cosmic microwave background was first detected in 1964 and since then space, ground, and balloon-based experiments have mapped temperature variations of this light left over from the very early universe, allowing cosmologists to see if theoretical models can reproduce the formation of objects that can be seen through cosmic history. The Planck satellite is three times more sensitive than the previous satellite, the Wilkinson Microwave Anistropy Probe (WMAP), and its unprecedentedly precise map depicts "how the universe was before its structure had time to develop," said Zaldarriaga. "We are seeing the initial conditions for this process of structure formation."

According to the standard cosmology model and the latest Planck data, the universe is made up of ordinary visible matter (less than 5 percent), dark matter (about 27 percent), and dark energy (about 68 percent). Dark matter, which emits no light but exerts a gravitational pull, is believed to be a particle that was left over from the Big Bang. It has not yet been produced in a laboratory, such as the Large Hadron Collider, nor have detectors on Earth detected it, even though it is believed to pass through the planet. Even less is known about the mysterious dark energy, a uniform background field that is credited with accelerating the expansion of the universe.

Through the effect of gravitational lensing (the bending of light due to the presence of matter curving spacetime), a method first proposed by Zaldarriaga and Uros Seljak in 1999, Planck was able to map the distribution of dark matter in the universe. Through the Sunyaev-Zeldovich effect (named in part for Rashid Sunyaev, Maureen and John Hendricks Visiting Professor in the School of Natural Sciences, it identifies hot-gas regions through distortions in the cosmic microwave background radiation), Planck mapped the distribution of hot gas in the universe and discovered new clusters of galaxies.

Recommended Viewing: A video of "The Latest News from the Cosmos," a lecture by Matias Zaldarriaga, may be viewed at http://video.ias.edu/zaldarriaga-lecture-5-13/.

In the 1980s, cosmologists developed inflation models of the very early universe that incorporated our current understanding of the laws of physics—the law of general relativity to understand how gravity works, and quantum mechanics to understand how matter behaves. To explain the universe's longevity and homogeneity, theorists introduced a period of inflation before the Big Bang. Without it, a universe, behaving according to the laws of general relativity, would collapse into a black hole or become completely empty within a period of a few fractions of a second. Inflation had a surprise bonus: due to the uncertainty principles of quantum mechanics, inflation had to last longer in different regions. These tiny differences could then act as the seeds for structure.



ESA AND THE PLANCK COLLABORATION

Planck's map of the cosmic microwave background, which depicts temperature variations of the light left over from the very early universe

According to inflation theory, as the universe expands exponentially fast, its geometry becomes flat—this geometry was confirmed experimentally around 2000. Theorists then had to use the laws of physics to solve the "graceful exit" problem of how to make the inflation stop so that the universe cools and structure starts to form. "In a sense, the material that filled the universe at the time of inflation had to act like a clock," said Zaldarriaga. "The universe was expanding and at some point it had to stop this inflationary period to start something new." Quantum mechanics then provides a source for fluctuations.

While Planck and WMAP have confirmed major details of inflation theory, in the coming months and years, cosmologists will try to explain some small anomalies in the Planck data, zero in on the correct prediction for identifying the physical system that stops the inflationary period, and develop better methods for detecting signatures of gravitational waves, which are believed to have been produced during inflation and could have shown up in the Planck data but haven't yet.

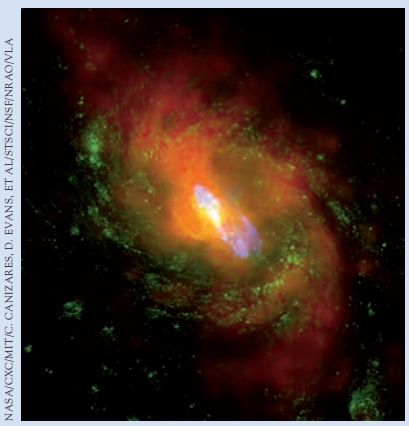
Only more data, more observations, and more thinking will help cosmologists resolve what Zaldarriaga described as cosmology's chicken (universe) and egg (inflation) problem, which leads to a range of possible solutions including the existence and collision of multiple chickens (and eggs) within the larger structure of a multiverse. "We have beautiful pictures of this chicken as it grows up," said Zaldarriaga. "Of course the first question everybody asks is 'Where does the chicken come from?' Our theory friends in the '80s came up with the idea that the chicken comes from an egg. If we say the chicken comes from an egg, where does the egg come from? It comes from a chicken. . . . Of course, we don't know exactly what eggs should look like. The obvious thing to do is to try to get better pictures of the early universe, of some property of this egg that we can compute, and then see if it matches what we are now saying an egg might look like. This is how we can make progress, and this is what we are trying to do."

—KDT

Black Holes and the Birth of Galaxies

Galaxies are the visible building blocks of the universe, astrophysical laboratories that have profoundly informed our knowledge of cosmology and nature. Black holes—once a bizarre mathematical consequence of Einstein's relativity theory—are now mainstream astronomy, thanks to studies of the centers of nearby galaxies where these exotic objects are routinely found. In May, Nadia Zakamska, former John N. Bahcall Fellow and Member (2005–10) in the School of Natural Sciences and Assistant Professor at Johns Hopkins University, described the basic theory of galaxy formation and shared insight into the complex interrelationship between supermassive black holes and their host galaxies.

According to basic galaxy formation theory, the universe consists of galaxies composed of billions to trillions of stars that form around a filamentary network of mysterious yet powerful dark matter that dominates gravity. It appears that almost every massive galaxy has a supermassive black hole in its center, which can be detected by its gravitational pull on surrounding stars in our own galaxy and by a characteristic stellar velocity pattern toward the center of other galaxies. The first galaxies formed from small quantum fluctuations that grew into gas that cooled and condensed into stars due to gravitational attraction; the first stellar black holes, which increase their size by consuming nearby stars and gas, were produced from explosions of massive stars (supernovae).



NASA/CXC/MIT/C. CANIZARES, D. EVANS, ET AL./STSCI/NSF/NRAO/VLA

A composite image showing wind from a supermassive black hole in the center of NGC 1068, a nearby galaxy

Aside from wanting to determine the composition of dark matter, cosmologists are also trying to understand why the current theory predicts the wrong number of galaxies. A possible explanation, which cosmologists are trying to match with observation, is that supermassive black hole winds, created when too much matter falls into the black hole, clear off the gas from its host galaxy and halt star formation. "This is a self-regulation process," said Zakamska. "If the galaxy tries to feed too much mass into the black hole, the black hole produces a powerful wind, shuts off the star formation, and stops its own feeding." A video of Zakamska's lecture "Gone with the Wind: Black Holes and their Gusty Influence on the Birth of Galaxies," sponsored by the Association of Members of the Institute for Advanced Study (AMIAS), may be viewed at http://video.ias.edu/zakamska-lecture-5-13/.

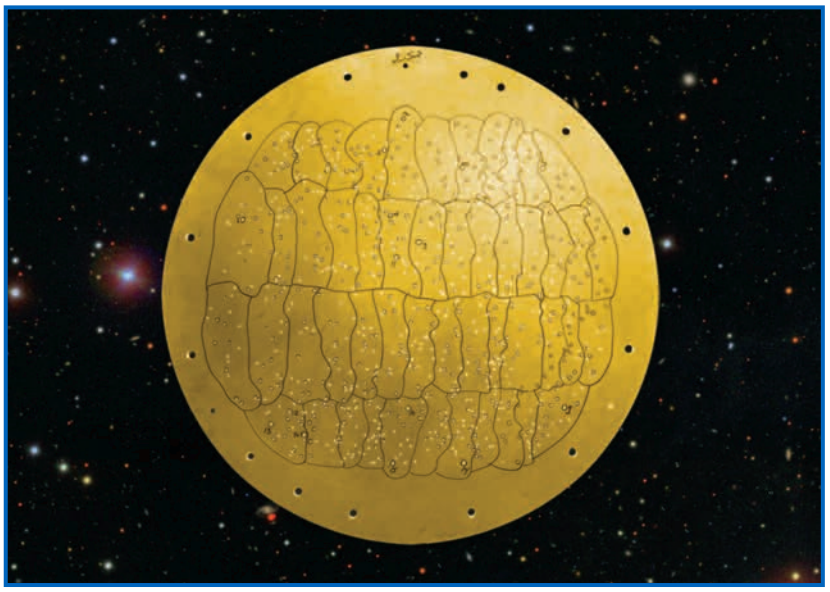
Measuring the Cosmos, Mapping the Galaxy, Finding Planets

By David H. Weinberg

Why is the expansion of the universe speeding up, instead of being slowed by the gravitational attraction of galaxies and dark matter? What is the history of the Milky Way galaxy and of the chemical elements in its stars? Why are the planetary systems discovered around other stars so different from our own solar system? These questions are the themes of SDSS-III, a six-year program of four giant astronomical surveys, and the focal point of my research at the Institute during the last year.

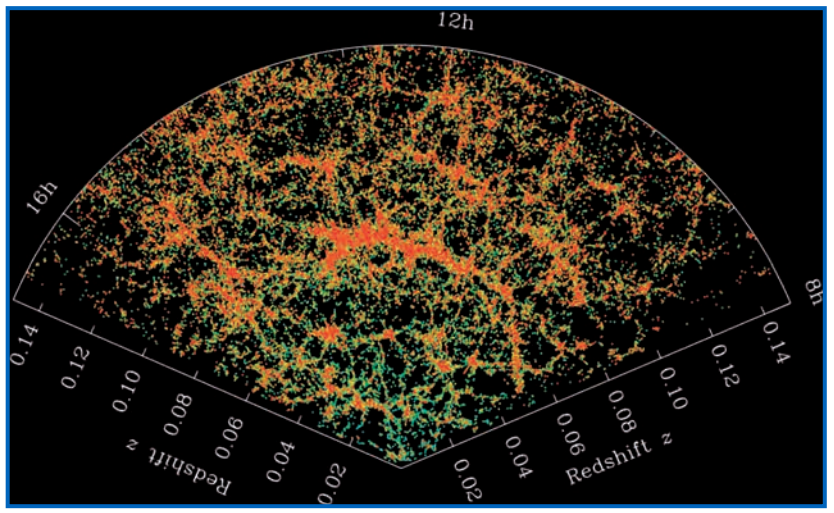
In fact, the Sloan Digital Sky Survey (SDSS) has been a running theme through all four of my stays at the Institute, which now span nearly two decades. As a long-term postdoctoral Member in the early 1990s, I joined in the effort to design the survey strategy and software system for the SDSS, a project that was then still in the early stages of fundraising, collaboration building, and hardware development. When I returned as a sabbatical visitor in 2001–02, SDSS observations were—finally—well underway. My concentration during that year was developing theoretical modeling and statistical analysis techniques, which we later applied to SDSS maps of cosmic structure to infer the clustering of invisible dark matter from the observable clustering of galaxies. By the time I returned for a one-term visit in 2006, the project had entered a new phase known as SDSS-II, and I had become the spokesperson of a collaboration that encompassed more than three hundred scientists at twenty-five institutions around the globe. With SDSS-II scheduled to complete its observations in mid-2008, I joined a seven-person committee that spent countless hours on the telephone that fall, sorting through many ideas suggested by the collaboration and putting together the program that became SDSS-III.

The SDSS uses a dedicated telescope (located in New Mexico) with a 2.5-meter-diameter mirror, similar in size to the Hubble Space Telescope's, but much smaller than those of the largest ground-based telescopes (whose mirrors are eight to ten meters across). What makes the SDSS special are the exceptionally powerful instruments on the *back* of the telescope. The first is a giant digital camera—the largest in the world at the time it was built—which has taken deep, multicolor images that cover more than half the northern-hemisphere sky, detecting over 100 million galaxies and 200 million stars. But to measure the distance to a galaxy or the velocity and chemical composition of a star, one has to disperse its light through a prism and identify the fine features etched on its spectrum by individual species of atoms, a kind of observation that astronomers have traditionally done one object at a time. The SDSS took this three-dimensional mapping into mass production by feeding its spectrographs with 640 optical fibers, plugged into 640 precision-drilled holes on a thirty-inch aluminum plate, each hole admitting the light from a single preselected galaxy, star, or quasar. After eight years of operations and more than 2,600 plates, SDSS-I and -II had measured spectra of nearly one million galaxies, more than one hundred thousand quasars, and half a million stars.



D. SCHLEGEL/SDSS-III

An SDSS-III plugplate, which admits light from preselected galaxies, stars, and quasars, superposed on an SDSS sky image



M. BLANTON/SDSS

Each dot on this slice through an SDSS map represents a galaxy, which is typically made up of about 100 billion stars. Blue dots mark younger and red dots mark older galaxies. The Earth is located at the vertex of the slice—the most distant galaxies in this map are 2 billion light years away from it.

The largest of the SDSS-III surveys (known as BOSS, the Baryon Oscillation Spectroscopic Survey) is aimed at the biggest mystery of contemporary cosmology: the accelerating expansion of the universe. While cosmic expansion was discovered eighty years ago by Edwin Hubble, it had generally been assumed that the expansion would slow down over time because of the gravitational attraction of matter in the universe. In the late 1990s, however, astronomers studying distant supernova explosions found that the expansion of the universe has been speeding up for the last five billion years. Either the universe is pervaded by an exotic form of energy that exerts repulsive gravity—perhaps the "vacuum energy" produced by quantum mechanical fluctuations in otherwise empty space—or else our prevailing theory of gravity itself breaks down on cosmological scales, maybe because gravity "leaks" into extra spatial dimensions that are hidden from our everyday experience.

