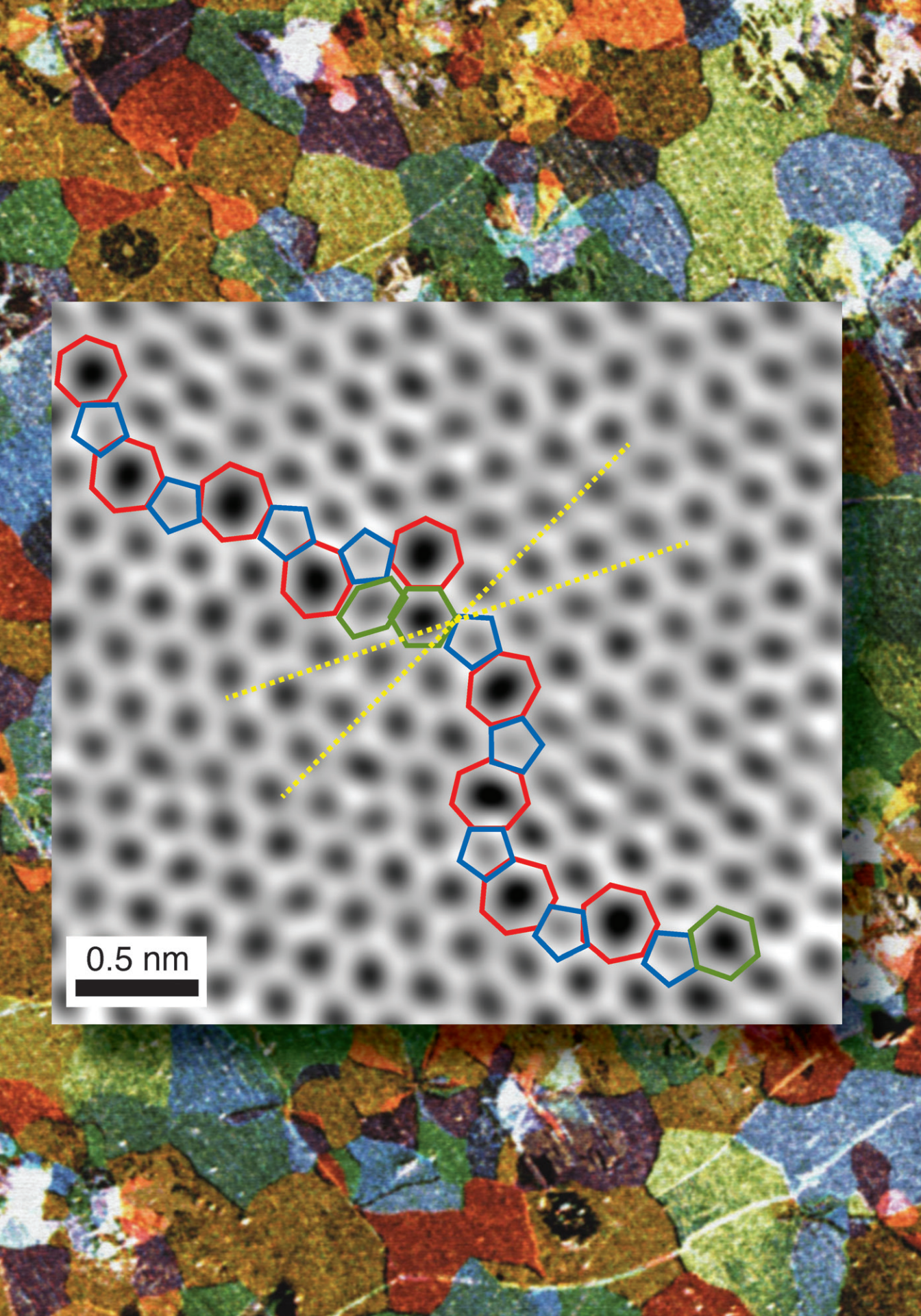
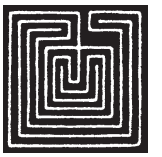
Dædalus

Journal of the American Academy of Arts & Sciences

Summer 2012

Science in the 21st Century


0.5 nm



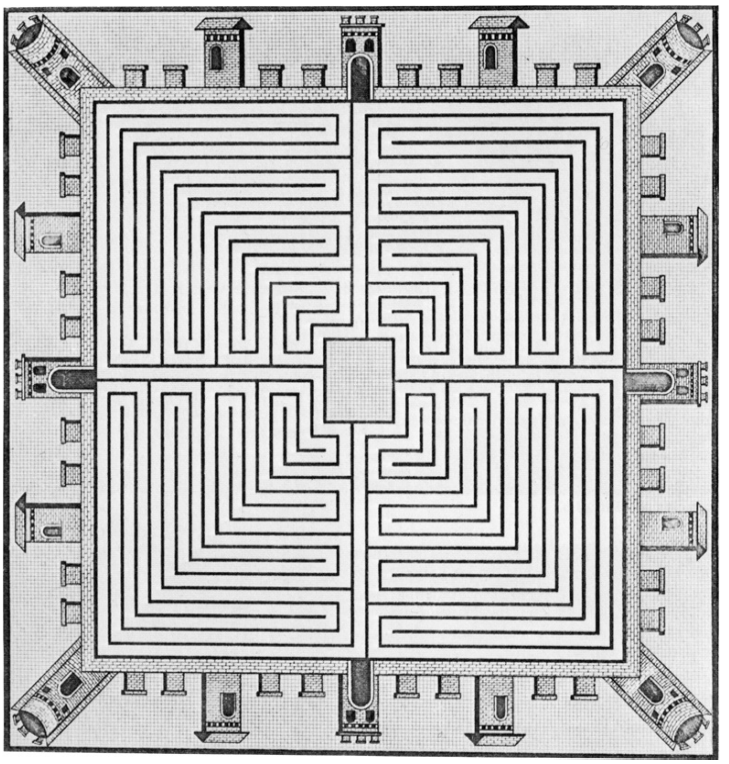
Inside front cover: The "patchwork quilt" shown in the background is actually an electron microscope image of a sheet of graphene, a film of conductive carbon that is no thicker than a single atom. Each patch of color is an atomically precise honeycomb of carbon atoms; where two patches with different rotations (different colors) touch, the perfect, six-sided honeycombs are stitched together in an imperfect line of five- and seven-member rings, as shown in the inset.

Researchers at Cornell University have found that these stitching defects make the film weaker but not less conductive. Understanding how to find and correct such imperfections is important for realizing graphene's potential applications in large-area electronics, such as touch screens or solar cells. This image first appeared in Pinshane Y. Huang et al., "Grains and Grain Boundaries in Single-Layer Graphene Atomic Patchwork Quilts," *Nature* 469 (January 2011), and is provided courtesy of the Cornell Center for Materials Research.

Dædalus is designed by Alvin Eisenman.

Dædalus

Journal of the American Academy of Arts & Sciences



Nineteenth-century depiction of a Roman mosaic labyrinth, now lost, found in Villa di Diomede, Pompeii

Dædalus was founded in 1955 and established as a quarterly in 1958. The journal's namesake was renowned in ancient Greece as an inventor, scientist, and unriddler of riddles. Its emblem, a maze seen from above, symbolizes the aspiration of its founders to "lift each of us above his cell in the labyrinth of learning in order that he may see the entire structure as if from above, where each separate part loses its comfortable separateness."

The American Academy of Arts & Sciences, like its journal, brings together distinguished individuals from every field of human endeavor. It was chartered in 1780 as a forum "to cultivate every art and science which may tend to advance the interest, honour, dignity, and happiness of a free, independent, and virtuous people." Now in its third century, the Academy, with its nearly five thousand elected members, continues to provide intellectual leadership to meet the critical challenges facing our world.

Dædalus Summer 2012
Issued as Volume 141, Number 3

Editorial offices: *Dædalus*, Norton's Woods,
136 Irving Street, Cambridge MA 02138.
Phone: 617 491 2600. Fax: 617 576 5088.
Email: daedalus@amacad.org.

Library of Congress Catalog No. 12-30299

ISBN 978-0-262-75148-3

Dædalus publishes by invitation only and assumes no responsibility for unsolicited manuscripts. The views expressed are those of the author of each article, and not necessarily of the American Academy of Arts & Sciences.

Dædalus (ISSN 0011-5266; E-ISSN 1548-6192) is published quarterly (winter, spring, summer, fall) by The MIT Press, Cambridge MA 02142, for the American Academy of Arts & Sciences. An electronic full-text version of *Dædalus* is available from The MIT Press. Subscription and address changes should be addressed to MIT Press Journals Customer Service, 55 Hayward Street, Cambridge MA 02142. Phone: 617 253 2889; U.S./Canada 800 207 8354. Fax: 617 577 1545. Email: journals-cs@mit.edu.

Printed in the United States of America by Cadmus Professional Communications, Science Press Division, 300 West Chestnut Street, Ephrata PA 17522.

Newsstand distribution by Ingram Periodicals Inc., 18 Ingram Blvd., La Vergne TN 37086, and Source Interlink Distribution, 27500 Riverview Center Blvd., Bonita Springs FL 34134.

Postmaster: Send address changes to *Dædalus*, 55 Hayward Street, Cambridge MA 02142. Periodicals postage paid at Boston MA and at additional mailing offices.

Subscription rates: Electronic only for nonmember individuals – $43; institutions – $119. Canadians add 5% GST. Print and electronic for nonmember individuals – $48; institutions – $132. Canadians add 5% GST. Outside the United States and Canada add $23 for postage and handling. Prices subject to change without notice.

Institutional subscriptions are on a volume-year basis. All other subscriptions begin with the next available issue.

Single issues: $13 for individuals; $33 for institutions. Outside the United States and Canada add $6 per issue for postage and handling. Prices subject to change without notice.

Claims for missing issues will be honored free of charge if made within three months of the publication date of the issue. Claims may be submitted to journals-cs@mit.edu. Members of the American Academy please direct all questions and claims to daedalus@amacad.org.

Advertising and mailing-list inquiries may be addressed to Marketing Department, MIT Press Journals, 55 Hayward Street, Cambridge MA 02142. Phone: 617 253 2866. Fax: 617 253 1709. Email: journals-info@mit.edu.

The typeface is Cycles, designed by Sumner Stone at the Stone Type Foundry of Guinda CA. Each size of Cycles has been separately designed in the tradition of metal types.

Prelude

Jerrold Meinwald

JERROLD MEINWALD, a Fellow of the American Academy since 1970, is the Goldwin Smith Professor of Chemistry Emeritus at Cornell University. His research has contributed to a wide range of chemical and chemical biological subjects, including organic photochemistry, reaction mechanisms, the synthesis of chiral inhalation anesthetics, natural product chemistry, and chemical ecology. His publications include the edited volumes *Chemical Ecology: The Chemistry of Biotic Interaction* (with Thomas Eisner, 1995) and *Science and the Educated American: A Core Component of Liberal Education* (with John G. Hildebrand, 2010). He is Secretary of the American Academy and Cochair of the Academy's Committee on Studies.

Scientific knowledge is cumulative and always open to revision that may be necessitated by the acquisition of new data or by theoretical advances. The ability of science to organize observations and make reliable predictions is constantly evolving, and its benefits to humankind abound. But given our present state of scientific sophistication, is there anything significant left to be learned? Before examining the status of and outlook for science in the twenty-first century, I would like to share a bit of personal history that provides some twentieth-century context and an important lesson.

From 1948 to 1952, as a student in R. B. Woodward's research group in the Department of Chemistry at Harvard University, I felt myself to be in organic chemical heaven. Woodward attracted some of the world's brightest and most ambitious graduate and postdoctoral students, and it was a fantastic privilege to have him as a mentor. Although still early in his career at the time, he had already transformed the art of molecular structure determination through his masterful examination of what information could be extracted from the ultraviolet and infrared absorption spectra of organic molecules. In addition, his brilliant planning and execution of the synthesis of complex natural products (quinine, cortisone, cholesterol, strychnine, and so on) was legendary. What more was there for organic chemists to do?

Sixty years ago, when I left Harvard to join the chemistry faculty at Cornell University, I considered

organic chemistry to be a fully mature subject. It had achieved its most essential theoretical insights with respect to molecular structure, stereochemistry, and reaction mechanisms. It had refined its most useful experimental techniques. Certainly, this beautifully developed body of knowledge would allow its practitioners to continue to solve many challenging problems, both those internal to the subject of organic chemistry and, increasingly, those related to biology and material science. But it seemed unlikely to undergo much in the way of further fundamental development. I could not have been more wrong. The state of our knowledge in the early 1950s might justifiably be regarded as almost primitive. The field made revolutionary, although to a large extent unforeseeable, advances in the second half of the twentieth century – so much so that most of the work done by organic chemists in the year 2000 and beyond has depended heavily on the application of experimental techniques and on theories that simply did not exist a half-century earlier. The lesson (well known, but easily forgotten) is that anticipating future events is difficult.

As an example of unexpected advances, the now-commonplace chemical analysis of complex mixtures containing hundreds of components (ranging from Chanel No. 5, to Château Margaux 2005, to urine samples examined for evidence of doping) relies on gas chromatographic or high-performance liquid chromatographic separations, two techniques developed only in the second half of the twentieth century. Nuclear magnetic resonance spectroscopy, which emerged during the same period, now makes possible complete molecular structure determinations of unknown compounds using (but not destroying) a sample of only a few micrograms (rather than the milligram quantities previously required to gain comparable information in a vastly slower and more complex fashion). This is a thousandfold gain in sensitivity alone. Recently developed mass spectrometric techniques are yet another millionfold more sensitive! Protein and nucleic acid structures are now determined almost routinely. New analytical procedures make it possible to solve problems that either were entirely beyond reach, or, at best, would have required years of effort, in a matter of days or even hours!

Our ability to synthesize both natural and non-natural materials has also improved dramatically. New methods for joining carbon atoms have greatly enriched the synthetic chemist's repertoire. The delicate art of constructing asymmetric molecules with the desired specific three-dimensional right- or left-handed shapes has benefited from the invention of powerful new synthetic strategies. Many of these advances will form the basis of a much-needed "green" chemical industry.

In this brave new molecular world, we have come to understand the basis for the differences between heat- and light-promoted chemical reactions. We can unravel the pheromonal courtship messages of insects and the chemical signaling ("quorum sensing") of bacteria. We can even visualize the activities of single molecules trapped in carbon nanotubes or confined within living cells. Overall, studies in the entire field of chemistry that could not have been realistically contemplated a half-century ago can now be undertaken with every hope of success. So is there really still more to do, or are contemporary chemists simply cleaning up a few remaining details? What is the outlook for the other natural sciences and mathematics? What have we learned from recent research, and what can we say about the future of scientific research more generally? These are the chief questions addressed in this issue.

We all are aware of new and potentially useful *applications* of science. We are constantly bombarded with advertising for novel electronic devices with amazing capabilities, for new drugs, and for other technology-based products with claims to improve the quality of our lives. Some of the claims may prove to be true. But most of these applications are not based on science that is particularly new. Moreover, while promoting "better living through chemistry" is not an unworthy aim, it is not the goal of chemistry itself. The purpose of studying chemistry is to understand, both qualitatively and quantitatively, the rules governing the properties and transformations of matter at the atomic, molecular, and supramolecular levels.

We are still very far from knowing all that we can know about our universe. There remains a large and important body of new knowledge waiting to be discovered in the ongoing pursuits of chemistry, physics, astronomy, geology, genetics, molecular biology, and evolutionary theory, among others. Interaction among these disciplines will be especially fruitful. In addition, the unique contribution of mathematics to this great intellectual endeavor is particularly powerful and interesting. We will continue to deepen our understanding of the universe, from particle physics to cosmology, from the simplest to the most complex systems. If we can manage to avoid total human disaster resulting from societal and environmental challenges (matters that in fact demand our most serious and immediate attention), we can be confident that many fundamental questions will be asked and answered in the decades to come. Both curiosity and necessity will ensure this outcome, although economic factors will play a significant role in determining when, where, and in what fields the most important advances will be made.

Jerrold Meinwald

The contributors to this volume have been selected not only because of their notable contributions to their own fields of research, but also because of their disciplinary judgment and vision. Much of what they have to say is intriguing and, in many instances, surprising. I am deeply grateful to each of them for bravely accepting the invitation to write an essay devoted to exploring the present and possible future of those areas of science with which they are most familiar. I would like particularly to thank my coeditor, May Berenbaum, for sharing the responsibility of editing this issue with me. I would also like to thank my Cornell colleagues Saul Teukolski, Steven Strogatz, and Melissa Hines for their invaluable advice. Although it was not possible to survey all the currently important areas of science in this collection of essays, I hope that our authors have made clear that scientific inquiry remains one of the most important and rewarding fields of intellectual endeavor.

Those fortunate enough to participate in twenty-first-century scientific discovery, invention, and analysis have an exciting and challenging time to look forward to. (It will be important that they take on the task of communicating the substance and significance of their results to the general public as well.) If the scientific progress made in the twentieth century provides any precedent, humankind can expect to occupy a much better-understood universe well before the twenty-first century is over!

The Search for Habitable Worlds: Planetary Exploration in the 21st Century

James F. Bell III

Abstract: The search for and detailed characterization of habitable environments on other worlds – places where liquid water, heat/energy sources, and biologically important organic molecules exist or could have once existed – is a major twenty-first-century goal for space exploration by NASA and other space agencies, motivated by intense public interest and highly ranked science objectives identified in recent National Academy decadal surveys. Through telescopic observations, terrestrial laboratory and field studies, and a "flyby, orbit, land, rove, and return" strategy for robotic exploration, particular emphasis will be placed on specific worlds already identified as potentially habitable: Mars, Jupiter's ocean moon Europa, and Saturn's icy and organic-bearing moons Titan and Enceladus. However, the potential abounds for surprising discoveries at many of our solar system's other planetary, satellite, and asteroidal destinations, as well as within newly discovered planetary systems around other stars.

JAMES F. BELL III is a Professor in the School of Earth and Space Exploration at Arizona State University in Tempe and President of The Planetary Society. He has been involved in many of NASA's recent robotic solar system exploration missions, including as lead scientist for the Pancam color stereo cameras on the Mars rovers Spirit and Opportunity and as a member of the science camera team on the Curiosity rover. He has published the space photography books *Postcards from Mars*, *Mars 3-D*, and *Moon 3-D*. He received the 2011 Carl Sagan Medal from the American Astronomical Society.

Modern astronomy and planetary science stand on the verge of making some of the most profound discoveries ever achieved in the exploration of the cosmos. Historian and author Stephen Pyne has called this era the "third great Age of Exploration." In the first age, Renaissance explorers such as Magellan, Columbus, and Vespucci discovered a so-called New World here on our own planet, mapping new continents and circumnavigating the globe in the sailing ships of the fifteenth and sixteenth centuries. In the second great Age of Exploration, seventeenth- to nineteenth-century Enlightenment explorers such as Cook, Lewis and Clark, and Powell pushed our exploration deep into those frontiers, discovering and documenting the details of new lands and peoples. At the same time, the scientific giants of the Enlightenment – Galileo, Brahe, Kepler, Copernicus, and Newton – helped us learn how we could eventually explore realms beyond our earthly home. The Scientific Revolution of the nineteenth to twentieth century propelled rocketry pioneers such as God-

dard and Tsiolkovsky to make modern space travel a reality.

Since its inception more than fifty years ago, the U.S. National Aeronautics and Space Administration (NASA) has been the primary engine of this third great Age of Exploration. Significant exploration firsts and scientific discoveries have also been made by other national and international governmental space organizations, including the Russian Federal Space Agency (known as Roscosmos, formerly the Soviet space program), the European Space Agency (ESA), and the space programs of Japan, China, Canada, and India, among other nations.

Over the past four hundred years, space exploration has primarily been conducted with telescopes. Telescopic observations have provided critical details on the orbital and physical properties (size, mass, composition) of planets, moons, asteroids, and comets; these details were invaluable for the more thorough follow-up observations performed by modern robotic space probes. Apart from initial telescopic observations, during the past half-century the primary philosophy of the world's combined space mission efforts is perhaps best embodied in NASA's recent exploration strategy slogan: "flyby, orbit, land, rove, and return."[1]

As Table 1 shows, exploration of the worlds around us has proceeded in a generally systematic fashion, beginning with relatively simple robotic missions that realized significant increases in resolution and measurement fidelity through close-up "flyby" encounters with solar system bodies. The next level of advancement has come from more complex (and expensive) space missions designed to orbit individual planetary bodies, providing even higher-resolution mapping data as well as systematic time histories of surface and/or atmospheric phenomena. Logical (and yet more complex) next steps have included surface landers or rovers and/or atmospheric probes or balloons for a smaller subset of planetary exploration targets; and sample return missions have ranged from robotic collection of milligrams to tens of grams of extraterrestrial materials to the hundreds of kilograms of lunar samples returned by the Apollo astronauts in the only human landing missions to another world yet conducted.

Table 1 reveals something remarkable: the initial flyby robotic space mission – that is, the reconnaissance of representative examples of all the major classes of objects in our solar system – is nearly complete. Once the New Horizons probe has flown by the Pluto-Charon system in mid-2015, planetary scientists will have obtained basic in situ data from all our solar system's planets and large moons, as well as important representatives of large and small asteroids, a variety of comet nuclei, and a few large, icy planetary bodies known as Kuiper Belt Objects (KBOs). Completing this initial reconnaissance of our solar system is a truly extraordinary achievement for our species.

The time seems appropriate, then, for the planetary science community to consider how best to build on this accomplishment, and how best to prioritize filling in the remainder of the exploration matrix presented in Table 1. Indeed, doing so has become a major goal of community-wide "decadal surveys" in both planetary sciences[2] and astrophysics.[3] From these surveys and other scientific, technical, and public interest considerations, it appears that at least in the near term, planetary science and the exploration of other worlds (in our solar system and beyond) will focus on the theme of *habitability*: that is, searching for past or present environments that may have been (or perhaps still are) conducive to the emergence and survival of life as we know it.

Table 1
The Solar System Exploration Matrix, as of late 2011

Method	*Mercury*	*Venus*	*Moon*	*Mars*	*Asteroids*	*Jupiter*	*Saturn*	*Outer Satellites*	*Uranus*	*Neptune*	*Comets*	*KBOs*	*Exo-planets*
Telescopic:	x	x	x	x	x	x	x	x	x	x	x	x	x
Spacecraft:													
Flyby	x	x	x	x	x	x	x	x	x	x	x	(9)	
Orbiter	x	x	x	x	x	x	x	(4)			(6)		
Lander or Probe		x	x	x	(2)	x		(5)			(6,7)		
Rover or Balloon		x	x	x	(3)								
Sample Return			x	(1)	x						(8)		
Human Landing			x										

(1) Initial planning for a possible joint NASA ESA Mars robotic sample return mission in the 2020s is under way.
(2) The NEAR-Shoemaker spacecraft orbited 433 Eros during 2000 and landed on its surface, surviving for a few days, in 2001.
(3) The Japanese Hayabusa mission to near-Earth asteroid 25143 Itokawa attempted unsuccessfully to deploy a small rover on the asteroid.
(4) This includes detailed orbital investigation of the Jovian satellites by the NASA *Galileo* mission from 1996 to 2003, and observations of the Saturnian satellites by the NASA/ESA *Cassini* orbiter from 2004 to the present.
(5) The ESA/NASA *Huygens* Titan entry probe descended to the surface of Titan in 2005 and survived for a few hours as a surface lander.
(6) The ESA Rosetta mission was launched in 2004 and will orbit and send a lander to comet 67P/Churyumov-Gerasimenko in 2014.
(7) The Deep Impact flyby mission intentionally crash landed a 370 kg copper/aluminum projectile on comet 9P/Tempel 1 in 2005.
(8) The Stardust mission returned a small cache of dusty grains shed off the nucleus of comet Wild-2 in 2006.
(9) The New Horizons mission was launched in 2005 and will fly past Pluto in 2015 and possibly one or more KBOs in the 2020s.

Source: Table created by author.

Planetary science is a relatively young, broadly interdisciplinary field that represents the merging of practices and practitioners from astronomy, physics, geology, geophysics, chemistry, geochemistry, atmospheric science, meteoritics, biology, mathematics, computer science, engineering, and other disciplines. Fundamentally, planetary science is founded on and overlaps significantly with Earth systems science, though examples and models based on the properties and behavior of our home planet must often be extrapolated significantly to cover a wider range of surface, atmospheric, and/or interior conditions observed or inferred for other planetary bodies. It is perhaps not surprising, then, that fundamental discoveries and new research directions in the Earth sciences often propagate into or enable new discoveries in planetary science as well.

A prime and highly relevant example is the recent discovery in the terrestrial biology and paleontology communities of the phenomenal diversity of life on our own planet. Specifically, both simple, single-celled organisms (for example, prokaryotes) as well as more complex multicellular life forms (eukaryotes) exist – and many thrive – in environments and ecosystems spanning an enormous range of temperature, pH, salinity, pressure, ionizing radiation, and other circumstances. For example, active prokaryotic ecosystems can be found in temperature ranges from -10° to 120°C, acidity levels of pH < 1.0 to 11, and salinity levels from fresh water to evaporite salt flats.[4] Eukaryotic systems are less tolerant of environmental extremes but can still survive and potentially thrive in temperatures from 0° to 50°C, acidity levels of pH 3 to 8, and in moderately salty lake environments.[5] Energy sources for life on Earth are not limited to sunlight, either. For example, some newfound organisms thrive on geothermal energy from deep-sea vents, while others eke out a meager existence deep in the crust from the small amount of energy released during oxidation of volcanic

rock. Some of these environmental variables are certainly intertwined; for example, increased salinity can significantly lower the freezing point of a solution, enabling some life forms to survive in temperatures well below the freezing point of fresh water. Nonetheless, the environmental range and diversity of life on our planet, recognized only in the last ten to fifteen years, is impressive.

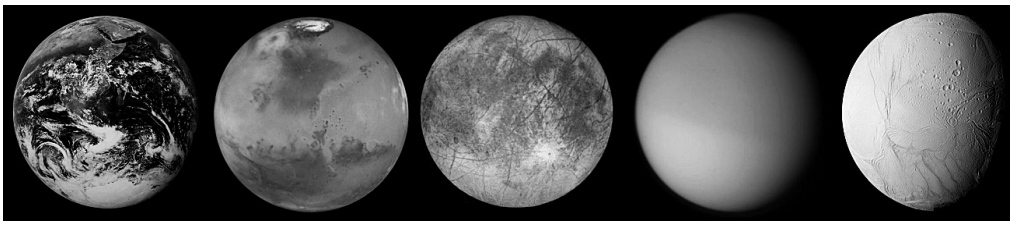
The discovery of the enormous array of habitable (and inhabited) environments on Earth has had a profound influence on research in evolutionary biology, paleontology, and the general study of the origin of life. It has also brought newfound respect and scientific credence to the nascent field of astrobiology – the search for evidence of life on other worlds. If life could have formed in or adapted itself to extreme environments on our own planet, then perhaps it could also have done so in other niches within our solar system (or in other systems) where conditions are or were similar.

Thus, the newfound discovery of a much wider range of habitable environments on our own planet than had previously been thought has expanded the range of potentially habitable environments thought to exist (or to have existed) on other worlds. While biologists and even philosophers have struggled to define specifically what is meant by "life as we know it," it has been easier to define a set of requirements that likely are essential (but not necessarily sufficient) for life as we know it to exist on another planetary body. These requirements include the presence of liquid water to act as a medium for organic chemistry; the presence of potential energy sources like sunlight, geothermal heat, redox reactions, or impact cratering events to fuel the biogeochemical cycles upon which life depends; and the presence of organic molecules made from at least carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur that can form simple to complex structures and participate in reactions capable of converting external energy sources into usable "internal" energy within cells.

The search for habitable worlds beyond our own has therefore become a search for liquid water, organic molecules, and potential sources of excess energy that are biologically useful. Identifying places or environments where these key requirements are met now infuses almost all the major goals of NASA and many other space agencies' science and exploration programs, trickling down into mission success requirements for many new robotic missions and even into success criteria for many individual researchers and laboratories working on government-funded research grants and contracts. As a unifying theme and strategy, searching for habitable worlds is also fairly popular with elected officials and the public, which is beneficial and likely not a coincidence, as these constituents are ultimately the source of most of the funding that enables the search to be conducted.

Planetary scientists have been searching in earnest for habitable environments elsewhere in the solar system for only a decade or two, yet we can already begin compiling a list of the "greatest hits" destinations, where the search has yielded important and surprising early results and where future missions and additional resources have the greatest chance of making the most profound discoveries. However, of the eight major planets, several dozen large moons, and several hundred large inner and outer solar system asteroids known to date, only four specific objects have yet made the list of the solar system's most exciting potential astrobiology hot spots: Mars, Europa, Titan, and Enceladus (see Figure 1).

Figure 1
Earth and the "Greatest Hits" Solar System Destinations



Earth (left) shown with (from left to right) Mars, Europa, Titan, and Enceladus (not to scale). The search for habitable environments at these destinations could yield dramatic and profound results. Source: NASA/Jet Propulsion Laboratory/Hubble Space Telescope/Galileo Project/Cassini Project.

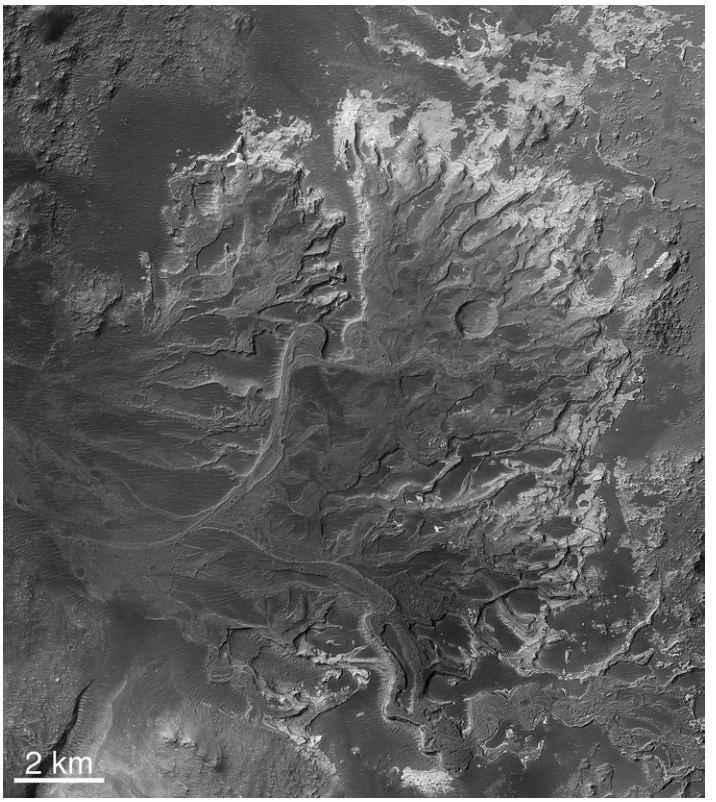
Astronomers have long suspected Mars of being an abode for life, but for most of history that view was held for the wrong reasons. The putatively artificial canals and other features attributed to intelligent aliens by late nineteenth- and early twentieth-century astronomers such as Percival Lowell turned out to be optical illusions. Better telescopic observations and early space flyby missions showed Mars to be covered with craters, not canals, providing evidence for an ancient surface more like the Moon's than the Earth's.

But subsequent orbital missions from the 1970s to the present have revealed the surface and atmosphere of the Red Planet in increasing detail. Mars is now known to have had a complex geological history comparable to our own planet's, involving volcanism, tectonism, and erosion as well as impact cratering.[6] Orbital data also reveal that Mars once had a global-scale magnetic field, indicating that the planet's metallic core may once have been partially molten (like Earth's) and providing evidence for the internal heat that may have driven early active geologic processes.

Several key pieces of evidence also suggest that the climate of Mars may have been much different in the past. Today, Mars has a thin atmosphere (with only about 1 percent of Earth's sea-level surface pressure) and a cold and extremely dry climate. Conditions at the surface today, and likely for at least the past few billion years, do not allow liquid water to be stable. Yet geologic features seen from orbit show evidence of large-scale networks of channels and valleys that appear to have been formed by flowing water, some in catastrophic flooding events, others in what must have been long-term, sustained periods of the higher atmospheric pressures and temperatures required to allow liquid water to remain stable on the surface. The discovery of what look like ancient, eroded river valleys was one of the earliest pieces of evidence for Mars once having had a warmer, wetter, more Earth-like climate.

More recently, higher-resolution orbiter data and measurements from a variety of Mars landers and rovers have revealed even stronger follow-on evidence for persistent liquid water flowing across the surface, in the form of small-scale gullies as well as deltas and other layered sedimentary deposits on macroscopic to microscopic scales (see Figure 2). This geologic evidence for a more watery past has been buttressed by geochemical evidence for the presence of water-formed rocks and minerals on Mars, such as clays and hydrated sulfate salts. Indeed, some of

Figure 2
Orbiter Photograph of Landforms within Mars's Eberswalde Impact Crater



This 2002 photograph from NASA's Mars Global Surveyor Mars Orbiter Camera shows a fan-shaped set of landforms that have been interpreted as the eroded remains of a shallow water delta within the Eberswalde impact crater, near 33°W, 24°S. Based on analogy with the formation of meandering channels and sediments formed in deltas on Earth, features like this suggest that liquid water was persistent on early Mars. Source: NASA/Jet Propulsion Laboratory/Malin Space Science Systems.

the ancient water that once flowed across the surface of Mars and eroded and weathered the planet's extensive deposits of volcanic lava flows and ash is still there today, locked up as OH- or H_2O within the mineral structures of the Red Planet's dust, salts, and clays.

The presence of liquid water, volcanic and geothermal heat sources, and a presumably steady supply of organic molecules from frequent asteroid and comet impacts all point to an ancient Martian environment that we would label "habitable" from a terrestrial life perspective. Indeed, in the absence of effective constraints on the planet's current geothermal heat gradient or its modern inventories of liquid water and organic molecules, it is possible that the Martian subsurface – from tens of meters to tens of kilometers depth – could still represent a habitable environment for life, just like Earth's crust.

The discovery of habitable past (and possibly even present) environments on Mars has made exploration of the Red Planet the highest priority for future NASA

and other space agency large-scale robotic missions.[7] For example, NASA's latest "flagship" class mission, the Mars Science Laboratory rover (dubbed "Curiosity") was launched in November 2011 and is scheduled for an August 2012 landing within Gale Crater, an ancient sedimentary basin containing a deep stratigraphy of layered, clay- and sulfate-bearing rocks. Unlike previous Mars rovers, this new one carries a sensitive mass spectrometer capable of sensing small quantities of organic molecules that may have been preserved in those ancient sediments. The spectrometer also has the novel ability to measure small changes in key isotopic ratios that could potentially indicate the presence of past biologic, as opposed to geologic, processes.

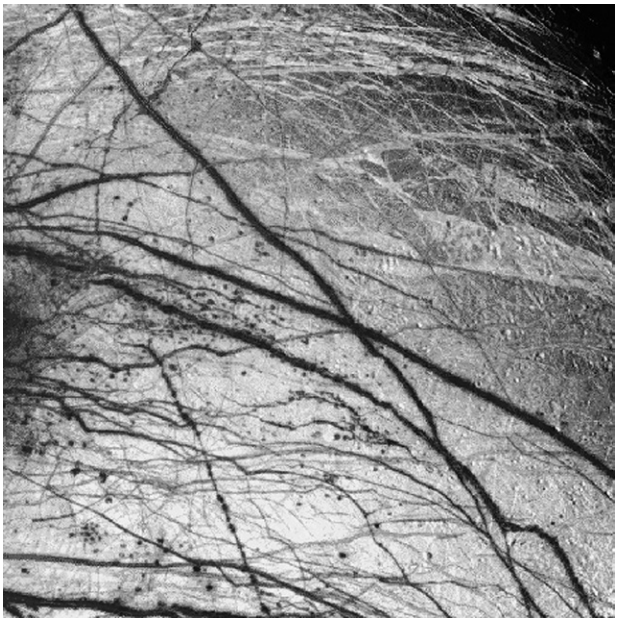
Jupiter's four large moons – Io, Europa, Ganymede, and Callisto – were discovered by Galileo in 1610, and their observed motions around Jupiter provided some of the strongest evidence against geocentric theories of the layout of our solar system. Telescopic exploration of these new worlds was limited, however, partly because of their vast distances (five times farther than the distance between the Earth and the Sun) and their relatively small sizes. The Voyager 1 and 2 flybys through the Jupiter system in the late 1970s thus provided a dramatic revolution in our understanding of these small planetary bodies, revealing, for example, active volcanic eruptions on Io and a thin, cracked, plate-like icy crust covering Europa.[8]

Europa quickly became the target of intense scrutiny because of its extremely smooth crust: the lack of any significant topography was consistent with the icy crust being very thin, perhaps "floating" on a liquid or sloshy ice-liquid "mantle" in the subsurface – a layer that many quickly began referring to as a possible ocean. Furthermore, the almost complete lack of impact craters on Europa suggested that the crust was extremely young, perhaps even being actively resurfaced by cryovolcanic processes (in which liquid water erupts through a water ice crust). Tidal forces from the periodic close encounters of Io, Europa, and Ganymede as they orbited Jupiter were thought to provide the energy responsible for Io's volcanoes and possibly for heating the interior of Europa above the melting point of ice.

The images and other data from the Voyager flybys provided intriguing but inconclusive evidence for an ocean under the thin, icy crust of Europa. Thus, NASA sent the Galileo robotic orbiter mission to Jupiter in the 1990s to study the giant planet and its rings and moons in even greater detail. Europa was a special focus of the Galileo mission, with high-resolution imaging, spectroscopy, magnetic fields, and gravity mapping campaigns conducted to try to confirm the existence of Europa's subsurface ocean.[9] Imaging data revealed evidence for the sea ice–like cracking and rotation of Europa's crust, suggesting that it is only a thin (a few to tens of kilometers in depth) icy shell. Spectroscopic observations detected salty evaporite minerals in the cracks between the icy plates (see Figure 3), as if a salty fluid had erupted or oozed from the subsurface and left mineral deposits behind when it evaporated upon exposure to the vacuum of space. Magnetic field observations exposed a conducting near-surface layer beneath the ice, with a volume and conductivity consistent with that of a deep (approximately 100 km) briny liquid water solution.

Galileo mission observations and interpretations provide compelling evidence – though still not proof – that there is an ocean-like "mantle" of liquid water beneath Europa's thin, cracked, icy crust. Some models even predict quantities of water in Europa's ocean in excess of twice

Figure 3
Orbiter Photograph of the Surface of Europa



An enhanced photograph taken in 1996 by the NASA Galileo Jupiter orbiter shows the cracked, icy surface of Europa. Europa's thin water-ice shell is divided into countless plates that warp and move relative to each other. Many ridges and cracks at plate boundaries reveal evidence of salty mineral deposits (indicated here by dark lines), which are likely the result of deeper ocean water oozing up from the subsurface and evaporating. The scene depicted here is about 1,300 km from side to side. Source: NASA/Jet Propulsion Laboratory.

the water in all of Earth's oceans. Further, density and gravity data suggest that the bottom of Europa's subsurface ocean is in direct contact with a rockier mantle, similar to the way Earth's oceans are in contact with crustal and mantle silicate rocks. Such contact would facilitate both heat transfer from Europa's interior and geochemical energy transfer through the aqueous weathering of rocky materials.

Powerful internal tidal and radiogenic heat sources, the likely presence of abundant liquid water, and the likely presence of abundant organic molecules delivered over time by asteroids and comets all point to Europa as being another of our solar system's most potentially habitable worlds. Indeed, a dedicated large-scale mission to orbit and/or land on Europa to test further the ocean hypothesis and assess the habitability of this distant "water world" has been ranked by the planetary science community as the second-highest priority (behind a Mars sample return mission).[10]

Further away still is Titan, yet another exciting, unusual, and potentially habitable world. Orbiting Saturn at almost ten times the Earth's distance from the Sun, Titan is larger than the planet Mercury but much colder, with a surface temperature of only about 90 degrees above absolute zero. Titan is special, though, because it is the only moon in our solar system with a thick atmosphere. Instruments on the Voyager flyby spacecraft revealed that atmosphere to be mostly

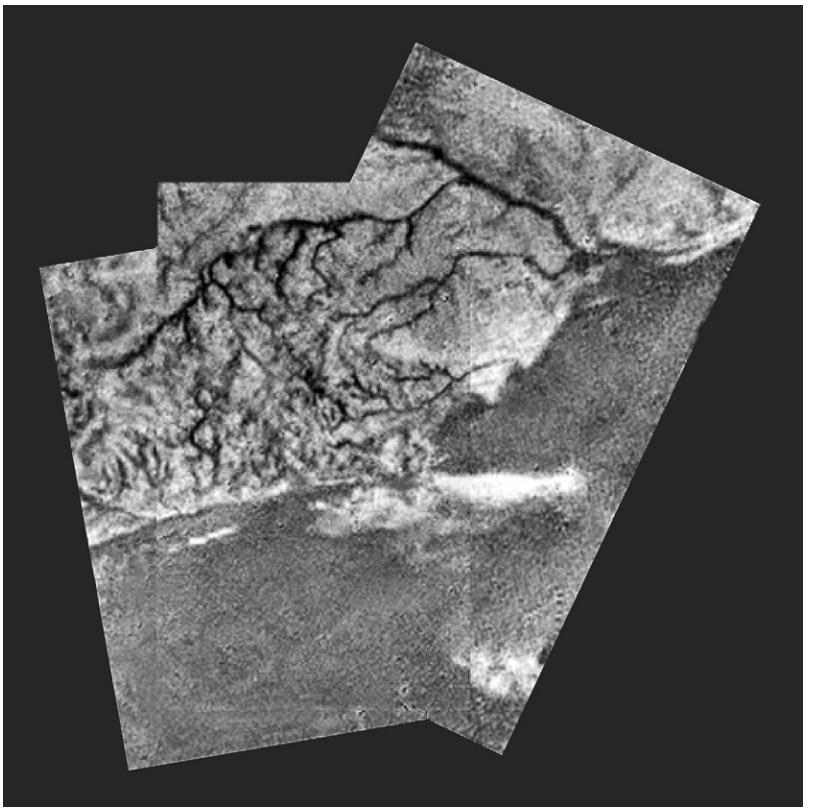
(slightly less than 80 percent) nitrogen, with a surface pressure 50 percent higher than Earth's. A hazy "smog" of hydrocarbons, created by the interaction of solar UV radiation and a small amount of atmospheric methane, hides Titan's surface from view at visible wavelengths. Similar to the Jupiter system, a dedicated orbiter was the next logical exploration step for the Saturn system. Thus, NASA and ESA combined efforts (and funding) to conduct the Cassini Saturn orbiter mission, which entered orbit around the ringed planet in 2004. Titan has been a special focus of Cassini observations, and the orbiter deployed Huygens, a descent probe and lander that successfully landed on Titan in early 2005.[11]

The Cassini orbiter's infrared and radar images (which can penetrate Titan's thick clouds and hazes), and the Huygens descent probe's imaging and compositional data, have revealed Titan to be a world of diverse and complex geologic landforms and atmospheric chemistry. Within Titan's atmosphere, sunlight breaks down methane molecules, parts of which can recombine into more complex hydrocarbons such as ethane, acetylene, and propane. At Titan's surface pressure and temperature, these hydrocarbons can exist as stable liquids. Radar and descent images show numerous river-like channel systems and shorelines (see Figure 4), likely formed from the erosion of the moon's icy "bedrock" by liquid hydrocarbons in much the same way that liquid water erodes the silicate bedrock of Earth's crust. Smooth, dark, pond- and lake-like features have also been mapped across Titan, thus making it potentially the only other planetary surface in our solar system where liquids can exist stably on the surface.

The emerging view of Titan as a world with a thick, highly reducing atmosphere and rivers, lakes, and an active "hydrologic" cycle of liquid hydrocarbons seems alien and, at first glance, not particularly habitable. Yet the environment turns out to be remarkably Earth-like in some respects, if we turn back the clock and think about what Earth may have been like early in its history. Prior to the emergence of oxygen-releasing microorganisms around 2.5 billion years ago, Earth's atmosphere was also likely to have been highly reducing – that is, dominated by large quantities of nitrogen and volcanically released gases such as methane, with little free oxygen and with major oxygen-bearing components like CO_2 and SO_2 potentially dissolved in the oceans and/or locked up in mineral deposits such as carbonates and sulfates. Life as we know it emerged and evolved in that environment – and that life was ultimately responsible for creating the highly oxidizing, disequilibrium atmosphere that exists and enables more complex life forms to thrive on our planet today.

While there is no good evidence for liquid water on Titan's surface today, the crust likely contains a significant component of water ice. It is thus possible that liquid water could exist in the warmer subsurface, especially if the water is mixed with salts or ammonia, which could substantially lower its melting temperature. Titan's density suggests the presence of a significant rocky interior composition, perhaps including a small rocky/metallic core, thus providing potential interior radiogenic heat similar to that created in the interiors of the terrestrial planets. That and other potential heat sources such as impacts or tidal energy, the presence of abundant and highly mobile organic molecules, and the potential presence of subsurface liquid water all combine to make Titan satisfy the basic requirements of a potentially habitable world, although one that might be trapped in a deep freeze so far away from the Sun's heat. Titan is per-

Figure 4
Orbiter Photograph of River Channels and Lake Shorelines on Titan



As it descended to the surface in 2005, the NASA/ESA Cassini Saturn orbiter's Huygens entry probe Descent Imager instrument acquired this stunning mosaic of dendritic river channels and lake shorelines on Titan. The geology and landforms are Earth-like in scale and eerily familiar, but the liquid causing these features is ethane or propane carving into "rock" made of solid water ice at a temperature of only around 90° above absolute zero. Source: NASA/ESA/Jet Propulsion Laboratory.

haps our solar system's only surviving natural laboratory for what conditions may have been like early in the history of our own planet. Therefore, even if conditions on Titan turn out not to be habitable today, studying that world in more detail could still provide us with important clues about the emergence and development of life on Earth.

The newest specific planetary body to make the list of potentially habitable worlds was a surprising entry. Saturn's tiny moon Enceladus is only about 500 km in diameter (about the distance from Philadelphia to Boston). Studied during the Voyager flybys of the early 1980s, it was not found to be particularly remarkable. Like all the other medium-sized satellites of Saturn, it has a bright surface composed of water ice and bears evidence of a complex history of impact cratering and probable tectonic stressing of the crust from internal and/or tidal forces.

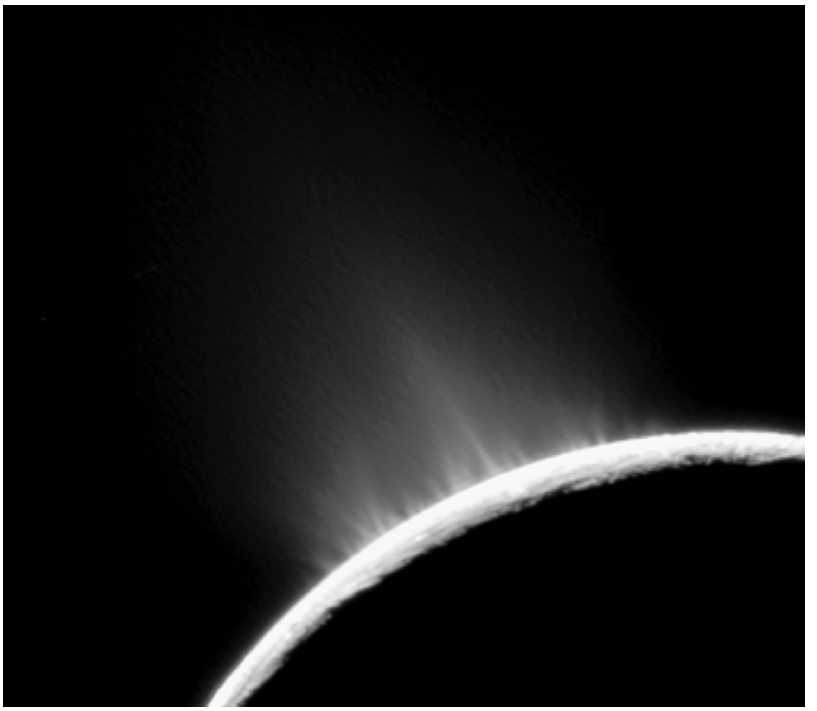
Since 2005, however, the closer and more numerous flyby encounters of the

Cassini Saturn orbital mission have uncovered some dramatic characteristics. Enceladus has an extremely thin atmosphere, and higher-temperature "hotspots" detected near the moon's south pole were found to be associated with a system of deep fractures in the icy crust now called "tiger stripes" because of their parallel linear nature and color contrasts with the surrounding icy terrains. More surprisingly, close-up imaging observations showed plumes of material streaming out of those fractures, creating the moon's thin atmosphere (see Figure 5). Enceladus is only the third known outer solar system body with active surface eruptions of material. (The others are Jupiter's Io and Neptune's Triton.) The Cassini team was able to modify the orbit of the spacecraft to fly the probe directly through those plumes on multiple encounters in order to sample their composition in situ – with exciting and dramatic results: the geysers spewing out of Enceladus's subsurface contain mostly water ice, but they also include simple and complex hydrocarbons such as propane, ethane, and acetylene, as well as ammonia.[12]

The discovery of ammonia in Enceladus's geysers is particularly important because it raises the possibility that a subsurface layer of liquid water (an ocean of sorts, like Europa's though much less voluminous) exists beneath this small moon's icy crust. This is a surprising discovery because such a small, distant, mostly icy planetary body is not expected to have an interior with enough internal heat from radiogenic, tidal, or other sources to support liquid water. Yet the geysers exist, providing strong evidence for liquid water within Enceladus. Even though planetary scientists are still actively debating the details of how such a small body could have such dramatic internal activity, the combination of liquid water, internal heat sources, and abundant organic molecules have vaulted enigmatic Enceladus onto the list of our solar system's astrobiologic hotspots, where habitable, or at least once-habitable, environments might exist.

The unexpected discovery of astrobiologically relevant environments on tiny Enceladus reminds us of how little we know about many of the solar system's other potential habitats. For example, additional evidence from the Galileo Jupiter orbiter suggests that there could be subsurface liquid water layers beneath the icy crust of Ganymede (the largest moon in the solar system) and perhaps even within Callisto as well. The Voyager 2 mission discovered evidence of internal activity on several of the medium-sized satellites of Uranus during its 1986 flyby of that system. Though none were observed to have eruptive activity, the flyby was remarkably brief, and the available imaging and other coverage is highly incomplete. Geyser-like activity observed on Neptune's large moon Triton during the Voyager 2 flyby in 1989 might indicate active internal processes and habitable subsurface environments, even at thirty times the Earth's distance from the Sun. Further still, a number of KBOs have recently been discovered, some likely even larger than the first known KBO, Pluto. Pluto has a thin atmosphere and a satellite system of its own. Are there internal tidal, radiogenic, or other heat-generating processes on these worlds that could enable the presence of liquid water? What about the larger main belt and Jupiter Trojan asteroids, many of which could have substantial internal heat sources and/or surface/subsurface organic molecule inventories? It has even been speculated that Venus may once have had a more Earth-like climate, including oceans, in the distant past. Could that hellish world once have harbored habitable environments? More detailed reconnaissance of these

Figure 5
Orbiter Photograph of Surface Eruptions on Enceladus



This enhanced image from NASA's Cassini Saturn orbiter shows jets of water ice, water vapor, and traces of organic compounds erupting from the south polar area of Enceladus. The photograph was acquired in 2005 during a special flyby designed to enhance the detectability of such plumes by pointing the cameras sunward toward the crescent-illuminated Enceladus. Source: NASA/Jet Propulsion Laboratory/Cassini Project.

destinations could further increase the number of potentially habitable environments in our solar system.

It could even be possible that by considering only solid planetary surfaces and interiors, we are relying too much on "inside the box" thinking about habitability. There are levels in the atmospheres of Venus and the giant planets, for example, where temperature and pressure conditions are more Earth-like and where liquid water could potentially exist. Could such environments be habitable in some way? Even so-called dead worlds, such as the Moon and Mercury, could offer possible habitable environments. For example, could the subsurface zones around permanently shadowed polar craters, containing organic-bearing ices from ancient comet and asteroid impacts, be "microniches" for complex biogeochemical or even biologic processes? Conventional wisdom says no, but recent discoveries of habitable environments on our own planet as well as others have shown that the conventional wisdom is often incorrect.

Finally, when assessing the future potential of the search for habitable worlds, one must take into account the exciting and profound discoveries (from the past decade or so) of extrasolar planets around nearby Sun-like stars.[13] The vast majority of the more than seven hundred extra-solar

planets discovered thus far are so-called hot Jupiters, gas giant planets orbiting in close proximity to their parent stars. These environments are probably not habitable to life as we know it because of their extremely high (1000°K to 1500°K) temperatures; but then again, we would not at first have expected many outer solar system environments in our own system to be potentially habitable given their extremely low (50°K to 100°K) temperatures.

Perhaps more exciting from an astrobiological standpoint has been the very recent discovery of more Earth-sized planets orbiting near or within the so-called habitable zone around nearby Sun-like stars. In our solar system, the habitable zone – the distance from a star at which an Earth-like planet could maintain liquid water on its surface – extends from about Venus to Mars, although the evidence for subsurface liquid water on Europa or Enceladus suggests that this definition is overrestrictive. Still, the discovery, for example, of several Earth-sized planets orbiting the Sun-like star Kepler-20[14] is profound. Even though those worlds orbit too close to their parent stars to be within the classically defined habitable zone, astronomers and planetary scientists all realize that it is only a matter of time (likely within the next year or two) that the first Earth-sized terrestrial planets are discovered orbiting within the "Goldilocks" region – not too hot, not too cold – where the convergence of liquid water, heat sources, and organic molecules would vault these distant planets onto the list of potentially habitable worlds.

The rate of discoveries related to the search for habitable worlds is likely to accelerate in the coming decades, as the research area remains scientifically interesting as well as compelling and exciting to the general public. In the near term, the Cassini Saturn mission will continue exploring the surface and atmosphere of Titan and the plumes of Enceladus during additional flybys. Three orbiters and one rover are actively exploring Mars on a daily basis, and the Curiosity rover[15] will explore ancient water-formed sedimentary rocks on Mars beginning in August 2012 – possibly providing the first definitive detections of organic molecules (or their preserved remnants) on the Red Planet. The Dawn mission[16] is completing the initial orbital reconnaissance of the large main belt asteroid Vesta and will soon begin the orbital investigation of the largest main belt asteroid, Ceres, partially to assess its potential for hosting past or present habitable environments. The New Horizons mission[17] will conduct a flyby of Pluto in July 2015, completing the initial spacecraft reconnaissance of the "classical" solar system (see Table 1), and perhaps further extending that mission by flying past one or more additional large KBOs in the 2020s. Discoveries of new extrasolar planets are announced weekly,[18] with the expectation that over time a diverse population of Earth-like worlds in our local galactic neighborhood will eventually emerge and that these potentially habitable worlds will be characterized in detail using existing ground-based and new space-based telescopic instruments.

In the longer term, NASA and other space agencies are developing plans for robotic missions to bring Martian soil and rock samples back to Earth for detailed laboratory study; to orbit and/or land on Europa, Ganymede, Titan, and Enceladus to conduct the next level of more robust exploration of those potentially habitable worlds; and to orbit and explore in more detail other destinations, such as Venus, the Uranus and Neptune systems, and examples of large small-body populations not yet explored by spacecraft, such as Jupiter's Trojan asteroids.

Planetary exploration plans are vague beyond the first half of this century, but viewed optimistically, they could include more capable (and complex) human exploration and sample return missions to many destinations on the solar system's list of potentially habitable environments, with Mars missions being the most likely to occur first. None of these new missions has officially been approved, however, and with government funding cuts to NASA's and other space agencies' budgets, many are in jeopardy of never being authorized. Nonetheless, the need for these kinds of missions will remain strong into the foreseeable future. In general, they will follow the principles of the "flyby, orbit, land, rove, and return" exploration strategy, and more specifically, they will continue and expand the search for habitable environments on the worlds around us.

ENDNOTES

1 See, for example, NASA, *Science Plan for NASA's Science Mission Directorate 2007 – 2016*, 2007, http://science.nasa.gov/about-us/science-strategy/Science_Plan_07.pdf.

2 National Research Council, *Vision and Voyages for Planetary Science in the Decade 2013 – 2022* (Washington, D.C.: National Academies Press, 2011), http://solarsystem.nasa.gov/docs/Vision_and_Voyages-FINAL.pdf.

3 National Research Council, *New Worlds, New Horizons in Astronomy and Astrophysics* (Washington, D.C.: National Academies Press, 2010), http://sites.nationalacademies.org/bpa/BPA_049810.

4 Kenneth H. Nealson, "The Limits of Life on Earth and Searching for Life on Mars," *Journal for Geophysical Research* 102 (1997): 23675 – 23686.

5 Ibid.

6 James F. Bell III, ed., *The Martian Surface: Composition, Mineralogy, and Physical Properties* (Cambridge: Cambridge University Press, 2008).

7 National Research Council, *Vision and Voyages for Planetary Science in the Decade 2013 – 2022*.

8 David Morrison, ed., *Satellites of Jupiter* (Tucson: University of Arizona Press, 1982).

9 Robert T. Pappalardo, William B. McKinnon, and Krishan K. Khurana, eds., *Europa* (Tucson: University of Arizona Press, 2008).

10 National Research Council, *Vision and Voyages for Planetary Science in the Decade 2013 – 2022*.

11 Robert Brown, Jean-Pierre LeBreton, and J. Hunter Waite, Jr., eds., *Titan from Cassini-Huygens* (Dordrecht, The Netherlands: Springer Press, 2009).

12 See, for example, J. Hunter Waite, Jr., et al., "Liquid Water on Enceladus from Observations of Ammonia and ^{40}Ar in the Plume," *Nature* 460 (2009): 487 – 490, http://dx.doi.org/10.1038/nature08153.

13 For an up-to-date census and details on the latest discoveries of extrasolar planets around nearby stars, see Paris Observatory, "Interactive Extra-Solar Planets Catalog," http://exoplanet.eu/catalog.php.

14 Francois Fressin et al., "Two Earth-sized Planets Orbiting Kepler-20," *Nature* online, 2011, http://dx.doi.org/doi:10.1038/nature10780.

15 For more details, see NASA's official mission website for the Mars Science Laboratory Curiosity rover, http://www.nasa.gov/mission_pages/msl/index.html.

16 See, for example, Mark D. Rayman et al., "Dawn: A Mission in Development for Exploration of Main Belt Asteroids Vesta and Ceres," *Acta Astronautica* 58 (2006): 605 – 616, http://dx.doi

.org/doi:10.1016/j.actaastro.2006.01.014; and James F. Bell III, "Dawn's Early Light: A Vesta Fiesta!" *Sky & Telescope*, November 2011, 32–37.

17 Details about the mission and spacecraft can be found on the official New Horizons website, http://pluto.jhuapl.edu/index.php.

18 See Paris Observatory, "Interactive Extra-Solar Planets Catalog."

E pluribus unum: From Complexity, Universality

Terence Tao

Abstract: In this brief survey, I discuss some examples of the fascinating phenomenon of universality in complex systems, in which universal macroscopic laws of nature emerge from a variety of different microscopic dynamics. This phenomenon is widely observed empirically, but the rigorous mathematical foundation for universality is not yet satisfactory in all cases.

TERENCE TAO, a Fellow of the American Academy since 2009, is Professor of Mathematics at the University of California, Los Angeles. His publications include *An Introduction to Measure Theory* (2011) and *Solving Mathematical Problems: A Personal Perspective* (2006). Three of his books, *Structure and Randomness* (2008), *Poincaré's Legacies* (2009), and *An Epsilon of Room* (2010), draw from the mathematical research blog he has maintained since 2007.

> Nature is a mutable cloud, which is always and never the same.
>
> –Ralph Waldo Emerson, "History" (1841)

Modern mathematics is a powerful tool to model any number of real-world situations, whether they be natural – the motion of celestial bodies, for example, or the physical and chemical properties of a material – or man-made: for example, the stock market or the voting preferences of an electorate.[1] In principle, mathematical models can be used to study even extremely complicated systems, with many interacting components. However, in practice, only very simple systems (ones that involve only two or three interacting agents) can be solved precisely. For instance, the mathematical derivation of the spectral lines of hydrogen, with its single electron orbiting the nucleus, can be given in an undergraduate physics class; but even with the most powerful computers, a mathematical derivation of the spectral lines of sodium, with eleven electrons interacting with each other and with the nucleus, is out of reach. (The *three-body problem*, which asks to predict the motion of three masses with respect to Newton's law of gravitation, is famously known as the only problem to have ever given Newton headaches. Unlike the two-body problem, which has a

simple mathematical solution, the three-body problem is believed not to have any simple mathematical expression for its solution, and can only be solved approximately, via numerical algorithms.) The inability to perform feasible computations on a system with many interacting components is known as the *curse of dimensionality*.

Despite this curse, a remarkable phenomenon often occurs once the number of components becomes large enough: that is, the aggregate properties of the complex system can mysteriously become predictable again, governed by simple laws of nature. Even more surprising, these macroscopic laws for the overall system are often largely *independent* of their microscopic counterparts that govern the individual components of that system. One could replace the microscopic components by completely different types of objects and obtain the same governing law at the macroscopic level. When this occurs, we say that the macroscopic law is *universal*. The universality phenomenon has been observed both empirically and mathematically in many different contexts, several of which I discuss below. In some cases, the phenomenon is well understood, but in many situations, the underlying source of universality is mysterious and remains an active area of mathematical research.

The U.S. presidential election of November 4, 2008, was a massively complicated affair. More than a hundred million voters from fifty states cast their ballots, with each voter's decision being influenced in countless ways by campaign rhetoric, media coverage, rumors, personal impressions of the candidates, or political discussions with friends and colleagues. There were millions of "swing" voters who were not firmly supporting either of the two major candidates; their final decisions would be unpredictable and perhaps even random in some cases. The same uncertainty existed at the state level: while many states were considered safe for one candidate or the other, at least a dozen were considered "in play" and could have gone either way.

In such a situation, it would seem impossible to forecast accurately the election outcome. Sure, there were electoral polls – hundreds of them – but each poll surveyed only a few hundred or a few thousand likely voters, which is only a tiny fraction of the entire population. And the polls often fluctuated wildly and disagreed with each other; not all polls were equally reliable or unbiased, and no two polling organizations used exactly the same methodology.

Nevertheless, well before election night was over, the polls had predicted the outcome of the presidential election (and most other elections taking place that night) quite accurately. Perhaps most spectacular were the predictions of statistician Nate Silver, who used a weighted analysis of all existing polls to predict correctly the outcome of the presidential election in forty-nine out of fifty states, as well as in all of the thirty-five U.S. Senate races. (The lone exception was the presidential election in Indiana, which Silver called narrowly for McCain but which eventually favored Obama by just 0.9 percent.)

The accuracy of polling can be explained by a mathematical law known as the *law of large numbers*. Thanks to this law, we know that once the size of a random poll is large enough, the probable outcomes of that poll will converge to the actual percentage of voters who would vote for a given candidate, up to a certain accuracy, known as the *margin of error*. For instance, in a random poll of a thousand voters, the margin of error is about 3 percent.

A remarkable feature of the law of large numbers is that it is *universal*. Does the

election involve a hundred thousand voters or a hundred million voters? It doesn't matter: the margin of error for the poll will remain 3 percent. Is it a state that favors McCain to Obama 55 percent to 45 percent, or Obama to McCain 60 percent to 40 percent? Is the state a homogeneous bloc of, say, affluent white urban voters, or is the state instead a mix of voters of all incomes, races, and backgrounds? Again, it doesn't matter: the margin of error for the poll will still be 3 percent. The only factor that makes a significant difference is the size of the poll; the larger the poll, the smaller the margin of error. The immense complexity of a hundred million voters trying to decide between presidential candidates collapses to just a handful of numbers.

The law of large numbers is one of the simplest and best understood of the universal laws in mathematics and nature, but it is by no means the only one. Over the decades, many such universal laws have been found to govern the behavior of wide classes of complex systems, regardless of the components of a system or how they interact with each other.

In the case of the law of large numbers, the mathematical underpinnings of the universality phenomenon are well understood and are taught routinely in undergraduate courses on probability and statistics. However, for many other universal laws, our mathematical understanding is less complete. The question of why universal laws emerge so often in complex systems is a highly active direction of research in mathematics. In most cases, we are far from a satisfactory answer to this question, but as I discuss below, we have made some encouraging progress.

After the law of large numbers, perhaps the next most fundamental example of a universal law is the *central limit theorem*. Roughly speaking, this theorem asserts that if one takes a statistic that is a combination of many independent and randomly fluctuating components, with no one component having a decisive influence on the whole, then that statistic will be approximately distributed according to a law called the *normal distribution* (or *Gaussian distribution*) and more popularly known as the *bell curve*. The law is universal because it holds regardless of exactly how the individual components fluctuate or how many components there are (although the accuracy of the law improves when the number of components increases). It can be seen in a staggeringly diverse range of statistics, from the incidence rate of accidents; to the variation of height, weight, or other vital statistics among a species; to the financial gains or losses caused by chance; to the velocities of the component particles of a physical system. The size, width, location, and even the units of measurement of the distribution vary from statistic to statistic, but the bell curve shape can be discerned in all cases. This convergence arises not because of any "low level" or "microscopic" connection between such diverse phenomena as car crashes, human height, trading profits, or stellar velocities, but because in all these cases the "high level" or "macroscopic" structure is the same: namely, a compound statistic formed from a combination of the small influences of many independent factors. That the macroscopic behavior of a large, complex system can be almost totally independent of its microscopic structure is the essence of universality.

The universal nature of the central limit theorem is tremendously useful in many industries, allowing them to manage what would otherwise be an intractably complex and chaotic system. With this theorem, insurers can manage the risk of, say, their car insurance policies without having to know all the complicated details of how car crashes occur; astronomers can

measure the size and location of distant galaxies without having to solve the complicated equations of celestial mechanics; electrical engineers can predict the effect of noise and interference on electronic communications without having to know exactly how this noise is generated; and so forth. The central limit theorem, though, is not completely universal; there are important cases when the theorem does not apply, giving statistics with a distribution quite different from the bell curve. (I will return to this point later.)

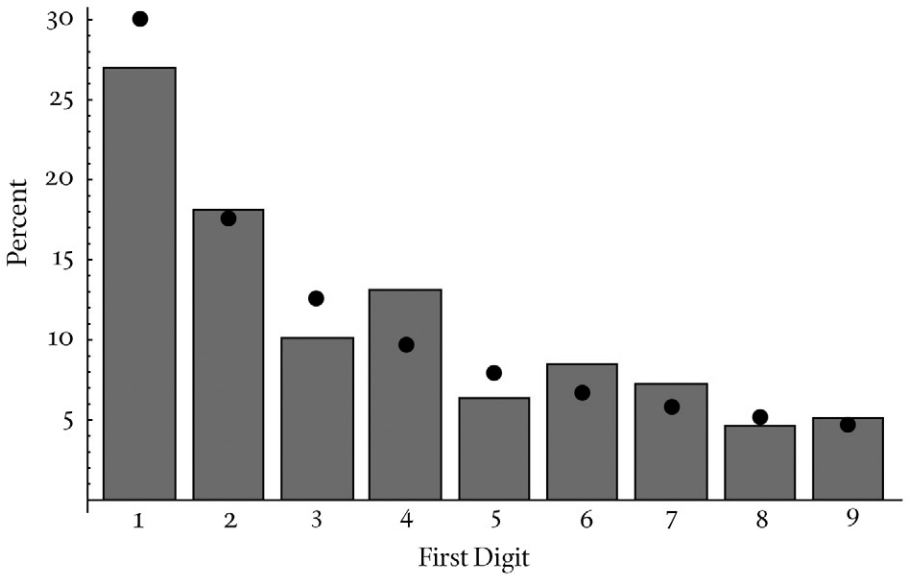
There are distant cousins of the central limit theorem that are universal laws for slightly different types of statistics. One example, *Benford's law*, is a universal law for the first few digits of a statistic of large magnitude, such as the population of a country or the size of an account; it gives a number of counterintuitive predictions: for instance, that any given statistic occurring in nature is more than six times as likely to start with the digit 1 than with the digit 9. Among other things, this law (which can be explained by combining the central limit theorem with the mathematical theory of logarithms) has been used to detect accounting fraud, because numbers that are made up, as opposed to those that arise naturally, often do not obey Benford's law (see Figure 1).

In a similar vein, *Zipf's law* is a universal law that governs the largest statistics in a given category, such as the largest country populations in the world or the most frequent words in the English language. It asserts that the size of a statistic is usually inversely proportional to its ranking; thus, for instance, the tenth largest statistic should be about half the size of the fifth largest statistic. (The law tends not to work so well for the top two or three statistics, but becomes more accurate after that.) Unlike the central limit theorem and Benford's law, which are fairly well understood mathematically, Zipf's law is primarily an empirical law; it is observed in practice, but mathematicians do not have a fully satisfactory and convincing explanation for how the law comes about and why it is universal.

So far, I have discussed universal laws for individual statistics: complex numerical quantities that arise as the combination of many smaller and independent factors. But universal laws have also been found for more complicated objects than mere numerical statistics. Take, for example, the laws governing the complicated shapes and structures that arise from *phase transitions* in physics and chemistry. As we learn in high school science classes, matter comes in various states, including the three classic states of solid, liquid, and gas, but also a number of exotic states such as plasmas or superfluids. Ferromagnetic materials, such as iron, also have magnetized and non-magnetized states; other materials become electrical conductors at some temperatures and insulators at others. What state a given material is in depends on a number of factors, most notably the temperature and, in some cases, the pressure. (For some materials, the level of impurities is also relevant.) For a fixed value of the pressure, most materials tend to be in one state for one range of temperatures and in another state for another range. But when the material is at or very close to the temperature dividing these two ranges, interesting phase transitions occur. The material, which is not fully in one state or the other, tends to split into beautifully fractal shapes known as clusters, each of which embodies one or the other of the two states.

There are countless materials in existence, each with a different set of key parameters (such as the boiling point at a given pressure). There are also a large number of mathematical models that physicists and chemists use to model these

Figure 1
Histogram of the First Digits of the Populations of the 237 Countries of the World in 2010



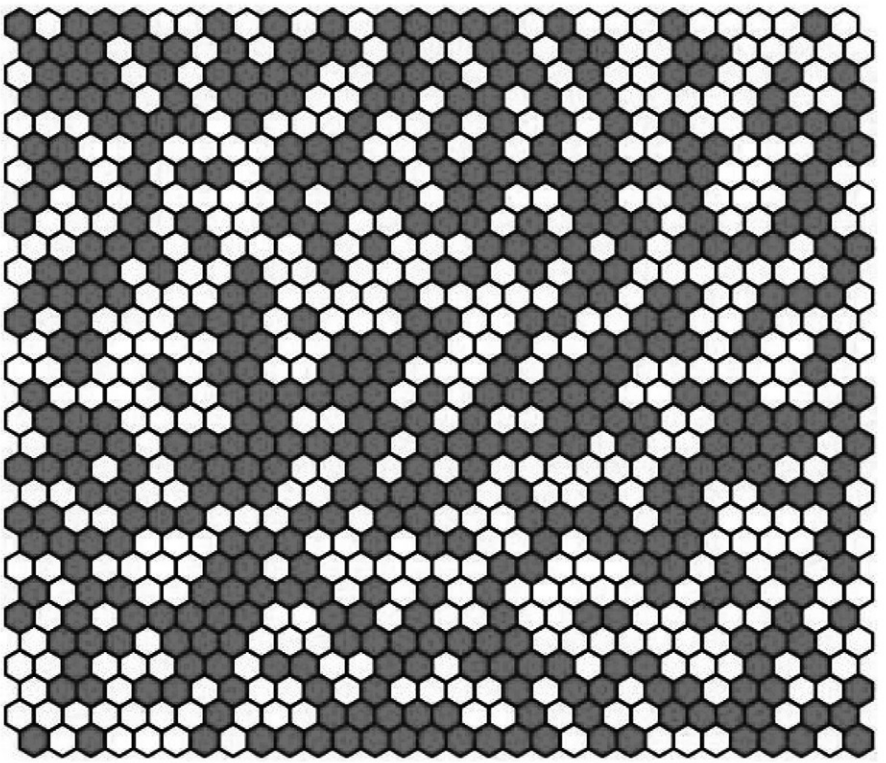
The black dots indicate the Benford's law prediction. Source: Wikipedia, http://en.wikipedia.org/wiki/File:Benfords_law_illustrated_by_world%27s_countries_population.png; used here under the Creative Commons Attribution-Share Alike license.

materials and their phase transitions, in which individual atoms or molecules are assumed to be connected to some of their neighbors by a random number of bonds, assigned according to some probabilistic rule. At the microscopic level, these models can look quite different from each other. For instance, the figures below display the small-scale structure of two typical models: a site percolation model on a hexagonal lattice (Figure 2), in which each hexagon (or site) is an abstraction of an atom or molecule randomly placed in one of two states, with the clusters being the connected regions of a single color; and a bond percolation model on a square lattice (Figure 3), in which the edges of the lattice are abstractions of molecular bonds that each have some probability of being activated, with the clusters being the connected regions given by the active bonds.

If one zooms out to look at the large-scale structure of clusters while at or near the critical value of parameters (such as temperature), the differences in microscopic structure fade away, and one begins to see a number of universal laws emerging. While the clusters have a random size and shape, they almost always have a fractal structure; thus, if one zooms in on any portion of the cluster, the resulting image more or less resembles the cluster as a whole. Basic statistics such as the number of clusters, the average size of the clusters, or the frequency with which a cluster connects two given regions of space appear to obey some specific universal laws, known as *power laws* (which are somewhat similar, though not quite the same, as Zipf's law). These laws arise in almost every mathematical model that has been put forward to explain (continuous) phase transitions

From Complexity, Universality

Figure 2
Site Percolation Model on a Hexagonal Lattice at the Critical Threshold

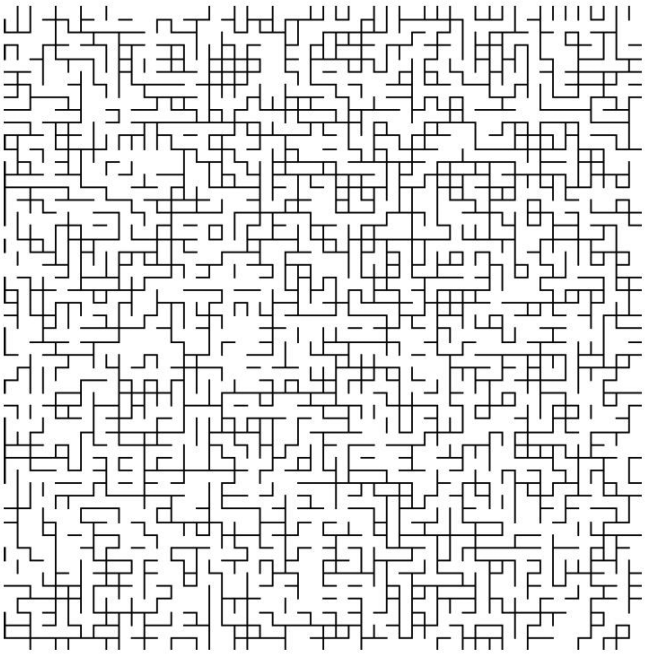


Source: Michael Kozdron, http://stat.math.uregina.ca/~kozdron/Simulations/Percolation/percolation.html; used here with permission from Michael Kozdron.

and have been observed many times in nature. As with other universal laws, the precise microscopic structure of the model or the material may affect some basic parameters, such as the phase transition temperature, but the underlying structure of the law is the same across all models and materials.

In contrast to more classical universal laws such as the central limit theorem, our understanding of the universal laws of phase transition is incomplete. Physicists have put forth some compelling heuristic arguments that explain or support many of these laws (based on a powerful, but not fully rigorous, tool known as the *renormalization group method*), but a completely rigorous proof of these laws has not yet been obtained in all cases. This is a very active area of research; for instance, in August 2010, the Fields medal, one of the most prestigious prizes in mathematics, was awarded to Stanislav Smirnov for his breakthroughs in rigorously establishing the validity of these universal laws for some key models, such as percolation models on a triangular lattice.

Figure 3
Bond Percolation Model on a Square Lattice at the Critical Threshold



Note the presence of both very small clusters and extremely large clusters. Source: Wikipedia, http://en.wikipedia.org/wiki/File:Bond_percolation_p_51.png; used here under the Creative Commons Attribution-Share Alike license.

As we near the end of our tour of universal laws, I want to consider an example of this phenomenon that is closer to my own area of research. Here, the object of study is not a single numerical statistic (as in the case of the central limit theorem) or a shape (as with phase transitions), but a discrete spectrum: a sequence of points (or numbers, or frequencies, or energy levels) spread along a line.

Perhaps the most familiar example of a discrete spectrum is the radio frequencies emitted by local radio stations; this is a sequence of frequencies in the radio portion of the electromagnetic spectrum, which one can access by turning a radio dial. These frequencies are not evenly spaced, but usually some effort is made to keep any two station frequencies separated from each other, to reduce interference.

Another familiar example of a discrete spectrum is the spectral lines of an atomic element that come from the frequencies that the electrons in the atomic shells can absorb and emit, according to the laws of quantum mechanics. When these frequencies lie in the visible portion of the electromagnetic spectrum, they give individual elements their distinctive colors, from the blue light of argon gas (which,

confusingly, is often used in neon lamps, as pure neon emits orange-red light) to the yellow light of sodium. For simple elements, such as hydrogen, the equations of quantum mechanics can be solved relatively easily, and the spectral lines follow a regular pattern; but for heavier elements, the spectral lines become quite complicated and not easy to work out just from first principles.

An analogous, but less familiar, example of spectra comes from the scattering of neutrons off of atomic nuclei, such as the Uranium-238 nucleus. The electromagnetic and nuclear forces of a nucleus, when combined with the laws of quantum mechanics, predict that a neutron will pass through a nucleus virtually unimpeded for some energies but will bounce off that nucleus at other energies, known as scattering resonances. The internal structures of such large nuclei are so complex that it has not been possible to compute these resonances either theoretically or numerically, leaving experimental data as the only option.

These resonances have an interesting distribution; they are not independent of each other, but instead seem to obey a precise repulsion law that makes it unlikely that two adjacent resonances are too close to each other – somewhat analogous to how radio station frequencies tend to avoid being too close together, except that the former phenomenon arises from the laws of nature rather than from government regulation of the spectrum. In the 1950s, the renowned physicist and Nobel laureate Eugene Wigner investigated these resonance statistics and proposed a remarkable mathematical model to explain them, an example of what we now call a *random matrix model*. The precise mathematical details of these models are too technical to describe here, but in general, one can view such models as a large collection of masses, all connected to each other by springs of various randomly selected strengths. Such a mechanical system will oscillate (or resonate) at a certain set of frequencies; and the Wigner hypothesis asserts that the resonances of a large atomic nucleus should resemble that of the resonances of a random matrix model. In particular, they should experience the same repulsion phenomenon. Because it is possible to rigorously prove repulsion of the frequencies of a random matrix model, a heuristic explanation can be given for the same phenomenon that is experimentally observed for nuclei.

Of course, an atomic nucleus does not actually resemble a large system of masses and springs (among other things, it is governed by the laws of quantum mechanics rather than of classical mechanics). Instead, as we have since discovered, Wigner's hypothesis is a manifestation of a universal law that governs many types of spectral lines, including those that ostensibly have little in common with atomic nuclei or random matrix models. For instance, the same spacing distribution was famously found in the waiting times between buses arriving at a bus stop in Cuernavaca, Mexico (without, as yet, a compelling explanation for why this distribution emerges in this case).

Perhaps the most unexpected demonstration of the universality of these laws came from the wholly unrelated area of *number theory*, and in particular the distribution of the prime numbers 2, 3, 5, 7, 11, and so on – the natural numbers greater than 1 that cannot be factored into smaller natural numbers. The prime numbers are distributed in an irregular fashion through the integers; but if one performs a spectral analysis of this distribution, one can discern certain long-term oscillations in the distribution (sometimes known as the music of the primes), the frequencies of which are described by a sequence of complex numbers known as

the (non-trivial) zeroes of the Riemann zeta function, first studied by Bernhard Riemann in 1859. (For this discussion, it is not important to know exactly what the Riemann zeta function is.) In principle, these numbers tell us everything we would wish to know about the primes. One of the most famous and important problems in number theory is the *Riemann hypothesis*, which asserts that these numbers all lie on a single line in the complex plane. It has many consequences in number theory and, in particular, gives many important consequences about the prime numbers. However, even the powerful Riemann hypothesis does not settle everything on this subject, in part because it does not directly say much about how the zeroes are distributed on this line. But there is extremely strong numerical evidence that these zeroes obey the same precise law that is observed in neutron scattering and in other systems; in particular, the zeroes seem to "repel" each other in a manner that matches the predictions of random matrix theory with uncanny accuracy. The formal description of this law is known as the Gaussian Unitary Ensemble (GUE) hypothesis. (The GUE is a fundamental example of a random matrix model.) Like the Riemann hypothesis, it is currently unproven, but it has powerful consequences for the distribution of the prime numbers.