BOSS will test the "vacuum energy" hypothesis with unprecedented precision, using a novel method that relies on a subtle feature in the clustering of galaxies and intergalactic matter. This feature, the imprint of "baryon acoustic oscillations" in the early universe, has a known physical scale, and after measuring its apparent size (e.g., as an angle on the sky) one can use simple trigonometry to infer the distances to objects that are billions of light years away. Precise determinations—accurate to 1 percent or better—require measuring cosmic structure over enormous volumes, which BOSS will do by mapping the spatial distribution of 1.5 million luminous galaxies and of absorbing gas along the lines of sight to 150,000 distant quasars. BOSS observes fainter objects than the original SDSS, so it required major upgrades to the spectrographs—more sensitive detectors, more efficient optical elements, 1,000 fibers instead of 640—which were installed and commissioned in fall 2009. The survey is now running full tilt and producing its first scientific results. However, the system is very complex, so a typical week still brings a software glitch or hardware problem that generates a cascade of email traffic and telecon discussion, and in rare cases an emergency trip to New Mexico by one of the instrument experts.

Closer to home, two SDSS-III surveys will map the structure and formation history of our own galaxy, the Milky Way. SEGUE-2 (whose acronymic history is too complicated to recount here) focuses on the outer galaxy, which observations and theory suggest was built largely via acts of galactic cannibalism, with the gravity of the Milky Way stretching and eventually destroying infalling satellite galaxies. The SEGUE maps (from SDSS-II and SDSS-III combined) contain about 350,000 stars, revealing partly digested strands of these galactic progenitors. The stellar motions measured by SEGUE also probe the mass and shape of the dark matter "halo" whose gravity holds the Milky Way together.

The inner galaxy is hidden from our view by interstellar dust, tiny smokelike particles that float between the stars and block visible light. APOGEE (the Apache Point Observatory Galactic Evolution Experiment) will map the inner galaxy using an innovative spectrograph that measures infrared light, which passes through interstellar dust nearly unscathed. With the exception of hydrogen, helium, and lithium, all atoms in the universe were forged in stars, then dispersed to the surrounding gas when the stars died. APOGEE spectra will allow separate measurements of a dozen chemical elements—carbon, oxygen, silicon, sulfur, iron, titanium, etc.—for each of the 100,000 stars that it observes. Because different elements form via different nuclear pathways in different kinds of stars, each of APOGEE's chemical "fingerprints" will encode information not just about the star being measured but about all of the preceding stars that contributed to its composition.

One of the biggest developments in astronomy over the last fifteen years has been the discovery of planets outside the solar system, most of them found via the slight wobble they induce as they orbit their parent stars. Many of the planetary systems discovered to date are very different from our own, with massive, Jupiter-like planets that loop around their parent stars in months or even days, often following elongated elliptical paths rather than the nearly circular orbits that prevail in the solar system. These oddities suggest that many planets "migrate" after birth or undergo chaotic gravitational battles with their siblings. The Sloan survey will, in characteristic fashion, attack this problem with large numbers, monitoring a total of 10,000 stars using a novel, fiber-fed instrument that can measure tiny motions (as small as a few meters per second) of sixty stars at a time. MARVELS (the Multi-object APO Radial Velocity Large-area Survey) hopes to detect between one and two hundred Jupiter-like planets in close orbits, allowing quantitative statistical tests of theories of planet formation and discovering rare systems that may reveal crucial short-lived phases in planetary evolution.

The Institute for Advanced Study helped start the SDSS with a critically timed financial contribution, but over the lifetime of the project its most important contributions have been human ones. Many Institute Members have done spectacular science with SDSS data over the years, and today four of the dozen scientists on the top-level SDSS-III management committee are former IAS postdocs. This is a remarkable statistic for a small institution focused largely on theoretical research. It speaks to the close interaction between theorists and observers in contemporary astronomy—with many individuals who straddle what was once a clear line of demarcation—and equally to the success of the Institute in inspiring its Members to pursue ambitious lines of research whose payoff may lie many years in the future. ■

David H. Weinberg is Professor of Astronomy and Distinguished Professor of Mathematical and Physical Sciences at Ohio State University. He was an AMIAS-supported Member in the School of Natural Sciences during the 2009–10 academic year and was a Member in the School in 1992–94, 2001–02, and 2006. He is the Project Scientist of SDSS-III.

Nova Aquilae and Extraterrestrial Intelligence that We May Not See

In the year 1918 a brilliant new star, called by astronomers Nova Aquilae, blazed for a few weeks in the equatorial sky. It was the brightest nova of this century. The biologist [J. B. S.] Haldane was serving with the British Army in India at the time and recorded his observation of the event:

> *Three Europeans in India looking at a great new star in the Milky Way. These were apparently all of the guests at a large dance who were interested in such matters. Amongst those who were at all competent to form views as to the origin of this cosmoclastic explosion, the most popular theory attributed it to a collision between two stars, or a star and a nebula. There seem, however, to be at least two possible alternatives to this hypothesis. Perhaps it was the last judgment of some inhabited world, perhaps a too successful experiment in induced radioactivity on the part of some of the dwellers there. And perhaps also these two hypotheses are one, and what we were watching that evening was the detonation of a world on which too many men came out to look at the stars when they should have been dancing.*

A few words are needed to explain Haldane's archaic language. He used the phrase "induced radioactivity" to mean what we now call nuclear energy. He was writing fifteen years before the discovery of fission made nuclear energy accessible to mankind. In 1924, scientifically educated people were aware of the enormous store of energy that is locked up in the nucleus of uranium and released slowly in the process of natural radioactivity. The equation $E=mc^2$ was already well known. But attempts to speed up or slow down natural radioactivity by artificial means had failed totally. The nuclear physicists of that time did not take seriously the idea that "induced radioactivity" might one day place in men's hands the power to release vast quantities of energy for good or evil purposes. Haldane had the advantage of being an outsider, a biologist unfamiliar with the details of nuclear physics. He was willing to go against the opinion of the experts in suggesting "induced radioactivity" as a possible cause of terrestrial or extraterrestrial disasters.

The example of Nova Aquilae raises several questions which we must answer before we can begin a serious search for evidence of intelligent life existing elsewhere in the universe. Where should we look, and how should we recognize the evidence when we see it? Nova Aquilae was for several nights the second brightest star in the sky. One had to be either very blind or very busy not to see it. Perhaps it was an artifact of a technological civilization, as Haldane suggested. How can we be sure that it was not? And how can we be sure that we are not now missing equally conspicuous evidence of extraterrestrial intelligence through not understanding what we see? There are many strange and poorly understood objects in the sky. If one of them happens to be artificial, it might stare us in the face for decades and still not be recognized for what it is.

—Freeman Dyson, Professor Emeritus in the School of Natural Sciences, in Disturbing the Universe *(Basic Books, 1979)*

Life on Other Planets

BY DAVID S. SPIEGEL

Until a couple of decades ago, the only planets we knew existed were the nine in our solar system. In the last twenty-five years, we've lost one of the local ones (Pluto, now classified as a "minor planet") and gained about three thousand candidate planets around other stars, dubbed exoplanets. The new field of exoplanetary science is perhaps the fastest growing subfield of astrophysics, and will remain a core discipline for the forseeable future.

The fact that any biology beyond Earth seems likely to live on such a planet is among the many reasons why the study of exoplanets is so compelling. In short, planets are not merely astrophysical objects but also (at least some of them) potential abodes.

VLADIMIR VOEVODSKY

Sighting a deer outside of Fuld Hall; how long did Earth have to wait before life "walked into sight"?

The highly successful Kepler mission involves a satellite with a sensitive telescope/camera that stares at a patch of sky in the direction of the constellation Cygnus. The goal of the mission is to find what fraction of Sun-like stars have Earth-sized planets with a similar Earth-Sun separation (about 150 million kilometers, or the distance light travels in eight minutes). During its half-decade mission lifetime, Kepler will be monitoring 150,000 stars, looking for slight periodic dips in starlight that occur if an exoplanet's orbital plane is oriented precisely along our line of sight. In this geometrical configuration, the planet moves directly between us and its parent star once per orbit, blocking a tiny fraction of the light from the star. Kepler has identified more than two thousand planet candidates so far, most of which are probably real. Early results suggest that somewhere between 5 percent and 50 percent of Sun-like stars probably have an approximate Earth-analogue!

So, we are starting to realize that potential homes for life are probably common in our galaxy. Among the several hundred billion stars, there might be tens of billions of rocky planets located in the "habitable zones" of their stars—the regions where they would have roughly Earth-like temperatures. With so many possible places where life might flourish, how much life can we expect is out there? This question might seem to invite nothing but wild speculation. However, there is a potential avenue for making an estimate.

As an analogy, consider someone who wants to know what fraction of the time there is a deer visible outside her window. One way to estimate this would be to sit by the window, looking out, and see how long she has to wait for the first deer to walk into sight. In Manhattan, the expected wait might be decades, and one could rightly infer that the fraction of the time that there is a deer in sight is very close to zero. In Fuld Hall at IAS, one probably wouldn't have to wait more than a few hours, and could rightly infer that deer are pretty frequently visible outside the window.

Similarly, we can look through the Earth's geological history to see when life appeared in Earth's history. How long, in other words, did Earth have to wait before life "walked into sight"? Earth was born about 4.5 billion years ago, but for the first half billion years of its existence, it was bombarded by impactors that probably sterilized it. For the past four billion years, though, the Earth has been essentially continuously habitable (meaning, it has had conditions suitable for liquid-water-based life). There is some evidence that the earliest living organisms had developed by 3.8 billion years ago, or within the first two hundred million years of the habitable history of the Earth. A common line of reasoning in the origin of life community argues that since abiogenesis (the process of life arising from abiotic conditions) occurred so early in the geological time scale of the Earth, it must be a reasonably probable process.

This argument is appealing, and it's certainly true that the early emergence of life on Earth provides some reason for optimism for an enthusiast of extrasolar life. However, together with Professor Edwin Turner of Princeton University, I recently critically reevaluated this argument (in the *Proceedings of the National Academy of Sciences*, vol. 109, issue 2) and found that we have less reason to expect that the galaxy is full of life than is sometimes assumed. One important reason is that there is a powerful selection effect in operation that is absent in the deer analogy. Specifically, what we know is more than simply that life showed up early; we also know that we are aware that life showed up early. Put differently, in order for us to exist, enough time had to have elapsed after abiogenesis occurred such that creatures could evolve who are capable of contemplating the frequency of inhabited planets. We don't know what the minimum evolutionary time scale is (in our case, it was about 3.8 billion years), but if, for instance, this minimum time scale is 3.5 billion years, then it would be impossible for us to find ourselves on a planet with late abiogenesis no matter how rare the abiogenesis process is.

What we know is more than simply that life showed up early; we also know that we are aware that life showed up early.

Thankfully, we will soon be able to empirically test the hypothesis that our galaxy is teeming with life. Even a single example of life that had a different abiogenesis event would count much more strongly toward the conclusion that life is common, given the right conditions. Possible examples of life with an independent origin include:

- so-called "shadow life" here on Earth (i.e., life that arose entirely separately from the tree of life, springing from a single root that is believed to encompass all currently known species);
- life elsewhere in our solar system (e.g., on Mars or Jupiter's moon Europa), if we were confident that there was no panspermia in either direction (i.e., that we didn't seed, for example, the Martian life via asteroids transporting material from Earth to Mars, nor they us);
- or life on an exoplanet.

Within the next several decades, with rapid progress in biology, with new space missions, and with large and sensitive new telescopes that are planned, it is conceivable that we might find any of these three kinds of independent life. In this way, we will be able to make a much more informed estimate of the frequency of life in the universe. ■

David S. Spiegel, Friends of the Institute for Advanced Study Member (2012–13) in the School of Natural Sciences, is focusing his research on theoretical studies of the climates of, and radiative transfer in, exoplanetary atmospheres; habitability models of terrestrial exoplanets; and radiation-dynamical models of gas giant planets.

How Open-Source Ideas Can Help Us Study Exoplanets

BY HANNO REIN

Pluto, the ninth planet in our solar system[1] was discovered in 1930, the same year the Institute was founded. While the Institute hosted more than five thousand members in the following sixty-five years, not a single new planet was discovered during the same time.

Finally, in 1995, astronomers spotted an object they called 51 Pegasi b. It was the first discovery of a planet in over half a century. Not only that, it was also the first planet around a Sun-like star outside our own solar system. We now call these planets extrasolar planets, or in short, exoplanets.

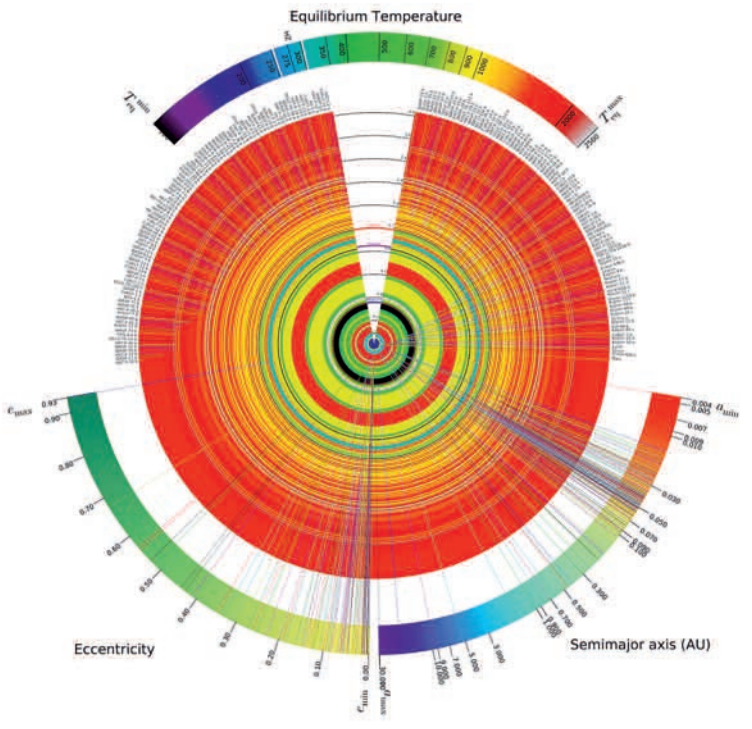
As it turns out, 51 Pegasi b is a pretty weird object. It is almost as massive as Jupiter, but it orbits its host star in only four days. Jupiter, as a comparison, needs twelve years to go around the Sun once. Because 51 Pegasi b is very close to the star, its equilibrium temperature is very high. These types of planets are often referred to as "hot Jupiters."

Since the first exoplanet was discovered, the technology has improved dramatically, and worldwide efforts by astronomers to detect exoplanets now yield a large number of planet detections each year. In 2011, 189 planets were discovered, approximately the number of visiting Members at the Institute every year. In 2012, 130 new planets were found. As of May 20 of this year, the total number of confirmed exoplanets was 892 in 691 different planetary systems.

Personally, I am very interested in the formation of these systems. We have so much information about every planet in our solar system, but little is known about all of these 892 exoplanets. Digging into this limited data set and trying to find out how exoplanets obtain their present-day orbits is very exciting. Many questions pop up by just looking at 51 Pegasi b. Why is it a hundred times closer to its star than Jupiter? Did it form farther out? Was it not too different from our own Jupiter in the past? For 51 Pegasi b, we think we know the answer. We believe that it formed at a farther distance from its star where conditions such as temperature are more favorable for planet formation, and then it moved inwards in a process called planet migration. For many of the other 891 planets, the story is more complicated, especially when multiple planets are involved. The diversity of planetary systems that have been found is tremendous. We haven't discovered a single system that looks remotely similar to our own solar system. This makes exoplanetary systems so exciting to study!

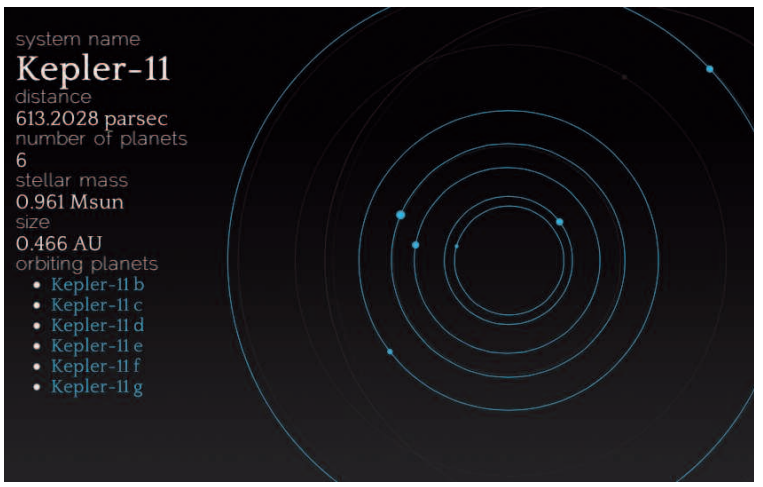
To do this kind of research, one needs a catalogue of all exoplanets. Several such databases exist, but they all share one fundamental flaw: they are not "open." These databases are maintained either by a single person or by a small group of scientists. It is impossible to make contributions to the database if one is not part of this inner circle. This bothered me because it is not the most efficient way, and it does not encourage collaboration among scientists. I therefore started a new project during my time at the Institute, the Open Exoplanet Catalogue. As the name suggests, this database, in comparison to others, is indeed "open." Everyone is welcome to contribute, make corrections, or add new data. Think of it as the Wikipedia version of an astronomical database.

The same idea has been extremely successful in the software world. With an open-source license, programmers provide anyone with the rights to study, modify, and distribute the software that they have written—for free. The obvious advantages are affordability and transparency. But maybe more importantly, perpetuity, flexibility, and interoperability are vastly improved by making the source code of software publicly available.



JORGE ZULUAGA

The customizable Comprehensive Exoplanetary Radial Chart illustrates the radii of planets according to colors that represent equilibrium temperatures, eccentricity, and other data relevant to assessing their potential habitability.



TOM HANDS

The winning submission to the Exoplanet Visualization Contest is ExoVis, an interactive website that allows one to visualize, search for, study, and compare all planetary systems in the Open Exoplanet Catalogue.

The success of the open-source movement is phenomenal. Every time you start a computer, open a web browser, or send an email, you are utilizing an open-source program, often in the background. The success story of open source is largely based on the wide adoption of distributed version-control systems.[2] These toolkits allow thousands of people to work and collaborate together on a single project. Every change ever made to any file can be traced back to an individual person. This creates a network of trust, based on human relationships. Initially, the concept of having thousands of people working on the same project may appear chaotic, risky, or plain impossible. However, studies have shown that this kind of large-scale collaboration produces software that is better[3] and more secure than using a traditional approach.

Recommended Reading: More information about the Open Exoplanet Catalogue, its workflow, and data format is available online at www.openexoplanetcatalogue.com/. Tom Hand's ExoVis website is hosted at www.tomhands.com/exovis/. High-resolution images of Jorge Zuluaga's Comprehensive Exoplanetary Radial Chart may be found at http://astronomia.udea.edu.co/iCERC/.

Astrophysics lags behind this revolution. While there are some software packages that are open source (and widely used), the idea of applying the same principles to data sets and catalogues is new. Extrasolar planets provide an ideal test case because the data set is generated by many different groups of observers from all around the world. Observations and discoveries are evolving so quickly that a static catalogue is not an option anymore.

To get people excited about the ideas and philosophy behind the Open Exoplanet Catalogue, I started a visualization competition, the "Exoplanet Visualization Contest," with the goal of coming up with stunning and creative ways to visualize exoplanet data. We set no restrictions to the kind of submission. The only requirement was that each submission had to use real data from the Open Exoplanet Catalogue. This led to an extremely diverse set of submissions. For example, we received publication-grade scientific plots, artistic drawings of potentially habitable exomoons, and an interactive website. One participant went so far as to design a wearable vest with built-in microcontrollers and displays that show exoplanet data. Thanks to a generous outreach grant from the Royal Astronomical Society in London, we were able to give out prizes to the best submissions. With the help of Scott Tremaine (Richard Black Professor in the School), Dave Spiegel (Member in the School), and Dan Fabrycky (Assistant Professor at the University of Chicago), two winners were chosen.

Second prize went to Jorge Zuluaga from Antioquia, Colombia. He designed a new way to present exoplanet data, such as planetary sizes and equilibrium temperatures. Those are of particular interest when it comes to determining whether a planet is potentially habitable or not. His submission, the Comprehensive Exoplanetary Radial Chart, illustrates the radii of exoplanets according to colors that represent their approximate equilibrium temperatures. The chart also shows information on planetary orbit properties, size of host stars, and potentially any other variable of interest.

The winner of the contest was Tom Hands, a Ph.D. student from Leicester. He wrote an interactive website, ExoVis, that visualizes all discovered planetary systems. The project makes use of HTML5, Javascript, jQuery, and PHP. One can search for planets, study their orbital parameters, and compare them to other systems, all within a web browser.

The Open Exoplanet Catalogue is a very new project. The crucial issue is to reach a large number of regular contributors; then, the quality of the data set will outperform all "closed" competitors in the long run in the same way Wikipedia is now much more widely used than the *Encyclopædia Britannica*. I am optimistic about the future. ■

Hanno Rein, Member (2010–13) in the School of Natural Sciences, is studying the formation and evolution of planetary systems. In addition to the Open Exoplanet Catalogue, he has developed a smartphone application called Exoplanet (http://exoplanetapp.com), which has attracted almost ten million users. The application lists all extrasolar planets and features three-dimensional visualizations and an interactive model of the Milky Way. Rein describes it as similar to Google Earth, but for the entire universe. He will be joining the University of Toronto as an Assistant Professor this fall.

1 Pluto was originally classified as the ninth planet in the solar system. In 2005, the International Astronomical Union decided to call Pluto a dwarf planet.

2 The most popular of those tools is Git, used by people who write the Linux kernel and many other major open-source projects.

3 In the software world, "better" is measured in units of bugs per line of code.

Univalent Foundations and the Large-Scale Formalization of Mathematics

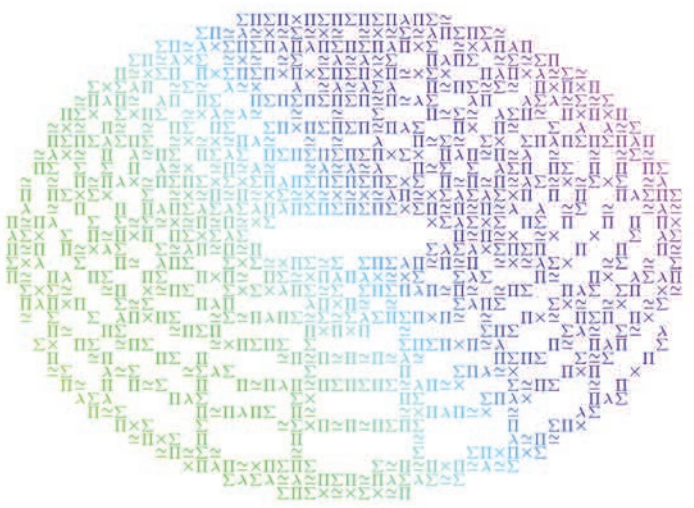
BY STEVE AWODEY AND THIERRY COQUAND

In 2012–13, the Institute's School of Mathematics hosted a special year devoted to the topic "Univalent Foundations of Mathematics," organized by Steve Awodey, Professor at Carnegie Mellon University, Thierry Coquand, Professor at the University of Gothenburg, and Vladimir Voevodsky, Professor in the School of Mathematics. This research program was centered on developing new foundations of mathematics that are well suited to the use of computerized proof assistants as an aid in formalizing mathematics. Such proof systems can be used to verify the correctness of individual mathematical proofs and can also allow a community of mathematicians to build shared, searchable libraries of formalized definitions, theorems, and proofs, facilitating the large-scale formalization of mathematics.

The possibility of such computational tools is based ultimately on the idea of logical foundations of mathematics, a philosophically fascinating development that, since its beginnings in the nineteenth century, has, however, had little practical influence on everyday mathematics. But advances in computer formalizations in the last decade have increased the practical utility of logical foundations of mathematics. Univalent foundations is the next step in this development: a new foundation based on a logical system called type theory that is well suited both to human mathematical practice and to computer formalization. It draws moreover on new insights from homotopy theory—the branch of mathematics devoted to the study of continuous deformations in space. This is a particularly surprising source, since the field is generally seen as far distant from foundations.

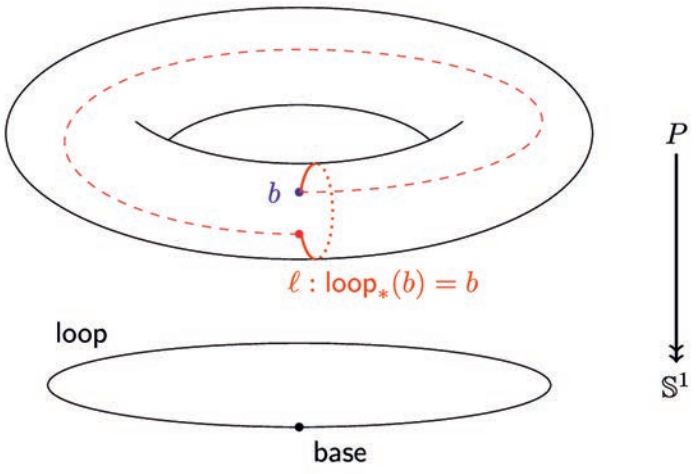
For the special year, a team of thirty-two leading researchers in the areas of computer science, logic, and mathematics from around the world was assembled at IAS to develop this new foundation of mathematics. An ambitious program of weekly seminars, lectures, working groups, tutorials, and other activities led to a lively interaction and a vigorous exchange of ideas, skills, and viewpoints, resulting in numerous collaborations among the participants. The program's goals were realized beyond expectations, producing a powerful and flexible new foundational system called homotopy type theory, based on earlier systems of type theory that were originally intended for constructive mathematics and computer programming, and augmented by new principles motivated by homotopy theory. In addition to a body of theoretical results pertaining to the foundations, a substantial amount of mathematics was developed in this new system, including basic results in homotopy theory, higher category theory, set theory, and the beginnings of real analysis. In parallel, efforts were focused on the development of new and existing computer proof assistants for the formalization of these and future results. An extensive library of code was established on which future work can be built, and formalized proofs of significant results in homotopy theory were given, such as computing many homotopy groups of spheres. In a remarkable, collaborative effort, a textbook was also written by the special-year participants, developing both the foundations and various specialized areas of mathematics in the new logical system. This book not only serves as a record of the results of the special year, but also as a useful introduction for future researchers entering the field.

This new conception of foundations of mathematics, so closely tied to the use of computers, seems so natural that future historians of mathematics may well wonder how Frege and Russell could have invented the idea of formal systems of foundations before there were any computers to run them on.



THE HOTT BOOK

Depicted here is a mathematical torus (a donut-shaped object that cannot be deformed into a sphere in a continuous way without tearing it) made of logical symbols. It represents homotopy type theory, a new branch of mathematics that connects homotopy theory (the study of continuous deformations) and type theory (a branch of mathematical logic).