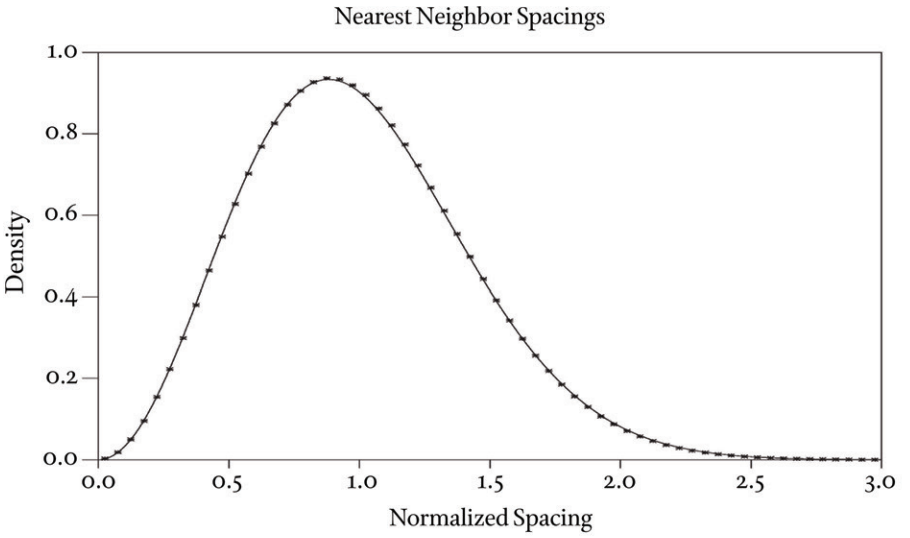
The discovery of the GUE hypothesis, connecting the music of the primes and the energy levels of nuclei, occurred at the Institute for Advanced Study in 1972, and the story is legendary in mathematical circles. It concerns a chance meeting between the mathematician Hugh Montgomery, who had been working on the distribution of zeroes of the zeta function (and more specifically, on a certain statistic relating to that distribution known as the pair correlation function), and the renowned physicist Freeman Dyson. In his book *Stalking the Riemann Hypothesis*, mathematician and computer scientist Dan Rockmore describes that meeting:

> As Dyson recalls it, he and Montgomery had crossed paths from time to time at the Institute nursery when picking up and dropping off their children. Nevertheless, they had not been formally introduced. In spite of Dyson's fame, Montgomery hadn't seen any purpose in meeting him. "What will we talk about?" is what Montgomery purportedly said when brought to tea. Nevertheless, Montgomery relented and upon being introduced, the amiable physicist asked the young number theorist about his work. Montgomery began to explain his recent results on the pair correlation, and Dyson stopped him short – "Did you get this?" he asked, writing down a particular mathematical formula. Montgomery almost fell over in surprise: Dyson had written down the sinc-infused pair correlation function.... Whereas Montgomery had traveled a number theorist's road to a "prime picture" of the pair correlation, Dyson had arrived at this formula through the study of these energy levels in the mathematics of matrices.[2]

The chance discovery by Montgomery and Dyson – that the same universal law that governs random matrices and atomic spectra also applies to the zeta function – was given substantial numerical support by the computational work of Andrew Odlyzko beginning in the 1980s (see Figure 4). But this discovery does not mean that the primes are somehow nuclear-powered or that atomic physics is somehow driven by the prime numbers; instead, it is evidence that a single law for spectra is so universal that it is the natural end product of any number of different processes, whether from nuclear physics, random matrix models, or number theory.

The precise mechanism underlying this law has not yet been fully unearthed;

Figure 4
Spacing Distribution for a Billion Zeroes of the Riemann Zeta Function, with the Corresponding Prediction from Random Matrix Theory



Source: Andrew M. Odlyzko, "The 10^{22}-nd Zero of the Riemann Zeta Function," in *Dynamical, Spectral, and Arithmetic Zeta Functions*, Contemporary Mathematics Series, no. 290, ed. Machiel van Frankenhuysen and Michel L. Lapidus (American Mathematical Society, 2001), 139–144, http://www.dtc.umn.edu/~odlyzko/doc/zeta.10to22.pdf; used here with permission from Andrew Odlyzko.

in particular, we still do not have a compelling explanation, let alone a rigorous proof, of why the zeroes of the zeta function are subject to the GUE hypothesis. However, there is now a substantial body of rigorous work (including some of my own work, and including some substantial breakthroughs in just the last few years) that gives support to the universality of this hypothesis, by showing that a wide variety of random matrix models (not just the most famous model of the GUE) are all governed by essentially the same law for their spacings. At present, these demonstrations of universality have not extended to the number theoretic or physical settings, but they do give indirect support to the law being applicable in those cases.

The arguments used in this recent work are too technical to give here, but I will mention one of the key ideas, which my colleague Van Vu and I borrowed from an old proof of the central limit theorem by Jarl Lindeberg from 1922. In terms of the mechanical analogy of a system of masses and springs (mentioned above), the key strategy was to replace just one of the springs by another, randomly selected spring and to show that the distribution of the frequencies of this system did not change significantly when doing so. Applying this replacement operation to each spring in turn, one can eventually replace a given random matrix model with a completely different model while keeping the distribution mostly unchanged – which can be used to show that large class-

es of random matrix models have essentially the same distribution.

This is a very active area of research; for instance, simultaneously with Van Vu's and my work from last year, László Erdös, Benjamin Schlein, and Horng-Tzer Yau also gave a number of other demonstrations of universality for random matrix models, based on ideas from mathematical physics. The field is moving quickly, and in a few years we may have many more insights into the nature of this mysterious universal law.

There are many other universal laws of mathematics and nature; the examples I have given are only a small fraction of those that have been discovered over the years, from such diverse subjects as dynamical systems and quantum field theory. For instance, many of the macroscopic laws of physics, such as the laws of thermodynamics or the equations of fluid motion, are quite universal in nature, making the microscopic structure of the material or fluid being studied almost irrelevant, other than via some key parameters such as viscosity, compressibility, or entropy.

However, the principle of universality does have definite limitations. Take, for instance, the central limit theorem, which gives a bell curve distribution to any quantity that arises from a combination of many small and independent factors. This theorem can fail when the required hypotheses are not met. The distribution of, say, the heights of all human adults (male and female) does not obey a bell curve distribution because one single factor – gender – has so large an impact on height that it is not averaged out by all the other environmental and genetic factors that influence this statistic.

Another very important way in which the central limit fails is when the individual factors that make up a quantity do not fluctuate independently of each other, but are instead correlated, so that they tend to rise or fall in unison. In such cases, "fat tails" (also known colloquially as "black swans") can develop, in which the quantity moves much further from its average value than the central limit theorem would predict. This phenomenon is particularly important in financial modeling, especially when dealing with complex financial instruments such as the collateralized debt obligations (CDOs) that were formed by aggregating mortgages. As long as the mortgages behaved independently of each other, the central limit theorem could be used to model the risk of these instruments; but in the recent financial crisis (a textbook example of a black swan), this independence hypothesis broke down spectacularly, leading to significant financial losses for many holders of these obligations (and for their insurers). A mathematical model is only as strong as the assumptions behind it.

A third way in which a universal law can break down is if the system does not have enough degrees of freedom for the law to take effect. For instance, cosmologists can use universal laws of fluid mechanics to describe the motion of entire galaxies, but the motion of a single satellite under the influence of just three gravitational bodies can be far more complicated (being, quite literally, rocket science).

Another instance where the universal laws of fluid mechanics break down is at the *mesoscopic* scale: that is, larger than the microscopic scale of individual molecules, but smaller than the macroscopic scales for which universality applies. An important example of a mesoscopic fluid is the blood flowing through blood vessels; the blood cells that make up this liquid are so large that they cannot be treated merely as an ensemble of microscopic molecules, but rather as mesoscopic agents with complex behavior. Other examples of materials with interesting mesoscopic behavior

include colloidal fluids (such as mud), certain types of nanomaterials, and quantum dots; it is a continuing challenge to mathematically model such materials properly.

There are also many macroscopic situations in which no universal law is known to exist, particularly in cases where the system contains human agents. The stock market is a good example: despite extremely intensive efforts, no satisfactory universal laws to describe the movement of stock prices have been discovered. (The central limit theorem, for instance, does not seem to be a good model, as discussed earlier.) One reason for this shortcoming is that any regularity discovered in the market is likely to be exploited by arbitrageurs until it disappears. For similar reasons, finding universal laws for macroeconomics appears to be a moving target; according to Goodhart's law, if an observed statistical regularity in economic data is exploited for policy purposes, it tends to collapse. (Ironically, Goodhart's law itself is arguably an example of a universal law.)

Even when universal laws do exist, it still may be practically impossible to use them to make predictions. For instance, we have universal laws for the motion of fluids, such as the Navier-Stokes equations, and these are used all the time in such tasks as weather prediction. But these equations are so complex and *unstable* that even with the most powerful computers, we are still unable to accurately predict the weather more than a week or two into the future. (By unstable, I mean that even small errors in one's measurement data, or in one's numerical computations, can lead to large fluctuations in the predicted solution of the equations.)

Hence, between the vast, macroscopic systems for which universal laws hold sway and the simple systems that can be analyzed using the fundamental laws of nature, there is a substantial middle ground of systems that are too complex for fundamental analysis but too simple to be universal – plenty of room, in short, for all the complexities of life as we know it.

ENDNOTES

[1] This essay benefited from the feedback of many readers of my blog. They commented on a draft version that (together with additional figures and links) can be read at http://terrytao.wordpress.com/2010/09/14/a-second-draft-of-a-non-technical-article-on-universality/.

[2] Dan Rockmore, *Stalking the Riemann Hypothesis: The Quest to Find the Hidden Law of Prime Numbers* (New York: Pantheon Books, 2005).

Small Machines

Paul L. McEuen

Abstract: Over the last fifty years, small has emerged as the new big thing. The reduction of information and electronics to nanometer dimensions has revolutionized science, technology, and society. Now scientists and engineers are creating physical machines that operate at the nanoscale. Using approaches ranging from lithographic patterning to the co-opting of biological machinery, new devices are being built that can navigate, sense, and alter the nanoscale world. In the coming decades, these machines will have enormous impact in fields ranging from biotechnology to quantum physics, blurring the boundary between technology and life.

> Look round the world, contemplate the whole and every part of it: you will find it to be nothing but one great machine, subdivided into an infinite number of lesser machines, which again admit of subdivisions to a degree beyond what human senses and faculties can trace and explain. All these various machines, and even their most minute parts, are adjusted to each other with an accuracy which ravishes into admiration all men who have ever contemplated them. The curious adapting of means to ends, throughout all nature, resembles exactly, though it much exceeds, the productions of human contrivance; of human design, thought, wisdom, and intelligence.
>
> –David Hume, *Dialogues Concerning Natural Religion*, 1779

PAUL L. McEUEN is the Goldwin Smith Professor of Physics at Cornell University, where he is also Director of the Kavli Institute at Cornell for Nanoscale Science and the Laboratory of Atomic and Solid State Physics. His research focuses on the science and technology of nanostructures, particularly carbon-based systems such as nanotubes and graphene. His work has appeared in *Nano Letters*, *Nature*, and *Science*, among other journals. His debut novel *Spiral*, a scientific thriller, was published in 2011.

How small can we make things? The physicist, scientific raconteur, and future Nobel Prize recipient Richard Feynman asked this question in December 1959. In a now-famous talk at the annual meeting of the American Physical Society, he took that very simple question and followed it to its end. His talk helped define the field of nanotechnology. With insight and precision, Feynman clearly outlined the promises and challenges of making things small. It was a clarion call, shaping the field of nanoscience over the next fifty years – a field that has, in turn, reshaped the world.

Small Machines

The program of miniaturization that Feynman outlined in 1959 can be divided into three main parts:

1) miniaturization of information;
2) miniaturization of electronics; and
3) miniaturization of machines.

Parts 1 and 2 of the revolution were already under way by the time Feynman gave his speech. (He was as much a reporter as a visionary.) Companies like Fairchild Semiconductor in the San Francisco Bay area and Texas Instruments in Austin were making transistors and assembling them into integrated circuits. Gordon Moore, founder and former CEO of Intel, would soon begin counting the number of transistors per chip, noticing the number was doubling every year or so. He predicted that the trend would continue for at least another decade.

After five decades of Moore's Law, we now have computers with 25 nanometer (nm) feature sizes. To put that in perspective, with a 25 nm pen one could draw a map of the world on a single Intel wafer, recording features down to the scale of an individual human being. Furthermore, running at a 2 gigahertz clock speed, a computer can perform as many operations in a second as a human can have thoughts in a lifetime.

As the integrated circuit shrank, a second revolution was taking place in data storage, with magnetic-core memories, hard drives, and flash memories steadily pushing down the size of a bit of data. Information is now pervasive, and it is small, with bit sizes measured in tens of nanometers. We are awash in digital information – zettabytes of it – adding the equivalent of a hundred thousand books every year to the global bookshelf for every man, woman, and child on the planet. In this age of big data, terabytes of information are available on everything from human genome sequences to astrophysical maps of stars and galaxies.

The miniaturization of computing and information storage is the most important technological development of the last half-century. But what about Part 3 of Feynman's program? Where are the nanomachines? And what exactly do we mean by *nanomachine*?

Simply stated, a machine is a device that accomplishes a task, usually with an input of energy and/or information. By this definition, a computer is a machine; so for our purposes, we'll narrow the definition to a device that operates when something *physical* moves. Such a machine can be as simple as a vibrating reed or as complex as an automobile. To be a nanomachine, it should be submicron, at the very edge of what can be resolved in a standard optical microscope.

In the last decade, we have seen critical advances in both the scientific underpinnings and the fabrication technologies needed to create, study, and exploit nanomachines. These miniature devices translate or vibrate, push or grab; they record, probe, and modify the nanoscale world around them. Progress has come from two very different approaches, each bringing its own ideas, themes, and materials to bear. The first approach, which I call the *lithographer's approach*, adopts the techniques of the microelectronics revolution discussed above. The second, the *hacker's approach*, seeks to appropriate the molecular machinery of life. I examine these approaches in turn and consider what is happening at the interface of these two disciplines. Finally, I look at emerging approaches that will take nanomachinery to the next level. My goal is not to offer a comprehensive survey, but rather to present a few snapshots that give a sense of the current state of the field as well as where it might be headed.

* * *

If quantum mechanics hasn't profoundly shocked you, you haven't understood it yet.

–Niels Bohr

In 2010, physicists at the University of California, Santa Barbara, created what some have called the first quantum machine: that is, a machine whose operation follows the laws of quantum mechanics (see Figure 1).[1] It is the latest breakthrough to emerge from the lithographer's approach to nanofabrication. This technology, usually called MEMs (MicroElectroMechanical systems) or NEMs (NanoElectroMechanical systems), exploits and extends the lithographic, thin-film deposition, and etching techniques of the microelectronics industry to make machines that move. The field of MEMs, which has roots stretching back to the 1980s, is now a $10 billion/year industry, with products ranging from accelerometers in airbags, microfluidic valves for lab-on-a-chip systems, and tiny mirrors for steering light in projectors. For example, Apple's iPhone 4 has two MEMs microphones and a three-axis MEMs gyroscope that detects when the phone rotates.

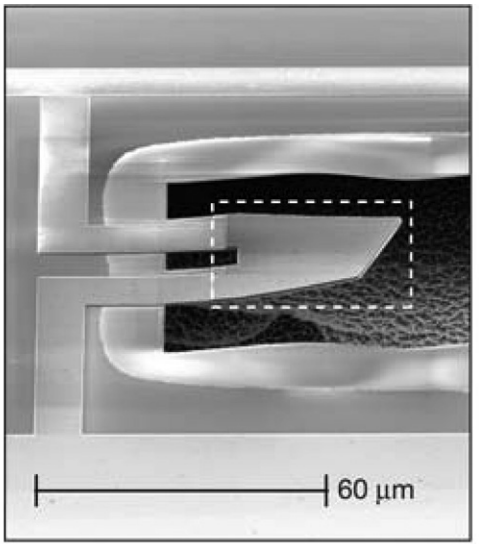
The field of NEMs pushes this approach to its limits, using advanced lithography to create mechanical devices with dimensions comparable to those found in the smallest integrated circuits. The reason is not simply miniaturization for its own sake: the rules of operation of nanomachines can be fundamentally different than their larger-scale counterparts. In particular, the possibility of seeing quantum behavior in nanoscale machines has tantalized and energized the field of NEMs for more than a decade. The counterintuitive rules of quantum mechanics – quantized energies, quantum tunneling, zero-point fluctuations, and the Schrödinger's cat paradox – have been tested with electrons and photons for nearly a hundred years. But can the rules of quantum mechanics also manifest in mechanical machines?

The UC-Santa Barbara group showed that the answer is yes. They used a vibrating nanoscale beam, one of the simplest possible nanomachines. According to the rules of quantum mechanics, the amplitude of the object's vibration is quantized, just like the orbits of electrons in an atom. Furthermore, even when cooled to absolute zero, the beam should exhibit quantum fluctuations in its position. A number of research groups have explored various geometries of resonators to try to reach these quantum limits, along with various techniques to detect the beam's miniscule motion.

The physicists at UC-Santa Barbara found success using a novel oscillating mode of the beam and a clever detection scheme involving a quantum superconducting circuit. They were able to cool the oscillator to its ground state, before adding a single vibrational quantum of energy and measuring the resulting motion. They were even able to put the device in a quantum superposition, where it was simultaneously in its ground state and also oscillating, the two possibilities interfering with each other like Schrödinger's dead-and-alive cat. *Science* magazine chose this experiment as its Breakthrough of the Year in 2010: the first demonstration of quantum behavior in a mechanical system.

So what? Why should we care about quantum machines? The first reason is pure curiosity, the drive to show that physical machines are also subject to quantum rules. But these devices may also have applications in the field of quantum information, where the digital bits of a computer are handled as quantum objects. They also stretch the boundaries of physical law, testing macro laws at the nanoscale and nano laws at the macro scale.

Figure 1
A Quantum Machine



Created by a team of physicists at the University of California, Santa Barbara, this nanomachine was the first to demonstrate quantum behavior in a mechanical system. Source: Aaron D. O'Connell et al., "Quantum Ground State and Single-Phonon Control of a Mechanical Resonator," *Nature* 464 (2010): 697; used here with permission from the authors.

For example, a number of theories have posited that gravity may act differently on small objects. Can these effects be detected in a nanoscale oscillator? Moving in the other direction, is there a scale at which the rules of quantum mechanics cease to work, thereby forcing a revision of the fundamental laws of quantum theory? These are scientific long shots; but a measurement that changed our notions about gravity or quantum mechanics would profoundly challenge our understanding of the workings of the universe.

* * *

> Over the next 20 years synthetic genomics is going to become the standard for making anything.
>
> –Craig Venter

The lithographer's nanomachines, for all their progress, pale in comparison to the machines of life. The MEMs devices in your iPhone cannot begin to match the complexity and sophistication of the simplest bacterium. So how far are we from building something more like a bacterium – say, a nanosubmarine that can travel through the bloodstream, searching out cancer cells? Or even better, how long until we have a machine that can make copies of itself, growing exponentially until there are trillions? We lithographers can only shake our heads in awe at the power of biology. Just one simple mechanical component – for example, the rotary motor that powers the flagellum of a bacterium – is far beyond our abilities. A lithography-based technology that would even remotely match the capabilities of life is decades away at best. But what if we don't want to wait for decades? What if we want our nanosubmarines now?

There is a shortcut, a hack, and one that humans have exploited before. Before hu-

mans could build tractors, they harnessed oxen to pull their ploughs. We already have one fully functioning nanotechnology: life. So why not hack it? Never mind decades of developing ever-more complex machines by lithographic processes and teaching them to work together in more sophisticated systems; just take control of the bacterium the way we did livestock. This is the dream of synthetic biology: to take unicellular life, harness it, reprogram it, and control its design down to the last amino acid.

Synthetic biology is the latest link in a long chain spanning from the domestication of animals, to the invention of farming, to animal and plant breeding, and on to genetic engineering. But synthetic biology aims to advance to the next level, albeit at the scale of the bacterium. The goal is to usurp cellular biological machines, to turn life into an engineering discipline. Instead of simply tinkering, one would create cells in the same way that an integrated circuit is assembled, mixing and matching motors and metabolic pathways, all programmed in the language of DNA. Sit at a computer, type out a genetic code, push a button, and see your dream of life come to be.

Researchers at the J. Craig Venter Institute in Maryland took a major step forward in 2010, creating what they called the first artificial organism.[2] The project was a tour de force; it took more than a decade and cost tens of millions of dollars. First, the group painstakingly built the genome of a known bacterium from scratch, synthesizing and stitching together the strands of DNA until they had reproduced the entire operating instructions for a cell. To prove ownership, they added a few genetic watermarks, including their names and quotes from James Joyce ("To live to err, to fall, to triumph, to recreate life out of life") and Richard Feynman ("What I cannot build I cannot understand"), all rendered in the ACGT (adenine, cytosine, guanine, and thymine) alphabet of life. Next, they put that genome into the shell of another bacterium, *Mycoplasma mycoides*, whose own genome had been removed. After some jiggering, they got the new organism to boot up, creating what they claimed as the first artificial life form. While many have criticized the work as overly hyped, it is nonetheless a landmark in synthetic biology, a demonstration of what is possible. It was voted as one of the runners-up for the 2010 Breakthrough of the Year by *Science* magazine, losing out to the quantum machine from the team at UC-Santa Barbara.

What are such artificial life forms good for? If one believes Craig Venter's quote at the start of this section, it appears that there is little they *wouldn't* be good for. The practitioners of synthetic biology are working to reprogram organisms to make everything from cheap malaria medications to biofuels. The promise is great, but the task is much harder than it seems. The giddy early days of the field are giving way to an appreciation of the complexity and finicky nature of biological organisms. Synthetic biologists have schemes to fix this, such as devolving life to a version simple enough for the genetic programmer to exert full control. They are stripping down simple bacteria to the minimal state needed to survive, where all the remaining parts and their interrelationships are fully understood.

Hacking life stirs up fear as well as hope. Venter's dictum of synthetic biology as the standard for making everything could also include new and dangerous pathogens. What's to stop a synthetic biologist from accidentally creating a dangerous organism, or a terrorist from knowingly creating one? On one hand, the danger may be overstated. The world is rife with nanomachines trying to kill us: viruses and bacterial infections have been per-

fecting their skills for eons. It is no trivial matter to come up with something truly new that our bodies could not handle. On the other hand, we do know that parasites and hosts coevolve, tuning their responses to each other in a delicate dance. The introduction of a known organism in a new setting could trigger dramatic rearrangements of these ecological relationships.

* * *

> The machine does not isolate man from the great problems of nature but plunges him more deeply into them.
>
> –Antoine de Saint-Exupéry

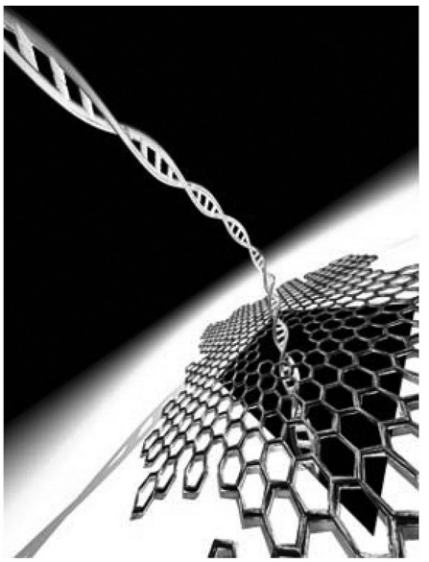
A new class of nanomachines is emerging that combines the best of the lithographer's and the hacker's approaches. These machines marry the speed and processing power of microelectronics and optics with the functionality of biological machines. The most dramatic examples are found in the field of DNA sequencing. The promise is easy to see: the human genome consists of three billion bases. What if we could read the genetic code at the speed of a modern microprocessor? A genome could be sequenced in seconds.

Nearly all next-generation sequencing techniques involve the careful integration of biological genomic machinery with electronic/optical elements to read and collect the data. One such technique is called SMRT, or Single Molecule Real Time sequencing; it was invented at Cornell University in 2003 and developed commercially by Pacific Biosciences.[3] A nanofabricated well is created that confines light to a very small volume containing a single DNA polymerase, the biological machine that constructs double-stranded DNA. As the polymerase adds new bases to the DNA strand, a characteristic fluorescence signal is emitted that indicates the identity of the genetic letter. This technique can be run in massively parallel fashion, with millions of wells simultaneously monitored. Other approaches under commercial development include the electronic detection of hydrogen ions released during synthesis, or the detection of the change in current measured when DNA threads itself though a nanopore.

These next-generation technologies have helped put DNA sequencing on a hyperspeed version of Moore's law. The first human genome cost a few billion dollars to sequence. As recently as the beginning of 2008, the cost was approximately $10 million. Now, four years later, the cost is closer to $10,000, a *thousandfold* drop in approximately four years! This remarkable drop in sequencing cost (and a steady drop in DNA synthesis cost) is redefining the possible, outperforming Moore's law for electronics by leaps and bounds. These hybrid organic-inorganic nanomachines are revolutionizing the biological sciences, turning genomic sequencing into what has been described as an information microscope with the ability to address questions in fields ranging from molecular biology to evolution and ecology. Next up is the $100 genome, which will likely set off a revolution in personal genomics. It would make whole genome sequencing as common as knowing your blood type. Medicines and procedures could then be tailored to fit specific genomic profiles.

In the last decade, new kinds of hybrid two-dimensional materials have appeared that bring electronics and biology closer together. These atomically thin materials combine many of the attributes of biological and microelectronic materials in a single platform. The most widely touted is graphene, the hexagonal arrangement of carbon atoms that is found in pencil lead. Individual sheets of graphene were isolated in 2004, and this work led to a Nobel Prize a scant six years later. These sheets combine electronic and optical

Figure 2
Artist's Rendition of a Graphene Nanopore DNA Sequencer



DNA passing through a hole in a one atom–thick graphene sheet changes the flow of ions around it in a way that can be used to determine the DNA's genetic sequence. Image courtesy of the Cees Dekker lab at TU Delft/Tremani.

properties that rival the best semiconductors and metals, but physically they are as flexible as a biological membrane. Scientists, including my own group, have already used them to create nanoscale resonators analogous to the quantum machines described above; other groups have created nanopores in these membranes that can be used to detect the translocation of DNA (Figure 2). Currently, the reach of lithography stops at approximately 20 nm, while a strand of DNA is approximately 2 nm across. These new materials help bridge that gap: the first atomic-scale materials that embody the full power of the microelectronics revolution but that can interact with biological molecules on their own terms.

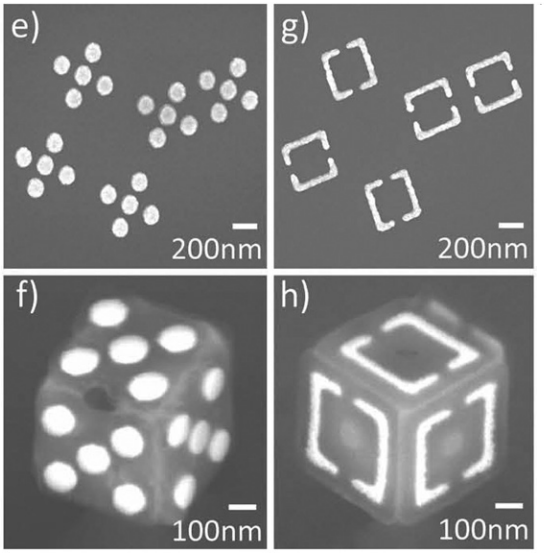
* * *

> The mechanic should sit down among levers, screws, wedges, wheels, etc. like a poet among the letters of the alphabet, considering them as the exhibition of his thoughts; in which a new arrangement transmits a new idea to the world.
>
> –Robert Fulton, 1796

Biology takes a very unusual approach to the creation of nanomachine parts: it folds them out of strings. When the ribosome makes a new part, it first creates a linear, one-dimensional amino acid biopolymer. That biopolymer, guided by the amino acid sequence encoded in its structure (and sometimes with the help of other machines), folds itself into useful forms. A conceptually similar approach is taken by the practitioners of the ancient art of origami, in which complex three-dimensional objects emerge from the folding-up of a two-dimensional sheet of paper.

To the nanotechnologist, this approach has much to recommend it. It allows a

Figure 3
Lithographic Origami

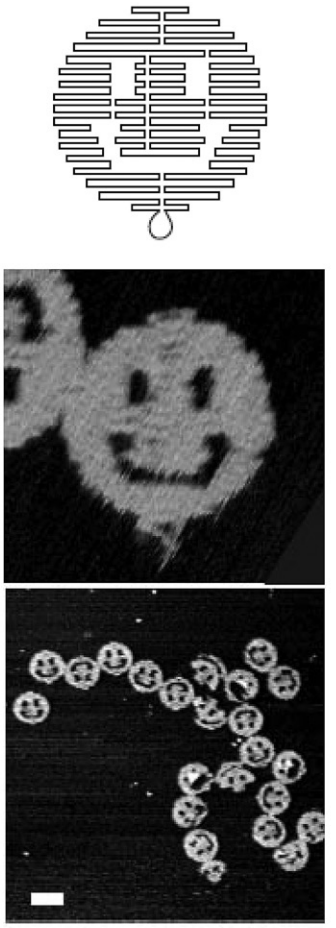


Panels are patterned using multistep lithography that subsequently folds into micron-sized cubes with novel physical, chemical, or electromagnetic properties. Source: Jeong-Hyun Cho et al., "3D Nanofabrication: Nanoscale Origami for 3D Optics," *Small* 7 (14) (2011): 1943; used here with permission from the authors.

planar fabrication technology like lithography, or a linear fabrication technology like DNA synthesis, to be the basis for constructing more complex three-dimensional systems. In the field of MEMS, engineers have been applying origami-like techniques for some time to create, for example, arrays of steerable mirrors that look like something from a miniature pop-up book. The Gracias research group at Johns Hopkins University is pushing this technique to the nanoscale, lithographically patterning 100-nm scale planar features connected with tin solder at the joints that, when heated, fold up the structure (Figure 3).[4] My own research group is attempting to perform similar origami tricks using atomically thin graphene membranes.

The hackers have also gotten into the origami game, with great success. The field was pioneered by New York University's Ned Seeman, who designed short DNA sequences that assembled themselves into interesting shapes. In 2006, Paul Rothemund at Caltech took DNA origami to the next level. He developed algorithms to form arbitrary two-dimensional patterns in DNA, from happy faces to maps of the world.[5] His approach begins with a single long "raster scan" DNA strand that forms the basic pattern, to which a number of short staple strands are designed to pin the structure together. Assembly involves mixing all the strands then heating and cooling the mix, after which it assembles itself. To demonstrate the technique, Rothemund made smiley faces by the billions, creating what has been called "the most concentrated happiness ever experienced on Earth" (see Figure 4). This set off a wave of research on DNA origami projects, including one project with robotic DNA spiders that travel a DNA origami

Figure 4
DNA Origami



DNA sequences can be designed so that they fold themselves into arbitrary two-dimensional shapes; here they are shown creating smiley faces 200 nm across. Source: Paul W.K. Rothemund, "Folding DNA to Create Nanoscale Shapes and Patterns," *Nature* 440 (2006): 297; used with permission from the author.

landscape.[6] More recent work by the Church research group at Harvard University's Wyss Institute for Biologically Inspired Engineering is taking this technology toward clinical applications. Church and colleagues have created a DNA origami robot that could passively carry a payload and then release it when certain molecular signatures are encountered on the surface of a cell, a kind of smart land mine that could be deployed to attack cancerous cells. It may not be the nanoscale submarine of the nano-technologist's fantasies, but it is certainly another step forward.

* * *

> Ever tried. Ever failed. No matter. Try again. Fail again. Fail better.
>
> –Samuel Beckett

Fifty years after Richard Feynman's speech to the American Physical Society, nanomachines are finally moving from

dream to reality. Young scientists are flocking to the field, drawn by the promise of a new technology that is progressing in leaps and bounds. In the area of synthetic biology, the International Genetically Engineered Machine (iGEM) competition, an undergraduate competition in synthetic biology, is entering its eighth year, with more than a hundred teams vying to build the coolest organism possible. The winner of the 2011 competition, a team from the University of Washington, created a strain of *E. coli* that could make the alkane components of diesel fuel. In the area of nano-bots, the National Institute of Science and Technology sponsors a Mobile Microrobotics Challenge, where dust mote–sized robots compete to push tiny soccer balls into goals.

So what about the ultimate: will we soon be building nanomachines that could genuinely be called a second form of life – machines that can build copies of themselves from raw materials, machines that can change and evolve? How will we do this? The obvious answer is to look to life for inspiration. But here we are confronted with a remarkable fact: we do not know how life did it. Life is a complex and interwoven technology, and no one yet understands how it bootstrapped itself into being. Life has two important parts, metabolism and replication. Did one part come first, or did they emerge together? Did life emerge from self-replicating RNA molecules, or did it develop metabolism and later add an information-containing molecule? It is one of the great outstanding questions in science; its answer may both shape and be shaped by advances in nanomachine design. Recall Feynman's quote: *What I cannot build I cannot understand*. The lessons learned as we try to build ever-more sophisticated nanomachines will almost certainly inform our understanding of the origins of life, and vice versa.

Back in the 1960s, the semiconductor industry joined together what had been disparate fields, marrying electronics to information/computation. Today, they are so closely connected in our minds that it is hard to disentangle them. We are currently crossing a similar threshold. Fifty years from now, nanomachines will likely be pervasive, with the boundary between the lithographic and hacked forms evermore difficult to distinguish. They will be inside us and outside of us. We will be studying their evolution and ecology. My guess is that we will have solved the riddle of the origin of life – and will have created a few more examples of life in the process. We'll have a hard time remembering that the fields of molecular biology and nanomachines were ever separate disciplines.

ENDNOTES

1 Aaron D. O'Connell et al., "Quantum Ground State and Single-Phonon Control of a Mechanical Resonator," *Nature* 464 (2010): 697.

2 Daniel G. Gibson et al., "Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome," *Science* 329 (2010): 52.

3 Michael J. Levene et al., "Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations," *Science* 299 (2003): 682.

4 Jeong-Hyun Cho et al., "3D Nanofabrication: Nanoscale Origami for 3D Optics," *Small* 7 (14) (2011): 1943.

5 Paul W.K. Rothemund, "Folding DNA to Create Nanoscale Shapes and Patterns," *Nature* 440 (2006): 297.

6 Kyle Lund et al., "Molecular Robots Guided by Prescriptive Landscapes," *Nature* 465 (2010): 207.

Can We Progress from Solipsistic Science to Frugal Innovation?

Daniel G. Nocera

Abstract: Energy demand in the twenty-first century will be driven by the needs of three billion people in the emerging world and three billion new inhabitants to our planet. To provide them with a renewable and sustainable energy supply is perhaps the greatest challenge for science in the twenty-first century. The science practiced to meet the energy needs of the twentieth century responded to a society of wealth, and energy systems were designed to be large and centralized. However, the inability of the emerging world to incur large capital costs suggests that a new science must be undertaken, one that does not rely on economy of scale but rather sets as its target highly manufacturable and distributed energy systems that are affordable to the poor. Only in this way can science provide global society with its most direct solution for a sustainable and carbon-neutral energy future.

DANIEL G. NOCERA, a Fellow of the American Academy since 2005, is the Henry Dreyfus Professor of Energy and Professor of Chemistry at the Massachusetts Institute of Technology, where he is also the Director of the Solar Revolutions Project and Director of the Eni Solar Frontiers Center. In Fall 2012, he moves to Harvard University and becomes the Patterson Rockwood Professor of Energy in the Department of Chemistry and Chemical Biology. He has conducted pioneering studies of the basic mechanisms of energy conversion in biology and chemistry, with a recent focus on the generation of solar fuels. He has published more than 325 articles and presented more than 650 invited lectures on topics relating to renewable energy.

Solipsism is "the view or theory that self is the only object of real knowledge or the only thing really existent."[1] Deriving from the Latin *sō-lus* ("alone") and *ipse* ("self"), solipsism – in its most extreme form – drives one to question whether an external world exists outside the mind. In many ways, the twentieth century was the century of solipsistic science. Science was practiced for the part of society that the scientist lived in and could observe. Because typical science is an expensive endeavor, that society was in large part one of affluence. Indeed, the knowledge and technology generated from the science of the twentieth century has served the developed world well. Any scientific or technological advance that immediately comes to mind inevitably originates from a handful of territories (for example, North America, Europe, Russia, Japan) with relatively high GDPs. But the benefits of this work have had little crossover to the poorer parts of the less-developed world.

Arguably, the inability of twentieth-century science to penetrate the underdeveloped world is rooted in cost. Research has been preoccupied with

invariably expensive targets: the "best" or the "most efficient" for the materials scientist, the "fastest" for the enzymologist, the "biggest" for the high energy physicist, or the "smallest" for the nanoscientist, to name a few. For this reason, science and technology in the last century served the needs defined by the voice of affluence. But a new voice is audible in the twenty-first century. It is a whisper now but soon will be a cacophony of overwhelming numbers. Will science respond to the needs of the underdeveloped and emerging world in the coming century, and will it do so with sufficient alacrity to address the most urgent issues affecting global society?

Nowhere has science wandered further from the world it needs to serve than in the field of energy. Today you hear about the smart grid (at least we now admit to having lived with a dumb energy system during the twentieth century); energy efficiency through materials design; engineering technologies to deliver natural gas; and grid storage. All are important science and technology targets for those with access to energy. But they have little to do with the energy needs of our future global society.

Consider the following energy equation[2]:

$$\dot{E} = N \times (GDP/N) \times (\dot{E}/GDP)$$

Here, $\dot{E}$ is energy consumption, N is the global population, GDP/N is the globally averaged GDP per capita, and $\dot{E}/GDP$ is the globally averaged energy intensity (that is, the energy consumed per unit of GDP). Carrying the numbers through the equation shows that our future global society will have an enormous appetite for energy. The rate of worldwide primary energy consumption will increase from 16.2 terawatts (TW; one TW equals one trillion watts, 1×10^{12} watts, or 1×10^{12} joules per second), as measured in 2007, to conservative estimates of 30 TW by mid-century and 45 TW by the end of the century.[3] Most, if not all, of this demand is driven by the growing world population, which is projected to increase from 6.2 billion at the beginning of this century to approximately 9.4 billion by 2050.[4] In addition to these three billion new inhabitants of the planet, three billion people in the emerging world will seek a rising standard of living. Because energy consumption scales directly with a country's GDP, energy use by developing nations will increase dramatically as they modernize.[5] Geopolitical, environmental, and economic security will likely be realized only if science in the twenty-first century can meet the energy demands of these six billion additional energy users by supplying them with a sustainable and carbon-neutral energy source.

To do so requires a different mindset from the solipsistic science of the previous century. First, the energy challenge of the twenty-first century cannot be addressed from the myopic viewpoint of a researcher in the isolated laboratory environment. The energy equation must be treated holistically. To begin, we can recast the above equation in a less mathematical form:

$$\dot{E} = (\text{society/culture}) \times (\text{economics/policy}) \times (\text{science/innovation})$$

Most energy research has emphasized the science/innovation part of this equation, with perhaps a modest nod by some in the direction of economics/policy; no approach has considered the first part of the equation. This fact is confounding, as the aspirations and needs of billions of people in emerging economies are what drive future global energy demand. If science is to have a timely and meaningful impact on society in this century, a number of questions must be considered,

including how does science: affect population growth? empower and educate the poor, especially women? become adopted within different cultures? contribute to an integrative energy and technology policy? translate within the constraints imposed by international law? affect the balance of wealth? impact human health? These are a few of the broad contextual questions that will need to be addressed as part of the society/culture and economics/policy components of the energy equation.