A figure from the book Homotopy Type Theory, *illustrating the principle of "circle induction." In homotopy type theory, basic geometric objects such as the circle are implemented using the same kind of inductive definitions typically used for the natural numbers, allowing for reasoning "by induction" in an entirely new form. In this case, if a property holds at the base point, and it holds "continuously in the loop," then it holds everywhere on the circle.*

The idea of logical foundations of mathematics goes back at least to Gottlob Frege's *Begriffsschrift* of 1879, which set out to show that arithmetic could be deduced entirely from logic in a way that was "free of gaps" and thus requiring no leaps of intuition. Frege's system of logical deductions—which looked a bit like complicated wiring diagrams—was soon discovered by Bertrand Russell to contain a contradiction: a disastrous logical inconsistency, which had the effect that mathematicians otherwise unconcerned with logic began to pay increased attention to logical precision. Russell himself proposed a solution based on what he called the theory of types, and Ernst Zermelo proposed another based on axioms for Georg Cantor's set theory. During the 1920s and '30s, mathematicians as prominent as David Hilbert, Hermann Weyl (Professor, 1933–51), and John von Neumann (Professor, 1933–57) worked on the foundations of mathematics, culminating in the famous discoveries of Kurt Gödel (Member, beginning in 1933; Professor, 1953–76) about the limits of logical formalization. Gödel showed namely that a complete and consistent logical formalization of even arithmetic was mathematically impossible; moreover, this result agreed with the practical experience of many mathematicians, that the formalization of even the most basic mathematical theories was impractically complicated and irrelevantly elaborate. Russell's system arrived at the result that 1 + 1 = 2 only after 362 pages of laborious formal deductions!

By the 1950s, a consensus had settled in mathematics that the program of logical foundations, while perhaps interesting in principle or as its own branch of mathematics, was going to be of little use to the general practice of mathematics as a whole. This view was only reinforced by the results of Gödel and Paul Cohen (Member, 1959–61, 67) on the formal undecidability of the famous continuum hypothesis. Much subsequent research in logical theory was related instead to the new field of computation; indeed, it was the early work in logic that had led to the development of the modern computer, and subsequent advances in theoretical and practical computation were closely tied to logical research.

But with recent advances in the speed and capacity of modern computers and theoretical advances in their programming has come a remarkable and somewhat ironic possibility: the use of computers to aid in the nearly forgotten program of formalization of mathematics. For what was once too complicated or tedious to be done by a human could now become the job of a computer. With this advance comes the very real potential that logical foundations, in the form of computer formalizations, could finally become a practical aid to the mathematician's everyday work, through verifying the correctness of definitions and proofs, organizing large-scale theories, making use of libraries of formalized results, and facilitating collaboration on the development of a unified system of formalized mathematics. Gödel may have shown that mathematics cannot be entirely formalized in principle, but in practice there are still great benefits to be had from formalization when sophisticated computer systems can be brought to bear.

This new conception of foundations of mathematics, so closely tied to the use of computers to provide the guarantee of formal rigor and to aid in handling the explosion of complexity, seems so natural that future historians of mathematics may well wonder how Frege and Russell could have invented the idea of formal systems of foundations before there were any computers to run them on. Nor is it a coincidence that foundations work so well in combination with computers; as already stated, the modern computer was essentially born from the early research in logic, and its modern programs and

(Continued on page 13)

Recommended Reading: For more information, visit http://homotopytypetheory.org, a website for the collection and dissemination of research, resources, and tools for the investigation of homotopy type theory.

Steve Awodey, Member (2012–13) in the School of Mathematics, is Professor of Philosophy at Carnegie Mellon University, specializing in logic and category theory. Awodey's membership is supported by the Friends of the Institute for Advanced Study and the Charles Simonyi Endowment. Thierry Coquand, Member (2012–13) in the School of Mathematics, is Professor of Computer Science at the University of Gothenburg in Sweden, specializing in logic and constructive mathematics. Coquand's membership is supported by the Ellentuck Fund and the Charles Simonyi Endowment.

Socio-Technological Aspects of Making the HoTT Book

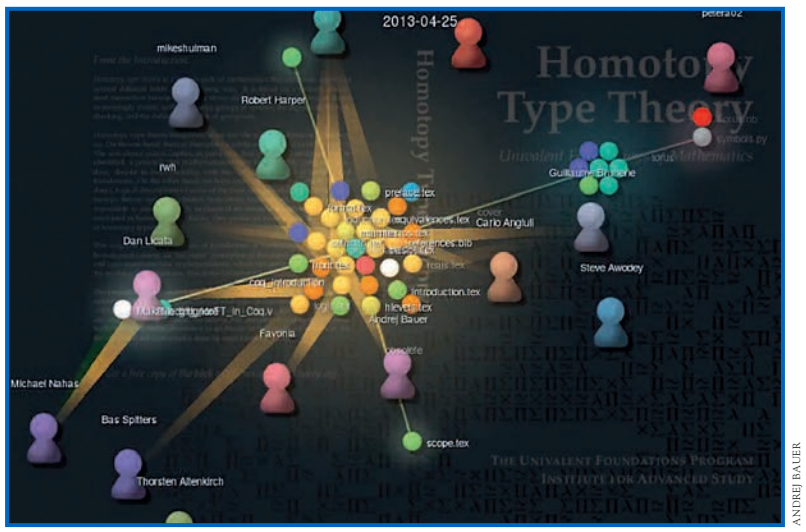
BY ANDREJ BAUER

Since spring, and even before that, I have participated in a great collaborative effort to write a book on homotopy type theory. It is finally finished and ready for public consumption. You can get the book freely at http://homotopytypetheory.org/book/. Mike Shulman has written about the contents of the book (http://golem.ph.utexas.edu/category/2013/06/the_hott_book.html), so I am not going to repeat that here. Instead, I would like to comment on the socio-technological aspects of making the book and in particular about what we learned from the open-source community about collaborative research.

We are a group of two dozen mathematicians who wrote a six-hundred-page book in less than half a year. This is quite amazing since mathematicians do not normally work together in large groups. A small group can get away with using obsolete technology, such as sending each other source LaTeX files by email, but with two dozen people even Dropbox or any other file synchronization system would have failed miserably. Luckily, many of us are computer scientists disguised as mathematicians, so we knew how to tackle the logistics. We used Git and GitHub.com. In the beginning, it took some convincing and getting used to, although it was not too bad. In the end, the repository served not only as an archive for our files but also as a central hub for planning and discussions. For several months, I checked GitHub more often than email and Facebook.

The spirit of collaboration that pervaded our group at the Institute for Advanced Study was truly amazing. We did not fragment. We talked, shared ideas, explained things to each other, and completely forgot who did what. The result was a substantial increase in productivity.

But more importantly, the spirit of collaboration that pervaded our group at the Institute for Advanced Study was truly amazing. We did not fragment. We talked, shared ideas, explained things to each other, and completely forgot who did what (so much in fact that we had to put some effort into reconstruction of history lest it be forgotten forever). The result was a substantial increase in productivity. There is a lesson to be learned here (other than the fact that the Institute is the world's premier research institution), namely that mathematicians benefit from being a little less possessive about their ideas and results. I know, I know, academic careers depend on proper credit being given and so on but really those are just the idiosyncrasies of our time. If we can get mathematicians to share half-baked ideas, not to worry who contributed what to a paper, or even who the authors are, then we will reach a new and unimagined level of productivity. Progress is made by those who dare to break the rules.



An animated visualization of the GitHub collaborations of over two dozen mathematicians working on the HoTT book over a period of six months may be viewed at http://vimeo.com/68761218/.

Truly open research habitats cannot be obstructed by copyright, publishers, patents, commercial secrets, and funding schemes that are based on faulty achievement metrics. Unfortunately, we are all caught up in a system that suffers from all of these evils. But we made a small step in the right direction by making the book source code freely available under a permissive Creative Commons license. Anyone can take the book and modify it, send us improvements and corrections, translate it, or even sell it without giving us any money. (If you twitched a little bit when you read that sentence, the system has gotten to you.)

We decided not to publish the book with an academic publisher at present because we wanted to make it available to everyone fast and at no cost. The book can be freely downloaded, as well as bought inexpensively in hardcover and paperback versions from Lulu.com. (When was the last time you paid under $30 for a six-hundred-page hardcover academic monograph?) Again, I can sense some people thinking, "Oh, but a real academic publisher bestows quality." This sort of thinking is reminiscent of Wikipedia vs. *Britannica* arguments, and we all know how that story ended. Yes, good quality research must be ensured. But once we accept the fact that anyone can publish anything on the internet for the whole world to see and make a cheap, professional-looking book out of it, we quickly realize that censure is not effective anymore. Instead we need a decentralized system of endorsements that cannot be manipulated by special interest groups. Things are moving in this direction with the recently established Selected Papers Network (https://selectedpapers.net) and similar efforts. I hope these will catch on.

However, there is something else we can do. It is more radical but also more useful. Rather than letting people only evaluate papers, why not give them a chance to participate and improve them as well? Put all your papers on GitHub and let others discuss them, open issues, fork them, improve them, and send you corrections. Does it sound crazy? Of course it does. Open source also sounded crazy when Richard Stallman announced his manifesto. Let us be honest, who is going to steal your LaTeX source code? There are much more valuable things to be stolen. If you are a tenured professor, you can afford to lead the way. Have your grad student teach you Git and put your stuff somewhere publicly. Do not be afraid; they tenured you to do such things.

We are inviting everyone to help us improve the HoTT book by participating on GitHub. You can leave comments, point out errors, or even better, make corrections yourself! We are not going to worry about who you are, how much you are contributing, and who shall take credit. The only thing that matters is whether your contributions are any good.

My last observation is about formalization of mathematics. Mathematicians like to imagine that their papers could in principle be formalized in set theory. This gives them a feeling of security, not unlike the one experienced by a devout man entering a venerable cathedral. It is a form of faith professed by logicians. Homotopy type theory is an alternative foundation to set theory. We too claim that ordinary mathematics can in principle be formalized in homotopy type theory. But you do not have to take our word for it! We have formalized the hardest parts of the HoTT book and verified the proofs with computer proof assistants. Not once but twice. And we formalized first, then we wrote the book, because it was easier to formalize. We win on all counts (if there is a race). I hope you like the book. It contains an amazing amount of new mathematics. ■

Andrej Bauer, Member (2012) in the School of Mathematics, is broadly interested in computation and the computational nature of mathematics, approaching the subject through logic, category theory, type theory, and constructive mathematics. He also works on mathematical foundations of programming languages with emphasis on their mathematical semantics. The text excerpted here first appeared at http://math.andrej.com/2013/06/20/the-hott-book/.

UNIVALENT FOUNDATIONS *(Continued from page 12)*

operating systems are still closely related to the logical systems from which they evolved. In a sense, modern computers can be said to "run on logic."

This is the starting point of the univalent foundations program: the idea that the time is now ripe for the development of computer proof assistants based on new foundations of mathematics. But it is not only the advance in technology that has made this program feasible today; recent breakthroughs in logical theory also play an important role. One such advance was the discovery of a connection between the system of type theory used by some modern proof systems and the mathematical field of homotopy theory, which usually requires a high level of mathematical abstraction to even get off the ground. This connection permits direct, logical formalizations of some important concepts having broad application in various fields of mathematics. An important example is the fundamental notion of a set, which in univalent foundations turns out to be definable from more primitive concepts, as was recently discovered by Voevodsky. A related discovery, also due to Voevodsky, is the univalence axiom, which states, roughly, that isomorphic mathematical objects may be identified. This powerful new principle of reasoning, which agrees with everyday mathematical practice but is not part of the traditional set-theoretic foundation, is fully compatible with the homotopical view, and indeed strengthens it, while greatly simplifying the use of type theory as a system of foundations. Finally, the discovery of direct, logical descriptions of some basic mathematical spaces, such as the n-dimensional spheres S^n, and various other fundamental constructions, has led to a system that is both comprehensive and powerful, while still being closely tied to implementation on a computer. ■

References

Steve Awodey, Alvaro Pelayo, and Michael A. Warren, "Voevodsky's Univalence Axiom in Homotopy Type Theory," *Notices of the American Mathematical Society* (forthcoming).

George Dyson, "Julian Bigelow: Bridging Abstract Logic and Practical Machines," *The Institute Letter* (Spring 2013): 14–15.

Thomas C. Hales, "Formal Proof," *Notices of the American Mathematical Society* 55, no. 11 (2008): 1370–80.

The Univalent Foundations Program at the Institute for Advanced Study, *Homotopy Type Theory: Univalent Foundations of Mathematics* (2013). http://homotopytypetheory.org/book/.

The Geometry of Random Spaces

BY MATTHEW KAHLE

I sometimes like to think about what it might be like inside a black hole. What does that even mean? Is it really "like" anything inside a black hole? Nature keeps us from ever knowing. (Well, what we know for sure is that nature keeps us from knowing *and* coming back to tell anyone about it.) But mathematics and physics make some predictions.

John Wheeler suggested in the 1960s that inside a black hole the fabric of spacetime might be reduced to a kind of quantum foam. Kip Thorne described the idea in his book *Black Holes & Time Warps* as follows (see Figure 1).

> *This random, probabilistic froth is the thing of which the singularity is made, and the froth is governed by the laws of quantum gravity. In the froth, space does not have any definite shape (that is, any definite curvature, or even any definite topology). Instead, space has various probabilities for this, that, or another curvature and topology. For example, inside the singularity there might be a 0.1 percent probability for the curvature and topology of space to have the form shown in (a), and a 0.4 percent probability for the form in (b), and a 0.02 percent probability for the form in (c), and so on.*

In other words, perhaps we cannot say exactly what the properties of spacetime are in the immediate vicinity of a singularity, but perhaps we could characterize their distribution. By way of analogy, if we know that we are going to flip a fair coin a thousand times, we have no idea whether any particular flip will turn up heads or tails. But we can say that on average, we should expect about five hundred heads. Moreover, if we did the experiment many times we should expect a bell-curve shape (i.e., a normal distribution), so it is very unlikely, for example, that we would see more than six hundred heads.