Second, cost must be a driver for energy research in the twenty-first century. The development of inexpensive solar energy technologies, while beneficial to the legacy world, is at odds with the needs of the nonlegacy world. In legacy nations, energy systems of the past and present operate at large scale, are centralized, and distribute energy to the masses via an expensive and complex network. Such infrastructure is not viable in the near-term future of nonlegacy states, where it is cost prohibitive to build centralized energy and distribution systems. In 2007, the total value of generation, transmission, and distribution infrastructure for regulated electric utilities in the United States was $440 billion; and capital expenditures exceeded $70 billion.[6] Reasonable recovery of capital expenditures requires designing energy systems that operate at large scale and high efficiency; consequently, energy systems of the legacy world come with significant balance-of-system (BOS) costs. Downscaling such technology for use in nonlegacy regions is not economically viable because the BOS costs do not scale commensurately. Thus, existing off-the-shelf technologies will be difficult to adapt to low cost.

Rather, in order to be adaptable in the nonlegacy world, the disruptive energy technologies of the twenty-first century will necessarily be *light* and *highly manufacturable* as well as *robust* and *low maintenance*. Simply put, new research is needed to develop what I have called the "fast food" equivalent of energy systems[7]; hence, the need for *frugal innovation* is at the fore.[8] Science that targets frugal innovation provides a win-win situation for the legacy world as well. Established energy markets are slow to turn over because they require significant capital investment. Thus, the quickest path to market penetration in the legacy world is one of low cost. To meet the objectives of renewable, low-cost, and highly manufacturable energy systems, science in the twenty-first century will develop a practical and realistic energy infrastructure for both nonlegacy *and* legacy worlds.

Third, the shift to frugal innovation sets a different design target for the research scientist. The most impactful scientific discoveries in the twenty-first century will involve working backward from a technology target. To this end, the research scientist must consider systems engineering from the outset. Discovery must cast an eye toward implementation under simply engineered conditions so that BOS costs remain low. Yet this objective is largely absent from current scientific practice. Consider the materials science of photovoltaics, that is, the process of converting sunlight into electricity. How often does one hear that solar panels are expensive and that there is a need for science to create a cheaper photovoltaic material? But analysis of the numbers suggests a disconnect in the logic of this statement. The cost of generating photovoltaic electricity from known semiconducting materials is plummeting; moreover, the cost of the photovoltaic semiconductor is less than 10 percent that of the module (the solar panel), its installation, and its maintenance combined.[9]

This does not mean that the search for new photovoltaic materials should cease. Indeed, it should continue, but not neces-

sarily with the target of a cheaper material. Rather, new photovoltaic materials must be discovered that would eliminate the fabrication process of the photovoltaic module as we now know it – thus yielding much more significant cost reductions. This is a different imperative for the researcher who wishes to make an impact in photovoltaics, especially one who wishes to help the poor. If scientists in the twentieth century expressed a desire to "make a cheaper photovoltaic material for solar modules," then scientists in this century must instead strive to "make a photovoltaic material that allows solar modules to be made more cheaply." This is only one example of many that could be enumerated. Unfortunately, most researchers do not understand such distinctions and consider them to be nuance. But in the twenty-first century, scientists must consider the "systems engineering" aspect of the technologies that their discoveries will target.

What are some of the research targets for energy science in the twenty-first century?

(1) *Solar energy*. Delivering an additional 16 TW to our world by 2050 is not a simple task. As has now been documented extensively, most energy sources are insufficient to keep pace with the growing global appetite for energy.[10] Conventional biomass is a limited energy source owing to the low energy efficiency of photosynthesis.[11] Nuclear energy requires a large number of sites that will be difficult to build fast enough to keep up with energy demand.[12] Moreover, a nuclear-based energy supply will require widespread public acceptance.[13] Finally, tides, wind, and other natural forces have too low an energy density to satiate demand.[14] These shortcomings, however, do not mean that continued research into such carbon-neutral energy supplies should be abandoned. If these resources are available, then they should be exploited. Indeed, they will be utilized by industry because the fundamental science for many of them has largely been developed.

But the fundamental research scientist should be at the frontier where industry is too nervous to venture. For this reason, and because of its enormous potential, solar energy will be the preeminent carbon-neutral energy source for research in the twenty-first century. Terrestrial solar insolation – that is, the solar radiation that reaches Earth's surface – exceeds the resource base of all other renewable energy sources combined. Additionally, it far exceeds what is necessary to support even the most technologically advanced society. The ability of solar radiation to meet future global energy demand is well documented.[15] But it needs to be developed strategically, with an emphasis on the nonlegacy world. An especially attractive approach is the idea of *personalized energy*, which would replace the centralized energy system of the twentieth century in much the same way that personal computers replaced the mainframes of the 1970s.[16]

Because energy use scales with wealth, point-of-use solar energy will put individuals, in the smallest village in the nonlegacy world and in the largest city of the legacy world, on a more level playing field. Moreover, personalized energy is secure because it is highly distributed and gives individuals control over the energy they rely on. And personalized energy can reach the six billion new energy users of this century via high-throughput manufacturing of distributed energy systems. However, major challenges confront the deployment of personalized solar energy on a large and distributed scale. The imperative to science is to develop new materials, reactions, and processes that enable solar energy to be sufficiently inexpensive

to penetrate global energy markets. Most, if not all, of these materials and processes entail a metallic or noncarbon-based main group element. Accordingly, the subject of inorganic chemistry will be especially germane to delivering personalized solar energy to our planet.

As discussed above, solar modules are indeed too expensive for the poor. New photovoltaic materials need to be developed that permit solar-to-current conversion with lower-cost modules, or with entirely new design approaches, such as nanoparticle-based systems, thus eliminating the BOS costs that plague current photovoltaic technology. Because society relies on a continuous energy supply and solar energy is diurnal as well as subject to intermittency arising from variable atmospheric conditions, an inexpensive storage mechanism is needed for solar energy to be a truly useful contributor to the primary energy supply. Unfortunately, most current methods of solar storage, including batteries, are characterized by energy densities that are too low for large-scale solar storage.[17] Conversely, the energy density of fuels is one hundred to one thousand times larger than that of any other storage method. Indeed, society has intuitively understood this disparity in energy density; all large-scale energy storage developed over the last century is in the form of fuels. But these fuels are carbon-based. The challenge for the discipline of chemistry (via new catalysis), and for science more generally, is to develop solar-to-fuels storage processes, such as creation of the artificial leaf,[18] that are low cost and adaptable to a distributed energy infrastructure.

Finally, although biomass is certainly a distributed solar-to-fuels energy source, any natural photosynthetic process is plagued by low efficiency (a theoretical maximal thermodynamic power conversion efficiency of roughly 10 percent, with the best-growing plants never exceeding 1 to 3 percent and algae at about 5 percent).[19] As is true of most living organisms, the plant or alga needs to use its stored energy for its cellular growth and maintenance. Synthetic biology provides exciting opportunities to radically redesign the photosynthetic apparatus to substantially improve the efficiency of natural photosynthesis.

(2) *Energy efficiency*. When industrialized society developed in the twentieth century, energy was not at a premium. Yet a tidal shift in energy costs is already apparent at the beginning of the twenty-first century. Industrial titans have challenged their employees to maintain production with 30, 50, and even 60 percent reductions in energy use in the coming decade. This challenge will require more energy-efficient materials and processes. Some avenues for research are obvious: for instance, the creation of better thermoelectric materials, which are able to convert temperature differences to electricity and vice versa.[20] But there are grander issues to tackle with regard to energy efficiency, and these will require a more profound change in the way we approach our professions.

Nowhere is this need for change better exemplified than in the petrochemical industry. Many of the products we use in modern society are derived from petroleum feedstocks. Plastics are exemplars. Very long chain hydrocarbons present in petroleum are broken down or "cracked" into C_2 and C_3 subunits (ethylene and propylene). An exquisite science in the twentieth century was developed to stitch the C_2 and C_3 subunits back together to furnish a variety of polymers with a desired property. Industry could pursue this approach because energy and petrochemicals were cheap. Nevertheless, the process is the model of inefficiency: a significant amount of energy is needed to

break the carbon-carbon bond, and still more energy is needed to put the carbon atoms back together into the long chains that compose the plastic. As energy and petrochemicals become more expensive, the cost of creating materials derived from them will rise. Thus, a new organic chemistry must develop around different feedstocks. This will require the creation of an enormous amount of fundamental new science.

As a case in point, despite amazing advances in synthetic methodologies that led to the birth of the pharmaceutical industry in the twentieth century, the organic chemist of today is hard-pressed to take the simplest of hydrocarbons (for example, hexane), activate their CH bonds, and join them with a defined connectivity. This is only one example of manufacturing inefficiency, which is the rule rather than the exception. Why are such inefficient practices so common? The answer is that energy in the twentieth century was simply too cheap. The twenty-first century will require a seismic shift in the way organic chemistry is executed. Thus, entire industries will have to rely on basic science in order to reinvent themselves in an energy-deprived world.

(3) *Materials sustainability*. Energy is the first resource to leak through the cracks in the dam of sustainability. The energy challenge of the twenty-first century is in large part due to overpopulation of our planet. And with population growth on the rise, many more cracks will appear in the sustainability dam – and a host of new challenges will soon follow. The criticality of water resources is coming into focus,[21] with food not far behind.[22] Yet there are more subtle materials sustainability challenges, too. How many realize that the availability of elements such as phosphorus will be a major problem in the twenty-first century? This biocritical element is in short supply because of geological and terrestrial variations in the phosphorus cycle, such as changes caused by the erosion that results from agriculture and human activity.[23] The net transfer of dissolved phosphorus from land to the oceans is 4 to 6 teragrams per year, which represents a doubling of prehuman input fluxes. Considering the importance of phosphorus in life cycles and fertilization, new chemistries must be developed for phosphorus recovery. In short, changes in Earth systems will drive a need for new basic science focused on element and materials recovery.

Some think that supplying energy to the poor will only exacerbate the sustainability problem. With more energy will come more demand. But we need to return to the energy equation. When people have access to energy, they are able to increase wealth. Study after study has shown that when people are empowered, they seek access to education. And education leads to declining birth rates. So by undertaking a science to provide energy to the poor, we establish a positive feedback loop that most directly addresses the sustainability issue.

In doing so, we avoid the possible future Kurt Vonnegut depicted shortly before his death. In words both chilling and consoling, Vonnegut described the planet as a living organism.[24] He reminded us that when an organism is sufficiently compromised, its immune system responds by eliminating irksome intruders. Viewing humans as the irksome intruder on our planet-organism, Vonnegut assured us that we need not worry about the planet. When we become sufficiently intolerable, the planet's immune system will respond, eliminating humans by not sustaining us in the dramatically altered environment that we have created.

Earth will continue to exist and flourish at high carbon dioxide levels and with a radically different environment, though

not as we know it. It is the human species, on the other hand, that is in a precarious state. When confronted about a solution, Vonnegut responded that the concerned among us should "get a gang" and do something about it. The "gang" of scientists who take up Vonnegut's call to arms must free themselves from the bonds of solipsism. They must consider all components of the energy equation in practicing their craft, leaving behind a withered brand of twentieth-century practice – solipsistic science – and embrace a new brand of science for the twenty-first century: frugal innovation. In this way, they will provide the technology needed to answer the greatest challenges confronting humanity in the twenty-first century, not the least of which will be to secure a renewable and sustainable energy supply for the nonlegacy world.

ENDNOTES

1 "Solipsism," in *The Oxford English Dictionary*, 2nd ed. (Oxford: Clarendon Press, 1989); online version updated March 2012, http://www.oed.com/view/Entry/184295.

2 Martin I. Hoffert, Ken Caldeira, Atul K. Jain, Erik F. Haites, L.D. Danny Harvey, Seth D. Potter, Michael E. Schlesinger, Stephen H. Schneider, Robert G. Watts, Tom M.L. Wigley, and Donald J. Wuebbles, "Energy Implications of Future Stabilization of Atmospheric CO_2 Content," *Nature* 395 (6705) (October 29, 1998): 881 – 884.

3 Nathan S. Lewis and Daniel G. Nocera, "Powering the Planet: Chemical Challenges in Solar Energy Utilization," *Proceedings of the National Academy of Sciences* 103 (43) (October 24, 2006): 15729 – 15735.

4 2009 *World Population Data Sheet* (Washington, D.C.: Population Reference Bureau, 2009), http://www.prb.org.

5 Daniel G. Nocera, "On the Future of Global Energy," *Dædalus* 135 (4) (Fall 2006): 112 – 115.

6 Marc Chupka and Gregory Basheda, "Rising Utility Construction Costs: Sources and Impacts," The Brattle Group (Washington, D.C.: The Edison Foundation, September 2007).

7 Daniel G. Nocera, "Fast Food Energy," *Energy & Environmental Science* 3 (8) (September 2010): 993 – 995.

8 Ratan Tata, "Out of India," *The Economist*, March 3, 2011.

9 Doug M. Powell, Mark T. Winkler, Hyunjoo Choi, Christie B. Simmons, David Berney Needleman, and Tonio Buonassisi, "Crystalline Silicon Photovoltaics: A Cost Analysis Framework for Determining Technology Pathways to Reach Baseload Electricity Costs," *Energy & Environmental Science* 5 (3) (March 2012): 5874 – 5883.

10 Lewis and Nocera, "Powering the Planet"; Derek Abbott, "Keeping the Energy Debate Clean: How Do We Supply the World's Energy Needs?" *Proceedings of the IEEE* 98 (1) (January 2010): 42 – 66; Richard E. Smalley, "Future Global Energy Prosperity: The Terawatt Challenge," *Materials Research Society Bulletin* 30 (6) (June 2005): 412 – 417.

11 Robert E. Blankenship, David M. Tiede, James Barber, Gary W. Brudvig, Graham Fleming, Maria Ghirardi, M. R. Gunner, Wolfgang Junge, David M. Kramer, Anastasios Melis, Thomas A. Moore, Christopher C. Moser, Daniel G. Nocera, Arthur J. Nozik, Donald R. Ort, William W. Parson, Roger C. Prince, and Richard T. Sayre, "Comparing Photosynthetic and Photovoltaic Efficiencies and Recognizing the Potential for Improvement," *Science* 332 (6031) (May 13, 2011): 805 – 809.

12 Derek Abbott, "Is Nuclear Power Globally Scalable?" *Proceedings of the IEEE* 99 (10) (October 2010): 1611 – 1617.

[13] Stephen Ansolabehere, John Deutch, Michael Driscoll, Paul E. Gray, John P. Holdren, Paul L. Joskow, Richard K. Lester, Ernest J. Moniz, and Neil E. Todreas, *The Future of Nuclear Power: An Interdisciplinary MIT Study* (Cambridge, Mass.: MIT Press, 2003).

[14] Abbott, "Keeping the Energy Debate Clean"; Abbott, "Is Nuclear Power Globally Scalable?"

[15] Lewis and Nocera, "Powering the Planet"; James Barber, "Photosynthetic Energy Conversion: Natural and Artificial," *Chemical Society Reviews* 38 (1) (January 2009): 185–196.

[16] Daniel G. Nocera, "Personalized Energy: The Home as a Solar Power Station and Solar Gas Station," *ChemSusChem* 2 (5) (May 25, 2009): 387–390; Daniel G. Nocera, "Chemistry of Personalized Solar Energy," *Inorganic Chemistry* 48 (21) (November 2, 2009): 10001–10017.

[17] Timothy R. Cook, Dilek K. Dogutan, Steven Y. Reece, Yogesh Surendranath, Thomas S. Teets, and Daniel G. Nocera, "Solar Energy Supply and Storage for the Legacy and Nonlegacy World," *Chemical Reviews* 110 (11) (November 2010): 6474–6502.

[18] Steven Y. Reece, Jonathan A. Hamel, Kimberly Sung, Thomas D. Jarvi, Arthur J. Esswein, Joep J.H. Pijpers, and Daniel G. Nocera, "Wireless Solar Water Splitting Using Silicon-Based Semiconductors and Earth Abundant Catalysts," *Science* 334 (6056) (November 4, 2011): 645–648; Daniel G. Nocera, "The Artificial Leaf," *Accounts of Chemical Research* 45 (April 4, 2012): 767–776.

[19] Blankenship et al., "Comparing Photosynthetic and Photovoltaic Efficiencies and Recognizing the Potential for Improvement."

[20] Jeffrey G. Snyder and Eric S. Toberer, "Complex Thermoelectric Materials," *Nature Materials* 7 (2) (February 2008): 105–114.

[21] Charles J. Vörösmarty, Pamela Green, Joseph Salisbury, and Richard B. Lammers, "Global Water Resources: Vulnerability from Climate Change and Population Growth," *Science* 289 (5477) (July 14, 2000): 284–288.

[22] H. Charles J. Godfray, John R. Beddington, Ian R. Crute, Lawrence Haddad, David Lawrence, James F. Muir, Jules Pretty, Sherman Robinson, Sandy M. Thomas, and Camilla Toulmin, "Food Security: The Challenge of Feeding 9 Billion People," *Science* 327 (5967) (February 12, 2010): 812–818.

[23] Gabriel M. Filippelli, "The Global Phosphorus Cycle: Past, Present, and Future," *Elements* 4 (2) (April 2008): 89–95.

[24] An interview with Kurt Vonnegut, by David Brancaccio, *PBS NOW*, October 7, 2005.

The Future of Fundamental Physics

Nima Arkani-Hamed

Abstract: Fundamental physics began the twentieth century with the twin revolutions of relativity and quantum mechanics, and much of the second half of the century was devoted to the construction of a theoretical structure unifying these radical ideas. But this foundation has also led us to a number of paradoxes in our understanding of nature. Attempts to make sense of quantum mechanics and gravity at the smallest distance scales lead inexorably to the conclusion that space-time is an approximate notion that must emerge from more primitive building blocks. Furthermore, violent short-distance quantum fluctuations in the vacuum seem to make the existence of a macroscopic world wildly implausible, and yet we live comfortably in a huge universe. What, if anything, tames these fluctuations? Why is there a macroscopic universe? These are two of the central theoretical challenges of fundamental physics in the twenty-first century. In this essay, I describe the circle of ideas surrounding these questions, as well as some of the theoretical and experimental fronts on which they are being attacked.

NIMA ARKANI-HAMED, a Fellow of the American Academy since 2009, is a Professor in the School of Natural Sciences at the Institute for Advanced Study. His interests range from quantum field theory and string theory to cosmology and collider physics. He has published his work in the *Journal of High Energy Physics*, the *Journal of Cosmology and Astroparticle Physics*, and *Nuclear Physics*, among other places.

Ever since Newton realized that the same force of gravity pulling down on an apple is also responsible for keeping the moon orbiting the Earth, fundamental physics has been driven by the program of *unification*: the realization that seemingly disparate phenomena are in fact different aspects of the same underlying cause. By the mid-1800s, electricity and magnetism were seen as different aspects of electromagnetism, and a seemingly unrelated phenomenon – light – was understood to be the undulation of electric and magnetic fields.

Relativity and quantum mechanics pushed the trend toward unification into territory far removed from ordinary human experience. Einstein taught us that space and time are different aspects of a single entity: space-time. Energy and momentum are united analogously, leading to the famous equivalence between mass and energy, $E = mc^2$, as an immediate consequence. Einstein further realized that space-time is not a static stage on which physics unfolds, but a dynamic entity that can curve and bend. Gravity is understood as a manifestation of

space-time curvature. This new picture of space-time made it possible to conceive of ideas that were impossible to articulate in the Newtonian picture of the world. Consider the most important fact about cosmology: we live in an expanding universe. The distance between two galaxies grows with time. But the galaxies are not rushing apart from each other into some preexisting space, as though blown out of an explosion from some common center. Rather, more and more space is being generated between the galaxies all the time, so from the vantage point of any one galaxy, the others appear to be rushing away. This picture, impossible to imagine in Newton's universe, is an inevitable consequence of Einstein's theory.

Quantum mechanics represented a more radical departure from classical physics, involving a completely new conceptual framework, both physically and mathematically. We learned that nature is not deterministic, and only probabilities can be predicted. One consequence is the famous uncertainty principle, by which we cannot simultaneously know the position and velocity of a particle to perfect accuracy. Quantum mechanics also allowed previously irreconcilable phenomena to be understood in a unified way: particles and waves came to be seen as limiting aspects of the underlying description where there are no waves at all, only quantum-mechanical particles.

The laws of relativity and quantum mechanics are the pillars of our current understanding of nature. However, describing physics in a way that is compatible with both of these principles turns out to be extremely challenging; indeed, it is possible only with an extremely constrained theoretical structure, known as *quantum field theory*. A quantum field theory is characterized by a menu of particles that interact with each other in various ways. The nature of the interactions is almost completely dictated by the rules of quantum mechanics, together with the requirement that the interactions take place at points in space-time, in compliance with the laws of special relativity. The latter requirement is known as the principle of *locality*.

One of the startling general predictions of quantum field theory is the existence of anti-particles such as the positron, which has the same properties as the electron but the opposite electric charge. This prediction has another striking consequence: namely, that even the vacuum has structure and dynamics.

Suppose we attempt to check that some small region of space-time is empty. Because of the uncertainty principle, we need higher energies to probe short distances. Eventually there is enough energy to make an electron and a positron, without violating either the conservation of energy or the conservation of charge. Instead of seeing nothing, probing the vacuum at small distances yields particle/anti-particle pairs. It is useful to think of the vacuum as filled with quantum fluctuations, with "virtual" particles and anti-particles popping in and out of existence on faster and faster timescales at shorter and shorter distances.

These quantum fluctuations give rise to measurable physical effects. For instance, the cloud of virtual electrons and positrons surrounding an electron is slightly perturbed by the electron's electric field. Any physical measurement of the electron's charge, then, will vary just slightly with distance, growing slowly closer in to the electron as more of the virtual cloud is pierced. These virtual effects can be calculated very precisely; in some circumstances, theoretical predictions and experimental observations can be compared to an astonishing level of precision. The virtual corrections to the magnetic proper-

ties of the electron, for example, have been theoretically computed to twelve decimal places, and they agree with experiment to that level of precision.

The second-half of the twentieth century saw a flurry of activity, on both experimental and theoretical fronts. These developments culminated in the 1970s with the construction of the Standard Model of particle physics, a specific quantum field theory that describes all known elementary particles and their interactions down to the smallest distances we have probed so far. There are four basic interactions: gravity and electromagnetism, which were familiar even to the ancients, as well as the weak and strong interactions that reveal themselves only on nuclear scales. Atomic nuclei consist of neutrons and protons. An isolated neutron is unstable, living for about fifteen minutes before disintegrating into a proton, electron, and an anti-neutrino. (This process is also responsible for radioactivity.) Fifteen minutes is enormously long compared to the typical timescales of atoms and nuclei, so the interaction responsible for triggering this decay must be very feeble – hence, *weak* interaction. The earliest incarnation of the strong interaction was noticed in the attraction keeping protons inside nuclei, counterbalancing their huge electrical repulsion.

Some familiar particles, such as electrons and photons, remain as elementary point-like entities in the Standard Model. Others, like the proton, are understood to be bound states, around 10^{-14} cm in diameter made of *quarks*, which are permanently trapped inside the proton through their interaction with *gluons*.

Strong, weak, and electromagnetic interactions seem completely different from each other at long distances, but we now know that these differences are a long-distance illusion. At short scales, these interactions are described in essentially the same quantum-field-theoretic language. Electromagnetism is associated with interactions between electrons and photons of a specific sort. Strong interactions arise from essentially identical interactions between quarks and gluons, while weak interactions connect particles like the electron and the neutrino in the same way, with massive cousins of the photon known as the W and Z particles.

Differences appear at long distances for subtle reasons. The electromagnetic interaction was the first to be detected and understood because the photon is massless and the interaction is long-ranged. The W and Z particles are massive, thus mediating an interaction with a short range of about 10^{-17} cm. The difference with quarks and gluons is more subtle still: the virtual effects of the cloud of gluons surrounding a quark make the "strong charge" of quarks slowly grow stronger at longer distances. At a distance of roughly 10^{-14} cm, the interaction is so strong as to permanently confine quarks inside protons and neutrons.

But from a fundamental short-distance perspective, these are details: the character of the laws is essentially identical. This fact illustrates the central reason why we probe short distances in fundamental physics. It is not so much because we care about the "building blocks of matter" and the associated set of particles we may discover, but because we have learned that the essential unity, simplicity, and beauty of the underlying laws manifest most clearly at short distances.

The Standard Model is one of the triumphs of physics in the twentieth century. It gives us a simple and quantitatively accurate description of everything we know about elementary particles and their interactions. Only one element of the theory has yet to be definitely confirmed by experiment. In the fundamental short-distance theory, where all the interactions are treat-

ed on a symmetrical footing, the particles are massless. The mass of particles, such as electrons or the *W* and *Z* particles, arises as a dynamic long-distance effect, known as the Higgs mechanism because of the particles' interactions with the so-called Higgs field. The typical length scale associated with these interactions is around 10^{-17} cm, which is, not coincidentally, also the range of weak interactions. As I discuss at greater length below, it is also fortuitously the distance scale we are now probing with the Large Hadron Collider (LHC), the particle accelerator located at the CERN laboratory just outside Geneva, Switzerland. Collisions at the LHC should put ripples in the Higgs field that manifest as the Higgs particle with very definite properties and experimental signatures. Indeed, last December, the LHC experiments reported preliminary evidence for events consistent with the production of the Higgs particle, with its expected properties. Analysis of the 2012 data should either yield a definitive discovery of the Higgs particle or definitively exclude the simplest realization of the Higgs mechanism within the Standard Model.

The success of the Standard Model gives us a strong indication that we are headed in the right direction in our understanding of fundamental physics. Yet profound mysteries remain, associated with questions that either lie outside the scope of the Standard Model or are addressed by it, but in a seemingly absurd way. Two of these questions stand out for both their simplicity and urgency, and will drive the development of fundamental physics in the twenty-first century.

The principle of locality – the notion that interactions take place at points in spacetime – is one of the two pillars of quantum field theory. It is therefore unsettling to realize that, due to the effects of both gravity and quantum mechanics, spacetime is necessarily an approximate notion that must emerge from more primitive building blocks.

Because of the uncertainty principle, we have to use high energies to probe short distances. In a world without gravity, we could resolve arbitrarily small distances in this way, but gravity eventually and dramatically changes the picture. At miniscule distances, so much energy has to be concentrated into such a tiny region of space that the region itself collapses into a black hole, making it impossible to extract any information from the experiment. This occurs when we attempt to probe distances around 10^{-33} cm, the so-called Planck length.

The Planck length is a ridiculously tiny distance scale – sixteen orders of magnitude smaller than the tiniest distances we are probing today at the LHC. Its tininess is a direct reflection of the extreme weakness of gravity compared to other forces of nature. The gravitational attraction between a pair of electrons is forty-two orders of magnitude smaller than their electrical repulsion. Classically, both the gravitational and electric forces vary with distance following an inverse-square law; however, at a distance of around 10^{-11} cm, this gets corrected in an important way: again because of the uncertainty principle, simply holding two electrons at shorter distances requires a huge amount of energy. The force of gravity increases with increasing mass, or with equivalently increasing energy, so the attraction between electrons begins to increase relative to the electrical repulsion. At around 10^{-31} cm, gravity surpasses the electric force, and at 10^{-33} cm, it dominates all interactions.

Thus, the combination of gravity and quantum mechanics makes it impossible to operationally probe Planckian distances. Every time we have encountered ideas in physics that cannot even in principle be observed, we have come to see such

ideas as approximate notions. However, this instance is particularly disturbing because the notion that emerges as approximate is that of space-time itself.

The description of the situation seems to relegate all the mysteries to tiny distances, and may suggest some sort of granular structure to space-time near the Planck scale. Much as the smooth surface of a table is resolved into discrete units made of molecules and atoms, one might imagine that "atoms of space-time" will replace space-time near the Planck length. This naive idea is very likely wrong. Any sort of granular structure to space-time picks a preferred frame of reference, where the size of the granularity is "small," in sharp conflict with the laws of relativity. But there is a deeper reason to suspect that something much more interesting and subtle than "atoms of space-time" is at play. The problems with space-time are not only localized to small distances; in a precise sense, "inside" regions of space-time cannot appear in any fundamental description of physics at all.

The slogan is that due to quantum mechanics and gravity, there are no "local observables." Indeed, before worrying about what a correct theory combining quantum mechanics and gravity ought to look like, it is worth thinking about what perfectly precise measurements can ever be made by experiments. These (in principle) exact observables provide a target for what the theory should predict.

Imagine trying to perform any sort of local measurement, by which I mean an experiment that can be done in a finite-sized room. To extract a perfectly precise measurement, we need (among other things) to use an infinitely large apparatus in order to avoid inaccuracies arising from the quantum fluctuations of the apparatus. If the apparatus has a large but finite number of components, on a huge but finite timescale, it suffers its own quantum fluctuations, and therefore cannot record the results of the experiment with perfect accuracy. Without gravity, nothing would stop us from conducting the experiment with an infinitely big apparatus to achieve perfect accuracy, but gravity obstructs this. As the apparatus gets bigger, it inevitably also gets heavier. If we are making a local measurement in a finite-sized room, at some large but finite size it becomes so heavy that it collapses the entire room into a black hole.

This means that there is no way, not even in principle, to make perfectly accurate local measurements, and thus local observables cannot have a precise meaning. There is an irreducible error associated with any local measurement that is made in a finite room. While this error is significant close to the Planck scale, it is negligible in ordinary circumstances. But this does not diminish the importance of this observation. The fact that quantum mechanics makes it impossible to determine precisely the position and velocity of a baseball is also irrelevant to a baseball player. However, it is of fundamental importance to physics that we cannot speak precisely of position *and* momentum, but only position *or* momentum. Similarly, the fact that gravity makes it impossible to have precise local observables has the dramatic consequence that the "inside" of any region of space-time does not have a sharp meaning, and is likely an approximate notion that cannot appear in a deeper underlying theory.

If we cannot speak precisely of local observables, what observables can we talk about? Instead of performing observations inside some region of space-time, we can push our detectors out to infinite distances, at the boundary of space-time, where we can make them infinitely big. We can then throw particles into the interior, where they interact and scatter with

each other in some way and emerge back out to infinity where they are measured. The results of these scattering experiments can be the perfectly precise observables that one might hope to calculate from a fundamental underlying theory.

String theory is our best attempt to make sense of the mysteries of quantum gravity, and it perfectly exemplifies this basic ideology. In its earliest incarnation, string theory computed the results of scattering processes and was thought of as a generalization of quantum field theory, with point-like particles replaced by extended loops of string. This idea miraculously passed several physical and mathematical consistency checks and spawned a huge amount of theoretical activity. The 1990s brought a steady stream of surprises revealing that string theory is not in fact a theory of strings, but contains both point-like particles as well as higher-dimensional objects as important ingredients.

By the late 1990s, these developments led to an amazing realization, widely considered to be the most important theoretical advance in the field in the past two decades. Early work in string theory focused on understanding scattering processes in flat space-time, where time marches uniformly from the infinite past to the infinite future and where space is not curved. But it is also possible to consider a different kind of geometry on very large scales, known as anti-de Sitter space. Here, time still marches uniformly from the infinite past to the infinite future, but space is curved. While the distance from a point on the interior to the boundary of space is infinite, due to the curvature, a light beam takes a finite amount of time to make it to the boundary. Thus, this geometry can be usefully thought of as the inside of a box.

There is a rich set of observables that we can talk about in this geometry: starting on the walls, we can throw particles into the interior of the box and watch them come back out to the walls at some finite time in the future. Because these experiments start and end on the walls, it is natural to wonder whether there is a way of describing the physics where the interior of the box makes no appearance.

Amazingly, such a description exists, and is given in terms of a completely ordinary quantum field theory living on the walls of the box, made from particles very much like the quarks and gluons of the strong interactions. When the interactions between the "gluons" are made very strong, the physics is completely equivalent to that of string theory living on the inside of the box. In a specific sense, gravity, strings, and an extra direction of space emerge from the strong interactions of a perfectly ordinary quantum field theory in one lower dimension, much like an apparently three-dimensional image can be encoded in a two-dimensional hologram.

At first sight, this holographic equivalence seems impossible. If we had a ball in the middle of the box, how could its position in the interior be encoded only on the walls? The presence of the ball in the interior is represented as some lump of energy in the description on the walls; as the ball moves around the interior, this lump correspondingly shrinks and grows in size. What about the force of gravity between two balls in the interior? The two corresponding lumps of energy modify the virtual cloud of gluons surrounding them, which in turn induces a net attraction between the lumps, precisely reproducing the correct gravitational force. In every physical sense, gravity and the extra direction of space making up the inside of the box do indeed emerge "holographically," from the dynamics of the theory that lives fundamentally on the walls. This correspondence gives us our first concrete clue as to how space-time may emerge from more primitive building blocks.

For the past hundred years, physics has been telling us that there are fewer and fewer observables we can talk about meaningfully. The transition from classical to quantum physics was the most dramatic in this regard: the infinite number of observables in a deterministic universe was reduced to merely computing probabilities. But this loss came with a silver lining: if there are fewer fundamental observables, seemingly disparate phenomena must be more closely related and unified than they appear to be. In this case, the loss of determinism was directly responsible for understanding waves and particles in a unified way. Adding gravity to the mix further eliminates all local observables and pushes the meaningful questions to the boundary of space-time, but this is also what allows gravity and quantum field theory to be holographically equivalent to each other. It is gratifying to see that all the major themes of theoretical physics over the past four decades, in quantum field theory and string theory, have been exploring different aspects of a single underlying structure. But can this theoretical discovery be applied to understanding quantum gravity in the real world? The box in which the gravitational theory lives can be arbitrarily large; indeed, if we did not know about cosmology, we might easily imagine that our universe is a box of this sort, with a size of about ten billion light years. Any questions about gravity and quantum mechanics on shorter scales, from the size of galaxies down to the Planck length, can be asked equally well in this toy box as in our own universe.

But a number of conceptual challenges must be overcome to describe the universe we actually live in, and most of them have to do with a deeper understanding of time. Indeed, the major difference between our universe and the "gravity in a box" toy model we have understood so well is that we do not live in a static universe. Our universe is expanding. Looking back in time, we eventually encounter Planckian space-time curvatures near the "big bang," where all our descriptions of physics break down along with the notion of time itself.

An equally profound set of questions is associated with understanding the universe at late times. Perhaps the most important experimental finding in fundamental physics in the past twenty years has been the discovery that the universe's expansion rate is accelerating and that the universe is growing exponentially, doubling in size every ten billion years or so. Due to this exponential growth, light from regions of space more than ten billion light years away will never make it to us: the finite part of the universe we now see is all we will ever have access to. This simple observation has huge implications. As discussed above, precise observables require a separation of the world into a) an infinitely large measuring apparatus and b) the system being studied. In our accelerating universe, with access to only a finite (though enormous) amount of material, it is impossible to make an infinitely large apparatus. Thus, we appear to have no precise observables to talk about. So what sort of fundamental theory should we be looking for to describe this situation? This is perhaps the deepest conceptual problem we face in physics today. Any progress on this question must involve some essentially new insight into the nature of time.

Having scaled these dizzyingly abstract heights, let us come back down to Earth and ask another set of far simpler seeming questions. One of the most obvious and important properties of the universe is that it is enormous compared to the tiny distance scales of fundamental physics, from atoms and nuclei all the way down to the Planck length. This big universe is also filled with interesting objects that are much larger than atoms. Why is there a macro-

scopic universe when the basic constituents of matter and all the fundamental distance scales are microscopic?

This question does not at first seem particularly profound: things are big because they are composed of a huge number of atoms. But this is not the whole story. In fact, things are big as a direct consequence of the extreme weakness of gravity relative to other forces in nature. Why is the Earth big? Its size is determined by competition between an attractive gravitational pressure that is counterbalanced by atomic pressures; planets can be so big precisely because gravity is an extremely weak force. Stars are big for a similar reason. If the Planck length were comparable to the scales of atomic and nuclear physics, gravity would be a vastly stronger force, and our planets and stars would all be crushed into black holes. Thus, instead of asking why there is a macroscopic universe, we could ask: why is Planck length so much smaller than all the other scales in physics?

This turns out to be a very deep question. One might think that the scales simply are what they are, and can easily be arranged to be vastly different from each other. But this is not the case. Huge quantum fluctuations near the Planck length seem to make it impossible for macroscopic phenomena to be coherent on larger distance scales.

The most dramatic puzzle arises from the energy carried by quantum fluctuations. Fluctuations in a box of Planckian size should carry Planckian energy, leading us to expect that the vacuum will have some energy density. This vacuum energy density is known as the *cosmological constant*, and we have estimated that it should be set by the Planck scale. Like all other forms of matter and energy, the vacuum energy curves space-time; if the cosmological constant is Planckian, the curvatures should also be Planckian, leading to the absurd conclusion that the universe should be crumpled up near 10^{-33} cm, or should be expanding at an explosive rate, doubling in size every 10^{-43} seconds. Obviously, this looks nothing like the universe we live in. As already mentioned, the expansion rate of our universe is in fact accelerating, but the universe is doubling in size every ten billion years or so. The simplest explanation for this acceleration is a small positive cosmological constant, with a size 120 orders of magnitude smaller than our Planckian estimate. This is the largest disagreement between a "back of the envelope" estimate and reality in the history of physics – all the more disturbing in a subject accustomed to twelve-decimal-place agreements between theory and experiment.

Before addressing more sophisticated questions, our description of nature given by the Standard Model must deal with the extremely basic question of why the universe is big. We have found a huge contribution to the cosmological constant from quantum fluctuations, but there can also be a purely classical part of the cosmological constant, whose size just so happens to delicately cancel the contributions from quantum fluctuations, to an accuracy of 120 decimal places. This is a deeply unsatisfying explanation, and for obvious reasons is referred to as *unnatural fine-tuning* of the parameters of the theory. The fine-tuning needed to understand why we have a big universe is known as the *cosmological constant problem*.

There is an analogous puzzle known as the *hierarchy problem*, related to the question of why atomic scales are so much larger than the Planck length. The relatively large size of the atom is a consequence of the small mass of the electron. As briefly reviewed above, an electron acquires its mass from bumping into the Higgs field, with a typical interaction length near 10^{-17} cm. But the Higgs field

itself should have enormous quantum fluctuations growing stronger toward the Planck scale, and so the typical length scale of its interactions with an electron should be closer to 10^{-33} cm. This outcome would make electrons sixteen orders of magnitude heavier than they are observed to be. To avoid this conclusion, we have to invoke another unnatural fine-tuning in the parameters of the theory, this time to an accuracy of one part in 10^{30}.

Unlike the difficulties with the ideas of space-time near the Planck length, these so-called naturalness problems do not represent a direct breakdown of our understanding of the laws of nature. But the extremely delicate adjustment of parameters needed to answer such basic questions seems incredibly implausible, suggesting that we are missing crucial new physical principles to provide a more satisfying explanation for why we have a macroscopic universe. It is as though we see a pencil standing on its tip in the middle of a table. While this scenario is not impossible, if we were confronted with this sight we would seek an explanation, looking for some mechanism that stabilizes the pencil and prevents it from falling over. For instance, we might look to see if the pencil is secretly hanging from a string attached to the ceiling.

The most obvious resolution to these fine-tuning problems would be to find an extension of the Standard Model that somehow removes large vacuum fluctuations. Because these fluctuations are an intrinsic feature of the unification of quantum mechanics and space-time, it stands to reason that any mechanism for removing them must change one of these two pillars of quantum field theory in some essential way; therefore, we can start by asking whether such modifications are even theoretically possible. Quantum mechanics is an extremely rigid theoretical structure, and in the past eight decades, no one has discovered a way to modify its principles even slightly. However, theorists have found an essentially unique theoretical structure – *supersymmetry* – that can extend our notion of space-time in a new way.