To get a feel for a random space, here is an example that you can make at home. All you need is a deck of playing cards, some paper, scissors, and tape.

Make a mini-deck of twelve playing cards: ace, two, three, four, five, six, seven, eight, nine, ten, jack, queen. Cut four paper triangles. Label their sides A–Q (to correspond to your deck of cards) as in Figure 2. Shuffle the cards, then take the top two cards from the deck. Say the cards are five and seven: tape the side labeled five to the side labeled seven, keeping the printed side of each triangle side up. (This ensures that you end up with an orientable surface.) Again, take the top two cards from the deck, tape the corresponding triangle sides, and repeat until you have used all twelve cards and all twelve sides are taped. As you get toward the end, you might have to really bend up your paper. But after gluing six pairs, you are mathematically certain to have a surface. What is uncertain is which surface.

One might end up with a surface homeomorphic (i.e., continuously deformable) to a sphere or a torus. But one might also end up with two spheres or a sphere and a torus, so the surface need not be connected. However, if one did this with many triangles, it would be very likely that the surface would be connected and the main question would be its genus—i.e., how many "handles" or "holes" does it have. It turns out that if one glues together n triangles randomly in this way, one should expect a surface of genus roughly $n/4$, on average. (This is a theorem of Nicholas Pippenger and Kristin Schleich, and independently of Nathan Dunfield and William Thurston.)

It turns out that this relatively simple model of a random space already encodes a lot of physics as n tends to infinity, and in fact one of the motivations to study it is that it serves as a two-dimensional discrete analogue of quantum gravity. So random spaces provide a mathematical model of something of fundamental interest in theoretical physics and cosmology.

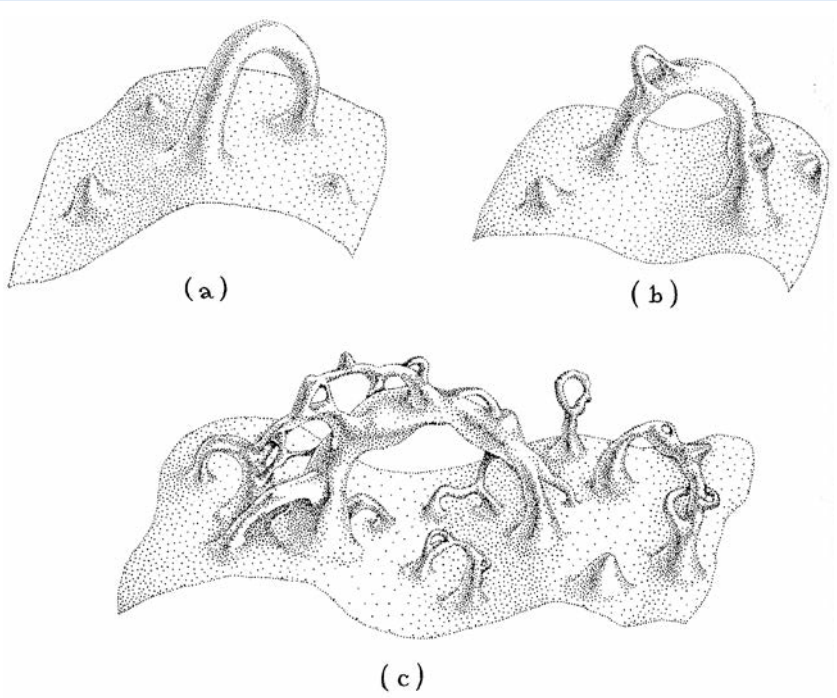


Figure 1. From Black Holes & Time Warps: Einstein's Outrageous Legacy *by Kip S. Thorne (W. W. Norton & Company, Inc., 1994)*

John Wheeler suggested in the 1960s that inside a black hole the fabric of spacetime might be reduced to a kind of quantum foam.... Perhaps we cannot say exactly what the properties of spacetime are in the immediate vicinity of a singularity, but perhaps we could characterize their distribution.

Random spaces also provide interesting models within mathematics itself, as well as useful constructions in theoretical computer science. To mathematicians and theoretical computer scientists, one of the important discoveries of the last fifty years is that random objects often have desirable, hard to come by otherwise, properties. There are many examples of this paradigm by now, but one of the first was in Ramsey theory.

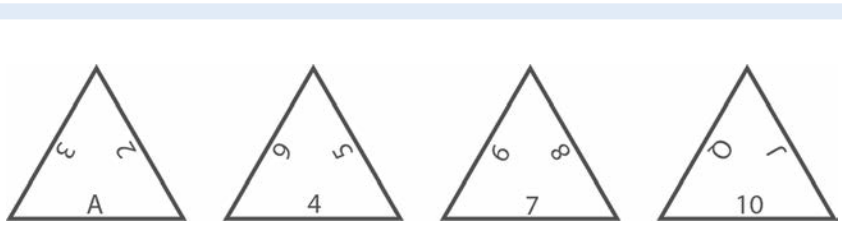


Figure 2

A combinatorial fact: among any party of six people, there must be either three mutual acquaintances or three mutual nonacquaintances. This isn't necessarily true for five people. Let $R(n)$ denote the smallest number of people that guarantees that if you have a party of $R(n)$ people there are either n mutual acquaintances or n mutual non-acquaintances. So, taking the example above, $R(3)=6$. It is also known that $R(4)=18$, i.e., among any eighteen people, there must be either four mutual acquaintances or four mutual non-acquaintances. But $R(n)$ isn't known for any larger n.

Paul Erdős suggested that if advanced aliens threaten the Earth, telling us they will blow us up unless we tell them $R(5)$ within a year, we should put together all the best minds and use all our computer resources and see if we can figure it out. But if they ask for $R(6)$, he warned, we should probably attack first.

When mathematicians can't compute something exactly, we often look for bounds or estimates. In the case of Ramsey theory, lower bounds come from somehow arranging a party with not too many mutual acquaintances or nonacquaintances. As the number of people gets large, to describe this kind of structure explicitly gets unwieldy, and after decades of people thinking about it, no one really knows how to do it very well. The best lower bounds we know come from the simple strategy of assigning acquaintanceship randomly.

This is a surprising idea at first, but it turns out to be powerful in a variety of settings. In many problems one wants to maximize some quantity under certain constraints. If the constraints seem to force extremal examples to be spread around evenly, then choosing a random example often gives a good answer. This idea is the heart of the probabilistic method.

Ramsey theory is one of many examples where the probabilistic method has been applied to combinatorics. This method has also been applied in many other areas of mathematics, including metric geometry. For example, Jean Bourgain (Professor in the School of Mathematics) showed that every finite metric space could be embedded in Euclidean space with relatively low distortion—his method was to carefully choose a random embedding and show that it has low distortion with high probability.

The probabilistic method has many applications in theoretical computer science as well. For example, a network made by randomly joining pairs of computers will be fairly robust, in the sense that everything might still be connected even if a few cables fail. Such networks are called expanders, and expanders are a very active area of research. Although random methods construct expanders easily, until recently the only explicit examples came from deep number-theoretic considerations. Peter Sarnak and Avi Wigderson (Professors in the School) have made fundamental contributions to the theory of expanders.

There has been recent interest in finding higher-dimensional analogues of expanders, and it has now been shown that certain random spaces, similar to those described above, have expander-like properties. It seems likely that these new higher-dimensional expanders will find applications in spectral clustering and topological data analysis, in sparsification of cell complexes, and probably in as yet unforeseen ways as well. ■

Matthew Kahle, a Member (2010–11) in the School of Mathematics, is interested in various interactions of probability and statistical mechanics with topology, geometry, and combinatorics. Beginning in the fall, he will be an Assistant Professor at the Ohio State University.

Finding Structure in Big Data

BY ANKUR MOITRA

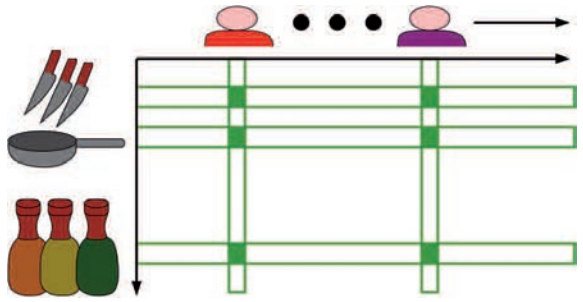
How do we navigate the vast amount of data at our disposal? How do we choose a movie to watch, out of the 75,000 movies available on Netflix? Or a new book to read, among the 800,000 listed on Amazon? Or which news articles to read, out of the thousands written everyday? Increasingly, these tasks are being delegated to computers—*recommendation systems* analyze a large amount of data on user behavior, and use what they learn to make personalized recommendations for each one of us.

In fact, you probably encounter recommendation systems on an everyday basis: from Netflix to Amazon to Google News, better recommendation systems translate to a better user experience. There are some basic questions we should ask: How good are these recommendations? In fact, a more basic question: What does "good" mean? And how do they do it? As we will see, there are a number of interesting mathematical questions at the heart of these issues—most importantly, there are many widely used algorithms (in practice) whose behavior we cannot explain. Why do these algorithms work so well? Obviously, we would like to put these algorithms on a rigorous theoretical foundation and understand the computational complexity of the problems they are trying to solve.

Here, I will focus on one running example and use this to explain the basic problems in detail, and some of the mathematical abstractions. Consider the case of Amazon. I have purchased some items on Amazon recently: a fancy set of cutting knives and a top-of-the-line skillet. What other products might I be interested in? The basic tenet of designing a recommendation system is that the more data you have available, the better your recommendations will be. For example, Amazon could search through its vast collection of user data for another customer (Alex) who has purchased the same two items. We both bought knives and a skillet, and Amazon can deduce that we have a common interest in cooking. The key is: perhaps Alex has bought another item, say a collection of cooking spices, and this is a good item to recommend to me, because I am also interested in cooking. So the message is: lots of data (about similar customers) helps!

Of course, Amazon's job is not so easy. I also bought a Kindle. And what if someone else (Jeff) also bought a Kindle? I buy math books online, but maybe Jeff is more of a Harry Potter aficionado. Just because we both bought the same item (a Kindle) does not mean that you should recommend Harry Potter books to me, and you certainly would not want to recommend math books to Jeff! The key is: What do the items I have purchased tell Amazon about my *interests*? Ideally, similar customers help us identify similar products, and vice-versa.

So how do they do it? Typically, the first step is to form a big table—rows represent items and columns represent users. And an entry indicates if a customer bought the corresponding item. What is the structure in this data? This is ultimately what we hope to use to make good recommendations. The basic idea is that a common interest is defined by a set of users (who share this interest) and a set of items. And we expect each customer to have bought many items in the set. We will call this a *combinatorial rectangle* (see image). The basic hypothesis is that the entire table of data we observe can be "explained" as a small number of these rectangles. So in this table containing information about millions of items and millions of users, we hope to "explain" the behavior of the users by a small number of rectangles—each representing a common interest.



ANKUR MOITRA

If two customers have a common interest in cooking, Amazon can use information about which items one of them has bought to make good recommendations to the other and vice-versa. Ankur Moitra is trying to develop rigorous theoretical foundations for widely used algorithms whose behavior we cannot explain.

The fundamental mathematical problem is: If the data can be "explained" by a small number of rectangles, can we find them? This problem is called *nonnegative matrix factorization*, and it plays a large role in the design of real recommendation systems.[1] In fact, there are many algorithms that work quite well in practice (on real data). But is there an efficient algorithm that *provably* works on every input? Recently, we showed that the answer is yes![2]

Our algorithm is based on a connection to a purely algebraic question: Starting with the foundational work of Alfred Tarski and Abraham Seidenberg, a long line of research has focused on the task of deciding if a system of polynomial inequalities has a solution. This problem can be solved efficiently provided the number of distinct variables is small.[3] And indeed, whether or not our table of data has a "good" nonnegative matrix factorization can be rephrased equivalently as whether or not a certain system of polynomial inequalities has a solution. So if our goal is to design fast algorithms, the operative question is: Can we reduce the number of variables? This is precisely the route we took, and it led us to a (much faster) provable algorithm for nonnegative matrix factorization whose running time is optimal under standard complexity assumptions.

Another fundamental mathematical question is: Can we give a theoretical explanation for why heuristics for these problems work so well in practice? There must be some property of the problems that we actually want to solve that makes them easier. In another work, we found a condition, which has been suggested within the machine learning community, that makes these problems much easier than in the worst case.[4] The crux of the assumption is that for every "interest," there must be some item that (if you buy it) is a strong indicator of your interest. For example, whoever buys a top-of-the-line skillet is probably interested in cooking. This assumption is known in the machine-learning literature as *separability*.[5] In many instances of real data, practitioners have observed that this condition is met by the parameters that their algorithm finds. And what we showed is that under this condition, there are simple, fast algorithms that *provably* compute a nonnegative matrix factorization.