Theories with supersymmetry are a special kind of quantum field theory that can be thought of as extending our usual four dimensions of space and time by four additional dimensions. The novelty is that distances in these extra dimensions are not measured by ordinary numbers, but by quantum variables: in a sense, supersymmetry makes space-time more intrinsically quantum-mechanical. Ordinary distances satisfy the basic multiplication law $a \times b = b \times a$, and are said to be *commuting* variables. However, distances in the quantum dimensions satisfy instead $a \times b = -b \times a$, with the crucial minus sign, and are said to be *anti-commuting*. In particular, $a \times a = -a \times a = 0$. Because of this, it is impossible to take more than a single "step" into the quantum dimensions. An electron can move around in our ordinary four dimensions, but it can also take this single step into the quantum dimensions. From the four-dimensional point of view, it will appear to be another particle, the *superpartner* of the electron, with the same mass and charge but different in its magnetic properties. The "symmetry" part of supersymmetry demands that the interactions respect a perfect symmetry between the ordinary and the quantum dimensions.

Supersymmetry is a deep idea that has played a major role in theoretical physics for the past forty years. It is an essential part of string theory, it has helped revolutionize our understanding of quantum field theory, and along the way it has opened up many new connections between physics and mathematics. Among its many remarkable properties, the one relevant to our discussion is that supersymmetry eliminates large vacuum quantum fluctuations

in a beautiful way. The inability to take more than a single step into the quantum dimensions means that there can be no wild fluctuations in the quantum dimensions; and because the quantum and ordinary dimensions must be treated symmetrically, there can be no large fluctuations in the ordinary dimensions either. More technically, the large fluctuations from the ordinary particles are perfectly canceled by their superpartners.

Of course, there is a catch: we haven't observed any of the superpartners for the ordinary particles! It is possible, however, that physics at short distances is supersymmetric, but that the perfect symmetry between ordinary and quantum dimensions is hidden by the same kind of long-distance illusion that hides the essential unity of strong, weak, and electromagnetic interactions. This long-distance "breaking" of supersymmetry has the effect of making the superpartners heavier than the ordinary particles we have seen, similar to how the W and Z particles are heavy while the photon is massless.

Can broken supersymmetry still address the fine-tuning problems? If nature becomes supersymmetric at around 10^{-17} cm, then the large quantum fluctuation in the Higgs field will be removed, yielding a completely natural resolution of the hierarchy problem. While there are a few other approaches to the hierarchy problem, supersymmetry is the most compelling, and there are some strong quantitative (though circumstantial) hints that it is on the right track. Whether it is supersymmetry or something else, a natural solution of the hierarchy problem demands *some* sort of new physics at around 10^{-17} cm. If nothing new happens until, say, 10^{-20} cm, then the quantum fluctuation of the Higgs field will be dragged to 10^{-20} cm. In order to make the actual interaction range of 10^{-17} cm natural, something new must show up at just this scale. This is why it is particularly exciting that we are probing exactly these distances at the LHC.

What about the much more severe cosmological constant problem? The cosmological constant is so tiny that its associated length scale is around a millimeter, and nature is clearly not supersymmetric at the millimeter scale. Supersymmetry does improve the fine-tuning problem for the cosmological constant from one part in 10^{120} to one part in 10^{60}, but this is small consolation. The difficulty is not just with supersymmetry: we have not seen any sort of new physics at the millimeter scale, so there is no hint that the cosmological constant problem is solved in a natural way.

This enormous challenge has led some theorists to imagine a different sort of explanation for fine-tuning problems, involving a radical change to our picture of space-time not at short distances, but at huge scales larger than the size of our observable universe. The idea takes some inspiration from developments in string theory over the last decade. String theory is a unique mathematical structure, but it has long been known that it has many different solutions, or *vacua*, each of which corresponds to a different possible long-distance world. The basic laws of nature are the same in all vacua, but the menu of particles and interaction strengths changes from vacuum to vacuum. The new realization is that the number of vacua with broken supersymmetry – the ones that might roughly resemble our world – is gargantuan: a rough estimate is that 10^{500} such vacua may exist. Furthermore, an important idea in cosmology, known as *eternal inflation*, makes it possible that all these vacua are actually realized somewhere in space-time. Many of these vacua have positive cosmological constants and are undergoing exponential expansion. Quantum mechanics enables bubbles of a new vacuum to form in this cosmology.

The bubble containing this "daughter" vacuum grows at nearly the speed of light and would naively appear to consume the "parent" vacuum. But this does not happen: because the parent is growing exponentially, it is never completely swallowed up, and it continues its exponential expansion forever. Thus, all possible daughter vacua are produced, giving rise to the picture of an infinite *multiverse* where all vacua are produced, infinitely often, somewhere in space-time.

In most of these vacua, the cosmological constant is enormous; but these vacua also undergo explosive accelerated expansion that would rip apart all structures, so in these regions the universe would be empty. However, there are so many vacua that, statistically speaking, some of them will have a small cosmological constant. It is only in those regions that the universe is not empty, and so it is not surprising that we should find ourselves there.

This picture is currently the only reasonable explanation that we have for the smallness of the cosmological constant, and it is not impossible that similar considerations may also be relevant for the hierarchy problem. So, is our universe just a tiny part of a vast and mostly lethal multiverse? If this picture is correct, it would be a further extension of the Copernican revolution. However, a number of major conceptual challenges must be overcome to determine whether these ideas make coherent sense, even on purely theoretical grounds. Because our own universe is accelerating, we can never see the other regions in the multiverse, and so it is not obvious that we can talk about these regions in a physically and mathematically meaningful way. But it is also not impossible to make proper sense of this picture. This has been an active area of research in the last decade, although serious theoretical progress on these problems still seems rather distant. Once again, the thorniest questions lie at the intersection of quantum mechanics, gravity, and cosmology.

Nima Arkani-Hamed

What might we expect to learn from experiments in the coming decade? The Large Hadron Collider is perhaps the most important experiment today, pushing the frontiers of fundamental physics. The accelerator itself is housed in a tunnel a hundred meters underground, with a circumference of twenty-seven kilometers. The tunnel contains a ring, where two sets of protons, moving in opposite directions, are accelerated to a speed 0.9999999 times the speed of light. The protons are made to collide head-on at two points around the ring, which are surrounded by enormous detectors. Two teams, each consisting of three thousand physicists, study the debris from these collisions, which give us a direct window into the laws of nature at distances of order 10^{-17} cm, an order of magnitude better than we have probed before.

As mentioned, a proton is not a point-like particle, but is a messy 10^{-14} cm bag containing quarks that are permanently trapped inside by gluons. When two of these messy bags collide at enormous energies, they usually break up into other messy collections of strongly interacting particles, zooming along in the initial direction of the beams. These typical interactions are not our main interest in probing short-distance physics. Rather, we are after the head-on collisions between the quarks and gluons in one proton and the quarks and gluons in the other. The telltale sign that a head-on collision has occurred is that particles scatter off at large angles relative to the initial direction of the beams. The collision can also produce energy enough to create new heavy particles and anti-particles.

Any new particles will typically be unstable, decaying on a timescale of order 10^{-27}

seconds into ordinary particles like electrons and positrons, quarks and antiquarks, and so on. These decay products will also spray off at large angles relative to the initial direction of the beam. Thus, studying all the debris from the high-energy collisions that come off at large angles is, in general, the best probe we have for studying short-distance physics. Having this rough means to discriminate "typical" and "interesting" events is crucial because the interesting events are exceedingly rare relative to the typical ones. There are about a billion typical collisions per second, whereas the timescale for producing, say, supersymmetric particles is expected to be in the range of a few per minute to a few per hour. The debris from these collisions are collected by the huge detectors and studied in great detail to look for the proverbial needle in the haystack.

The first order of business at the LHC is the search for the Higgs particle. As noted, analysis of the 2012 data should either definitively confirm or definitively exclude its existence. (Most physicists expect the former, especially following the solid hint reported in December 2011.) Assuming that the existence of the Higgs particle is confirmed, an accurate measurement of the rate at which it is produced, and the way it interacts with other particles, will shed light on whether it behaves as expected in the Standard Model, or whether it has modified properties that would indicate new physics.

The search for supersymmetry, or some other natural mechanism that would solve the hierarchy problem, is another central goal of the LHC program. The collision between quarks can have sufficiently high energy to pop the quarks into quantum dimensions and produce *squarks*, which rapidly decay to ordinary particles and other superpartners. In the simplest versions of the theory, the lightest of all the superpartners is a stable, electrically neutral particle that is so weakly interacting it sails through the detectors without leaving a trace. Thus, supersymmetric events should have the distinctive feature of seeming to have "missing" energy and momentum. No evidence for superpartners has yet emerged in the data, and the searches are beginning to encroach on the territory where superpartners must show up, if the supersymmetry indeed naturally solves the hierarchy problem.

After running through 2012, the LHC will stop and restart operations in 2014 – 2015 with twice its current energy. What might we know by 2020? The discovery of supersymmetry would represent the first extension of our notion of spacetime since Einstein and would confirm one of the most important theoretical ideas of the past forty years. We would also find a completely satisfactory understanding of the question, why is gravity weak? On the other hand, if neither supersymmetry nor any other sort of natural solution to the hierarchy problem appears in the data, the situation will be much more confusing. We will have solid experimental evidence for fine-tuning in the parameters that determine elementary particle masses, something we have never seen in such dramatic fashion. This would strongly resonate with the apparently enormous fine-tuning problems associated with the cosmological constant, and would give theorists a strong incentive to take the ideas of the multiverse more seriously.

It should be clear that we have arrived at a bifurcatory moment in the history of fundamental physics, a moment that has enormous implications for the future of the subject. With many theoretical speculations pointing in radically different directions, it is now up to experiment to render its verdict!

The twentieth century was dominated by the ideas of relativity and quantum me-

chanics, and their synthesis is quantum field theory. As I have discussed, there are strong reasons to think that some essentially new ideas are needed in the twenty-first century. The LHC is poised to shed significant light on the question of why a macroscopic universe exists, but the questions having to do with the deeper origin of space-time seem tied to the Planck scale, offering little hope for direct clues from experiment in the near future. Even so, the requirements of physical and mathematical consistency have provided a strong guide to the theoretical investigation of these questions. Indeed, the spectacular progress in string theory over the last four decades, which has time and again surprised us with unanticipated connections between disparate parts of physics and mathematics, has been driven in this way. Today, however, we confront even deeper mysteries, such as coming to grips with emergent time and the application of quantum mechanics to the entire universe. These challenges call for a bigger shift in perspective. Is there any hope for taking such large steps without direct input from experiment?

We can take some inspiration by looking at the path that led from classical physics to relativity and quantum mechanics. Some of the crucial clues to future developments were lying in plain sight, in the structure of existing theories. Einstein's motivations for developing both special and general relativity were rooted in "obvious" properties of classical physics. Newton's laws already had a notion of Galilean relativity. However, Galilean relativity allowed for arbitrarily large signal velocities and thus action at a distance. This was in conflict with Maxwell's laws of electromagnetism, in which the interactions involving electromagnetic fields were local. Einstein resolved this purely theoretical conflict between the two pillars of classical physics by realizing that the Galilean notion of relativity had to be deformed to one that was compatible with a maximal speed for signal propagation and thus with locality.

The loss of determinism in passing from classical to quantum mechanics was a much more radical change in our picture of the world, and yet even this transition was presaged in classical physics. Newton's laws are manifestly deterministic; given the initial position and velocity of a particle, together with all the forces acting on it, the laws of motion tell us where the particle goes in the next instant of time. However, in the century after Newton, physicists and mathematicians discovered a reformulation of Newton's laws that led to exactly the same equations, but from a completely different philosophical starting point. Of all the possible trajectories a particle can take from A to B, it chooses the one that minimizes the average value of difference between the kinetic and potential energies along the path, a quantity known as the *action* of the path. This law does not look deterministic: the particle seems to sniff out all possible paths it could take from A to B and then chooses the one that minimizes the action. But it turns out that the paths that minimize the action are precisely the ones that satisfy Newton's laws.

Why should it be possible to talk about Newton's laws in such a different way, which seems to hide their most essential feature of deterministic evolution in time? We now know the deep answer to this question is that the world is quantum-mechanical. As Richard Feynman pointed out in the mid-1940s, a quantum-mechanical particle takes all possible paths from A to B; in the classical limit, the dominant contributions to the probability are peaked on the trajectories that minimize the action, which are, secondarily, the ones that satisfy Newton's laws. Since quantum mechanics is not deterministic, the clas-

sical limit of the theory could not land on Newton's laws, but instead lands on a different formulation of classical physics in which determinism is not manifest but rather is a secondary, derived notion.

If there are any clues hiding in plain sight today, they are lurking in the many astonishing properties of quantum field theory and string theory that have been uncovered over the past two decades. The founders of quantum field theory could never have imagined that it might describe a theory of gravity in a higher-dimensional curved space, and yet it does. We have learned that theories that seem completely different from the classical perspective are secretly identical at the quantum-mechanical level. Many of these developments have uncovered deep connections between physics and modern mathematics. Even "bread and butter" calculations in field theory, needed to understand the strong interaction processes at the LHC, have revealed major surprises. Textbook calculations for the rates of these processes quickly lead to hundreds of pages of algebra, yet in recent years we have understood that the final expressions can fit on a single line. These simplifications are associated with a new set of completely hidden symmetries enjoyed by ordinary quantum field theories. They have been sitting under our noses undetected for sixty years, and now they are exposing connections to yet another set of new structures in mathematics.

Thus, while we may not have experimental data to tell us about physics near the Planck scale, we do have an ocean of "theoretical data" in the wonderful mathematical structures hidden in quantum field theory and string theory. These structures beg for a deeper explanation. The standard formulation of field theory hides these amazing features as a direct consequence of its deference to space-time locality. There must be a new way of thinking about quantum field theories, in which space-time locality is not the star of the show and these remarkable hidden structures are made manifest. Finding this reformulation might be analogous to discovering the least-action formulation of classical physics; by removing space-time from its primary place in our description of standard physics, we may be in a better position to make the leap to the next theory, where space-time finally ceases to exist.

Microbes as Menaces, Mates & Marvels

Bonnie L. Bassler

Abstract: The conventional understanding of microbes as causative agents of disease has led us to fear them and to consider them our deadly enemies. Much less appreciated are the central roles microbes play in shaping the environment and in maintaining plant, animal, and human health. All metazoan organisms – organisms that we can see with the naked eye – exist in lifelong partnerships with vast microbial communities. These "microbiomes" supply metazoans with essential life processes that are not encoded in nonmicrobial genomes. Recent work in microbiology has revealed that microbes, like metazoans, have specific body plans and sensory systems, that they can communicate with each other, and that they orchestrate collective behaviors. Investigations of these ancient yet enduring processes are uncovering the fundamental design principles of life. Microbes are also storehouses of new molecules, biochemical pathways, and materials with medical, industrial, and agricultural relevance. Scientists are harnessing these microbial products in efforts to confront humanity's most pressing problems. This essay explores the wonder, complexity, power, and utility of microbes in the twenty-first century.

BONNIE L. BASSLER, a Fellow of the American Academy since 2007, is the Squibb Professor of Molecular Biology at Princeton University and a Howard Hughes Medical Institute Investigator. Her research focuses on the molecular mechanisms that bacteria use for intercellular communication. Her work has appeared in such journals as *Molecular Microbiology*, *Genes & Development*, and *Cell*. She edited the volume *Chemical Communication among Bacteria* (with Stephen C. Winans, 2008).

You may think you've seen everything, but in fact, most of the world is invisible to us. It is composed of, and was constructed by, microbial organisms. Microbes are single-celled organisms too small to see with the naked eye. They include archaea, fungi, and protists, but overwhelmingly, they are bacteria. For billions of years these invisible critters, our forefathers, have been shaping the Earth and making it a suitable place for us to live. Higher organisms – all plants, invertebrates (including insects), and vertebrates (including humans) – occupy only a sliver of the world. We live exclusively within narrow ranges of temperature, air pressure, atmosphere, pH, and nutrient sources. Microbes, by contrast, have adapted to, inhabited, exploited, and tamed every niche on Earth. They thrive under extreme pressure and at high temperatures miles below the ocean surface, enveloped in the sulfurous smoke spewing from thermal vents. They live encrusted in the sediments of lakes beneath ice caps that have not thawed in more than ten thousand years. There, they exist in a world that has not seen

the light of day or taken a breath from the atmosphere for millennia. Microbes flourish at the pH of battery acid, and in this hostile environment, they produce the stunning, gemstone-colored pools found in Yellowstone National Park. Unlike the human diet, which is narrow and invariant, microbes consume, well, everything. They eat the regular stuff – proteins and carbohydrates – but also radioactivity, rocks, crude oil, rust, paint, dirt, and wood. Every plant, animal, and human requires a vast community of commensal microbes working 24/7 to keep it alive and healthy.

As the foundation of nineteenth- and twentieth-century biological science, microbes enabled scientists to dismiss the possibility of spontaneous generation. They inspired Robert Koch's postulates, revealing disease causation for the first time. Microbes gave scientists a way to deduce the famous DNA double helix; they showed that heritable information is encoded in units of genes and that the essential parts of life are DNA, RNA, proteins, and lipids. Ironically, these paradigm-changing discoveries also made the microbial sciences passé. By the mid-twentieth century, microbes were generally thought to have been picked clean. Science instead focused on more "complex" problems: namely, embryonic development, neurocircuitry, and cancer.

However, much more remains to be gleaned from examining the lives and personalities of microbes. Microbiologists recently demonstrated that microbes possess the features of metazoans: they have precise body plans and sophisticated sensory mechanisms; they communicate with each other; and they amass their numbers in order to act collectively. Investigations of life's oldest organisms are teaching us how these processes operate in metazoans. Further, microbiologists are mining microbes for new genes, molecules, and biochemical pathways – work that holds great promise for medical, agricultural, industrial, and technological applications. This exciting research has coincided with the alarming rise of antibiotic-resistant microbial pathogens across the planet. New research directions, coupled with the recurrence of the microbial threat, have propelled microbiology once again to the forefront of science. This essay explores the wonder, complexity, power, and utility of microbes in the twenty-first century.

We divide all life into two categories: prokaryotes and eukaryotes. Every eukaryotic cell has a nucleus; prokaryotic cells do not. Prokaryotes are further divided into archaea and bacteria.[1] We suspect that Earth's first organisms were archaea because they thrive in habitats mirroring those of early Earth: hot, salty, anoxic, and metal-rich. Bacteria live in more temperate places. Eukaryotes include fungi, protists, plants, invertebrates, and vertebrates. Eukaryotic species are also predominantly microbial, which means that you haven't seen most of them, either.

The Earth is about 4.5 billion years old.[2] At present, the oldest known prokaryotic fossil dates back 3.5 billion years. The oldest known eukaryotic fossil, a microbe, is about 1.7 billion years old. By contrast, the most ancient macroscopic fossil is only about 650 million years old.[3] Thus, the better part of the world's living history is exclusively microbial, and it is predominantly prokaryotic. Archaea had their day back when the Earth was in chaos. They have trouble living in the presence of oxygen, so when the planet became oxygenated roughly 2.5 billion years ago, archaea were relegated to the most hostile locations remaining on the planet. Bacteria, on the other hand, flourished because they adapted to both oxygenated and anoxic environments. They prevailed and made the world we know. Indeed, the vast majority of biological change and diversifica-

tion has occurred and continues to occur in the unseen bacterial world.

The Gates Foundation, the National Academy of Engineering, the Plant Sciences and Environmental Sciences divisions of the National Research Council, and other groups spanning diverse scientific disciplines have proposed sets of "grand challenges" facing humanity. Strikingly, the clear consensus among leading scientists is that the microbial sciences are key to confronting the majority of these challenges.[4] In the recent National Academies report *A New Biology for the 21st Century*, for example, all four of the suggested grand challenges have significant microbiological components.[5] The overarching scientific imperative is to ensure that microbes' potential contributions to a healthy planet are fully realized. There are three main reasons why the science world is directing its focus toward microbes.

Microbes are Menaces. The top killers of humans are microbes. Chief among these are HIV (a virus), malaria (a eukaryotic microbe), and tuberculosis (a bacterium), which collectively claim more than five million lives every year. The next biggest killers are microbes that cause diarrheal diseases and pneumonias, which lead to another two million human deaths per year.[6] At present, these microbial scourges are restricted to underdeveloped regions. If you live in a developed country, there is little chance (at least for now) that you will die of an infectious disease unless you are elderly and frail. Residents of developed countries will overwhelmingly die from heart disease or cancer. Antibiotic control of bacterial pathogenicity is one of the greatest advances ever made to improve human health. This success came seventy-five years ago, and the bacterial problem was thought to be solved. Classic bactericidal drugs are excellent: they have few side effects and are broad-spectrum. Because of these qualities, only incremental improvements to existing drugs have been made in the past forty years. During this same time frame, pathogenic bacteria emerged across the globe, and many of them are resistant to conventional antibiotics. This predicament has been fueled by the widespread use of bactericidal drugs. Multidrug-resistant bacteria are now prevalent around the world, and they contribute greatly to the increasing microbial threat to human health.

Microbes are Mates. Every plant and animal (including insects and humans) depends on microbial partners for life because they provide essential functions that are not encoded in metazoan genomes.[7] Consider the human-microbe relationship: there are ten times more microbial cells than human cells and one hundred times more microbial genes than human genes in and on every person. These metrics suggest that, at best, you are 10 percent human (in terms of cells), and more realistically, you are only 1 percent human (in terms of genes).[8] The remainder of "you" is microbial. Not only do you inherit your human genome from your parents, you also inherit, during the first few months of life, your microflora. These vast communities of microbes – your "microbiome" – have effects on metabolism, immunity, and behavior[9]; and properly balanced, invisible interactions with your microbes are essential to your well-being. Microbes cover you in a thin biofilm that serves as invisible body armor protecting you from environmental insults. Microbes help digest your food, provide your vitamins, aid in the development of your blood vessels, and educate your immune system. Microbes also occupy every empty nook and cranny in and on you to make that space unavailable to pathogens that try to gain access. A profound microbial component has been shown in a growing num-

ber of ailments including allergies, asthma, obesity, cancer, and autoimmune disorders.[10]

Microbes are Marvels. Microbes hold answers to the world's most challenging issues: food, health, energy, and the environment. They represent a virtually inexhaustible source of biodiversity, metabolic ingenuity, and natural products. They are the workhorses for the production of industrial catalysts and pharmaceuticals ranging from insulin, to antibiotics, to vaccines, to probiotics. Microbes are the most promising source for the next generation of environmentally and politically neutral fuels. They provide the chassis and parts sets for synthetic biology, a new field of science devoted to developing robust, industrial-scale biological machines and processes. They are required partners for all plant growth, making microbes an untapped resource for adapting crops to grow in more places with fewer inputs. Microbes are critical drivers of Earth's biogeochemical cycles and are therefore important players in climate change, both as sentinels and, potentially, as mitigators.[11]

To be fair, microbes also contribute enormously to greenhouse gas emissions stemming from the reactions they carry out in the industrialized livestock industry.[12] For example, cattle require microbes to digest their food, and the consequence of these microbial fermentation reactions is production of methane, a greenhouse gas.[13] Further, the possibility exists that, as Earth's temperature rises, the Arctic permafrost could melt, giving microbes access to the vast stores of organic carbon trapped beneath. If this occurs, microbes could release huge amounts of greenhouse gases.[14] Frankly, however, both of these predicaments were caused by humans; and because microbes cannot resist easy access to food, the microbes are only indirectly to blame. This smudge on their reputation does not make microbes any less marvelous because microbes also have the ability to consume, sequester, and degrade greenhouse pollutants.

Finally, microbes are the ultimate model organisms for molecular and cellular biology, making possible the spectacular advances in health and biotechnology that have come from those disciplines.

Microbes are the research frontier for the twenty-first century. Two burgeoning bacterial research fields – bacterial cell biology and bacterial cell-to-cell communication – are exemplars of the progress microbiologists are making in understanding biological complexity and in generating promising possibilities for technological applications.

Keeping House: Unseen Orderliness and Complexity. For nearly all of the four hundred years that humans have known about bacteria, they have been considered amorphous bags of goop. Eukaryotes have long been known to possess a subcellular architecture in which DNA, RNA, and proteins are localized to the right place at the right time. Bacteria, on the other hand, were thought too small to be significantly organized. Consequently, cell biology, the scientific discipline that aims to understand cellular organization, was generally restricted to eukaryotes. Remarkable recent advances in imaging technology, however, have made it possible for scientists to peer into bacterial cells as they traditionally peered into bigger eukaryotic cells.[15] We see that bacteria are decidedly organized, and they are loaded with molecular machinery that is breathtaking in its design, complexity, and efficiency.[16]

Bacteria move by means of miniature motors that operate like boat engines complete with propellers.[17] These motors use the molecule ATP as fuel, and this contraption allows cells to swim through liquid at a pace that, given their size, would out-lap Olympian Michael Phelps ten to

one. Bacteria propel themselves more than ten body-lengths per second.[18] Phelps, during his world-record-setting Olympic freestyle swim, moved at about one body-length per second. When bacteria settle out of their liquid world onto a surface – including ocean sediments, oil slicks, and the bodies of larger organisms – they sense that they need another form of transportation. They then sprout thousands of appendages allowing them to crawl spider-like across surfaces. When they leave the surface to return to the liquid environment, these legs fall off and the boat propeller reengages.[19] They are the perfect amphibious vehicles.

Bacteria possess similarly stunning equipment to control information flow, inheritance, and reproduction. Marvelous multipart machines are constructed from component protein building blocks. One such apparatus can rapidly copy the millions of nucleotide bases that compose the genome.[20] Accuracy in this task is essential, and so in the midst of copying, this biological machine proofreads every letter. How accurate is this proofreader? If a typist transcribed at a reasonable rate – say, forty words per minute – and that typist worked continuously eight hours a day, five days a week (with two weeks off for vacation), it would be as if he or she made one mistake every forty years.

Accurately copying the genetic code from generation to generation is one requirement for life to happen. Additionally, the biological machinery of a cell must convert the one-dimensional information embedded in the DNA molecule into a complete three-dimensional organism. The cellular components are not swishing around willy-nilly. Rather, there is a precise spatial organization to each part that provides asymmetry, another feature that we now understand is essential for life, even in bacteria. Without asymmetry, no embryo could develop, our neurons could not convey information, our intestines could not absorb nutrients, and bacteria could neither swim nor crawl. Indeed, all cells, even bacterial cells, possess subcellular architectures in which the component parts are put in very specific places at specific times. Dare I say, molecular *feng shui*?

How do bacteria accomplish such feats? Bacteria do it like eukaryotes do it; or, to put things in proper perspective, eukaryotes perform these tasks in the same way as bacteria. In eukaryotes, cellular organization is largely established through a so-called cytoskeletal network – your molecular skeleton – made up of three different types of filament proteins. We have recently learned that bacteria have their own versions of each filament that together organize the cell.[21] In addition, bacteria construct miniature assembly lines containing ordered arrays of enzymes that function sequentially in biochemical pathways to ensure the stepwise and efficient production of products.[22] (Perhaps Henry Ford got the idea for mass production of automobiles from his microbiome.) This molecular assembly-line capability, which was discovered only a few years ago, appears conserved from bacteria to humans.

From moment to moment, bacteria assess and adapt to what is happening around them. If a nutrient or a poison wafts by, bacteria move in the direction of a food source while they swim away from a toxin. If the environment is too salty, putting the bacteria in danger of shriveling up and dying, they install pumps in their membranes to actively squirt ions out of the cytoplasm. If the pH becomes acidic, bacteria synthesize buffer molecules that bind to and sop up excess protons. To monitor the environment, bacterial membranes are decorated with antenna-like apparatuses called receptors, which connect the outside world to the inside of the cell.[23] When

bacteria encounter an environmental change, the receptors sound a sensory alert in the cytoplasm where information in the DNA can be extracted to instantaneously make appropriate biomolecules to deal with the situation. Remarkably, bacterial receptors look and work essentially the same way as receptors in plants and animals. Indeed, our cell surfaces are likewise adorned with membrane-spanning receptors that monitor and react to hormones, that tell our cells to go to sleep or wake up, and that alert our cells to react to heat, cold, and pain. In all cases, bacterial or human, the receptor's job is to monitor changes in the outside world and shuttle that information internally so that cells can do the right thing at the right time. Bacteria invented this solution for adapting to a changing world billions of years ago. Eukaryotic cells subsequently co-opted these biochemical mechanisms and biological design principles to robustly convey extracellular information into cells.

We have long known that some bacterial parts (DNA, RNA, lipids, and some important proteins) are essential for bacterial life. The new field of microbial cell biology has demonstrated that viability depends on these bits being precisely located within the bacterial cell in space and in time.[24] This revelatory finding offers scientists exciting new ways to imagine combating pathogenic bacteria. The hope is to identify new antibacterial compounds that do not target the essential molecules of life, such as DNA and RNA, but rather, that disrupt how bacteria distribute particular components to defined subcellular destinations at the correct times. We can exploit this new knowledge of cell biology for beneficial and industrial purposes. We are becoming able to logically string together preexisting biological units and induce them to perform new functions. We can now engineer mini-assembly lines to build organisms that do or produce useful things. We can construct miniature scaffolds to use as new materials or tissues or building platforms. Such research holds tremendous promise for the future of renewable energy, new material synthesis, environmental sustainability, food production, and medicine.

Mob Psychology: Bacteria Talk to Each Other and Orchestrate Group Activities. Bacteria are miniscule. They are five hundred times smaller than human cells, which are also microscopic. Most of the interesting things that bacteria encounter are comparatively enormous. How then is it possible for bacteria to influence their environment? New research shows that the principal reason bacteria are so successful is that they rarely act alone. Rather, bacteria keep track of the number of cells in the vicinity; when there are enough present, they cooperate to synchronously undertake tasks that would be unproductive if carried out by an individual bacterium acting alone. We call this phenomenon *quorum sensing*.[25]

Quorum-sensing bacteria count their cell numbers by communicating with one another. They use chemical molecules, called autoinducers, as their "words."[26] As quorum-sensing bacteria multiply, each cell releases autoinducer molecules into the surroundings. Because each bacterium contributes a share of autoinducer to the environment, the quantity of the extracellular autoinducer increases in step with the increasing bacterial cell number. When the autoinducer accumulates above a threshold level, receptors on the bacterial cell surfaces become capable of efficiently detecting the molecules' presence. The receptors relay information into the cytoplasm, specifying, in effect, that the number of neighbors present has exceeded the minimum needed to accomplish some task. The bacterial group then responds with a population-wide, synchronous change in activity. In essence, quorum sensing is a bacterial voting procedure.

The bacteria cast chemical votes, they tally the vote, and all the members of the community go along with the outcome.

Quorum sensing controls collective behaviors including the release of toxins and other virulence factors, biofilm formation, and DNA exchange. These types of tasks are effective only when undertaken en masse. Say, for example, you eat contaminated food and a few bacterial pathogens are lucky enough to come along for the ride. The amount of a toxin that a few bacteria could manage to dribble out would be inconsequential in terms of establishing an infection. The more cunning strategy is for the bacteria to wait, to multiply, and to recognize when they have enough cells that can, together, launch an attack that will successfully overcome your host defenses. Thus, bacteria require quorum sensing to be virulent. The good news is that they also require quorum sensing to perform the many beneficial jobs they routinely do for us to keep us healthy.

Quorum sensing is the norm in the bacterial world, and in each bacterial species, hundreds of traits are controlled by this chemical discourse.[27] Bacteria count to different numbers depending on the task at hand. Some collective jobs require the concerted effort of only a modest number of cells, while other duties require significantly more team members. Evolution has apparently optimized quorum sensing so that correctly sized battalions undertake appropriate missions.

Every quorum-sensing bacterium has multiple quorum-sensing circuits. That is, bacteria are multilingual, and they converse using a rich chemical lexicon.[28] Beyond simply counting, bacteria use different quorum-sensing molecules to distinguish between self and non-self, and they decode blends of autoinducer molecules to extract information about the ratios of different species present. In turn, they tailor their collective behaviors depending on who is in the majority and who is in the minority of a mixed-species bacterial consortium (that is, friend or foe). To accomplish this feat, bacteria employ a chemical vocabulary composed of molecules that identify self, non-self but closely related, and non-related.[29] In essence, they can determine "you are my sibling," or "you are my cousin," or "you are not family."

Research into the social lives of bacteria is providing the first understanding of the origins of cell-to-cell communication, self versus non-self recognition, how synchronicity is achieved in collective processes, and the evolution of cooperation.[30] Given that life began with bacteria, there is a high likelihood that discovering the principles underpinning nature's earliest command-and-control centers will provide insight into analogous processes in eukaryotic organisms. The second outcome of this research is practical: we now know that quorum sensing is intimately linked to biofouling and pathogenesis. Interfering with quorum sensing opens up new ways to control bacteria that erode industrial processes, and it offers a fundamentally new approach to antibiotics to battle virulent bacteria.

Anti-quorum-sensing antibiotic strategies should be especially difficult for bacteria to bypass by mutation. If collective action is required for virulence to be effective, then a single cell that fortuitously acquires a mutation making it blind to a quorum-sensing drug does not gain an advantage from the mutation. The "resistant" bacterium will switch into quorum-sensing mode at the appropriate time; however, other nearby bacteria that remain susceptible to the drug will not. In all likelihood, the resistant mutant will have decreased fitness because it will undertake energy-expensive quorum-sensing behaviors without reaping the benefits of col-

lective action.[31] This quorum-sensing resistance scenario is fundamentally unlike the development of resistance to traditional antibiotics, in which the resistant mutant and its offspring receive an immediate growth advantage. The use of traditional antibiotics fuels the growth and spread of antibiotic-resistant bacteria whereas anti-quorum-sensing therapies are not predicted to contribute to resistance.

Equally interesting in terms of future quorum-sensing research is the development of pro-quorum-sensing strategies that make bacterial "chit chat" more effective. The goal is to enhance quorum sensing to encourage beneficial collective behaviors and improve production of compounds or processes of value to society. For example, we could intensify microbial conversations to encourage bacteria that produce modest amounts of interesting natural products to make considerably larger amounts. We could convert bacteria that are inefficient at bioremediation or at biofuel production into powerhouses. Developing potent pro-quorum-sensing molecules could enable these and other applications to occur at industrial scales.

Microbes are the repositories of 3.5 billion years of evolutionary secrets that have enabled life and shaped a planet. They can kill us and they can rescue us. To ensure the latter fate, we need a concerted effort toward understanding and harnessing the power of microbes. A vanguard of scientists from biology, chemistry, physics, computation, evolution, and engineering have entered the fray. Our objective is to deliver to society a comprehensive understanding of microbial complexity so that humans can effectively conquer the bad microbes, enslave the useful microbes, and reward the good microbes that devote their tiny lives to keeping us alive.

Microbial diversity surpasses everything else on the planet. Scientists have studied only a handful of microbial species; we know there are millions more. That means there are thousands of millions of microbial genes that produce molecules, ingenious pathways, biological machines, and structures that can be discovered and exploited for medical, industrial, and agricultural purposes.[32] Microbes are our planet's only limitless renewable resource, and this cache remains virtually untapped. We can look to the accumulated smarts of eons of evolution, expertly preserved inside microbes, for timely approaches and solutions to problems of global significance.

ENDNOTES

[1] Carl R. Woese, "Bacterial Evolution," *Microbiology Review* 51 (2) (1987): 221.

[2] G. Brent Dalrymple, *The Age of the Earth* (Stanford, Calif.: Stanford University Press, 1991).

[3] J. William Schopf, *Cradle of Life: The Discovery of Earth's Earliest Fossils* (Princeton, N.J.: Princeton University Press, 2001).

[4] David Hooper, Roberto Kolter, and Bonnie Bassler, "ASM Comments on Bioeconomy Blueprint Request for Information," American Society for Microbiology, December 13, 2011, http://www.asm.org/index.php/policy/bioec-12-11.html.

[5] National Research Council, *A New Biology for the 21st Century* (Washington, D.C.: National Academies Press, 2009).

[6] *The Global Burden of Disease: 2004 Update* (Geneva, Switzerland: World Health Organization, 2008).

[7] Fredrik Bäckhed, Ruth E. Ley, Justin L. Sonnenburg, Daniel A. Peterson, and Jeffrey I. Gordon, "Host-Bacterial Mutualism in the Human Intestine," *Science* 307 (5717) (2005): 1915.

[8] Steven R. Gill, Mihai Pop, Robert T. DeBoy, Paul B. Eckburg, Peter J. Turnbaugh, Buck S. Samuel, Jeffrey I. Gordon, David A. Relman, Claire M. Fraser-Liggett, and Karen E. Nelson, "Metagenomic Analysis of the Human Distal Gut Microbiome," *Science* 312 (5778) (2006): 1355.

[9] June L. Round and Sarkis K. Mazmanian, "The Gut Microbiota Shapes Intestinal Immune Responses during Health and Disease," *Nature Reviews Immunology* 9 (5) (2009): 313; Lora V. Hooper, Tore Midtvedt, and Jeffrey I. Gordon, "How Host-Microbial Interactions Shape the Nutrient Environment of the Mammalian Intestine," *Annual Review of Nutrition* 22 (2002): 283.

[10] Les Dethlefsen, Margaret McFall-Ngai, and David A. Relman, "An Ecological and Evolutionary Perspective on Human-Microbe Mutualism and Disease," *Nature* 449 (7164) (2007): 811; Ruth E. Ley, Peter J. Turnbaugh, Samuel Klein, and Jeffrey I. Gordon, "Microbial Ecology: Human Gut Microbes Associated with Obesity," *Nature* 444 (7122) (2006): 1022; Peter J. Turnbaugh, Ruth E. Ley, Michael A. Mahowald, Vincent Magrini, Elaine R. Mardis, and Jeffrey I. Gordon, "An Obesity-Associated Gut Microbiome with Increased Capacity for Energy Harvest," *Nature* 444 (7122) (2006): 1027.

[11] Hooper, Kolter, and Bassler, "ASM Comments on Bioeconomy Blueprint Request for Information."

[12] Food and Agricultural Organization of the United Nations, "Livestock a Major Threat to Environment," FAO Newsroom, November 29, 2006, http://www.fao.org/newsroom/en/news/2006/1000448/index.html.

[13] Kristen A. Johnson and D. E. Johnson, "Methane Emissions from Cattle," *Journal of Animal Science* 73 (8) (1995): 2483.

[14] Eric A. Davidson and Ivan A. Janssens, "Temperature Sensitivity of Soil Carbon Decomposition and Feedbacks to Climate Change," *Nature* 440 (7081) (2006): 165.

[15] Zemer Gitai, "New Fluorescence Microscopy Methods for Microbiology: Sharper, Faster, and Quantitative," *Current Opinion in Microbiology* 12 (3) (2009): 341.

[16] Lucy Shapiro, Harley H. McAdams, and Richard Losick, "Why and How Bacteria Localize Proteins," *Science* 326 (5957) (2009): 1225.

[17] Howard C. Berg, "The Rotary Motor of Bacterial Flagella," *Annual Review of Biochemistry* 72 (2003): 19.

[18] Nicholas C. Darnton, Linda Turner, Svetlana Rojevsky, and Howard C. Berg, "On Torque and Tumbling in Swimming *Escherichia coli*," *Journal of Bacteriology* 189 (5) (2007): 1756.

[19] Linda McCarter and Michael Silverman, "Surface-Induced Swarmer Cell Differentiation of *Vibrio parahaemolyticus*," *Molecular Microbiology* 4 (7) (1990): 1057.

[20] Zvi Kelman and Mike O'Donnell, "DNA Polymerase III Holoenzyme: Structure and Function of a Chromosomal Replicating Machine," *Annual Review of Biochemistry* 64 (1995): 171.

[21] Matthew T. Cabeen and Christine Jacobs-Wagner, "The Bacterial Cytoskeleton," *Annual Review of Genetics* 44 (2010): 365.

[22] Michael Ingerson-Mahar, Ariane Briegel, John N. Werner, et al., "The Metabolic Enzyme CTP Synthase Forms Cytoskeletal Filaments," *Nature Cell Biology* 12 (8) (2010): 739.

[23] James A. Hoch and Thomas J. Silhavy, eds., *Two-Component Signal Transduction* (Washington, D.C.: ASM Press, 1995).

[24] John N. Werner, Eric Y. Chen, Jonathan M. Guberman, et al., "Quantitative Genome-Scale Analysis of Protein Localization in an Asymmetric Bacterium," *Proceedings of the National Academy of Sciences* 106 (19) (2009): 7858; Ingerson-Mahar, Briegel, Werner, et al., "The Metabolic Enzyme CTP Synthase Forms Cytoskeletal Filaments."

[25] Christopher M. Waters and Bonnie L. Bassler, "Quorum Sensing: Cell-to-Cell Communication in Bacteria," *Annual Review of Cell and Developmental Biology* 21 (2005): 319; Wai-Leung Ng and Bonnie L. Bassler, "Bacterial Quorum-Sensing Network Architectures," *Annual Review of Genetics* 43 (2009): 197.

[26] Andrew Camilli and Bonnie L. Bassler, "Bacterial Small-Molecule Signaling Pathways," *Science* 311 (5764) (2006): 1113.

[27] Steven T. Rutherford, Julia C. van Kessel, Yi Shao, et al., "AphA and LuxR/HapR Reciprocally Control Quorum Sensing in Vibrios," *Genes & Development* 25 (4) (2011): 397.

[28] Bonnie L. Bassler and Richard Losick, "Bacterially Speaking," *Cell* 125 (2) (2006): 237.

[29] Xin Chen, Stephan Schauder, Noelle Potier, et al., "Structural Identification of a Bacterially Quorum-Sensing Signal Containing Boron," *Nature* 415 (6871) (2002): 545; Douglas A. Higgins, Megan E. Pomianek, Christina M. Kraml, et al., "The Major *Vibrio cholerae* Autoinducer and Its Role in Virulence Factor Production," *Nature* 450 (7171) (2007): 883.

[30] Carey D. Nadell and Bonnie L. Bassler, "A Fitness Trade-Off between Local Competition and Dispersal in *Vibrio cholerae* Biofilms," *Proceedings of the National Academy of Sciences* 108 (34) (2011): 14181; Carey D. Nadell, João B. Xavier, Simon A. Levin, et al., "The Evolution of Quorum Sensing in Bacterial Biofilms," *PLoS Biology* 6 (1) (2008): e14.

[31] Lee R. Swem, Danielle L. Swem, Colleen T. O'Loughlin, et al., "A Quorum-Sensing Antagonist Targets Both Membrane-Bound and Cytoplasmic Receptors and Controls Bacterial Pathogenicity," *Molecular Cell* 35 (2) (2009): 143.

[32] Bonnie L. Bassler, "Small Cells – Big Future," *Molecular Biology of the Cell* 21 (22) (2010): 3786.

Fossils Everywhere

Neil H. Shubin

Abstract: History is omnipresent in the natural world, from inside rocks on the continents to the genes, cells, and organs of each creature on the planet. Linking the historical records of rocks, fossils, and genes has been a boon to understanding the major events in evolution. We use these seemingly different lines of evidence as tools for discovery: analyses of genes can predict likely places to find fossils, and new fossils can provide the means to interpret insights from genetics. Viewed in this way, every living thing on Earth is the extreme tip of a deeply branched tree of life that extends three billion years into the past. Genes and fossils reveal how deeply connected our species is to the rest of the living world and the planet itself.

NEIL H. SHUBIN, a Fellow of the American Academy since 2009, is the Robert R. Bensley Distinguished Service Professor of Organismal Biology and Anatomy and Associate Dean of Biological Sciences at the University of Chicago. He has performed expeditionary research programs in Canada, Africa, the continental United States, Asia, and Greenland that have led to new insights into the origin of major groups of vertebrates. He is the author of *Your Inner Fish: A Journey Into the 3.5-Billion-Year History of the Human Body* (2008), and his work has appeared in *Nature* and the *Journal of Vertebrate Paleontology*, among other publications.

More than a century of discovery has led us to the realization that the descendants of fish now walk on land, those of dinosaurs fly in the air, and the evolutionary offspring of arboreal primates fly in space and have left footprints on the moon. One hundred years ago these evolutionary transitions would have seemed utterly impossible, or worse, unthinkable. For example, most fish reproduce, feed, and breed in water; for their relatives to invade land, almost every system of their bodies would apparently need to change. If the same conceptual challenges hold for every major step in the history of life, how could we ever come to terms with ancient events, let alone understand their relevance to our lives today? We must look to the genes, cells, and organs of every creature alive today to understand the more than 3.5 billion years of the history of life. Each new piece of evidence that emerges helps reveal how the past has shaped us and our world.

We live in an age of invention; new technology changes what we can do, how we live, and what kinds of questions we can ask about our world. The doubling time of computer chip speeds is surpassed by the rate at which we can sequence whole genomes at ever-decreasing cost. The genome of any species can now be identified and compared among crea-

tures as different as yeast and humans.[1] Genes can even be swapped between species, moving basic bits of DNA between flies, worms, and mice. The exponential rate of technological change in biology once prompted a colleague of mine to admit that he could have collected all the data for his Ph.D. thesis – written in the early 1990s – in a single week. My colleague made that comment about five years ago; he could now execute the dissertation in an afternoon.

In the face of this technological revolution, fossil bones seem almost quaint. Most of us encounter these relics in museums, where the dinosaurs, ground sloths, and mammoths stand motionless in neoclassical buildings of marble and granite. Both the subject and the field appear frozen in time; paleontologists digging in rocks to find remnants of long-lost life is a far cry from a roomful of humming computers and gene sequencers. But these are special times; profound insights into the great transformations of life have come from linking new genetic, developmental, and computational tools with approaches that date from the days of Leonardo da Vinci.

Fossils are a kind of window into our perceptions of life, the planet, and our historical connection to them.[2] Most of us take their meaning entirely for granted, so much so that it is hard to envision how strange these objects seemed when first encountered by philosophers centuries ago. Our own conception of them – as evidence of creatures that inhabited long-lost worlds – arose in parallel with an entirely novel way of thinking about the natural world.

In 1541, Conrad Gessner, then twenty-five years old, landed in Zurich as a lecturer in physics. His path to his new post was anything but easy: having lost his father in battle at a young age, he found himself in an unfortunate marriage. Friends bailed him out of the union, ultimately helping him travel to France to study medicine. Returning to Zurich, Gessner had many loves, not least of which were the mountains of Switzerland. His passions were the beauty of nature and the physical exercise of climbing. Lured by the majesty of the snow line, he declared a goal of reaching the summit of a different challenging Swiss mountain each year.

Gessner developed an ardor for describing nature – first the plants, then the animals. His four-volume opus *Historiae Animalium*, published between 1551 and 1558, was remarkable for its richly detailed descriptions and illustrations of the world's living things. In this tome, Gessner did something that relatively few had done before him: he specifically compared entities inside rocks to bones and shells of contemporary organisms. He illustrated crabs, clamshells, and urchins, revealing that a number of rocks contained similar entities.

How could rocks look like living creatures? In the 1500s, answers to this question took several forms.[3] One common explanation was that the rocks held monstrosities that were destroyed in the great flood during Noah's time. Another theory, common in Gessner's day, was that life-like objects were produced by the same forces that made the rocks: some stones contained things that only coincidentally looked like wood, bones, and teeth. These objects were not associated with creatures alive or dead because they were considered natural outgrowths of the rock itself. The other explanation was that fossils reflected a kind of Loch Ness phenomenon: perhaps they were mysterious animals that could be found alive in remote or unexplored regions of the planet.

All these conceptions changed in 1666, when fishermen working on the coast of Italy caught a giant shark. This monster

from the deep drew the attention of the Grand Duke of Tuscany, who was a great patron of science. He ordered it sent to Niels Stensen, one of the young scientists he was supporting at the time. Stensen (known in his publications by the latinized *Steno*) studied medicine and had originally earned the Duke's favor for his extraordinary knowledge of anatomy and his clever use of experimentation to understand how bodies function. Steno described the bones, muscles, and nerves of the shark head, but his most memorable observation came from studying teeth.

A long-standing puzzle, dating from before Steno's time, was the presence of oddly shaped objects, known as "tongue stones," commonly found in clumps on the ground or still embedded in rocks. These stones had an uncanny resemblance to shark teeth. The Roman natural historian Pliny the Elder declared that they either fell out of the sky or dropped from the moon. Others followed the standard interpretation, viewing them as natural outgrowths of rocks. Steno changed everything. He looked not only at the stones, comparing them to teeth, but also at the rocks in which they were found. He noted that the stones were recovered from cliffs made up of layer after layer of rock, one on top of another. To Steno, tongue stones were actually shark teeth, and not just any shark teeth – they were ancient shark teeth, buried under layers of sediment. Steno developed a theory about what the stones were and how they were preserved. The completion of his shark monograph in 1667 was an important moment in the birth of paleontology as a discipline.[4]

Viewing fossils as the remnants of past life, and layers of rock as reflecting a succession of ages, is a relatively new way of understanding the planet and life on it. The rocks of the world are a library of sorts, whose pages and chapters record eons of time. Layer after layer reflect changes in the atmosphere, climate, and geography of the planet, and the fossils inside provide a window into the succession of living things.

This approach is not just a new way of thinking; it also is a means of discovering. Since the days of Steno and his contemporaries, paleontologists have used a growing knowledge of the world to identify places likely to yield fossil discoveries. While paleontological discovery is often accidental – for example, by construction or road crews hitting fossil bones as they unearth rock – most discoveries are planned. That is, to examine a question or problem, such as determining links between fish and amphibians, we begin by narrowing down the mapped regions of Earth to small sites where fossils might be found. The approach is straightforward: find places where rocks of the right age and the right type to preserve fossils of interest are exposed at the surface. Economic incentive fuels part of this search: geological surveys spurred by the potential of oil, gas, and mineral development often prompt states and private industries to map the rocks within their purview. Geological maps, commonly made at a very fine scale, are often easy to come by, as are aerial photographs that reveal the exposures in any given area.

Knowing a few relatively simple features about the geological landscape greatly enhances the odds of finding new fossils. The first is rock type. Sedimentary rocks are best: unlike igneous or metamorphic rocks, they have not been superheated or transformed by the tremendous pressures that exist within the Earth. They may have been laid down in ancient oceans, streams, soils, or even sand dunes.

Understanding how grains sort inside a rock, as well as how different layers of rock change relative to one another, can give clues to the kind of environment in which

the rock was deposited. Stream beds, for example, can have a lenticular shape, almost like the cross section of a channel bed. Within ancient channels are often grains that range from rounded cobbles to fine particles. The size of these sediments and the way they are sorted within the deposit reveal not only if the rock was deposited in a stream, but also whether the channel was big or small and if the water was running fast or slowly. By studying these features, fossil hunters can predict where they might find fossils in the field. The likelihood of finding a complete articulated skeleton in the middle of an ancient channel bed is vanishingly small: moving water may have scattered and broken bones, particularly smaller and fragile ones. When looking for high-quality bones in freshwater settings, we tend to focus on the margins of streams, in the eddies and banks where matter would settle out at different times of the year. Field paleontologists develop a catalog of rock occurrences like these and, when on the rocks, will typically make a beeline for their favorites.

Finding new places to look also means predicting the right age of rock to investigate. Here, the full suite of biological information comes into play. To appreciate this approach, we must look back to a time in biology when the study of DNA and the study of fossils were utterly separate approaches to science.

Biology is a vast field encompassing multiple levels of organization, from molecules, to genes and cells, to entire ecosystems. Work on each of these levels has its own empirical approach: different tools of microscopy, spectroscopy, and field analysis underlie disciplines as varied as structural biology and ecology. In the 1970s, biology became increasingly fragmented according to level and approach, with a number of prominent, comprehensive biology departments splitting into the more narrowly focused departments of molecular biology, cell biology, organismal biology, and so on. But at the same time that the biological disciplines were becoming more divided, the questions and conceptual tools to unite them began to emerge with increasing rapidity.

One of the greatest boons to paleontology originally appeared to be the most significant threat to its existence as a productive field of inquiry. The approach began with a simple notion of Darwin's: descent with modification.[5] If there is a common history to life on Earth, then descendants should be modified versions of their ancestors. Just as each individual is a modified descendant of its parents, so, too, should species be descendants of their ancestors. This pattern of descent with modification should yield a pattern in the history of life: that is, the features that creatures share with one another should reflect their history. For example, if we wanted to know how clams, fish, mice, and people are all related, we would compare their characteristics and discover that these creatures share DNA, cells, and other features, but that fish, mice, and people share characteristics absent in clams – such as a backbone, skull, and centralized brain. Mice and people are even more similar; unlike fish, they share hair, warm bloodedness, and mammary glands. We could even look at the genes and proteins of each of these animals and come to the same conclusion: fish, mice, and people are more closely related to one another than any are to clams. Moreover, mice and people have more in common with each other than they do with fish.[6] By adding species and features to this analysis, we could ultimately develop a hypothesis of a complete tree of life.

The important point to draw from this exercise is that we do not need a single fossil to infer the relationships among living

things. By applying the "descent with modification" approach to the relationships discussed above, we can infer that mice and people share a more recent common ancestor with each other than they do with fish. Armed with a knowledge of genes, tissues, and organs of living creatures, we can infer the hierarchy of life – a tree of relatedness that shows the relative recency of common ancestry.

Far from removing fossils from the picture, however, this approach defines their importance. Plotting the relationships between living vertebrates helps us construct the family tree, for example, by demonstrating that turtles and lizards are more closely related to mammals than birds are. But this method alone cannot tell us what the ancestors of mammals looked like. When we explore the fossil record from rocks 230 million years old, we find a number of creatures with reptilian jaws and skulls, but with a dog-like posture. These creatures have features of the ear, teeth, and skull that reveal intermediate conditions between so-called reptilian bodies and those of mammals. Fossils bring to light transitional features, ancient environments, and ecosystems that have been lost in time.[7]

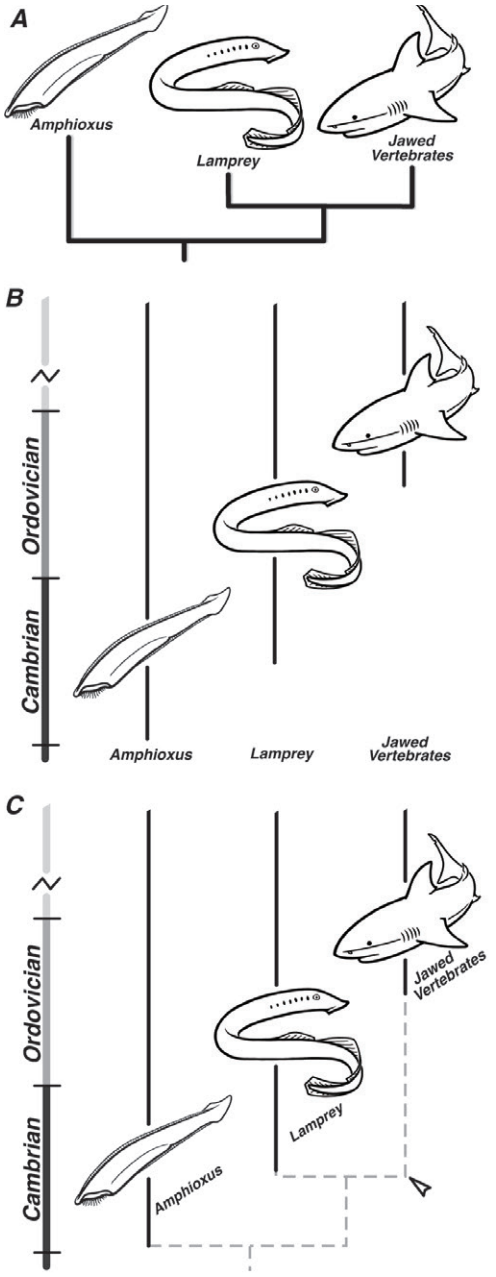
Together, genes and fossils provide information that each alone cannot. If you take a tree of relatedness developed from genes, or from any kind of data (Figure 1a), and map known fossil occurrences onto it (Figure 1b), the end result is a clear picture of what is unknown (Figure 1c). By looking for these so-called ghost taxa[8] – extinct species we infer should be present but are absent – we can concentrate our field efforts to fill huge gaps in the fossil record with transitional forms. There is a deep beauty to the idea that comparisons of DNA in different species can give us clues about where to discover new fossils inside rocks.

In the 1940s, no approach helped paleontologists understand the origin of whales. With nostrils modified to become blowholes, no hind limbs, and extreme modifications of the brain case, whales were a complete enigma. They were so odd that they could not easily be compared to any creature – living or extinct. The problem was so great that the paleontologist George Gaylord Simpson inserted the group arbitrarily into his classic 1945 classification of mammals, saying that cetaceans are "the most peculiar and aberrant of mammals," and adding that "there is no proper place for them in a scale naturae."[9]

A few years later, in the late 1940s, two scientists used a crude test to look at the similarity of proteins in the blood of different mammals.[10] Using an assay that criminologists employed to discern human from animal blood at a crime scene, they tested the plasma of different species by looking at how they interacted with antibodies. Closely related species should have more similar antibody reactions than more distantly related ones. The scientists found that the proteins in the blood of whales were more similar to those of even-toed ungulates – including deer, hippos, and goats – than to anything else. But this was a puzzling discovery. These creatures, artiodactyls, have a very distinctive ankle bone, consisting of a double pulley joint that helps them with their running and bounding gaits. Extant whales not only had no ankle bones with a pulley joint, but they had no hind limbs whatsoever. So how could there be a connection? Simpson's problem remained.

As new techniques to compare genes and proteins emerged in the ensuing decades, scientists gained a bonanza of new data to compare whales with other mammals. By the mid-1990s, mitochondrial genes,[11] milk casein genes,[12] and others not only strengthened the artiodactyl idea

Figure 1
Filling Gaps



Using genes to explore for fossils, (a) an evolutionary tree can be constructed for living creatures, such as sharks, jawless fish, and their closest invertebrate relatives; (b) the fossil representatives of each form can be mapped in time; and (c) merging the tree and the fossil occurrences reveals places in the geological record where fossils are likely missing (indicated by the dotted line and arrow). Source: Figure created by John Westlund, University of Chicago; used here with permission from Westlund.

but led to the proposition that one group, hippos, are the closest living relative of whales. Yet fossil data spoke to a different theory of whale relationships, although not conclusively. Comparison of the teeth and skulls of whales to other mammals suggested a relationship to an extinct group of terrestrial, four-legged creatures known as mesonychids. In fact, everything known of the anatomy suggested that artiodactyls are only distantly related to whales. As the authors of one of the genetic studies noted, "paleontological information is grossly inconsistent with [the artiodactyl] hypothesis."[13]

About ten years before this flurry of molecular work, Philip Gingerich and his colleagues were investigating fossil exposures in Pakistan. Gingerich had followed the paleontological rulebook: the rocks, at about forty-seven million years old, were the right age (they reflected the interval when the diverse orders of mammals came about); were the right type (they were mapped as ancient stream deposits); and were well exposed. Gingerich, however, was working from an inaccurate map, and once on-site, he realized that instead of stream beds, he was standing on an ancient ocean. That setback did not stop him from looking for fossils anyway. He and his team found a number of new fossils, including pelvic bones that the team jokingly called "walking whales." A few years later, the rocks yielded whale fossils in the form of some isolated skulls.[14] But those pelvic bones remained enigmatic.

With whale origins now on his mind, Gingerich shifted his focus to Egypt, the home of well-exposed marine rocks from a slightly younger age. Sure enough, the team discovered whales. In addition, and fittingly for the Darwinian theory, these whales had hind limbs. This was a gratifying and important discovery, but not entirely unexpected under Darwinian thinking. Because whales share a common ancestor with other mammals, their close relatives must have been quadripeds.

Then, as one of Gingerich's graduate students was cleaning a fossil whale skeleton in preparation for its extraction back to the lab, a small, pulley-shaped bone appeared to poke out of the rock. Once removed and cleaned, the bone was clearly identifiable as an ankle bone. And this was not just any ankle bone, but one from the double-pulleyed ankle of an artiodactyl. Armed with the new fossils showing the transformational character of evolution, we are now in a position to understand how the whale's unique body plan arose and what the ecosystems it lived in looked like during the change. A prediction, born of blood samples and extended to proteins and genes, was confirmed inside ancient rocks.[15]

We are accustomed to thinking of a revolution in gene sequencing and molecular technology, but we are also experiencing one in the field of paleontology. Whales with legs are one of a number of creatures that tell us of the great transformations in the history of life. Using the paleontological playbook, expeditions have discovered worms with heads,[16] fishes with elbows, wrists, and necks,[17] feathered dinosaurs,[18] and human precursors,[19] to name only a few. Indeed, in the last twenty years, we have discovered more creatures informative of evolutionary transitions than in the previous millennium.

Exploratory paleontologists such as Phil Gingerich use knowledge of evolutionary history and the geological record to find evidence of ancient transitions. Another record altogether can provide clues. In the late 1990s, David Kingsley and Katie Peichel began a hunt for the ideal species to study the way traits and genes evolve in natural populations.[20] Ever since the days of T. H. Morgan, biologists have used so-called model organisms, such as fruit flies,

house mice, and African clawed frogs, to provide insights into basic questions of genetics, cell biology, and development. Laboratory species have features that make studying their basic biology accessible: they typically breed rapidly and easily, have anatomical or behavioral features that might provide general insights, and are tractable to study using molecular, microscopic, and cellular methods. Kingsley and Peichel had an additional goal: they wanted to find a creature that would allow them to trace the genes involved in the origin of new organs, physiological processes, and behaviors. Their search revealed the potential of a fish, ranging from one to four inches long, called the threespine stickleback.

The threespine stickleback is an ordinary-looking fish with a long history of study. The famed Dutch ethologist Niko Tinbergen won a Nobel Prize in part for his work on them. Ecologists and paleontologists have had their turn at the species, too, producing a vast literature that contains thousands of scientific papers and analyses. To Kingsley and Peichel, the stickleback had all the characteristics of an excellent genetic system: the creatures breed easily and develop relatively rapidly. But most interesting was the tremendous variety of subspecies of threespine sticklebacks that have evolved since the glaciers retreated fifteen thousand years ago. As the ice gave way, new lakes and streams emerged. From their ancestral marine range, migratory ocean sticklebacks invaded or became restricted to different streams and lakes, often becoming isolated and evolving a number of important characteristics. The ecological and physiological environment of freshwater forms is so different from the denizens of the oceans that the freshwater sticklebacks evolved a number of new features – losing their protective armor, changing their feeding structures, and sometimes reducing or losing hind limbs, among a host of other new traits.

The key point is that the differences between freshwater and marine sticklebacks are so large that, for all intents and purposes, they could be characterized as different species. However, although they are often reproductively isolated in the wild, these very different kinds of sticklebacks can still be coaxed to interbreed under the right conditions in the laboratory. Thus, the team could interbreed the animals and identify the chromosomal regions responsible for the differences among natural populations. By breeding the different kinds of fish and analyzing their genetic structure, Kingsley, Peichel, and their colleagues could trace how changes at the genetic level were associated with dramatic changes in the body and physiology of the new kinds of stickleback.

One of the novelties that distinguishes many freshwater from marine sticklebacks is a reduction in the pelvis and the pelvic spines that attach to it. Marine sticklebacks live with a number of predators, and the presence of big pelvic spines is one defense to avoid being eaten. Freshwater sticklebacks, on the other hand, frequently evolve in environments that lack the soft-mouthed predators found in the ocean. Moreover, because fin skeletons are metabolically expensive to develop, the freshwater fish often have smaller pelvises and spines, or they lose these features altogether. With this information as their inspiration, Kingsley, Peichel, and colleagues set off to collect sticklebacks for breeding experiments that would identify the genetic region responsible for the loss of the pelvis and spines in different populations. Much like Gingerich homing in on sites to find fossil whales, Kingsley and colleagues chose the right places on Earth to obtain the sticklebacks.

The resulting genetic analysis revealed a number of chromosomes involved in the

reduction of the pelvic appendage. But in terms of relevant data, one site reigned supreme: the region responsible for most of the limb loss[21] contained the famous Pitx1 gene. Pitx1 was known in mammals and fish to be involved in the development of tissues across the body, from heads to appendages.

When the team looked at the differences in the gene itself, they found that the DNA sequence of Pitx1 was largely unchanged between marine and freshwater fish.[22] At first glance this finding seems utterly strange: how can Pitx1 be involved with a major anatomical change like loss of the pelvis if the gene itself does not have any recognizable differences among the different fish? If the structure for the gene is not the culprit, perhaps the loss of pelvic fins relates to a change in the elements that control the activity of the gene. Genes often have one or more outside elements that serve as a kind of switch determining when and where the gene is active. Some of these regions, known as regulatory elements, are highly specific to one organ or tissue. Changes to the regulatory elements can bring about a modular change specific to one region. By contrast, a change in the sequence in the gene itself could have an effect everywhere the gene is active. Imagine a house with one furnace but different thermostats in each room. A change to the furnace would affect the entire house, a change to a thermostat only a single room. The same is true with the genes and their regulatory elements.

Detecting regulatory elements is difficult and involves fusing DNA sequences with visible labels in order to determine where particular sequences are active, manipulating the DNA to see what happens when a region is deleted or changed, and sometimes swapping DNA between species and individuals. With this tool kit, Kingsley's group identified a relatively short stretch of DNA that serves as a regulatory switch controlling Pitx1's expression only in hind fins.[23] A mutation in this region – the "thermostat" for a single location in the body – leads to loss of fin development in the hind fin while preserving other functions of Pitx1. The difference between freshwater fish that lack pelvic fins and their marine cousins that retain them lies largely in the stretches of DNA that control gene activity. This makes sense: a change in the structure or sequence of the gene would affect every tissue in which the gene is active. Given that Pitx1 has global effects, a change is likely to be harmful, if not lethal. The tissue-specific changes in gene activity mean that fins can change independently of the rest of the body.

Not only can this area of regulatory DNA be identified and its function mapped, but it can be swapped between different kinds of stickleback.[24] When the Kingsley group took the Pitx1 regulatory element from a stickleback with a complete pelvis and inserted it into an individual from a population that had lost the pelvis, something remarkable happened: like a ghost from the past, the pelvis appeared.[25] Kingsley and his colleagues swapped genes to make a fossil of sorts.

Little sticklebacks may open a window to great transformations, perhaps even to the hind-limb loss we see in fossils like those discovered by Gingerich. The more we look, the more we find similarities in the regulatory genes that underlie the development of tissues, organs, and the architecture of the bodies of diverse animals. Pitx1 is no different; it is seen in mammals, fish, lizards, and birds. And the gene leaves a signature of its activity in the limb, showing a preference for one side of the body over the other. Given this clue, Kingsley and his coauthors suppose that modifications of the regulation of this gene may underlie limb reduction in

many other creatures, including aquatic mammals such as manatees.[26] Indeed, mutations in Pitx1 activity cause a range of limb malformations in mammals, such as clubfoot in human infants.[27]

This story is more general than Pitx1, manatees, whale fossils, or even human skeletons. By leveraging the genetic and geological record to discover fossils, and moreover, by using molecular biology to isolate genes underlying evolutionary change and test their effects in the laboratory, the study of great transformations in the history of life can look forward to a future as a predictive science. In the coming years, it is not unlikely that we will be able to study evolution in the distant past both by finding fossils with increasing precision and by reconstructing evolution's effects, either in part or in full, in the laboratory.

The layers of crust on Earth, like the genes, cells, and DNA of every living thing, are chronicles of history. But rocks, bodies, and genes are not independent records of time; they are linked by billions of years of planetary and biological evolution. Every living thing is the most extreme tip of a branch of an almost boundless tree of life; and all living creatures contain artifacts of a history nearly as ancient as the planet. There is something almost poetic to the notion that 3.5 billion years of change has brought one of these species to a moment when it can see its own past and grasp the deep interconnections embedded in the world around it.

ENDNOTES

1 Barbara R. Jasny and Laura M. Zahn, "A Celebration of the Genome, Part IV"; Eric S. Lander, "The Accelerator"; Peter Donnelly, "Making Sense of the Data"; David Botstein, "Fruits of Genome Sequencing for Biology"; Yijun Ruan, "Presenting the Human Genome: Now in 3D!"; Steven E. Hyman, "The Meaning of the Human Genome Project for Neuropsychiatric Disorders"; Mary-Claire King, "A Healthy Son"; Vololona Rabeharisoa, "Socializing Genetic Diseases"; Liz Lerman, "The Genome Dances"; and Steve Gano and Ro Kinzler, "Bringing the Museum into the Classroom," all in "Essays on Science and Society," *Science* 331 (2011): 1024–1029.

2 Martin J.S. Rudwick, *The Meaning of Fossils: Episodes in the History of Paleontology* (Chicago: University of Chicago Press, 1985).

3 Ibid.

4 Ibid.

5 Willi Henning, *Phylogenetic Systematics* (Champaign: University of Illinois Press, 2000).

6 Ian J. Kitching, Peter L. Forey, Christopher J. Humphries, and David M. Williams, *Cladistics: Theory and Practice of Parsimony Analysis* (Oxford: Oxford University Press, 1998).

7 Michael J. Donoghue, James A. Doyle, Jacques Gauthier, Arnold G. Kluge, and Timothy Rowe, "The Importance of Fossils in Phylogeny Reconstruction," *Annual Review of Ecology and Systematics* 20 (1) (1989): 431–460. See also, Philip C.J. Donoghue and M. Paul Smith, eds., *Telling the Evolutionary Time: Molecular Clocks and the Fossil Record* (Boca Raton, Fla.: CRC Press, 2003), chap. 5.

8 Ibid.

9 George G. Simpson, "The Principles of Classification and a Classification of Mammals," *Bulletin of the American Museum of Natural History* 85 (1945): 1–350.

[10] Alan A. Boyden and Douglas G. Gemeroy, "The Relative Position of the Cetacea Among the Orders of Mammalia as Indicated by Precipitin Tests," *Zoologica* 35 (1950): 145 – 151.

[11] Michel C. Milinkovitch, Guillermo Ortí, and Axel Meyer, "Revised Phylogeny of Whales Suggested by Mitochondrial Ribosomal DNA Sequences," *Nature* 361 (6410) (1993): 346 – 348.

[12] John Gatesy, Cheryl Hayashi, Mathew A. Cronin, and Peter Arctander, "Evidence from Milk Casein Genes that Cetaceans are Close Relatives of Hippopotamid Artiodactyls," *Molecular Biology and Evolution* 13 (7) (1996): 954 – 963.

[13] Ibid.

[14] Philip D. Gingerich, Neil A. Wells, Donald E. Russell, and S. M. Ibrahim Shah, "Origin of Whales in Epicontinental Remnant Seas: New Evidence from the Early Eocene of Pakistan," *Science* 220 (4595) (1983): 403 – 406.

[15] Philip D. Gingerich, Munir ul Haq, Iyad S. Zalmout, Intizar H. Khan, and M. Sadiq Malkani, "Origin of Whales from Early Artiodactyls: Hands and Feet of Eocene Protocetidae from Pakistan," *Science* 293 (5538) (2001): 2239 – 2242.

[16] Jun-Yuan Chen, Di-Ying Huang, and Chia-Wei Li, "An Early Cambrian Craniate-Like Chordate," *Nature* 402 (6761) (1999): 518 – 522.

[17] Edward B. Daeschler, Neil H. Shubin, and Farish A. Jenkins Jr., "A Devonian Tetrapod-Like Fish and the Evolution of the Tetrapod Body Plan," *Nature* 440 (7085) (2006): 757 – 763.

[18] Qiang Ji and Shu-An Ji, "On Discovery of the Earliest Bird Fossil in China and the Origin of Birds," *Chinese Geology* 233 (1996): 30 – 33; Qiang Ji, Mark A. Norell, Ke-Qin Gao, Shu-An Ji, and Dong Ren, "The Distribution of Integumentary Structures in a Feathered Dinosaur," *Nature* 410 (6832) (2001): 1084 – 1108; Philip J. Currie, Qiang Ji, Mark A. Norell, and Shu-An Ji, "Two Feathered Dinosaurs from Northeastern China," *Nature* 393 (6687) (1998): 753 – 761; Alan H. Turner, Peter J. Makovicky, and Mark A. Norell, "Feather Quill Knobs in the Dinosaur Velociraptor," *Science* 317 (5845) (2007): 1721; Xing Xu, Xiaoting Zheng, and Hailu You, "Exceptional Dinosaur Fossils Show Ontogenetic Development of Early Feathers," *Nature* 464 (7293) (2010): 1338 – 1341.

[19] Tim D. White, Berhane Asfaw, Yonas Beyene, Yohannes Haile-Selassie, C. Owen Lovejoy, Gen Suwa, and Giday WoldeGabriel, "*Ardipithecus ramidus* and the Paleobiology of Early Hominids," *Science* 326 (5949) (2009): 64, 75 – 86; Michel Brunet, Franck Guy, David Pilbeam, Daniel E. Lieberman, Andossa Likius, Hassane T. Mackaye, Marcia S. Ponce de León, Christoph P.E. Zollikofer, and Patrick Vignaud, "New Material of the Earliest Hominid from the Upper Miocene of Chad," *Nature* 434 (7034) (2005): 752 – 755; Martin Pickford, Brigitte Senut, Dominique Gommery, and Jacques Treil, "Bipedalism in *Orrorin tugenensis* Revealed by its Femora," *Comptes Rendus Palevol* 1 (4) (2002): 191 – 203.

[20] Catherine L. Peichel, "Fishing for the Secrets of Vertebrate Evolution in Threespine Sticklebacks," *Developmental Dynamics* 234 (4) (2005): 815 – 823.

[21] Michael D. Shapiro, Melissa E. Marks, Catherine L. Peichel, Benjamin K. Blackman, Kirsten S. Nereng, Bjarni Jónsson, Dolph Schluter, and David M. Kingsley, "Genetic and Developmental Basis of Evolutionary Pelvic Reduction in Threespine Sticklebacks," *Nature* 428 (6984) (2004): 717 – 723.

[22] Ibid.

[23] Yingguang Frank Chan, Melissa E. Marks, Felicity C. Jones, Guadalupe Villarreal, Jr., Michael D. Shapiro, Shannon D. Brady, Audrey M. Southwick, Devin M. Absher, Jane Grimwood, Jeremy Schmutz, Richard M. Myers, Dmitri Petrov, Bjarni Jónsson, Dolph Schluter, Michael A. Bell, and David M. Kingsley, "Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer," *Science* 327 (2010): 302 – 305.

[24] Ibid.

[25] Ibid.

[26] Michael D. Shapiro, Michael A. Bell, and David M. Kingsley, "Parallel Genetic Origins of Pelvic Reduction in Vertebrates," *Proceedings of the National Academy of Sciences USA* 103 (37) (2006): 13753 – 13758.

[27] Christina A. Gurnett, Farhang Alaee, Lisa M. Kruse, David M. Desruisseau, Jaquline T. Hecht, Carol A. Wise, Anne M. Bowcock, and Matthew B. Dobbs, "Asymmetric Lower-Limb Malformations in Individuals with Homeobox PITX1 Gene Mutation," *The American Journal of Human Genetics* 83 (2008): 616 – 622.

Deciphering the Parts List for the Mechanical Plant

Chris Somerville

Abstract: The development of inexpensive DNA sequencing technologies has revolutionized all aspects of biological research. The proliferation of plant genome sequences, in conjunction with the parallel development of robust tools for directed genetic manipulation, has given momentum and credibility to the goal of understanding several model plants as the sum of their parts. A broad inventory of the functions and interrelationships of the parts is currently under way, and the first steps toward computer models of processes have emerged. These approaches also provide a framework for the mechanistic basis of plant diversity. It is hoped that rapid progress in this endeavor will facilitate timely responses to expanding demand for food, feed, fiber, fuel, and ecosystem services in a period of climate change.

CHRIS SOMERVILLE is the Philomathia Professor of Alternative Energy and Director of the Energy Biosciences Institute at the University of California, Berkeley. His current research focuses on plant cell-wall polysaccharide synthesis and conversion to liquid fuels. His work has appeared in *Science*, *Proceedings of the National Academy of Sciences*, and *Current Biology*, among other publications.

As the end of the previous millennium drew near, I accepted an invitation from a leading biomedical journal to summarize the major advances in knowledge of mechanistic plant biology during the preceding one hundred years.[1] By *mechanistic*, I mean an intellectual framework that seeks to understand an organism as the sum of its parts: that is, in terms of the chemical structures and reactions that support life. One of the challenges of summarizing this vast topic was disentangling the many advances that revolutionized our understanding of plants nonspecifically – a rising tide of knowledge that lifted all boats. In spite of their long separation from a common ancestor, plants and animals (and fungi and bacteria) contain many proteins and genes with significant structural similarity. Thus, much of what is known about the function of plant proteins and genes, and the molecular and cellular processes they participate in, has been inferred from knowledge of the function of homologous genes in other types of organisms. In this sense, the study of plant biology is a subset of a broad campaign to understand all life forms. However, plants and animals are thought to have separated about 1.6 billion years ago from a

common unicellular ancestor; so in addition to the fundamental differences in how they obtain energy, there are many interesting differences in how multicellularity and adaptive responses evolved.

After surveying the previous century, I concluded that early biologists were astute observers, and that succeeding generations had not obviously improved on that aspect of inquiry but had better experimental tools, analytical devices, and context. One key factor underlying major advances was technological improvements that facilitated compelling experiments. Thus, for instance, Calvin's elucidation of the path of carbon during photosynthesis was enabled by the availability of the newly discovered ^{14}C isotope from the nearby Berkeley accelerator. Similarly, advances in plant biology over the past ten years were also largely attributable to improvements in one technology – DNA sequencing – and one sociological phenomenon: enthusiasm for model species. In 2000, an international consortium completed the first full genome sequence of a higher plant, *Arabidopsis thaliana*, and placed it in the public domain via a graphical Web interface that allowed users to browse the genome and connect individual genes to the scientific papers describing their functions. Because it was completed before the development of the very high-throughput sequencing technologies that are widely used today, the DNA sequence is estimated to have cost more than $75 million to produce. By comparison, an essentially complete DNA sequence of an Arabidopsis plant can be obtained today for about $5,000 – a dramatic testament to the development of DNA sequencing technology during the last decade.

The second major factor underlying progress in mechanistic plant biology was the widespread adoption of Arabidopsis as a model organism. Interest in Arabidopsis began in the early 1980s, when the first generation of plant biologists to engage in gene cloning and characterization recognized the virtues of an easily cultivated organism with a small diploid genome and a short life cycle. In conjunction with some parallel developments in recombinant DNA technologies, the availability of the Arabidopsis genome sequence changed mechanistic plant biology more profoundly than any plant-specific discovery of the previous century.

To understand why the genome sequence combined with the widespread use of model species is enabling, it is useful first to reflect on how the mechanistic aspects of plants, and most other model organisms, are currently understood. In brief, all organisms can be viewed as adaptive machines that exist to make copies of themselves, or hybrid copies of themselves and their sexual partners. DNA encodes a parts list and some instructions for how many parts to make under various circumstances. Some parts – proteins and RNA molecules – have characteristic lifetimes that may vary by several orders of magnitude and according to information about where to locate themselves within a cell, which other parts to interact with, and how to carry out those interactions. Proteins (and to a lesser extent some RNAs) have the ability to make or modify other parts – usually simple chemicals such as lipids, amino acids, nucleic acids, and sugars that are the building blocks of cells – or to carry excited electrons that are used to power life.

A long-term goal of many plant biologists is to understand what each of the roughly 33,600 genes in Arabidopsis and other plants does and to integrate that information into a predictive mechanistic model. Ideally, this would be a computer model: the cyberplant. Given that at least sixteen thousand scientists worldwide use Arabidopsis as a model species

for research, we could collectively obtain detailed experimental information about the function of every gene in a plant like Arabidopsis within a decade. We could meet this objective by organizing the community to eliminate duplication of effort and maximize information sharing. I believe that a first pass at a complete description of a plant is well within reach. Because all flowering plants (angiosperms) evolved from a common ancestor within the past 125 million years, the mechanisms underlying many aspects of growth and development have been conserved at the molecular level. Thus, a detailed analysis of all genes in several strategically selected angiosperms will provide a broad base of knowledge that is applicable to all higher plants. Although I have emphasized Arabidopsis here because it is the most advanced model plant, I anticipate that similar approaches will be implemented in several other species – for example, rice (*Oryza sativa*) – that represent divergent nodes of angiosperm diversity, and that knowledge of all plants will ultimately involve interpolation from deep knowledge of strategically placed nodes.

The standard method for interrogating gene function is to increase or decrease the activity of the gene and observe how that change affects relevant processes or the organism as a whole. Activity levels can be altered by genetically increasing or decreasing the amount of the mRNA or protein encoded by the gene or, frequently, by altering the catalytic activity of the protein, its ability to bind other proteins, or its location. In one technique, the so-called reverse genetics approach, the investigator replaces an endogenous gene with an altered copy and observes the effect. The challenge of this powerful approach is knowing which aspect of the organism to test for an effect. Indeed, the investigator may find it necessary to become an expert in all aspects of biology in order to design useful tests. Thus, the more broadly useful approach has been to randomly mutagenize the genome, then screen for mutants in which a process of interest is altered. This approach exploits the investigator's deep knowledge of a specific aspect of biology and facilitates the testing of many variants of a gene, revealing not only more or less of the gene product but also more subtle effects, such as loss of regulatory factors. The challenge in this case is to identify the corresponding gene. Additionally, both approaches may be complicated by the presence of duplicated genes that mask the effects of a mutation in only one of the genes. The two approaches are complementary and are frequently used simultaneously.