In fact, this is just one instance of a broader agenda: I believe that exploring these types of questions will be an important step in building bridges between theory and practice. Our goal should not be to find a theoretical framework in which recommendations (and learning, more generally) are computationally hard problems, but rather one in which learning is easy—one that explains (for example) why *simple* recommendation systems are so good. These questions lie somewhere between statistics and computer science, because the question is *not*: How much data do you need to make good recommendations (e.g., the statistical efficiency of an estimator)? Algorithms that use the bare minimum amount of data are all too often very hard to compute. The emerging question is: What are the best tradeoffs between making the most of your data, and running in some reasonable amount of time? The mathematical challenges abound in bringing these perspectives into not just recommendation systems—but into machine learning in general. ■

Ankur Moitra is an NSF Computing and Innovation Fellow in the School of Mathematics at the Institute. His primary research interests are in algorithms, learning, and convex geometry. Prior to joining IAS, he received his Ph.D. in 2011 and his M.S. in 2009 from the Massachusetts Institute of Technology, both in theoretical computer science.

1 "Learning the Parts of an Object by Nonnegative Matrix Factorization," Daniel Lee and H. Sebastian Seung, *Nature* 401, October 21, 1999

2 "Computing a Nonnegative Matrix Factorization—Provably," Sanjeev Arora, Rong Ge, Ravi Kannan, and Ankur Moitra, *Symposium on Theory of Computing*, 2012

3 "On the Computational Complexity and Geometry of the First-Order Theory of the Reals," James Renegar, *Journal of Symbolic Computation* 13: 3, March 1992

4 "Learning Topic Models—Going Beyond SVD," Sanjeev Arora, Rong Ge, and Ankur Moitra, http://arxiv.org/abs/1204.1956, 2012

5 "When does Nonnegative Matrix Factorization give a Correct Decomposition into Parts?" David Donoho and Victoria Stodden, *Advances in Neural Information Processing Systems* 16, 2003

Randomness and Pseudorandomness

BY AVI WIGDERSON

The notion of randomness has intrigued people for millennia. Concepts like "chance," "luck," etc., play a major role in everyday life and in popular culture. In this article, I try to be precise about the meaning and utility of randomness. In the first part, I describe a variety of applications having access to *perfect* randomness, some of which are undoubtedly familiar to the reader. In the second part, I describe *pseudo*randomness, the study of random-looking phenomena in nonrandom (or weakly random) structures, and their potential uses.

Perfect randomness and its applications

The best way to think about perfect randomness is as an (arbitrarily long) sequence of coin tosses, where each coin is *fair*—has a 50-50 chance of coming up heads (H) or tails (T)—and each toss is *independent* of all others. Thus the two sequences of outcomes of twenty coin tosses,

HHHTHTTTHTTHHTTTTTHT and
HHHHHHHHHHHHHHHHHHHH,

have exactly the same probability: $1/2^{20}$.

Using a binary sequence of coin tosses as above, one can generate other random objects with a larger "alphabet," such as tosses of a six-sided die, a roulette throw, or the perfect shuffle of a fifty-two-card deck. One of the ancient uses of randomness, which is still very prevalent, is for gambling. And indeed, when we (or the casino) compute the probabilities of winning and losing in various bets, we implicitly assume (why?) that the tosses/throws/shuffles are perfectly random. Are they? Let us look now at other applications of perfect randomness,

(Continued on page 16)

RANDALL HAGADORN

Avi Wigderson, Herbert H. Maass Professor in the School of Mathematics

 (Continued from page 15)

and for each you should ask yourself (I will remind you) where the perfect randomness is coming from.

Statistics: Suppose that the entire population of the United States (over three hundred million) were voting on their preference of two options, say red and blue. If we wanted to know the *exact* number of people who prefer red, we would have to ask each and every one. But if we are content with an *approximation*, say up to a 3 percent error, then the following (far cheaper procedure) works. Pick *at random* a sample of two thousand people and ask only them. A mathematical theorem, called "the law of large numbers," guarantees that with probability 99 percent, the fraction of people in the sample set who prefer red will be within 3 percent of that fraction in the entire population. Remarkably, the sample size of two thousand, which guarantees the 99 percent *confidence* and 3 percent *error* parameters, does not depend on the population size at all! The same sampling would work equally well if all the people in the world (over six billion) were voting, or even if all atoms in the universe were voting. What is crucial to the theorem is that the two thousand sample is completely random in the entire population of voters. Consider: numerous population surveys and polls as well as medical and scientific tests use such sampling—what is their source of perfect randomness?

Physics and chemistry: Consider the following problem. You are given a region in a plane, like the one in Figure 1. A domino tiling of this region partitions the region into 2x1 rectangles—an example of such a tiling is given in Figure 2. The question is: how many different domino tilings does a given region have? Even more important is counting the number of partial tilings (allowing some holes). Despite their entertaining guise, such counting problems are at the heart of basic problems in physics and chemistry that probe the properties of matter. This problem is called the "monomer-dimer problem" and relates to the organization of diatomic molecules on the surface of a crystal. The number of domino tilings of a given region determines the thermodynamic properties of a crystal with this shape. But even for small regions this counting problem is nontrivial, and for large ones of interest, trying all possibilities will take practically forever, even with the fastest computers. But again, if you settle for an estimate (which is usually good enough for the scientists), one can obtain such an estimate with high confidence via the so-called "Monte-Carlo method" developed by Nicholas Metropolis, Stanislaw Ulam, and John von Neumann. This is a clever probabilistic algorithm that takes a "random walk" in the land of all possible tilings, but visits only a few of them. It crucially depends on perfect random choices. In the numerous applications of this method (and many other probabilistic algorithms), where is the randomness taken from?

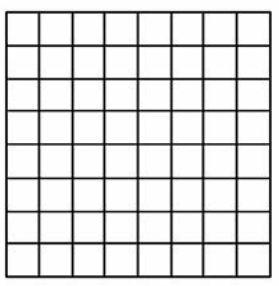


Figure 1

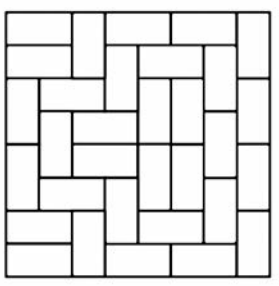


Figure 2

Congestion in networks: Imagine a large network with millions of nodes and links—it can be roads, phone lines, or, best for our purpose, the internet. When there is a large volume of traffic (cars/calls/email messages), congestion arises in nodes and links through which a lot of traffic passes. What is the best way to route traffic so as to minimize congestion? The main difficulty in this problem is that decisions as to where cars/calls/emails go are individual and uncoordinated. It is not hard to see that (in appropriate networks) if the many source-destination pairs were random, congestion would, almost surely, be quite small in *every* node. However, we don't tend to choose where we go or whom we call randomly—I call my friends and you call yours, and in such cases high congestion is bound to arise. To fix this problem, Leslie Valiant proposed the following ingenious idea, which is used in practice. Whenever A wants to send an email to B, she will actually choose a random intermediate point C, send the email to C, and ask C to forward it to B (forget privacy and compliance issues —they are beside the point here). While doubling the number of email messages, Valiant proved that (in appropriate networks) the congestion drops by huge factors with very high probability. Again, perfect randomness and independence of different decisions are essential for this solution to work.

Game theory: Sometimes the need for perfect randomness arises not for improved efficiency of some task (as in the previous examples), but for the very understanding of fundamental notions. One such notion is "rational behavior," a cornerstone of economics and decision theory. Imagine a set of agents (e.g., people, companies, countries, etc.) engaged in a strategic interaction (e.g., traffic, price competition, cold war) in which each agent influences the outcome for everyone. Each agent has a set of optional strategies to choose from, and the choices of everyone determine the (positive or negative) value for each. All agents have this information—what set of actions then would constitute rational behavior for them all? John Nash formulated his (Nobel Prize–winning) notion of "Nash equilibrium" sixty years ago, which is widely accepted to this day. A set of strategies (one for each agent) is said to be a Nash equilibrium if no player can improve its value by switching to another strategy, given the strategies of all other agents (otherwise, it would be rational for that player to switch!). While this is a natural stability notion, the first question to ask is: which games (strategic situations as above) possess such a rational equilibrium solution? Nash proved that *every* game does, regardless of how many agents there are, how many strategies each has, and what value each agent obtained given everyone's choices . . . provided that agents can toss coins! Namely, allowing *mixed* strategies, in which agents can (judiciously) choose at random one of their optional strategies, makes this notion universal, applicable in every game! But again, where do agents in all these situations take their coin tosses?

Cryptography: This field, which underlies all of computer security and e-commerce today, serves perhaps as the best demonstration of how essential randomness is in our lives. First and foremost, in cryptographic situations, there are *secrets* that some know and others don't. But what does that mean? "Secret" is another fundamental notion whose very definition requires randomness. Such a definition was given by Claude Shannon, the father of information theory, who quantified the amount of uncertainty (just how much we *don't* know about it) using another fundamental notion, *entropy*, which necessitates that the objects at hand be random.

For example, if I pick a password completely randomly from all decimal numbers of length ten, then your chances of guessing it are precisely $1/10^{10}$. But if I choose it randomly from the set of phone numbers of my friends (also ten-digit numbers), then your uncertainty is far smaller: your probability of guessing my secret is larger, namely 1/the number of my friends (and yes, cryptography assumes that my adversaries know everything about me, except the outcomes of my coin tosses). But secrets are just the beginning: *all* cryptographic protocols like public-key encryption, digital signatures, electronic cash, zero-knowledge proofs, and much more, rely completely on randomness and have no secure analogue in a deterministic world. You use such protocols on a daily basis when you log in, send email, shop online, etc. How does your computer toss the coins required by these protocols?

Pseudorandomness

A computational view of randomness: To answer the repeatedly asked question above, we have to carefully study our ubiquitous random object—the coin toss. Is it random? A key insight of theoretical computer science is that the answer depends on who (or which application) uses it! To demonstrate this, we will conduct a few (mental) experiments. Imagine that I hold in my hand a (fair) coin, and a second after I toss it high in the air, you, as you are watching me, are supposed to guess the outcome when it lands on the floor. What is the probability that you will guess correctly? 50-50 you say? I agree! Now consider a variant of the same experiment, in which the only difference is that you can use a laptop to help you. What is the probability that you will guess correctly now? I am certain you will say 50-50 again, and I will agree again. How can the laptop help? But what if your laptop is connected to a super computer, which is in turn connected to a battery of video recorders and other sensors around the room? What are your chances of guessing correctly now? Indeed, 100 percent. It would be trivial for this machinery to calculate in one second all the required information: speed, direction, and angular momentum of the coin, the distance from my hand to the floor, air humidity, etc., and provide the outcome to you with certainty.

The coin toss remained the same in all three experiments, but the observer changed. The uncertainty about the outcome depended on the observer. Randomness is in the eye of the beholder, or more precisely, in its computational capabilities. The same holds if we toss many coins: how uncertain the outcome is to a given observer/application depends on how they process it. Thus a phenomenon (be it natural or artificial) is deemed "random enough," or *pseudorandom*, if the class of observers/applications we care about cannot distinguish it from random! This viewpoint, developed by Manuel Blum, Shafi Goldwasser, Silvio Micali, and Andrew Yao in the early 1980s, marks a significant departure from older views and has led to major breakthroughs in computer science of which the field of cryptography is only one. Another is a very good understanding of the power of randomness in probabilistic algorithms, like the "Monte-Carlo method." Is randomness actually needed by them, or are there equally efficient deterministic procedures for solving the monomer-dimer problem and its many siblings? Surprisingly, we now have strong evidence for the latter, indicating the weakness of randomness in such algorithmic settings. A theorem by Russell Impagliazzo and Wigderson shows that, assuming *any* natural computational problem to be intractable (something held in wide belief and related to the P≠NP conjecture), randomness has no power to enhance algorithmic efficiency! Every probabilistic algorithm can be replaced by a deterministic one with similar efficiency. Key to the proof is the construction of pseudorandom generators that produce sequences indistinguishable from random ones by these algorithms.

Random-like behavior of deterministic processes and structures: What can a clever observer do to distinguish random and nonrandom objects? A most natural answer would be to look for "patterns" or properties that are extremely likely in random objects, and see if the given object has them. The theorem mentioned above allows the observer to test *any* such property, as long as the test is efficient. But for many practical purposes, it suffices that the object has only *some* of these properties to be useful or interesting. Examples in both mathematics and computer science abound. Here is one: A property of a random network is that to sever it (break it into two or more large pieces), one necessarily has to sever many of its links. This property is extremely desirable in communication networks and makes them fault-tolerant. Can one construct objects with such a random-like property deterministically and efficiently?

This question has been addressed by mathematicians and computer scientists alike, with different successful

(Continued on page 17)

 (Continued from page 16)

constructions, e.g., by Gregory Margulis, Alexander Lubotzky, Ralph Philips, and Peter Sarnak on the math side and by Omer Reingold, Salil Vadhan, and Wigderson on the computer science side. An even more basic fault-tolerant object is an *error-correcting code*—a method by which a sender can encode information such that, even if subjected to some noise, a receiver can successfully remove the errors and determine the original message. Shannon defined these important objects and proved that a random code is error-correcting. But clearly for applications we need to construct one efficiently! Again, today many different deterministic constructions are known, and without them numerous applications we trust every day, from satellites to cell phones to CD and DVD players, would simply not exist!

Proving that deterministic systems and structures possess random-like properties is typically approached differently by mathematicians and computer scientists. In mathematics the processes and structures are organic to the field, arising from number theory, algebra, geometry, etc., and proving that they have random-like properties is part of understanding them. In computer science, one typically starts with the properties (which are useful in applications) and tries to efficiently construct deterministic structures that have them. These analytic and synthetic approaches often meet and enhance each other (as I will exemplify in the next section). A National Science Foundation grant to further explore and unify such connections in the study of pseudorandomness was recently awarded to Jean Bourgain, Sarnak, Impagliazzo, and Wigderson in the Institute's School of Mathematics.