The availability of a complete genomic DNA sequence makes it relatively easy to identify the mutation corresponding to any genetic difference between two accessions of Arabidopsis. In the simplest cases, a researcher might identify an interesting new mutant and, by genetically mapping the mutation underlying the phenotype to the DNA sequence, can identify the corresponding gene. Similarly, natural variation between accessions can be resolved by genetic crosses to single genetic loci and mapped onto the DNA sequence to identify the basis for differences. Thus, any aspect of plant growth and development that can be marked by a mutation can be linked to a change in one or more specific genes. Because it has now become so inexpensive to obtain the complete genome sequence of an Arabidopsis plant, some researchers have resequenced the entire genome of the mutant – approximately 150 million base pairs – to find the change in DNA sequence corresponding to a mutation.[2] The relatively small size of the Arabidopsis genome compared to most other higher plants makes genome resequencing particularly easy. However, similar approaches are possible for many

species of higher plants, although some species are easier than others because of genome size, ploidy, self-incompatibility, and related features.

The benefits of having a large number of scientists working on one or a few model organisms are manifold. First, the community benefits from the availability of research tools and reagents of broad utility. Indeed, the mere fact that a large user community exists encouraged some scientists to invest large amounts of time and effort in building tools that would benefit the community. An important example is the "Arabidopsis insertion collection": several hundred thousand accessions of Arabidopsis in which a fragment of exogenous DNA is randomly inserted in the genome, frequently within a gene. To create the collection, several groups produced large numbers of transgenic plants with random insertions of exogenous DNA. Then, for each of the hundreds of thousands of DNA insertions, the DNA flanking the insertion was recovered and sequenced so that the location of the insertion in the genome could be determined. Seeds of each of the insertion mutants were made freely available at several stock centers around the world. Thus, when a researcher wishes to investigate the function of a gene, he or she can log into an electronic database and request seeds from one or more lines that specifically lack a functional copy of that gene. Variation in the properties of the gene can be explored by genetically transforming the mutant with natural or synthetic variants of the gene or its relatives.

As another benefit of the community approach, researchers studying different phenomena frequently discover that they are observing different aspects of a common process. Thus, for instance, a researcher who discovers a protein that catalyzes a specific chemical reaction might find that the protein was previously found to be required for some other process, such as disease resistance, by a colleague with no knowledge of the catalytic function of the protein. This kind of second-order knowledge creation, which is essential to the eventual development of a comprehensive understanding, is accelerating through the prevalence of electronic data resources and the expanding ability of the community to link biological information to specific genes.

Another DNA sequence-based approach that has revolutionized plant biology exploits very high-throughput RNA or DNA sequencing technologies or DNA hybridization methods in order to measure simultaneously the abundance of mRNAs for each gene in a tissue sample. At the most basic level, this method catalogs which genes contribute to the state of a tissue sample under the condition in which the sample was taken. Investigators can compare samples taken from different tissues or conditions to observe how the organism reshapes gene expression in order to respond to different developmental states or environmental conditions. Perhaps more important, by using computational methods to compare the data compiled from large numbers of experiments, it is possible to identify genes that are coregulated (that is, expressed at the same time and place). Searching for highly coregulated genes frequently allows investigators to identify previously unknown components of processes. For instance, by searching for genes that were highly coregulated with the known subunits of the enzyme complex that synthesizes cellulose, my colleagues and I recently identified genes that had not previously been implicated in the process.[3]

At present, this gene-centric approach sheds light primarily on isolated mechanisms rather than the operation of the organism as a whole. The field of plant

biology is to a large extent still in an inventory phase, in which the genes that contribute to all aspects of growth and development are being identified and ordered into networks and pathways. Thus, for instance, the genes that contribute to flower development or disease tolerance have been identified by searching for mutations that alter these processes, cloning the corresponding genes, assigning probable function to the genes by comparing the gene sequences to databases of all previously known genes, identifying the genes that are coregulated, analyzing mutations in those genes, and so on. Ultimately, these analyses allow placement of genes into pathways that sequentially carry out specific tasks. Listing the parts and placing them in pathways and networks will presumably be followed by a phase in which the many mechanisms that comprise a whole organism will be conceptually integrated. That phase seems likely to take place on computers that will generate testable hypotheses and identify where experimental measurements are needed to populate models and simulations. I expect that phase to mark the arrival of theoretical biology as a mainstream activity.

Mechanistic research on plant biology can be artificially subdivided into five major, intersecting topics: evolution, development, adaptation, biotic interactions, and molecular and cellular mechanisms. Developmental biologists are systematically describing the networks of genes and the cellular processes that underlie the seemingly miraculous development of a multicellular plant from a single cell. For the time being, much of the work focuses on describing how each tissue or cell type develops. For instance, the surface of plants is usually punctuated by the presence of large numbers (for example, thousands per cm^2) of pairs of cells (stomata) that open and close, like a pair of lips, to regulate the flow of gases and water vapor in and out of the leaf. Mutant analysis appears to have identified all the genes that are involved in this process. Specifically, careful examination of what fails to take place when each gene is altered has enabled biologists to observe the sequence of events and describe causes and effects in molecular detail. This effort has provided insight into how the differentiation of stomata from leaf epidermal cells takes place.[4] The current hope is that such knowledge will provide a road map for how other specialized cell types might develop. However, I think this paradigm is a stopgap measure based on the fact that we are in the midst of a large discovery process. I believe that in the longer term, scientists will describe in detail how every cell type develops in many plant species. In other words, unless we assume that research on plants will conclude at some point, the current use of models and examples will gradually be supplanted by complete descriptions of enough model species to allow predictive manipulation of any plant species.

As I use it here, *adaptation* refers to the ability of an organism to modify its morphology or composition in response to environmental cues. Because plants are sessile, most have a large repertoire of adaptive responses. For instance, in response to attacks by pests and pathogens, plants may activate pathways for the production of toxins. In response to low temperature or drought, some plants undergo a wide variety of changes in chemical composition that facilitate survival under those conditions. If exposed to toxic minerals, plants induce the expression of factors that sequester the toxins. Because light quality varies at different locations in the canopy, plants may alter the composition of the photosynthetic apparatus to make better use of light, or they may

stimulate growth to facilitate better access to light. The ability to measure the expression of all genes simultaneously has greatly facilitated a description of underlying mechanisms involved in these and many other adaptive responses. At the same time, the development of genetic methods has similarly facilitated the identification of the genes that control and participate in such responses. These discoveries have triggered the rational development of genetically modified plants that are better able to withstand stressful conditions. The first generation of transgenic crop plants engineered to better withstand drought conditions recently obtained regulatory approval, and the approval of freezing-tolerant trees is pending.

One of the richest areas of discovery during the past decade has been the elucidation of many of the mechanisms plants use to survive biotic interactions. As any gardener knows, there is a large number of organisms that can devastate plants: viruses, nematodes, insects, fungi, bacteria, slugs, and vertebrates of many kinds. It is remarkable that any plants survive in the natural world. Indeed, approximately 40 percent of agricultural productivity in Africa and Asia is reportedly lost to pests and pathogens.[5] Thus, one of the most promising avenues to increasing the availability of food and fiber is to explore ways to reduce such losses. One of the first transgenic crops grown commercially employed an insecticidal protein to reduce losses to insects that were not controlled adequately by other methods, such as the application of insecticide. Interestingly, most plants are resistant to most pests and pathogens. Crop damage is largely caused by highly specialized pests and pathogens that have become adapted to only a few host species. Understanding the factors that allow pests and pathogens to identify their hosts might facilitate the development of molecular cloaks of invisibility. This is essentially how "mosquito repellents" containing DEET provide protection: the active ingredient blocks one or more of several receptors used by mosquitoes to identify a host.[6] Additionally, most plants have a suite of defensive responses to pests and pathogens. In the most extreme case, infection by a pathogen triggers the death of cells in the vicinity of the infection, thereby starving the pathogen. In this and most other kinds of defensive reactions, there is a cost to the host, so the defensive mechanisms are not activated until necessary. Many aspects of the mechanisms by which plants sense and respond to these events have been discovered during the past ten years, presenting new opportunities to breed or engineer pest and pathogen resistance.

Beyond the defense mechanisms of plants, much recent progress has contributed to a broad understanding of basic mechanisms that operate at the cellular and molecular levels. Some advances entail comparative biology, in which the details of a molecular process are first worked out in an organism that has advantages for basic research, and then the homologous mechanism in several plant species is described. However, many important aspects of plant biology are unique to plants and, therefore, cannot be approached by using convenient model species such as yeast, nematodes, flies or mice. Thus, understanding light-mediated signaling, phytohormone-mediated responses, some aspects of pathology, and aspects of development have been hot topics during the past decade. Some plant-specific subjects, such as photosynthesis and cell wall biosynthesis, attract relatively small followings at present; but trends shift, and the recent interest in lignocellulosic fuels has ignited new interest in plant cell-wall biochemistry.

Essentially all aspects of knowledge about plant biology have surged in the

ten years since the completion of the Arabidopsis genome sequence and the subsequent DNA sequencing of many other plant species. One important outcome of the proliferation of genome sequences has been insight into the mechanistic basis of diversity. For instance, following the completion of the poplar genome, poplar trees and Arabidopsis were found to have remarkably similar gene content.[7] The significant differences in morphology and life cycle are manifestations of differences in the regulation of a very similar suite of genes. Likewise, progress in understanding the molecular basis of flower development has provided fascinating insights into why the Linnaean system of plant classification, which is based on flower morphology, has been so broadly useful. Knowledge of how gene action leads to floral morphology explains how a small number of changes in key genes can lead to very large differences in morphology.[8] The present is an exciting time for evolutionary biologists, who can now trace with precision the DNA rearrangements that accompanied or gave rise to the formation of new species. Plants are characterized by a high degree of polyploidy, and the remnants of ancient genome duplications can be seen in their genome sequences. The dynamic nature of plant genomes has become clear.

Looking forward, I predict that research in plant biology will be shaped by several major trends that seem certain to have broad impacts and by some peculiarities of the field. A central fact of studying plants is that there are many species we care about (roughly 180 are used by humans for nondecorative purposes). In contrast to biomedical research, in which most resources are directed toward one species, much of the effort and resources in plant biology are used to translate knowledge gained from models into other species or to generate insights without the technical benefits of working on a model. Support for work on model species is under pressure both from agricultural commodity groups that favor devoting research funds to species of economic importance, and from a large community of biologists for whom the important questions in biology are related to diversity, ecosystem function, and levels of explanation far removed from molecular processes. Thus, for the foreseeable future, a lack of financial support for research means that we will not discover the function of all genes in the model species or integrate such knowledge into a coherent understanding of how the organisms function. Ideally, this important quest will trickle along to completion later this century. I recognize that it is also important to understand the mechanistic basis of plant diversity and to apply knowledge to the plants of utility; but it is unfortunate that national priorities, particularly the enthusiasm for military force, limit progress toward knowledge that has enormous potential to affect human well-being. Indeed, I think that we are entering an era in which we will recognize a pressing need for deep knowledge about all aspects of plant biology.

We cannot know the future, but we can make some predictions based on the intersection of four key trends: the expanding human population, the growing economies of some less-developed nations, the impact of climate change, and the declining rate of discovery of new petroleum reserves. Put simply, the expansion of the human population to nine or ten billion will require production of more food, feed, and fiber. Effects of climate change on the distribution of rainfall will make it more difficult to produce crops in some regions. Economic expansion is associated with increased demand for animal protein, which, in turn, strongly increases demand for animal feed. Finally, both climate change and

declining petroleum reserves have created interest in production of energy from biomass, creating potential competition for land that might otherwise be used for food and feed production. A large amount of land worldwide can be used to expand agriculture and forestry, including as much as a billion acres that was farmed in the past and later abandoned to agriculture. However, because of market failures and poverty, even if that land is brought back into production, it may not be enough to prevent expansion of agriculture or forestry onto land that has never been cultivated. In tension with this fact is the widespread awareness that natural ecosystems are an important resource that must be preserved in large contiguous blocks in order to maintain the biological diversity contained therein. In the case of the Amazon, it is also possible that a large region of forest must be preserved in order to maintain the climate on which the forest depends for existence – particularly in the face of concurrent climate change. These trends will create incentives to intensify and optimize the production of most types of domesticated plants, including trees, so as to restrain the expansion of agricultural land.

At present, most of the knowledge required to intensify and optimize plant production is not of the mechanistic variety. Indeed, some of the most important opportunities may arise from broad knowledge of which types of plants can be used to produce food, feed, fiber, or fuel on marginal land. However, I believe that when we are able to understand in detail how several plant species operate as machines, we will be able to predict how those machines can be modified through breeding or genetic engineering to optimize production in the many climatic zones, photoperiods, and soil types that are available around the world. We will know how to accelerate breeding from the current eight- or ten-year cycles that may involve tens of thousands of test plots and that are too expensive to facilitate development of improved cultivars of most crops. We will know how to breed or engineer plants to resist key pests and pathogens, or to withstand drought, or to grow on saline soils. We will know how to generate hybrid vigor to achieve the greatest possible yield every year, rather than only occasionally. Plant-derived foods will contain optimal balances of nutrients to maximize feed efficiency and human nutrition. Perhaps we will be able to convert important annual species into perennials so that the energy and environmental costs of annual tilling will be reduced. Ultimately, we will realize the inherent potential of each of the many domesticated plant species, and we will use mechanistic knowledge to accelerate the domestication of many more. Mechanistic knowledge will liberate us from the thousands of years of trial and error that have produced the domesticated species we rely on today.

ENDNOTES

[1] Chris Somerville, "The 20th Century Trajectory of Plant Biology," *Cell* 100 (2000): 13 – 25.

[2] Ryan S. Austin, Danielle Vidaurre, George Stamatiou, Robert Breit, Nicholas J. Provart, Dario Bonetta, Jianfeng Zhang, Pauline Fung, Yunchen Gong, Pauline W. Wang, Peter McCourt, and David S. Guttman, "Next-Generation Mapping of Arabidopsis Genes," *The Plant Journal* 67 (2011): 715 – 725.

[3] Staffan Persson, Hairong Wei, Jennifer Milne, Greer Page, and Chris Somerville, "Large-Scale Coexpression Analysis Reveals Novel Genes Involved in Cellulose Biosynthesis," *Proceedings of the National Academy of Sciences* 102 (2005): 8633 – 8638.

[4] Juan Dong and Dominique Bergmann, "Stomatal Patterning and Development," *Current Topics in Developmental Biology* 91 (2010): 267 – 297.

[5] George N. Agrios, *Plant Pathology*, 5th ed. (New York: Academic Press, 2005).

[6] Mathias Ditzen, Maurizio Pellegrino, and Leslie B. Vosshall, "Insect Odorant Receptors are Molecular Targets of the Insect Repellant DEET," *Science* 319 (2008): 1838 – 1842.

[7] Gerry A. Tuskan et al., "The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray)," *Science* 313 (2006): 1596 – 1604.

[8] Przemyslaw Prusinkiewicz, Yvette Erasmus, Brendan Lane, Lawrence D. Harder, and Enrico Coen, "Evolution and Development of Inflorescence Architectures," *Science* 316 (2007): 1452.

The Coming Epidemic of Neurologic Disorders: What Science Is – and Should Be – Doing About It

Gregory A. Petsko

Abstract: The Earth's population is aging fast, and the coming sharp increase in the number of people over age sixty-five will bring with it an epidemic of age-related neurodegenerative diseases, such as Alzheimer's and Parkinson's diseases. Currently, no cures exist for the major neurologic disorders. Unless cures can be found, by 2050 the cost of these diseases will exceed $1 trillion annually in the United States, and the burden for other countries will scale with their populations. Despite exciting advances in our understanding of these diseases, both government research funding and the efforts of industry have failed to keep pace with this unmet medical need. Private philanthropy has done better, but the total dollars spent on developing diagnostics and therapeutics for neurologic disorders still lags far behind that spent on much less prevalent diseases. The challenge for biomedical research in the next forty years is to identify markers that would allow early detection of high-risk cohorts, and to develop therapies that either will prevent the diseases from starting at all in susceptible populations or will arrest their progression before severe damage to the central nervous system has occurred.

GREGORY A. PETSKO, a Fellow of the American Academy since 2002, is Professor of Neurology at Weill Cornell Medical College and Tauber Professor of Biochemistry and Chemistry Emeritus at Brandeis University. His research interests include enzyme structure and function and the development of treatments for Alzheimer's, Parkinson's, and Lou Gehrig's diseases. For the past twelve years he has written a column on science and society that first appears monthly in the journal *Genome Biology*. A compilation of his columns has been published as *Gregory Petsko in Genome Biology: The First Ten Years* (2010).

What would you think if you were told that the entire population of the four largest cities in the United States had suddenly come down with an incurable, fatal disease? You would probably suspect a terrorist biowarfare attack or else the emergence of some horrible new strain of bird flu or Ebola-like virus. Barring a medical miracle, something very like that is certain to happen in about forty years' time. By 2050, the United States is predicted to have thirty-two million people over the age of eighty, and unless something is done to prevent it, about sixteen million of them will have Alzheimer's disease. That's more than the populations of New York, Los Angeles, Chicago, and Houston put together.

The United States is not alone in this explosion of the elderly and potentially infirm. Throughout the world today, there are more people aged sixty-five and older than the entire populations of Russia, Japan, France, Germany, and Australia – combined.

That seems like a lot, but it is miniscule compared with what is coming. From 1950 to 2050, the world population will have increased by a factor of 3.6; but those sixty and older will have increased by a factor of 10, and those eighty and older will rise by a factor of 27.

Figure 1 shows how the world's demographics are driving this outcome. Right now, only a handful of countries have 20 percent of their population over the age of sixty-five; by 2050, most countries will. That has never happened before in human history. For hundreds of thousands of years, the age distribution of the human population has been a pyramid, with a large number of healthy, productive young people at the bottom supporting a much smaller number of sicker, nonworking old people at the top. (That's right: life is a Ponzi scheme.) But a perfect storm of low birthrate in the developed world combined with increased life expectancy in most countries is causing the pyramid to invert. At the moment, the fastest-growing demographic group in the United States consists of octogenarians and above.

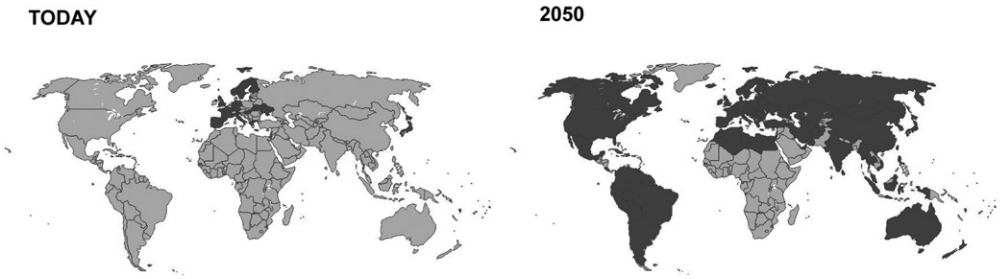
If older people were generally healthy, this trend would not be alarming. Unfortunately, age is a risk factor for most of Hamlet's thousand natural shocks that flesh is heir to: the incidence of most cancers, stroke, and heart disease rises with age, for example. But it is the major neurologic disorders that show the most dramatic age-dependence, with an incidence for most of them that increases exponentially after about age sixty. If you are fortunate enough to live to your mid-eighties, you have about a one in two chance of being unfortunate enough to suffer from either Alzheimer's or Parkinson's disease.

When we consider the financial and human costs of a future in which tens of millions of people will be afflicted with these devastating disorders, it becomes clear that we cannot afford to let that future materialize. Today, the economic cost of untreatable neurologic diseases exceeds $350 billion per year in the United States alone. By 2050, it is estimated that a single one of these diseases, Alzheimer's, will cost the country $1 trillion annually (the total Gross Domestic Product is currently about $15 trillion). Statistics for other developed countries scale by their populations.

The human cost is almost incalculable. Most Alzheimer's and Parkinson's patients are cared for at home, and the combined psychological and financial burden on their caregivers is crippling. It is estimated that there are fourteen to fifteen million unpaid Alzheimer's caregivers devoting seventeen *billion* hours to looking after a relative or friend with the disease, and that their combined economic loss is greater than $200 billion per year. These numbers will, of course, only increase as the number of patients increases during the coming decades. It used to be said during the Cold War years that you couldn't understand the Russian psyche until you appreciated the fact that, statistically, every family in the Soviet Union lost a close relative in World War II. In forty years' time, statistically, almost every family in the United States will have a relative with Alzheimer's or Parkinson's disease, and much of the rest of the developed world will be in the same situation.

How are governments and biomedical research funding agencies responding to this looming crisis? One has to say, not all that well. In the United States, for example, federal support for Alzheimer's research is, on a per-patient basis, about thirty times less than federal funding for HIV/AIDS research. (That is not to imply that AIDS funding is too great, but to suggest that Alzheimer's funding is woefully inadequate.) Encouragingly, President Obama recently announced a war on

Figure 1
Worldwide Increase in the Share of Population over the Age of Sixty-Five



Black shading denotes countries where those over the age of sixty-five make up more than 20 percent of the population today and are projected to in 2050. Data for this figure were taken from various sources, principally the World Health Organization and the United Nations. Although the demographic information for different countries is of varying reliability, the trend is clear. Source: Figure created by author.

Alzheimer's disease, with a goal of finding a cure by 2025. This is an admirable objective, and he has called for about $50 million in new research funds to support it in the coming budget year. Still, that is far below the more than $2 billion that would be needed to bring funding for this disorder into line with its present and future impact on the United States.

Nor has that darling of political conservatives, the private sector, exactly stepped up to the plate. Many drug companies, for reasons that will be explained below, have actually been quitting the central nervous system disease sector recently, despite the gigantic profits that would accrue to anyone first-in with an effective Alzheimer's or Parkinson's drug.

Ironically, this bailing out and inadequate funding is occurring at a time when unprecedented advances have been achieved in our understanding of these hitherto mysterious diseases, with new discoveries coming at an ever-increasing rate. To understand why the prospect of finding treatments has never been better, and yet the very companies that are needed to bring such treatments to the clinic appear to be scared off, it is useful to look at what we now know about neurodegenerative illnesses.

There are many disorders of the brain that are characterized by the slow dying of neurons, the nerve cells that send electrochemical signals through the brain's circuitry. The major disorders, in terms of prevalence, are (1) the dementias, which include Alzheimer's disease, Lewy Body dementia, Huntington's disease, and the so-called frontotemporal dementias; (2) the movement disorders, the chief of which is Parkinson's disease; and (3) the motor neuron diseases, of which the most common is amyotrophic lateral sclerosis, also known as ALS or Lou Gehrig's disease, after its most famous victim. Each of these diseases presents with symptoms that can be understood, at least in part, by the specific death of neurons in a particular region of the brain (although generally many other regions are also affected, particularly in the latter stages of the disease). Progression can be rapid, as in the case of ALS, where death typically occurs three to five years after the appearance of symp-

toms, or very slow, as in most cases of Parkinson's disease, which may progress to fatality over a twenty-year period. There is no cure for any of these diseases, and the few existing treatments either only relieve symptoms for a period of time or have very little effect on progression.

Despite these apparent differences in presentation and primary site of degeneration, it is now clear that there are two striking commonalities among these disorders. The first is that in or around the damaged neurons, one can often observe dense aggregates of tangled-up protein molecules, with a specific, primary protein component that is largely characteristic of each disease. For example, the senile plaques that are the histopathological hallmark of Alzheimer's disease contain large amounts of a misfolded small protein called Abeta. Lewy bodies, characteristic of Parkinson's disease and Lewy Body dementia, are primarily composed of aggregates of the protein alpha-synuclein.[1] This observation suggests that protein misfolding and aggregation may play a role in the pathology of these diseases.

The second commonality is that the major neurodegenerative disorders come in two flavors: a sporadic form that is not inherited and accounts for the majority of cases, and a rarer genetic form that runs in families in a predictable, Mendelian pattern. (The exception is Huntington's disease, which appears always to be genetic.) With the complete sequencing of the human genome earlier in this century and advances in genetic mapping techniques, it has now become possible to identify many, though not yet all, of the genes whose mutation is responsible for the inherited forms of some of these disorders; and one of the first observations made was that the gene coding for the major protein component of the specific aggregates was usually one of them. Moreover, when the mutant forms of the proteins were made and studied, it was typically found that the mutations made the protein unstable and prone to aggregation, both in the test tube and in cultured nerve cells. Occasionally, it was also possible to link mutations in some of the other genes that did not code for that protein but were responsible for cases of the disease, with increased aggregation of the same protein.[2]

Thus, over the past dozen or so years, a model has emerged for the rare, familial forms of most of the major neurodegenerative diseases that can be summarized as follows: in order to function, protein molecules must fold properly, like an origami bird with wings that flap. But some mutations, either in the gene that codes for the bird itself or in other genes that somehow help the bird to fold (or affect its stability when it is folded, or in some cases its location inside the cell), lead to a misfolded bird – a crumpled wad of paper, if you will. Over time, these wads accumulate and clump together until, for reasons still not understood, the neurons in which the aggregates build up begin to die.

The challenge then became to relate this model to the more common, sporadic forms of the disease, which tend to be idiopathic (that is, to have no known cause). And the exciting development has been that in some cases, it has been possible to do that. For example, environmental exposure to certain pesticides is a risk factor for Parkinson's disease, and in animal models exposure to such pesticides not only produces some Parkinson's-like symptoms but also occasionally gives rise to Lewy body-like aggregates containing alpha-synuclein.[3] In another example, genetic risk factors for some of the sporadic diseases have also been identified; that is, we have located genes whose mutation increases the chance that one will get the non-inherited form of the disease but does not guarantee that one will. And in a few cases, these risk-associated mutations turn

out to be *in the same genes* where different mutations produced the familial form of the disease.[4]

There is now general agreement that, for the most part, the detailed molecular mechanisms of most, if not all, of these disorders are likely to be broadly similar, and that they involve the demented origami of protein misfolding and aggregation. To be sure, different diseases affect different proteins, each of which folds into quite a different paper animal when it functions properly, but each of which wads up and aggregates in a similar way in its disease. There is less agreement on whether the dense aggregates one can see easily are the cause of the disorder or its consequence. One theory that is gaining followers is that the actual toxic species are smaller clumps of protein that are not visible. Why this process should be age-dependent is not clear; but considering the known fact that all cells have some quality-control machinery that recognizes wadded up protein as potentially dangerous and either refolds it properly or disposes of it, a logical assumption might be that if this process is imperfect, over time the untreated misfolded protein could slowly accumulate to toxic levels. This assumption is important to test, because if it is true it has profound implications for how the diseases might be diagnosed and treated.

Continuing the origami analogy, this model suggests several strategies for therapy. If a particular protein's misfolding is causing the disease, then drugs that prevent misfolding – acting, if you will, as a kind of molecular Scotch tape to hold the origami animal in its proper shape – might prevent or delay the onset of the disease. The Scotch tape method has given promising results in clinical trials for some quite different, rare genetic diseases that are also caused by unstable mutant proteins.[5] Another approach would be to increase the efficiency of the quality-control system that deals with misfolded proteins so that it can handle more of them over a longer period of time. This idea is being tested in a number of start-up companies. Yet another approach is based on the assumption that breaking up the large aggregates that one can see – the senile plaques in Alzheimer's disease and the Lewy bodies in Parkinson's disease, for example – would be of therapeutic benefit. Several companies and academic laboratories are pursuing this strategy, typically by using antibodies to dissolve the dense aggregates, although in some cases drugs are also being sought that would do the same thing. Of course, if the large aggregates are not the toxic species, this approach may fail. To date, there is no convincing evidence that the strategy is sound, but no real evidence that it isn't, either.

Besides these advances in our understanding of the mechanisms of the diseases and the new therapies they suggest, there is another reason for optimism: it may actually be easier to treat neurodegenerative diseases than it is to treat, say, cancer. If you want to cure a cancer patient, whether by surgery or radiation or chemotherapy or anti-tumor antibodies, you had better be near perfect, because if you leave even a few rogue cells alive, the cancer is likely to return, often more aggressively than before. And it is very hard to be perfect. But neurodegenerative diseases generally develop late in life, and most progress slowly. If the average age of onset of Parkinson's disease could be raised from sixty-five to eighty, we would count that as a victory. If Parkinson's took forty years to progress to fatality instead of twenty, most people with the disease would die of something else long before its symptoms became untreatable. One doesn't need perfection: it might be enough just to tweak the system a bit –

and that should be a lot easier, at least in theory.

Given all these positive developments, why are pharmaceutical companies deserting the arena, and why are the funding agencies not pouring even more money into such promising lines of research? One reason, of course, is that times are hard, and finding additional research dollars is neither easy for corporations nor politically popular for governments. But even if we assume that the economy will eventually recover, there are other major obstacles to translating these new discoveries into cures.

The first is, ironically, the size of the market. Suppose you had a drug that was likely to prevent one of the major neurodegenerative diseases. In that scenario, we wouldn't be looking to give that drug to the five million people who have Alzheimer's or the one million who have Parkinson's; we're talking about potentially giving it to everyone over the age of, maybe, fifty – in other words, about a hundred million people in the United States alone. You might think drug companies would hyperventilate over a market that big, and they do, but out of fear, not greed. If you give a drug to a hundred million apparently healthy people, no matter how safe that drug seemed to be in clinical trials involving a much smaller number of individuals, there are likely to be tens of thousands who will suffer severe side effects, possibly including death, because no clinical trial can adequately control for the enormous genetic, dietary, and environmental diversity of the human race. An example is the recent scandal over the pain drug Vioxx, which seemed to be relatively safe in the limited patient population for which it was developed. It was later marketed to a much broader population, where it caused severe cardiovascular problems in some people, leading to estimated thousands of deaths by heart attack and stroke.[6]

Because of the Vioxx catastrophe, no drug company will bring an Alzheimer's or Parkinson's prevention drug to market unless there were some way of determining who was likely to get the disease eventually, with a high degree of certainty, or who already had it, even if they had not yet begun to show symptoms. Unfortunately, there is no reliable way to do that for the approximately 90 percent of the cases that are not genetic – at least not yet. New developments in imaging techniques such as MRI and PET scanning are promising, but they are too expensive to be used for screening large numbers of people. In a number of companies and laboratories, efforts are under way to find what are called biomarkers – changes in gene expression or protein levels or metabolism – that would indicate a disease is in its earliest stages or is almost certainly going to start soon. The hope of being able to *prevent* neurodegenerative diseases rests on the success of these efforts.

So may the hope of *treating* such diseases in people who already have them. When a patient presents with symptoms of the so-called early stage of Alzheimer's or Parkinson's disease, for example, huge numbers of neurons have already died, and a number of secondary processes, such as inflammation, have set in. Arresting a disease process that far advanced may be almost impossible; even if possible, it is unlikely to result in restoration of the functions (memory, smooth movement) that have already been lost. Once again, if treatment to halt progression is to have a chance of working, it may be essential to identify that an individual has the disease before significant symptoms appear – another job for imaging methods and biomarkers that we do not yet have.

The second major obstacle is the lack of good animal models for nearly all the neurodegenerative diseases. Therapeutic development for any disease proceeds in

well-established stages, one of which is determining if the treatment is safe in certain animals. While not perfect, our animal models for toxicity are not bad, so many therapies that are safe in animals are later shown (at much greater cost, of course) also to be safe in people (so-called Phase I clinical trials). Most failures of treatments that are tested in human trials occur in the next stage, Phase II, which is designed to determine if the therapy is efficacious against the disorder in question. Consider what that means: after spending hundreds of millions of dollars to develop a drug that treats disease X in mice and dogs and even monkeys, almost half the time that drug will not produce the desired clinical result in humans. The obvious conclusion from this depressing fact is that our animal models for disease are often not very good, and among the worst are the animal models for age-related neurodegenerative diseases, probably because the lifespan of any model animal simply does not approach the threescore years and ten that are typically required for these diseases to emerge in people. Consequently, to generate any animal model at all, abnormal gene expression or some form of chemical stress, or both, is necessary, and the resulting model never shows the same slow progression (or even, in many cases, the same aggregates in the same neurons) as the human disease. Absent decent animal models, drug companies understandably shy away from a field in which they stand a good chance of spending a fortune developing a treatment that will have an even higher than normal failure rate in Phase II clinical trials.

The third obstacle is the clinical trials themselves – or, more precisely, their design. Until recently, for example, all the clinical trials of potential Alzheimer's drugs were carried out on late-stage Alzheimer's patients, and the readout for efficacy was improvement in cognition. Yet cognition is not an easy thing to measure with precision, and most Alzheimer's patients do not show a steady cognitive decline anyway, but rather a choppy one. Even an apparent improvement may not be due to the drug at all. Moreover, the brain of a late-stage Alzheimer's patient is so severely ravaged by the disease that it actually has holes where healthy tissue used to be. To expect a therapeutic to show a measurable benefit in such cases is naive at best, and perhaps even foolhardy.

These, then, are the challenges that must be overcome by biomedical research – and soon, if the coming epidemic is to be averted. Considering the current state of our knowledge, and the barriers to therapeutic development, I offer the following suggestions for improved progress in the fight against neurodegenerative diseases:

1) Biomarkers must be found that either signal the start of the disease long before symptoms appear, or that identify people whose risk of developing the disease soon is so high that it justifies treating them even though they do not yet have the disorder.

2) We must answer the question of whether the macroscopic aggregates are themselves toxic to neurons, or are a reservoir for the (smaller) toxic species, or are actually an attempt by the organism to protect itself from protein misfolding by sequestering the clumps. Absent this information, the optimal strategy for prevention and treatment is unclear. Clinical trials of plaque-clearing antibodies now under way may shed light on this issue: if they show efficacy, then the third possibility can probably be ruled out.

3) Recent exciting data that suggest that the toxic species may be able to spread from neuron to neuron, like an infec-

tious agent, need to be confirmed.[7] If they hold up, then new avenues for therapy to retard disease progression are suggested, including blocking the release of the toxic species from an "infected" cell or blocking its uptake by a neighboring cell.

4) Strategies for stabilizing the proteins that misfold in these diseases – the molecular Scotch tape approach – need to be pursued vigorously and evaluated carefully. This line of attack has a number of appealing features but also many uncertainties, such as how early in the progression of the disease must such a drug be administered in order to be effective.

5) Clinical trials for all these diseases need to be designed more intelligently and creatively. Surrogate diseases may represent an attractive approach: for example, carriers for the inherited, recessive lysosomal storage disorder Gaucher disease are at greatly elevated risk for Parkinson's disease,[8] raising the intriguing possibility that a treatment for the small number of people who actually have Gaucher disease could be used to reduce the risk for Parkinson's disease in the much larger number of Gaucher carriers. Clinical trials for Gaucher disease are much easier to design and carry out than trials for neurodegenerative diseases.

6) The traditional silos that balkanize medicine – and biomedical research – by organs and phenotypic presentation need to be broken down so that the connections between different diseases can be exploited to find new approaches for therapy. As an example, Alzheimer's patients are at much lower risk for all forms of cancer than age-matched controls; and now that people are living longer with cancer, it can be seen that the reverse is also true: cancer survivors are at significantly lower risk of Alzheimer's disease.[9] Could the cure for, say, Alzheimer's disease be sitting on some drug company's shelf as a failed cancer drug that was never tried on the right disease? More than just improved communication between oncologists and neurologists is needed to follow up fascinating leads like this one. What is needed is an entirely new way of thinking about disease: not in traditional terms of organs and tissues but in terms of pathways and processes inside the cell.

7) The current inadequate funding for neurologic disorder research must be redressed. One of the few problems that can often be solved by throwing more money at it is our understanding and ability to treat human disease: consider the successes of the war on cancer and the fight against HIV/AIDS. True, large increases in funding do lead to a lot of mediocre science being funded, but they also lead to a lot of great science being funded. And they also lead to cures – not immediately, to be sure (the time from basic research discovery to approval of a clinical treatment can easily be fifteen to twenty years), but eventually. One thing is certain: if increases in funding draw first-rate scientists to a field, inadequate funding drives many of them away. Private philanthropy can make a big impact here, and it may not be an exaggeration to say that without private philanthropy, we would be in much worse shape than we are in terms of progress, not only on neurodegenerative diseases but in medicine in general. Sanford I. Weill, the chairman of the Board of Overseers of Weill Cornell Medical College, puts it this way: "The life-saving benefits now possible thanks to biomedical research

happen largely through philanthropy, and the commitment of donors who recognize that we are on the cusp of a revolution in medicine. They are the ones who step into a leadership role and make it possible. It is the greatest legacy we can leave to our families, friends, and future generations."

Companies are right that drugs targeted for the central nervous system are painfully difficult to create and very difficult to win approval for. It might be time for governments to assume some of the risk of doing so, not by trying to develop such drugs themselves – pharmaceutical development is no job for amateurs – but by underwriting some of the cost. A $2 billion/year fund, available by peer-reviewed competition to companies that have a promising clinical candidate and a sensible clinical trial design (see #5), might bring those companies back into the sector or take some biotech-developed drugs deep into human trials without costly partnering.

Neither the private nor the public sector will give this area the attention it needs without pressure from the lay public. Taking a page from the HIV/AIDS activists' book, the patients afflicted with these disorders need to speak loudly, and with one voice. Up to now, each disease has existed largely in its own universe, with foundations and patient-oriented groups focused on their particular disorder. If we realize that many of these seemingly different diseases have similar underlying causes, often present together, and that their seemingly distinct pathologies may mask their interrelatedness, then progress in any one disease may legitimately be seen as progress in many, if not all.

All these things are possible if we have the will and resources to do them. We cannot afford to fail, because if we do, the future is beyond bleak. A child born today will be about forty years old when, projections say, sixteen million of his or her fellow Americans will have Alzheimer's or Parkinson's disease (and a significant number will have both). That child will also be living in a world where the health care system is bankrupt, the social fabric is rotting, and nearly every family knows the despair and horror of watching a loved one slowly succumb to a hideous illness.

But trend is not destiny; the future is what we make it. The challenge for biomedical research in an aging world is to help create a future where both young and old can prosper. We must take up this challenge now, because the clock is ticking.

ENDNOTES

[1] The differences between these diseases are probably at least as important as their commonalities, but in the interest of clarity, the similarities are emphasized here. Reviews tend to be disease-specific, but one that treats the subject more generally is Eszter Herczenik and Martijn F. Gebbink, "Molecular and Cellular Aspects of Protein Misfolding and Disease," *The FASEB Journal* 22 (7) (July 2008): 2115 – 2133.

[2] For examples from Parkinson's disease, see Joshua M. Shulman, Philip L. De Jager, and Mel B. Feany, "Parkinson's Disease: Genetics and Pathogenesis," *Annual Review of Pathology* 6 (February 28, 2011): 193 – 222. A good treatment of the Alzheimer's case is Rita J. Guerreiro, Deborah R. Gustafson, and John Hardy, "The Genetic Architecture of Alzheimer's Disease: Beyond APP, PSENs and APOE," *Neurobiology of Aging* 33 (3) (March 2012): 437 – 456.

3 An excellent recent review is Shin Hisahara and Shun Shimohama, "Toxin-Induced and Genetic Animal Models of Parkinson's Disease," *Parkinson's Disease* 2011 (2011): 951709.

4 Ian Martin, Valina Dawson, and Ted M. Dawson, "Recent Advances in the Genetics of Parkinson's Disease," *Annual Review of Genomics and Human Genetics* 12 (September 22, 2011): 301 – 325.

5 For an overview of the method, see Dagmar Ringe and Gregory A. Petsko, "What are Pharmacological Chaperones and Why are They Interesting?" *Journal of Biology* 8 (9) (October 13, 2009): 80. Genetic disorders called lysosomal storage diseases have been treated with particular success by this method. A recent review is Giancarlo Parenti, "Treating Lysosomal Storage Diseases with Pharmacological Chaperones: From Concept to Clinics," *EMBO Molecular Medicine* 1 (5) (August 2009): 268 – 279.