Randomness purification: Returning to the question of providing perfect randomness to *all* (as opposed to specific) applications, we now put no limits on the observers' computational power. As true randomness cannot be generated deterministically, one cannot help but assume some, possibly imperfect, source of random coin tosses. Can one deterministically and efficiently convert an imperfect random source to a perfect one? How should we model imperfect randomness?

Experience with nature gives some clues. Without getting into (the interesting) philosophical discussion of whether the universe evolves deterministically or probabilistically, many phenomena we routinely observe seem at least partly *unpredictable*. These include the weather, stock market fluctuations, sun spots, radioactive decay, etc. Thus we can postulate, about any such phenomena, that their sequence of outcomes possesses some entropy (but where this entropy resides we have no clue). Abstractly, you can imagine an adversary who is tossing a sequence of coins, but can choose the bias of each in an arbitrary way—the probability of heads may be set to 1/2, 1/3, .99 or even $1/\pi$, so long as it is not 0 or 1 (this would have zero entropy). Moreover, these probabilities may be correlated arbitrarily—the adversary can look at past tosses and accordingly determine the bias of the next coin. Can we efficiently use such a defective source of randomness to generate a perfect one? The (nontrivial) answer is no, as shown twenty years ago by Miklos Santha and Umesh Vazirani, who defined these sources, extending a simple model of von Neumann. But while dashing hope in one direction, they also gave hope in another, showing that if you have *two* (or more) such sources, which are independent of each other, then *in principle* one can utilize them together to deterministically generate perfect randomness. So if, for example, the weather, stock market, and sun spots do not affect each other, we can hope to combine their behavior into a perfect stream of coin tosses. What was missing was an efficient construction of such a randomness purifier (or *extractor* in computer science jargon).

The solution of this old problem was recently obtained using a combination of analytic and synthetic approaches by mathematicians and computer scientists. Some time ago David Zuckerman suggested the following idea: suppose A, B, and C represent the outcome of our samples of (respectively) the weather, the stock market, and sun spots (think of them as integers[1]). He conjectured that the outcome of the arithmetic $A \times B + C$ would have more entropy (will be more random) than any of the inputs. If so, iterating this process (with more independent weak sources) will eventually generate a (near) perfect random number! Zuckerman proved that this concept follows from a known mathematical conjecture. While this mathematical conjecture is still open, recent progress was made on a completely different conjecture by Bourgain, Nets Katz, and Terence Tao (extending the work of Paul Erdős and Endre Szemerédi). They studied properties of random *tables*, and tried to find such properties in specific, arithmetic tables, namely the familiar addition and multiplication tables. Here is an intuitive description of the property they studied. Consider a small "window" in a table (see Figure 3).

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	21	32	111	74	5	16	5	66	198	101	43	91	1
2	97	66	208	148	62	132	185	27	37	127	74	115	193
3	45	209	179	204	124	10	202	89	212	39	75	26	6
4	129	1	134	45	8	156	224	14	162	130	96	143	35
5	113	53	69	81	41	109	68	130	21	51	140	73	180
6	182	216	142	22	105	206	22	22	175	88	66	9	127
7	173	33	26	120	30	221	33	69	25	207	188	36	31
8	111	163	179	28	112	79	210	195	216	24	197	39	138
9	90	161	171	88	79	27	222	170	130	94	58	55	61
10	117	119	133	206	64	19	155	27	94	186	99	118	151
11	161	112	1	28	124	109	217	16	152	108	7	191	222
12	161	43	45	187	208	152	155	130	216	34	193	184	55
13	197	53	1	18	195	120	39	109	143	82	87	210	11
14	192	53	124	57	171	113	177	128	155	64	8	178	18
15	10	163	7	95	26	6	140	117	86	148	24	203	25

Figure 3: A random table and a typical window

Call such a window *good* if only a "few" of the numbers in it occur more than once. It is not hard to prove that in a random table, *all* small windows will be good. Now what about the addition and multiplication tables? It is very easy to see that each has bad windows![2] However, Bourgain, Katz, and Tao showed that when *taken together* these two tables are good in the following sense (see Figure 4): for every window, it is either good in the multiplication table or in the addition table (or both)! Boaz Barak, Impagliazzo, and Wigderson gave a statistical version of this result, and used it to prove that Zuckerman's original extractor works!

+	1	2	3	4	5	6	7	8	9	10	11	12	13
1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	4	5	6	7	8	9	10	11	12	13	14	15	16
4	5	6	7	8	9	10	11	12	13	14	15	16	17
5	6	7	8	9	10	11	12	13	14	15	16	17	18
6	7	8	9	10	11	12	13	14	15	16	17	18	19
7	8	9	10	11	12	13	14	15	16	17	18	19	20
8	9	10	11	12	13	14	15	16	17	18	19	20	21
9	10	11	12	13	14	15	16	17	18	19	20	21	22
10	11	12	13	14	15	16	17	18	19	20	21	22	23
11	12	13	14	15	16	17	18	19	20	21	22	23	24
12	13	14	15	16	17	18	19	20	21	22	23	24	25
13	14	15	16	17	18	19	20	21	22	23	24	25	26
14	15	16	17	18	19	20	21	22	23	24	25	26	27
15	16	17	18	19	20	21	22	23	24	25	26	27	28

×	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	2	3	4	5	6	7	8	9	10	11	12	13
2	2	4	6	8	10	12	14	16	18	20	22	24	26
3	3	6	9	12	15	18	21	24	27	30	33	36	39
4	4	8	12	16	20	24	28	32	36	40	44	48	52
5	5	10	15	20	25	30	35	40	45	50	55	60	65
6	6	12	18	24	30	36	42	48	54	60	66	72	78
7	7	14	21	28	35	42	49	56	63	70	77	84	91
8	8	16	24	32	40	48	56	64	72	80	88	96	104
9	9	18	27	36	45	54	63	72	81	90	99	108	117
10	10	20	30	40	50	60	70	80	90	100	110	120	130
11	11	22	33	44	55	66	77	88	99	110	121	132	143
12	12	24	36	48	60	72	84	96	108	120	132	144	156
13	13	26	39	52	65	78	91	104	117	130	143	156	169
14	14	28	42	56	70	84	98	112	126	140	154	168	182
15	15	30	45	60	75	90	105	120	135	150	165	180	195

Figure 4: The addition and multiplication tables

The above story is but one example. Fundamental results from number theory and algebraic geometry, mainly on the "random-like" behavior of rational solutions to polynomial equations (by André Weil, Pierre Deligne, Enrico Bombieri, and Bourgain) were recently used in a variety of extractor constructions, purifying randomness in different settings.

1 Actually they should be taken as numbers modulo some large prime p, and all arithmetic below should be done modulo p.

2 If rows and columns of a window form an arithmetic progression, the addition table will be bad. If they form a geometric progression, the multiplication table will be bad.

Avi Wigderson, Herbert H. Maass Professor in the School of Mathematics, is a widely recognized authority in the diverse and evolving field of theoretical computer science. His main research area is computational complexity theory, which studies the power and limits of efficient computation and is motivated by fundamental scientific problems. Since being appointed to the Faculty in 1999, Wigderson has overseen the Institute's activities in theoretical computer science, which began in the 1990s, initially organized by visiting professors with the involvement of Enrico Bombieri, IBM von Neumann Professor in the School.

The European Association for Theoretical Computer Science and the Association for Computing Machinery Special Interest Group on Algorithms and Computation Theory awarded the 2009 Gödel Prize for outstanding papers in theoretical computer science to Wigderson and former Visitors Omer Reingold (1999–2003) and Salil Vadhan (2000–01). The three were selected for their development of a new type of graph product that improves the design of robust computer networks and resolves open questions on error correction and derandomization.

Million-dollar questions on pseudorandomness: Two of the most celebrated open problems in mathematics and computer science, the Riemann Hypothesis and the P vs. NP question, can be stated as problems about pseudorandomness. These are two of the seven Clay Millennium problems, each carrying a $1 million prize for a solution (see www.claymath.org/millennium-problems for excellent descriptions of the problems as well as the terms for the challenge). They can be cast as problems about pseudorandomness despite the fact that randomness is not at all a part of their typical descriptions. In both cases, a concrete property of random structures is sought in specific deterministic ones.

For the P vs. NP question the connection is relatively simple to explain. The question probes the computational difficulty of natural problems. It is simple to see that *random* problems[3] are (almost surely) hard to solve, and P vs. NP asks to prove the same for certain *explicit* problems, such as "the traveling salesman problem" (i.e., given a large map with distances between every pair of cities, find the shortest route going through every city exactly once).

Let's elaborate now on the connection of the Riemann Hypothesis to pseudorandomness. Consider long sequences of the letters L, R, S, such as

SSRSLLLLLSLRRLSRRRRRSLSLSLL...

Such a sequence can be thought of as a set of instructions (L for Left, R for Right, S for Stay) for a person or robot walking in a straight line. Each time the next instruction moves it one unit of length Left or Right or makes it Stay. If such a sequence is chosen at *random* (this is sometimes called a random walk or a drunkard's walk), then the moving object would stay relatively close to the origin with high probability: if the sequence was of n steps, almost surely its distance from the starting point would be close to $\sqrt{n}$. For the Riemann Hypothesis, the explicit sequence of instructions called the Möbius function is determined as follows for each step t. If t is divisible by any prime more than once then the instruction is Stay (e.g., t=18, which is divisible by 3^2). Otherwise, if t is divisible by an *even* number of distinct primes, then the instruction is Right, and if by an odd number of distinct primes, the instruction is Left (e.g., for t=21=3×7 it is Right, and for t=30=2×3×5 it is Left). This explicit sequence of instructions, which is determined by the prime numbers, causes a robot to look drunk, if and only if the Riemann Hypothesis is true! ■

3 This has to be formally defined.

Studying the Shape of Data Using Topology

BY MICHAEL LESNICK

The story of the "data explosion" is by now a familiar one: throughout science, engineering, commerce, and government, we are collecting and storing data at an ever-increasing rate. We can hardly read the news or turn on a computer without encountering reminders of the ubiquity of big data sets in the many corners of our modern world and the important implications of this for our lives and society.

Our data often encodes extremely valuable information, but is typically large, noisy, and complex, so that extracting useful information from the data can be a real challenge. I am one of several researchers who worked at the Institute this year in a relatively new and still developing branch of statistics called topological data analysis (TDA), which seeks to address aspects of this challenge.

In the last fifteen years, there has been a surge of interest and activity in TDA, yielding not only practical new tools for studying data, but also some pleasant mathematical surprises. There have been applications of TDA to several areas of science and engineering, including oncology, astronomy, neuroscience, image processing, and biophysics.

The basic goal of TDA is to apply topology, one of the major branches of mathematics, to develop tools for studying *geometric features of data*. In what follows, I'll make clear what we mean by "geometric features of data," explain what topology is, and discuss how we use topology to study geometric features of data. To finish, I'll describe one application of TDA to oncology, where insight into the geometric features of data offered by TDA led researchers to the discovery of a new subtype of breast cancer.

In this article, by "data" I simply mean a finite set of points in space. In general, the space in which our points lie can have many dimensions, but for now the reader may think of the points as sitting in two or three dimensions. For a concrete example, each point in a data set in three-dimensional space might correspond to a tumor in a cancer study, and the x, y, and z coordinates of the point might each correspond to the level of expression of a different gene in a tissue sample of the tumor.

What, then, do I mean by "geometric features of data?" Rather than offer a formal definition, I'll give three representative examples of the sorts of geometric features of data we study in TDA. I'll take the data in each of the examples to lie in two-dimensional space.

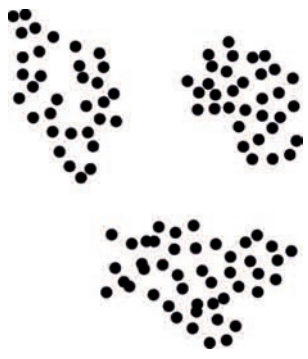


Figure 1: A data set with three clusters

As a first example, consider the data set in Figure 1. We see that the data breaks up into three distinct clusters. Clusters like these are a first type of geometric feature of data we study in TDA. We'd like to count the number of distinct clusters in the data and partition the data into its clusters. We'd like to be able to do this even when the cluster structure of the data is corrupted by noise, as in Figure 2.

The problem of detecting clusters in data is in fact an old and well-studied problem in statistics and computer science, but TDA has recently introduced some new ideas and tools to the problem.[1]

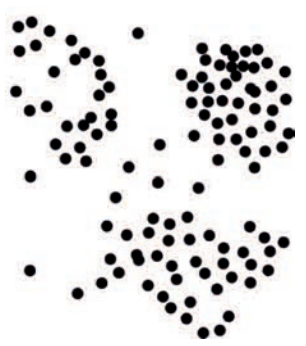


Figure 2: A data set with three noisy clusters

A second kind of geometric feature of data we study in topological data analysis is a "loop." Figure 3 gives an example of a loop in a data set. Again, we'd like to be able to detect a loop in a data set even when it is corrupted by noise, as in Figure 4.

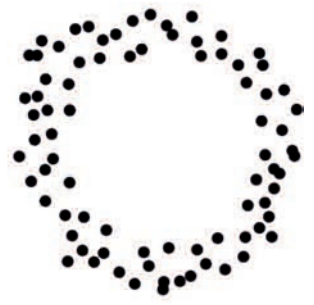


Figure 3: A data set with a loop

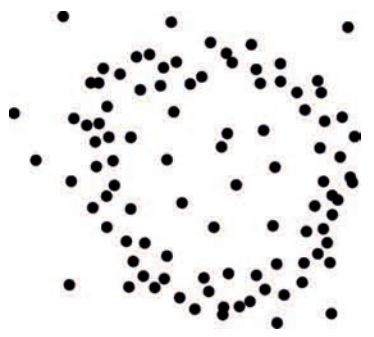


Figure 4: A data set with a noisy loop

A third kind of geometric feature we study in TDA is a "tendril." Figure 5 depicts a data set with three tendrils emanating from a central core. In a data set with this sort of structure, we'd like to detect the presence of the tendrils, count the tendrils, and partition the data into its different tendrils.

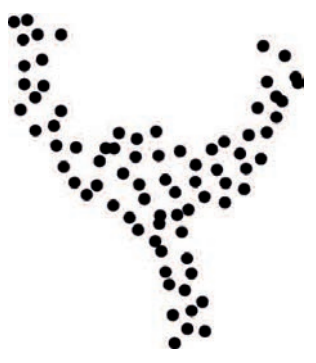


Figure 5: A data set with three tendrils emanating from a central core

The objective of research in TDA is to develop tools to detect and visualize these kinds of geometric features, and to develop methodology for quantifying the statistical significance of such features in randomly sampled data. Because much of the data arising in scientific applications lives in high-dimensional spaces, the focus is on developing tools suitable for studying geometric features in high-dimensional data.

Why, though, should we be interested in studying such features of data in the first place? The key premise behind this line of research is that *insight into the shape of scientifically relevant data has a good chance of giving insight into the science itself*.

Experience has shown that this premise is a reasonable one. Cluster analysis is used as a matter of course throughout the experimental sciences to extract scientific information from data;[2] the study of loops and their higher-dimensional analogues has recently offered insight into questions in biophysics[3] and natural-scene statistics;[4] and, as I will describe in the last section of this article, the study of tendrils has recently offered insight into oncology.[5]

As noted above, TDA studies the geometric features of data using topology. Topology is the study of the properties of a geometric object that are preserved when we bend, twist, stretch, and otherwise deform the object without tearing it. The primary example of such a property is the presence of *holes* in the object; as such, topology is concerned largely with the formal study of holes. (Homotopy theory, discussed in the article about the Institute's univalent foundations program, is a central part of topology. However, homotopy theory also admits an axiomatic formulation that abstracts away from the topological setting and provides a framework for the adaption of topological ideas to settings outside of topology.)

To anyone who's ever eaten a slice of swiss cheese or a doughnut, the notion of a hole in a geometric object is a familiar and intuitive one; the idea that the number of holes in a geometric object doesn't change when we bend, twist, and stretch the object is similarly intuitive.

In topology, we distinguish between several different kinds of holes. A hole at the center of a donut is an example of a first kind of hole; the hollow space inside an inflated, tied ballon is an example of a second kind of hole. In geometric objects in more than three dimensions, we may also encounter other kinds of holes that cannot appear in objects in our three-dimensional world.

As intuitive as the notion of a hole is, there is quite a lot to say about holes, mathematically speaking. In the last century, topologists have put great effort into the study of holes, and have developed a rich theory with fundamental connections to most other areas of modern mathematics. One feature of this theory is a well-developed set of formal tools for computing the number of holes of different kinds in a geometric object. TDA aims to put this set of tools to use in the study of data. Computations of the number of holes in a geometric object can be done automatically on a computer, even when the object lives in a high-dimensional space and cannot be visualized directly.

Besides the number of holes in an object, another (very simple) property of a geometric object that is preserved under bending, twisting, and stretching is the number of components (i.e., separate pieces) making up the object. For example, a plus sign + is made up of one component, an equals sign = is made up of two components, and a division sign ÷ is made up of three components. Deforming any of these symbols without tearing does not change the number of components in the symbol. We regard the problem of computing the number of components that make up a geometric object as part of topology. In fact, in a formal sense, this problem turns out to be closely related to the problem of computing the number of holes in a geometric object, and topologists think of these two problems as two sides of the same coin.

How do we use topology to study the geometric features of data? Without pretending to give a full answer to this question, I'll mention some of the basic ideas. To begin, I'll describe a primitive strategy for studying data using topology that, while unsatisfactory for most applications,

(Continued on page 19)

is the starting point for what is done in practice.

As mentioned above, topology offers tools for computing numbers of holes and components in a geometric object; we would like to apply these tools to our study of data. However, a data set X of n points in space has n components and no holes at all, so directly computing the numbers of holes and components of X will not tell us anything interesting about the geometric features of X.

To study X using topology then, we will consider not the topological properties of X directly, but rather the topological properties of a "thickening" of X.

I'll explain this in detail. Assume that X is a finite set of points in the plane (two-dimensional space). Let δ be a positive number, and let $T(X, \delta)$ be the set of all points in the plane within distance δ from some point in X; we think of $T(X, \delta)$ as a "thickening" of the data set X.

For example, let X_1 be the data set of Figure 1. Figure 6 shows $T(X_1, \delta_1)$ in red for some choice of positive number δ_1, together with the original data X_1 in black. For a second example, let X_2 be the data set of Figure 3. Figure 7 shows $T(X_2, \delta_2)$ in red, for some choice of positive number δ_2, together with X_2 in black. For especially nice data sets X and good choices of δ, the clusters in X will correspond to components of $T(X, \delta)$ and the loops in X will correspond to holes in $T(X, \delta)$. For instance, in Figure 6 the clusters in X_1 correspond to the components of $T(X_1, \delta_1)$, and in Figure 7 the loop in X_2 corresponds to the hole in $T(X_2, \delta_2)$.

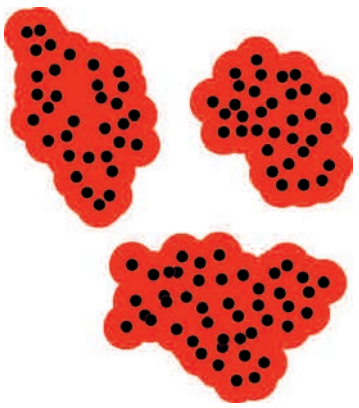


Figure 6: $T(X_1,\delta_1)$, for some choice of δ_1, is shown in red; X_1 is shown in black.

Figure 7: $T(X_2,\delta_2)$, for some choice of δ_2, is shown in red; X_2 is shown in black.

Thus, for nice data sets X, we can get insight into the geometric features of X by studying the topological properties of $T(X, \delta)$. The same strategy also works for studying the geometric features of a data set sitting in a high-dimensional space, in which case the data cannot be visualized directly.

Most data sets we encounter in practice are not as nice as those of Figures 1 and 3, and though the primitive TDA strategy we have described does extend to data in high-dimensional spaces, for typical data sets X in any dimension, the strategy has several critical shortcomings. For one, the topological properties of $T(X, \delta)$ can depend in a very sensitive way on the choice of δ, and a priori it is not clear what the correct choice of δ should be, or if a correct choice of δ exists at all, in any sense. Also, the topological properties of $T(X, \delta)$ are not at all robust to noise in X, so that this strategy will not work for studying the geometric features of noisy data sets, such as those in Figures 2 and 4. Moreover, this approach to TDA is not good at distinguishing small geometric features in the data from large ones.

Thus, for dealing with most data one encounters in practice, more sophisticated variants of this basic strategy are required. Much of the recent research in TDA has been focused on developing such variants. One central idea in this direction is that it is much better to consider at once the topological properties of the entire family of objects $T(X, \delta)$ as δ varies than it is to consider the topological properties of $T(X, \delta)$ for a single choice of δ. This is the idea behind *persistent homology*, a key technical tool in TDA.

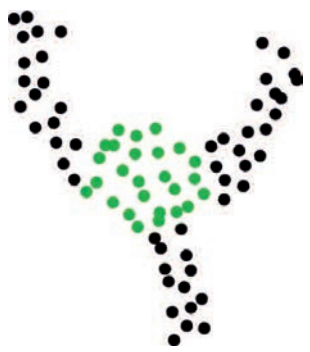


Figure 8: The central core of the data set of Figure 5

The problem of studying tendrils in data is closely related to the problem of studying clusters. To see this, consider Figure 8, where the points in the central core of the data in Figure 5 are shown in green. If we were to have a principled way of identifying the central core of the data, then by removing that central core, we would obtain a data set with three distinct clusters, as in Figure 9, where each cluster corresponds to a tendril in the original data set. It is natural to expect, then, that some of the topological tools that are useful for studying clusters can be extended to the study[5] of tendrils, and in fact this is the case.

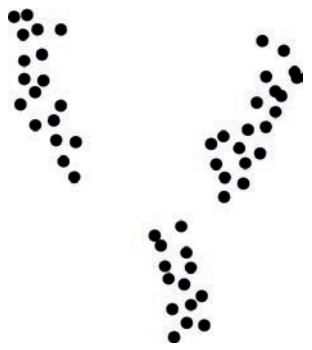


Figure 9: When we remove the central core of the data set of Figure 5, we get a data set with three clusters.

In work published in 2011 by Monica Nicolau, Gunnar Carlsson, and Arnold Levine (Professor Emeritus in the School of Natural Sciences),[5] insight offered by TDA into the geometric features of data led the authors to the discovery of a new subtype of breast cancer.

The authors studied a data set describing the gene expression profiles of 295 breast cancer tumors, each from a unique patient. The data set consists of 295 points sitting in a 24,479-dimensional space: each point corresponds to one tumor and, roughly speaking, each of the 24,479 coordinates of the point specifies the level of expression of one gene in a tissue sample of the corresponding tumor.

To begin their analysis of the data, the researchers mapped the data from the 24,479-dimensional space into a 262-dimensional space in a way that preserved aspects of the geometric structure of the data relevant to cancer, while eliminating aspects of that structure not relevant to cancer.

The researchers then studied the geometric features of the data in 262-dimensional space using a TDA tool called Mapper.[6] They discovered a three-tendril structure in the data loosely analogous to that in the data of Figure 5. In addition, they found that one of these tendrils decomposes further, in a sense, into three clusters. One of these three clusters, they observed, corresponds to a distinct subtype of breast cancer tumor that had hitherto not been identified. This subtype, which the authors named *c-MYB*$^+$, comprises 7.5 percent of the data set (22 tumors). Tumors belonging to the *c-MYB*$^+$ subtype are genetically quite different than normal tissue, yet patients whose tumors belonged to this subtype had excellent outcomes: their cancers never metastasized, and their survival rate was 100 percent.

The work of Michael Lesnick, Member (2012–13) in the School of Mathematics, focuses on the theoretical foundations of topological data analysis.

A standard approach to the classification of breast cancers, based on clustering, divides breast cancers into five groups. The *c-MYB*$^+$ subtype does not fit neatly into this classification scheme: the *c-MYB*$^+$ tumors divide among three of the five groups. The results of Nicolau, Carlsson, and Levine thus suggest a nuance to the taxonomy of breast cancer not accounted for in the standard classification model.

These results illustrate how the tools of TDA can be useful in helping researchers tease out some of the scientific information encoded in their high-dimensional data. They are just one of a growing number of examples where TDA has facilitated the discovery of interesting scientific information from data. Still, in spite of good progress in the field over the last several years, there's still much to be done in terms of fleshing out the mathematical and statistical foundations of TDA, and in terms of algorithm and software development. The shared hope among researchers in the field is that by advancing the theory and tools of TDA, we can lay the groundwork for the discovery of new applications of TDA to the sciences.

For further details about TDA, see any of the several surveys available on TDA,[7-9] or the book.[10] ■

1 Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba, "Persistence-Based Clustering in Riemannian Manifolds," in *Proceedings of the Twenty-Seventh Annual ACM Symposium on Computational Geometry* (Association for Computing Machinery, 2011), 97–106.

2 Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys (CSUR)* 31, no. 3 (1999): 264–323.

3 Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, and Vidit Nanda, "Topological Measurement of Protein Compressibility via Persistence Diagrams," MI Preprint Series 2012-6, Faculty of Mathematics, Kyushu University, 2012.

4 Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian, "On the Local Behavior of Spaces of Natural Images," *International Journal of Computer Vision* 76, no. 1 (2008): 1–12.

5 Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson, "Topology-Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival," *Proceedings of the National Academy of Sciences* 108, no. 17 (2011): 7265–70.

6 Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson, "Topological Methods for the Analysis of High-Dimensional Data Sets and 3-D Object Recognition," in *Eurographics Association Symposium on Point-Based Graphics* 22 (The Eurographics Association, 2007).

7 Gunnar Carlsson, "Topology and Data," *Bulletin of the American Mathematical Society* 46, no. 2 (2009): 255–308.

8 Herbert Edelsbrunner and John L. Harer, "Persistent Homology: A Survey," in *Surveys on Discrete and Computational Geometry: Twenty Years Later: AMS-IMS-SIAM Joint Summer Research Conference, June 18–22, 2006, Snowbird, Utah* 453 (American Mathematical Society, 2008), 257.

9 Robert Ghrist, "Barcodes: The Persistent Topology of Data," *Bulletin of the American Mathematical Society* 45, no. 1 (2008): 61.

10 Herbert Edelsbrunner and John L. Harer, *Computational Topology: An Introduction* (American Mathematical Society, 2010).

Institute for Advanced Study

Institute for Advanced Study
Einstein Drive
Princeton, New Jersey 08540
(609) 734-8000
www.ias.edu
publications@ias.edu

www.facebook.com/InstituteforAdvancedStudy
www.twitter.com/the_IAS
www.youtube.com/videosfromIAS

A digital version of this publication is available at www.ias.edu/letter.

Articles from the *Institute Letter* are available online at www.ias.edu/about/publications/ias-letter/articles.

To receive monthly updates on Institute events, videos, and other news by email, subscribe to *IAS eNews* at www.ias.edu/news/enews-subscription.