6 Janice Hopkins Tanne, "Merck Pays $1bn Penalty in Relation to Promotion of Rofecoxib," *BMJ* 343 (November 28, 2011): d7702.

7 Although this might sound like a virus, or like Mad Cow Disease, there is no evidence that these proteins are infectious from person to person. What is now widely believed is that the neurodegeneration starts in some particular set of neurons and then can be propagated to other parts of the brain along the connections those neurons make with others. For Alzheimer's, the newest observations are Li Liu, Valerie Drouet, Jessica W. Wu, Menno P. Witter, Scott A. Small, Catherine Clelland, and Karen Duff, "Trans-Synaptic Spread of Tau Pathology *In Vivo*," *PLoS One* 7 (2) (2012): e31302; and Alix de Calignon et al., "Propagation of Tau Pathology in a Model of Early Alzheimer's Disease," *Neuron* 73 (4) (February 23, 2012): 685 – 697. Similar findings have been reported for the pathogenic molecules in ALS – for example, Magdalini Polymenidou and Don W. Cleveland, "The Seeds of Neurodegeneration: Prion-like Spreading in ALS," *Cell* 147 (3) (October 28, 2011) – and Parkinson's disease: see Christopher J.R. Dunning, Juan F. Reyes, Jennifer A. Steiner, and Patrik Brundin, "Can Parkinson's Disease Pathology be Propagated from One Neuron to Another?" *Progress in Neurobiology* 97 (2) (May 2012): 205 – 219.

8 Wendy Westbroek, Ann Marie Gustafson, and Ellen Sidransky, "Exploring the Link between Glucocerebrosidase Mutations and Parkinsonism," *Trends in Molecular Medicine* 17 (9) (September 2011): 485 – 493. See also S. Pablo Sardi, Priyanka Singh, Seng H. Cheng, Lamya S. Shihabuddin, and Michael G. Schlossmacher, "Mutant GBA1 Expression and Synucleinopathy Risk: First Insights from Cellular and Mouse Models," *Neurodegenerative Diseases* 10 (April 2012): 195 – 202.

9 This striking inverse correlation has been shamefully neglected, both in research and in research funding. It also applies to Parkinson's disease (except for melanoma, where having Parkinson's *increases* the risk for melanoma, and vice versa) and schizophrenia. A good, brief review (albeit from a particular perspective) is Rafael Tabarés-Seisdedos, Nancy Dumont, Anaïs Baudot, Jose M. Valderas, Joan Climent, Alfonso Valencia, Benedicto Crespo-Facorro, Eduard Vieta, Manuel Gómez-Beneyto, Salvador Martínez, and John L. Rubenstein, "No Paradox, No Progress: Inverse Cancer Comorbidity in People with Other Complex Diseases," *The Lancet Oncology* 12 (6) (June 2011): 604 – 608. For an excellent discussion of the schizophrenia case, see Yang Wang, Guang He, Lin He, and John McGrath, "Do Shared Mechanisms Underlying Cell Cycle Regulation and Synaptic Plasticity Underlie the Reduced Incidence of Cancer in Schizophrenia?" *Schizophrenia Research* 130 (1 – 3) (August 2011): 282 – 284.

Biodiversity & Environmental Sustainability amid Human Domination of Global Ecosystems

David Tilman

Abstract: Concern about the loss of Earth's biological diversity sparked two decades of research of unprecedented intensity, intellectual excitement, and societal relevance. This research shows that biodiversity is among the most important factors determining how ecosystems function. In particular, the loss of biodiversity decreases the productivity, stability, and efficiency of terrestrial, freshwater, and marine ecosystems. These research findings come at a time of rapidly increasing threats to global biodiversity resulting from agricultural land clearing, climate change, and pollution caused by globally accelerating demand for food and energy. The world faces the grand, multifaceted challenge of meeting global demand for food and energy while preserving Earth's biodiversity and the long-term sustainability of both global societies and the ecosystems upon which all life depends. The solutions to this challenge will require major advances in, and syntheses among, the environmental and social sciences.

DAVID TILMAN, a Fellow of the American Academy since 1995, is Regents Professor in the Department of Ecology, Evolution and Behavior at the University of Minnesota; he is also a Professor in the Bren School of Environmental Science and Management at the University of California, Santa Barbara. His publications include *The Functional Consequences of Biodiversity: Empirical Progress and Theoretical Extensions* (edited with Ann P. Kinzig and Stephen W. Pacala, 2002), *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (edited with Peter Kareiva, 1997), *Resource Competition and Community Structure* (1982), and more than two hundred articles in scientific journals.

The existence of life is the defining feature of Earth, and diversity is the most striking aspect of life on Earth. Since the origins of life three billion years ago, the biological diversity of life, or its *biodiversity*, has been on an upward but jagged trajectory along which the formation of new species has exceeded, with but few exceptions, the loss of existing species. The exceptions were major extinction events, attributable to catastrophic occurrences such as meteor impacts and globally massive volcanic activity. Earth now has on the order of five million species, all descended from the same ancestor. This biodiversity has been an enduring source of wonderment and scientific mystery from the era of great naturalist-explorers, such as Darwin and Wallace, to the present.

Earth also has seven billion people who, in meeting their needs for food and energy, have become a globally dominant force affecting Earth's ecosystems and threatening their biological diversity. The

rapid acceleration in human global impacts through land-clearing and the destruction of natural habitats, fossil fuel combustion and climate change, nutrient pollution, and other activities has led to projections that humanity may be in the process of causing species extinctions at a rate rivaling some of the largest extinction events found in the fossil record.[1] These projections have raised a series of questions and concerns, the most fundamental of which are: What caused the evolution of Earth's amazingly great biodiversity? Does the loss of this biodiversity matter? Are there practices that could consistently and stably sustain the habitability of Earth while meeting the food, energy, and other needs and demands of the nine to ten billion people who will populate the planet by mid-century? If such practices are discovered, what policies, ethics, or approaches could lead to their global adoption?

The first question, on the origins of biodiversity, has been a mainstay of evolutionary research at least since Darwin. The question of whether biodiversity loss really matters was first raised in the 1980s. By the mid-1990s, it had ignited a wave of ecological research of unprecedented intensity, intellectual excitement, controversy, and societal relevance. In so doing, it helped transform the discipline of ecology into a more mechanistic and predictive science in which hypotheses are tested against the outcomes of multiple field experiments; observations in multiple ecosystems, both natural and those experiencing human impacts; and the predictions of alternative mathematical theories.

In this essay, I consider the biodiversity revolution and its aftermath by summarizing its major discoveries, controversies, and resolutions. The scientific revolution that has occurred, as remarkable as it has been, is only the initial step toward the discoveries that are needed if society is to achieve greater environmental sustainability as well as ensure full and equitable lives for all peoples. Some of the major mysteries that remain are discussed later in the essay; as is always the case in science, many more mysteries await their discovery.

I also address social and cultural issues that arise from the application of new scientific knowledge to society. These are perhaps the greatest challenges because advances in scientific knowledge can contribute to achieving societal goals only if the knowledge is accepted and adopted by society. How, though, is this done? The ethical precepts, laws, and customs of a society are the result of hundreds to thousands of years of often-slow cultural evolution. What would or could motivate the rapid changes in customs, laws, and ethics that may be needed now that human activities have become the dominant global force affecting how ecosystems function?

In 1958, Charles Elton, the great Oxford ecologist, hypothesized that stability is greater in ecosystems containing a diverse set of species. He worked in the style of ecological research that was popular in his day, undertaking qualitative comparisons of habitats that differed in their diversity: species-rich meadows versus nearby monocultures of crop plants, for example, or isolated and depauperate islands versus highly diverse mainland communities. He further suggested that habitats with high biodiversity are less susceptible to invasion by exotic species. A century earlier, Darwin had indicated that greater plant diversity is associated with greater primary productivity, but this insight lay dormant until rediscovered in 1993 by Sam McNaughton.[2]

At about the same time that Elton was carrying out his research, G. E. Hutchinson, the noted aquatic ecologist at Yale, observed how paradoxically high the

diversity of many ecosystems seems to be.[3] Then-current mathematical theory predicted that the number of coexisting species should be no greater than the number of distinct resources for which the species compete. In contrast, even in seemingly simple habitats such as the well-mixed open waters of lakes and the oceans, the number of coexisting species of algae was often an order of magnitude greater than the number of limiting nutrients for which they competed. Hutchinson's paradox sparked my fascination with biodiversity. I dedicated the first two decades of my career to understanding how species compete with each other, and how and why such competitive interactions so often lead to the coexistence of many species rather than to domination by one or a few species.

Elton's ideas flourished for a decade or two, only to be put aside as the discipline began to develop a tradition of experimental, observational, and theoretical research. Scientists now sought the mechanistic construction of a species-based understanding of the dynamics of multispecies communities and ecosystems. New theory played a pivotal role in this transition. Robert May, the brilliant physicist-turned-ecologist from Princeton and, later, Oxford, presented elegant mathematics showing that the stability of communities of competing species declines as communities become more diverse.[4] May's mathematical demonstration that the dynamics of individual species become less stable at higher biodiversity led to debate on how diversity affects the stability of natural ecosystems. After reviewing more than two hundred papers on the issue, Daniel Goodman criticized the superorganismal perspectives then in vogue in ecosystem ecology and concluded that Elton's diversity-stability hypothesis was not supported by a preponderance of evidence.[5]

By 1973, when the second edition of his book *Diversity and Stability in Model Ecosystems* was published, May suggested an alternative resolution to the debate – that ecosystem properties might grow more stable with diversity even as population stability declined – but his insight was overlooked. Indeed, for the next two decades, most ecologists, myself included, considered diversity of little relevance to stability or other ecosystem processes. Then-current research, much of it performed with well-replicated field experiments, focused on the mechanisms of interaction among a few species and on how the traits of each species influence the dynamics and outcome of interactions among those species. Higher-level questions about how the number of interacting species might have an impact on the functioning of ecosystems were set aside while ecologists worked to transform the field into a more mechanistic and predictive science.

During this period, a few scholars continued to study biodiversity. In 1981, Paul and Anne Ehrlich, evolutionary ecologists at Stanford, published their book *Extinction: The Causes and Consequences of the Disappearance of Species*. They raised concern about how human activities are threatening global biodiversity and how loss of this biodiversity could harm the functioning of ecosystems and the services they provide to society. Edward O. Wilson, the Harvard evolutionary biologist, also wrote extensively on this issue and was awarded a Pulitzer Prize for his 1992 book, *The Diversity of Life*. The work of the Ehrlichs, Wilson, Peter Raven, Sam McNaughton, Stuart Pimm, and others had so elevated global concerns about the loss of biodiversity that the United Nations convened an "Earth Summit" in Rio de Janeiro in 1992. This gathering led to the international Convention on Biological Diversity.

Shortly afterward, Hal Mooney and Detlef Schulze organized a small meeting of ecologists to synthesize and evaluate how the functioning of ecosystems might depend on biodiversity. The supporting evidence, though scattered and scant by the standards of the discipline, was sufficient to reignite my interest in this question as well as the interest of almost everyone else who attended the meeting. The resulting edited book presented intriguing concepts suggesting that ecosystem functioning could be linked to biodiversity.[6]

Once rejected, an idea rarely regains traction in science because its seeming flaws are well known. Yet two papers published in *Nature* in 1994 reopened debate over the diversity-stability hypothesis. The first, "Biodiversity and Stability in Grasslands," explored how the stability of grassland plant communities in response to a major drought depends on plant diversity.[7] John Downing and I had approached these data with more than our usual level of scientific skepticism. We tried hundreds of different analyses, each aimed at rejecting the hypothesis that greater plant diversity leads to greater ecosystem stability. We instead found that every analysis supported that hypothesis. Results from more than two hundred plots showed that stability, measured as resistance to the effects of a major disturbance, is a sharply and significantly increasing function of plant diversity. In particular, during the drought, the productivity of grassland plots containing one to three species fell to about one-tenth of predrought levels, whereas plots containing fifteen to twenty-five species had their productivity fall to only about half of their predrought levels. Three months later, Shahid Naeem and collaborators published the paper "Declining Biodiversity Can Alter the Performance of Ecosystems," which reported that simpler and less diverse laboratory food webs are less productive than those that are more diverse.[8] By 1997, several papers had reported similar effects of plant diversity on primary productivity based on well-replicated biodiversity experiments performed in field conditions.[9]

As should occur in science whenever evidence seems to challenge current ideas, this growing body of evidence was met with skepticism. In the first paper to question the apparent effects of plant biodiversity on primary productivity, Michael Huston raised doubts about the ability of the experiments to reject an alternative cause called *sampling effects*.[10] He presented the intriguing hypothesis that the effects come not from diversity per se, but from the greater probability that a highly productive plant species would be present at higher diversity. If the productivity of a plot is determined mainly by the growth of its most productive species, Huston reasoned, then the greater productivity observed in higher diversity plots might merely mean that they have a greater chance of containing a highly productive species. That same year, David Wardle and collaborators published a study of a set of small islands showing that island productivity is more dependent on fire frequency and other factors than on plant biodiversity.[11] Next, in their paper "The Statistical Inevitability of Stability-Diversity Relationships in Community Ecology," Dan Doak and collaborators offered an alternative hypothesis to explain the apparent effect of diversity on ecosystem stability.[12]

The biodiversity revolution was under way, and what ensued was more than a decade of discovery characterized by numerous rounds of debate and resolution driven by the interplay of experimental results, novel analyses, theoretical predictions, and observations in natural ecosystems. Ecology, as a science, had come of age. Of all the grand debates that have occurred in ecology, the biodiversity debate

was the first to be so thoroughly tested via the interplay of numerous focused experiments, new theory, and quantitative field observations. One after another, novel hypotheses were proposed, tested, modified, and synthesized as more than one hundred different biodiversity experiments were performed around the world. This large number of experiments opened up ecology to meta-analysis, a new tool that greatly contributed to the biodiversity synthesis.[13] As this occurred, it became increasingly clear that the loss of biodiversity has many more and larger impacts, some via newly discovered pathways, on ecosystem functioning than had ever been envisioned. An idea cast aside in the 1970s had turned into one of the most highly studied and well-understood concepts of ecology.

The evidence that led to a new biodiversity paradigm came from a confluence of results of experiments and theory.[14] To provide a flavor of this work, and especially its findings, I briefly summarize five types of ecological processes that are now known to be affected by the loss of biodiversity.

Productivity. The growth of plants provides the "primary productivity" that is the basis of all ecosystem functions. Experiments have shown that, on average, plots planted with highly diverse mixtures of plant species annually produce about 70 to 100 percent more aboveground biomass – that is, they have greater primary productivity – than plots planted with monocultures of these same species.[15] The positive effect of plant diversity on ecosystem productivity has been observed in ecosystems ranging from temperate grasslands[16] to tropical, Mediterranean, and boreal ecosystems.[17] In experiments in which fish species diversity was manipulated, treatments with greater numbers of fish species produced more fish biomass.[18]

Most of these biodiversity experiments, though, lasted only one or two years. The few long-term experiments that have been done reveal that the initial effects of biodiversity on productivity increase through time.[19] For instance, results from the longest-running biodiversity experiment, which my collaborators and I established in Minnesota in 1994, show that the annual biomass production of the highest-diversity plots (those planted with sixteen species) increased through time much more than the average biomass of the same sixteen species growing in monocultures. In particular, in the third and fourth years of the experiment, the high-diversity plots had, on average, 92 percent greater production than the monocultures. This figure increased to 157 percent by years 8 and 9, and to 190 percent by years 17 and 18, which are the source of our most recent data.

Stability. The stability of an ecosystem process is a measure of the constancy of the process in response to disturbance. Greater ecosystem stability thus means that the process is better buffered and less variable. Natural and managed ecosystems experience a wide variety of intensities and types of perturbations: climatic variation (cool or warm and/or wet or dry periods); disease or pest outbreaks, fire, erosion, landslides and other physical disturbances; and shifts in the structure of food chains, such as from loss of top predators. Long-term biodiversity experiments have provided direct tests of the dependence of ecosystem stability on plant diversity. For instance, year-to-year variation in annual biomass production was lower in higher-diversity plots in both a European grassland experiment and our Minnesota grassland biodiversity experiment,[20] showing that higher diversity leads to greater stability of primary productivity in systems experiencing year-to-year climate variation. Similarly,

observational studies show that the stability of the productivity of marine fisheries is higher in those fisheries that have greater fish species diversity, and that greater numbers of genetic varieties of wheat lead to less variation in yields as well as higher yields.[21]

Disease. Most pathogens and disease are specific to one or a few species. As a result, the rate of disease transmission from an infected individual to a susceptible individual of the host species is proportional to the population density of the host. This fundamental principle of epidemiology suggests that the incidence of disease for a given host species should decline when the host species is living in a more diverse community. Because each plant species would be less abundant than in monoculture, the incidence of species-specific plant diseases should, on average, decline as plant diversity increases. This expectation is supported by results of numerous biodiversity experiments. For instance, fungal pathogens that grow on the surfaces of leaves are much less abundant at higher plant diversity.[22] Transmission rates for diseases of many other types of species, including amphibians, corals, fish, and birds, are similarly lower when the biodiversity of the host community is greater.[23]

Resistance to Invasion. Charles Elton's observations (discussed above) led him to suggest that more diverse ecosystems are less easily invaded by exotic species. Biodiversity experiments have provided broad support for this hypothesis. An experiment in which seeds of numerous nonresident plant species were added to plots in a biodiversity experiment that differed in both their species numbers and their functional group compositions showed two marked effects. First, the added species were less likely to invade not only when the diversity of the established plant community was high but also when the invading species were functionally similar to established abundant species. Additional work shows that the major factor inhibiting invasion is low availability of the limiting soil nutrient, and that more diverse plant communities reduce soil nutrient concentrations to lower levels.[24]

Biodiversity and Agriculture. Four crops – maize, wheat, rice, and soybeans – provide about 80 percent of the food calories consumed globally. Because of the rapidity with which crop pathogens and pests evolve and overcome plant defenses, the sustainability of these crop yields is highly dependent on continued breeding for resistance to the latest varieties of pathogens and pests. For instance, "IR8," the rice variety that began the Green Revolution in Asia in 1967, had its yield fall 24 percent over the subsequent thirty years because of pathogens and pests. Nine subsequent rice varieties had their yields decline by similar rates after their introductions.[25] For crop breeding to stay ahead of pests and pathogens, breeders must have an immense storehouse of genetic variants, at least some of which are resistant to emerging pathogens and pests. Even more genetic diversity is needed to find new genetic combinations that increase crop yields. Thus, although most such crops are grown as monocultures (and perhaps particularly because they are grown in such a way), genetic diversity is of great economic and societal value.

Biodiversity can be used as a tool to increase crop yields in some situations.[26] For instance, Youyong Zhu and collaborators found that growing two varieties of rice in alternating sets of rows (a practice called *intercropping*) greatly decreases incidence of a significant fungal pathogen that attacks a highly valued variety but to which the second variety is resistant. Long Li and collaborators observed that intercropping of faba beans and maize increases maize yields by 40 percent and

faba bean yields by 25 percent; they also found that this over-yielding is caused by differences in the rooting depths and seasonality of growth of two crops, as well as by faba beans' ability to mobilize otherwise unavailable phosphorus. While many crop combinations do not over-yield, Li reports that combinations that do over-yield are planted on 28 million hectares in China. Intercropping is rarely practiced in Western nations today, but its ability to increase yields in particular cases might offer benefits (although intercropping also introduces a number of challenges for mechanized agriculture that need to be solved).

After the initial experimental demonstrations that biodiversity affects numerous aspects of ecosystem functioning, attention shifted to why biodiversity matters. The first discussions centered on the role that sampling effects might play versus the importance of niche differences among species. Application of a variety of increasingly sophisticated analytical techniques has shown that sampling effects (later called *selection effects*) are generally unimportant and that niche differentiation effects (also called *complementarity*) are the predominant cause. This finding was especially evident in instances where species had several years to interact and thus the effects of their interactions were well established.

The understanding of why biodiversity matters was also illuminated by mathematical theory. In particular, a sequence of papers showed that when competing species have trade-offs in their traits that allow them to coexist stably, the net result is that ecosystem stability and productivity increase with diversity.[27] In addition, more diverse ecosystems reduce limiting resources to lower levels, both contributing to their greater productivity and reducing the abilities of other species to invade, as an invading species would have to survive and grow on the resources left unconsumed by established species.

Two hypotheses have received increasingly robust support from biodiversity experiments. First, species coexist with other competing species precisely because the species have trade-offs; any trait that increases the ability of individuals in a species to deal with one limiting factor must necessarily make them less able to deal with some other limiting factor.[28] Second, changes in biodiversity have consistent and predictable impacts on many aspects of ecosystem functioning; the trade-offs among the species that share a habitat mean that larger numbers of these species will, on average, be better at dealing with limiting factors in that habitat. Thus, the very processes that have allowed Earth to accumulate such a large number of species also mean that greater diversity would affect ecosystem functioning in exactly the ways that have now been observed experimentally.

An important corollary of "biodiversity matters" is that "species matter." This point has arisen repeatedly, both from results of biodiversity experiments and from ecological theory. An important demonstration of the "if diversity matters, species must matter" hypothesis was offered by Anthony Ives, Kay Gross, and Jennifer Klug, who showed theoretically that ecosystem stability depends not on the number of species per se, but on the differences among the species.[29] Because multiple competing species can stably coexist only if they have trade-offs in their traits, the biodiversity of an ecosystem (when enumerated by the simple metric of the number of species present) affects ecosystem processes precisely because the species differ from each other.

Human well-being is highly dependent on nature. The total land surface of Earth

is 13 billion hectares, of which 4 billion hectares are in the Arctic or Antarctic, or are desert or tundra. The vast majority of the remaining 9 billion hectares is heavily used by people, with about 5 billion hectares serving as agricultural lands, roughly one-fourth of which is farmed and the rest used for livestock production. Much of the remaining land is forested, with about 1.5 billion hectares being actively managed for tree production globally. Thus, two of humanity's most essential needs – food and shelter – are directly dependent on the productivity and stability of about 75 percent of Earth's usable lands. Moreover, because greenhouse gases are released from fossil fuel combustion, there is increased interest in also using land to produce biomass for conversion into biofuels with low greenhouse gas emissions.

Society depends on nature not only for goods such as food, timber, and energy, but also for a variety of ecosystem services.[30] We need potable water, a resource that is produced by intact grassland and forest ecosystems, and that is harmed by some agricultural and industrial activities. Intact ecosystems minimize flooding; they are a major storehouse of organic carbon that would otherwise be released into the atmosphere as carbon dioxide (CO_2) if the land were cleared; they create the fertile soils on which agricultural productivity depends.

One of the great challenges facing humanity is to find ways to meet its needs for food, timber, energy, and other goods while maintaining the ability of managed and natural ecosystems to provide vital ecosystem services. Discovery and adoption of better management practices will be essential to optimizing the production of goods and ecosystem services from managed lands, and thus increasing the long-term sustainability of the full range of goods and services that people need. It seems likely that biodiversity may play a central role in achieving greater sustainability, but this is a hypothesis in its infancy. The fate of this hypothesis – and of global biodiversity – is at present uncertain.

The next fifty years are likely the final period of rapid expansion in human population and consumption. Global population, which had increased 270 percent in the twentieth century, is likely to increase from its current seven billion people to about nine or ten billion, a 35 percent change, by the middle of this century, at which point global population growth may halt. This astounding population increase, though, is small compared to the increases in per-capita global consumption (measured as per-capita Gross Domestic Product) across this same time period. During the twentieth century, the real (inflation adjusted) buying power of a typical person increased by 360 percent, and it is projected to increase by about 150 percent during the next fifty years as the peoples of many developing nations gain "middle class" incomes.

The double-whammy of greatly increased population and even more greatly increased consumption per individual has already turned humans from being one of many species on Earth to being the dominant force affecting all ecosystems. Moreover, human environmental impacts are likely to double or triple by mid-century because of the anticipated global increases in both per-capita consumption and population size.[31] To meet an estimated doubling in global demand for food may require that about 1 billion hectares of tropical forest and grasslands be cleared for crop production and that agricultural fertilization, which can cause serious water pollution, increase about 170 percent. Land-clearing leads to the loss of biodiversity and is a major source of greenhouse gas release. Moreover,

agriculture itself accounts for about 37 percent of total human-caused greenhouse gas releases, and such releases would more than double as food demand doubled. In comparison, all forms of transportation combined account for only 20 percent of global greenhouse gas emissions.

Global energy demand is increasing at least as rapidly as is food demand, and most of this increased demand is being met by the combustion of fossil fuels. The net result of these food and energy trajectories will be major climate changes, the irreversible loss of a significant portion of Earth's biodiversity, and greatly decreased provisioning of numerous vital ecosystem services. Although there are insights to be gained from articulating the environmental problems that human activities are causing, it is even more important to find solutions.

The science, social science, and business of sustainability are all in their infancy. What we see now is the embryo of an unknown organism. Its development will be guided by the creativity and careers of the next generation. In that spirit, I offer a few thoughts about the challenges and possibilities ahead as we seek viable solutions.

Efficiency. The expanding human domination of the globe will affect biodiversity, climate, and numerous ecosystem services largely in terms of increased demand for food and energy. There are two equally important types of solutions to this problem. The first type focuses on decreasing demand for food and energy, and the second on meeting these demands in ways that lessen environmental impacts. Demand can be reduced by increases in efficiency. Energy efficiency is a familiar topic, but food efficiency is not. About a quarter to a third of global food production is wasted, with the causes of this wastage differing among societies. The major reason why a projected 35 percent increase in global population is expected to cause a 100 to 110 percent increase in global demand for crops is that per-capita meat consumption increases with income, and each kilogram of meat protein requires that livestock be fed from 3 to 20 kilograms of crop protein. This range in values occurs because animals differ greatly in the efficiency with which they convert grain protein into edible animal protein, with farm-raised fish being about eight times more efficient than cattle, and poultry being about four times more efficient than cattle. Direct human consumption of grain protein is even more efficient. Dietary shifts toward non-livestock proteins would provide environmental benefits, as would advances in the efficiencies with which livestock convert feed into meat. Thus, there are contributions to be made toward achieving greater environmental sustainability across disciplines as divergent as the culinary arts (via the creation of delicious but environmentally efficient entrées) and animal nutrition.

Lessening the Impacts. Modern societies are highly dependent on energy. The three greatest impacts of fossil fuel combustion come from the release of greenhouse gases (which cause climate change), of fine particulate matter (which causes respiratory problems and increases mortality), and of mercury (which causes health problems). Wind and solar power are alternatives that reduce these impacts, but adoption of these technologies has been slow because of challenges related to cost and reliability. Almost all current vehicles require liquid fuels. Electric vehicles may be the solution, but in order to achieve meaningful deployment, advances in battery technology must be made that would increase mileage range to somewhere between three hundred and five hundred miles.

Air transport may always depend on liquid fuels. The challenge is to create liquid fuels that are greenhouse gas neutral and that do not compete with food crops

for fertile land. If biofuels did compete for fertile land, their greenhouse gas benefits would likely be eliminated because of the greenhouse gas emissions associated with additional land-clearing to meet global food demand. Or, even worse, escalating food prices might harm the diets of the world's poorest people, in effect having airplanes and vehicles outcompete the already malnourished poor for food.

Using, Not Losing, Biodiversity. Biodiversity might provide a solution for this problem. Consider an as-yet untested possibility: the production of carbon-negative biofuels. As already mentioned, biomass production can be increased by 70 to 200 percent when highly diverse mixtures of species are planted. The greatest reported yield increases are from the diverse mixtures of native plants that my collaborators and I have grown in Minnesota on highly degraded soils that were no longer suitable for agriculture.[32] Although we observed no detectable increase in soil carbon and nitrogen stores for the monoculture plots, the highest diversity plots removed from the atmosphere and stored as soil organic carbon about 4.4 tons of CO_2 per hectare per year. As we reported in the paper "Carbon-Negative Biofuels from Low-Input High-Diversity Grassland Biomass," this biomass could be used to produce liquid transportation fuels that are carbon-negative. Because of carbon sequestration in the degraded soils, we calculated that the net effect of growing the biomass, making the fuels, and combusting the fuels would be a reduction in atmospheric CO_2. This possibility, though, is yet to be pursued.

Biodiversity might also help us better meet growing demand for food and forest products. The research showing that more diverse fisheries are more productive suggests that we might be able to harvest more seafood from aquaculture operations that have the right combination of competing or facilitating species. Similarly, the 1.5 billion hectares of managed forests may yield more timber and other forest products if they are planted to the right mixtures of tree species. Again, these possibilities have never yet been pursued commercially, much less globally.

Mysteries and Paradoxes. The path toward achieving environmental sustainability is filled with mysteries and paradoxes. Mysteries motivate science, leading to advances in fundamental science and technological breakthroughs. Although I have focused on such scientific advances in this essay, we also need fundamental advances in our understanding of ourselves. Humans are unique among all species in how dependent our welfare is on culturally transmitted knowledge. Our lives depend on knowledge accumulated during the ten thousand–year history of agriculture and on advances in public health, civil engineering, and medicine. We are now so highly dependent on knowledge that many people dedicate the first twenty-five to thirty years of their lives to obtain the knowledge needed for a professional career. The pursuit of such training by women and men is perhaps the most important force that is causing global population growth to slow.

Human behavior, though, is often confused or even paradoxical. Why do so many members of the most knowledge-dependent species on Earth act in ways that ignore, or even deny, knowledge? Why do individuals refuse to accept modern scientific knowledge as relevant or even as valid? At the same time that medical science has shown that healthy diets and active lifestyle can extend lives by a decade, people around the world are becoming more overweight and more inactive than ever before. People complain about the high cost of gasoline and yet preferentially buy expensive vehicles that have low fuel efficiency. People whose lives

have been saved by novel antibiotics that overcame drug resistance that had evolved in a pathogen often deny the existence of evolution. Others deny climate change.

I do not make these points to disparage anyone, and I deeply value intellectually honest skepticism on any topic. Rather, I mention them because the environmental issues that Earth faces are problems generated by humanity. So, too, must it be humanity that discovers and embraces the solutions. This effort will require that we learn more not just about the environment but also about ourselves. I can imagine no issue more worthy of pursuit than the grand, multifaceted challenge of helping society live sustainably on Earth.

ENDNOTES

1 Paul Ehrlich and Anne Ehrlich, *Extinction: The Causes and Consequences of Disappearance* (New York: Random House, 1981), 294; Stuart L. Pimm and Peter Raven, "Extinction by Numbers," *Nature* 403 (2000): 843–845; Peter M. Vitousek et al., "Human Alteration of the Global Nitrogen Cycle: Sources and Consequences," *Ecological Applications* 7 (1997): 737–750.

2 Samuel J. McNaughton, "Biodiversity and Function of Grazing Systems," in *Biodiversity and Ecosystem Functioning*, ed. Ernst-Detlef Schulze and Harold A. Mooney (Heidelberg, Germany: Springer-Verlag, 1994), 525.

3 G. Evelyn Hutchinson, "The Paradox of the Plankton," *The American Naturalist* 95 (1961): 137–145.

4 Robert May, *Diversity and Stability in Model Ecosystems*, 2nd ed. (Princeton, N.J.: Princeton University Press, 1973), 265.

5 Daniel Goodman, "The Theory of Diversity-Stability Relationships in Ecology," *The Quarterly Review of Biology* 50 (1975): 237–266.

6 *Biodiversity and Ecosystem Functioning*, ed. Schulze and Mooney.

7 David Tilman and John A. Downing, "Biodiversity and Stability in Grasslands," *Nature* 367 (1994): 363–365.

8 Shahid Naeem et al., "Declining Biodiversity Can Alter the Performance of Ecosystems," *Nature* 368 (1994): 734–737.

9 David Tilman, David Wedin, and Johannes Knops, "Productivity and Sustainability Influenced by Biodiversity in Grassland Ecosystems," *Nature* 379 (1996): 718–720; David U. Hooper and Peter M. Vitousek, "The Effects of Plant Composition and Diversity on Ecosystem Processes," *Science* 277 (1997): 1302–1305; David Tilman et al., "The Influence of Functional Diversity and Composition on Ecosystem Processes," *Science* 277 (1997): 1300–1302.

10 Michael A. Huston, "Hidden Treatments in Ecological Experiments: Re-evaluating the Ecosystem Function of Biodiversity," *Oecologia* 110 (1997): 449–460.

11 David A. Wardle, Olle Zackrisson, Greger Hornberg, and Christiane Gallet, "The Influence of Island Area on Ecosystem Properties," *Science* 277 (1997): 1296–1299.

12 Daniel F. Doak et al., "The Statistical Inevitability of Stability-Diversity Relationships in Community Ecology," *The American Naturalist* 151 (1998): 264–276.

13 Bradley J. Cardinale et al., "Effects of Biodiversity on the Functioning of Trophic Groups and Ecosystems," *Nature* 443 (2006): 989–992; Patricia Balvanera et al., "Quantifying the Evidence for Biodiversity Effects on Ecosystem Functioning and Services," *Ecology Letters* 9 (2006): 1146–1156; Bradley J. Cardinale et al., "The Functional Role of Producer Diversity in Ecosystems," *American Journal of Botany* 98 (2011): 572–592.

[14] See ibid. as well as David Tilman, Clarence L. Lehman, and Kendall T. Thomson, "Plant Diversity and Ecosystem Productivity: Theoretical Considerations," *Proceedings of the National Academy of Sciences* 94 (1997): 1857–1861; David Tilman, "The Ecological Consequences of Changes in Biodiversity: A Search for General Principles," *Ecology* 80 (1999): 1455–1474; Michel Loreau, "Biodiversity and Ecosystem Functioning: Recent Theoretical Advances," *Oikos* 91 (2000): 3–17; Clarence L. Lehman and David Tilman, "Biodiversity, Stability and Productivity in Competitive Communities," *The American Naturalist* 156 (2000): 534–552; Michel Loreau, "Linking Biodiversity and Ecosystems: Towards a Unifying Ecological Theory," *Philosophical Transactions of the Royal Society-Biological Sciences* 365 (2010): 49–60.

[15] Cardinale et al., "Effects of Biodiversity on the Functioning of Trophic Groups and Ecosystems"; Balvanera et al., "Quantifying the Evidence for Biodiversity Effects on Ecosystem Functioning and Services"; Cardinale et al., "The Functional Role of Producer Diversity in Ecosystems."

[16] Ibid. as well as Tilman, Wedin, and Knops, "Productivity and Sustainability Influenced by Biodiversity in Grassland Ecosystems"; Hooper and Vitousek, "The Effects of Plant Composition and Diversity on Ecosystem Processes"; Tilman et al., "The Influence of Functional Diversity and Composition on Ecosystem Processes."

[17] Monserrat Villa et al., "Species Richness and Wood Production: A Positive Association in Mediterranean Forests," *Ecology Letters* 10 (2007): 241–250; Daniel Piotto, "A Meta-Analysis Comparing Tree Growth in Monocultures and Mixed Plantations," *Forest Ecology and Management* 255 (2008): 781–786; Alain Paquette and Christian Messier, "The Effect of Biodiversity on Tree Productivity: From Temperate to Boreal Forests," *Global Ecology and Biogeography* 20 (2011): 170–180.

[18] Michael P. Carey and David H. Wahl, "Determining the Mechanism by which Fish Diversity Influences Production," *Oecologia* 167 (2011): 189–198.

[19] David Tilman et al., "Diversity and Productivity in a Decade-Long Grassland Experiment," *Science* 292 (2001): 843–884; Bradley J. Cardinale et al., "Impacts of Plant Diversity on Biomass Production Increase through Time because of Species Complementarity," *Proceedings of the National Academy of Sciences* 104 (2007): 18123–18128; Peter B. Reich et al., "Impacts of Biodiversity Loss Escalate through Time as Redundancy Fades," *Science* 336 (May 4, 2012): 589–592.

[20] David Tilman, Peter B. Reich, and Johannes Knops, "Biodiversity and Ecosystem Stability in a Decade-Long Grassland Experiment," *Nature* 441 (2006): 629–632; Andrew Hector et al., "General Stabilizing Effects of Plant Diversity on Grassland Productivity through Population Asynchrony and Overyielding," *Ecology* 9 (2010): 2213–2220.

[21] Nathan R. Franssen, Michael Tobler, and Keith B. Gido, "Annual Variation of Community Biomass is Lower in More Diverse Stream Fish Communities," *Oikos* 120 (2011): 582–590. Salvatore Di Falco, Jean-Paul Chavas, and Melinda Smale, "Farmer Management of Production Risk on Degraded Lands: The Role of Wheat Variety Diversity in the Tigray Region, Ethiopia," *Agricultural Economics* 36 (2007): 147–156.

[22] Charles E. Mitchell, David Tilman, and James V. Groth, "Effects of Grassland Plant Species Diversity, Abundance, and Composition on Foliar Fungal Disease," *Ecology* 83 (2002): 1713–1726.

[23] Felicia Keesing et al., "Impacts of Biodiversity on the Emergence and Transmission of Infectious Diseases," *Nature* 468 (2010): 647–652.

[24] Joseph E. Fargione and David Tilman, "Diversity Decreases Invasion via Both Sampling and Complementarity Effects," *Ecology Letters* 8 (2005): 604–611.

[25] Kenneth G. Cassman et al., "Meeting Cereal Demand while Protecting Natural Resources and Improving Environmental Quality," *Annual Reviews of Environmental Resources* 28 (2003): 315–358.

[26] Youyong Zhu et al., "The Use of Rice Variety Diversity for Rice Blast Control," *Scientia Agricultura Sinica* 36 (2003): 521 – 527; Long Li et al., "Diversity Enhances Agricultural Productivity via Rhizosphere Phosphorus Facilitation on Phosphorus-Deficient Soils," *Proceedings of the National Academy of Sciences* 104 (2007): 11192 – 11196.

[27] Tilman, Lehman, and Thomson, "Plant Diversity and Ecosystem Productivity"; Tilman, "The Ecological Consequences of Changes in Biodiversity"; Loreau, "Biodiversity and Ecosystem Functioning"; Lehman and Tilman, "Biodiversity, Stability and Productivity in Competitive Communities"; Loreau, "Linking Biodiversity and Ecosystems."

[28] Ibid.

[29] Anthony R. Ives, Kay Gross, and Jennifer L. Klug, "Stability and Variability in Competitive Communities," *Science* 286 (1999): 542 – 544.

[30] Gretchen C. Daily, ed., *Nature's Services* (Washington, D.C.: Island Press, 1997), 375.

[31] Jonathan A. Foley et al., "Solutions for a Cultivated Planet," *Nature* 478 (2011): 337 – 342; David Tilman, Christian Balzer, Jason Hill, and Belinda L. Befort, "Global Food Demand and the Sustainable Intensification of Agriculture," *Proceedings of the National Academy of Sciences* 108 (2011): 20260 – 20264.

[32] David Tilman, Jason Hill, and Clarence Lehman, "Carbon-Negative Biofuels from Low-Input High-Diversity Grassland Biomass," *Science* 314 (2006): 1598 – 1600.

Postlude

May R. Berenbaum

MAY R. BERENBAUM, a Fellow of the American Academy since 1996, is Professor and Head of the Department of Entomology at the University of Illinois at Urbana-Champaign. She is interested in the chemical interactions between herbivorous insects and their host-plants, and the implications of such interactions on the organization of natural communities and the evolution of species. She has published more than two hundred articles in refereed scientific journals as well as six books about insects for the general public, including *The Earwig's Tail: A Modern Bestiary of Multi-Legged Legends* (2009), *Buzzwords: A Scientist Muses on Sex, Bugs, and Rock 'n' Roll* (2000), and *Bugs in the System: Insects and Their Impact on Human Affairs* (1994).

The essays in this issue are tributes to the ability of their scientist-authors to highlight and summarize the exciting developments within their own areas of expertise, and then to explain those developments in a clear and accessible manner. Moreover, they are tributes to the courage of these authors; predicting the future, even that of a scientific field, is an inherently unscientific enterprise. The conduct of science depends on cold, hard, verifiable facts, and forecasting the future is necessarily rife with uncertainty. This may partly explain why predicting the future of science has often been the province of non-scientists. For example, it was a French professor of Hebrew, Syriac, and Chaldaic at the Collège de France, one Ernest Renan, who in 1890 penned the essay, "The Future of Science."[1] (This was a man initially trained for the priesthood, no less, and whose best-known work was *Life of Jesus*, one volume in a series on the history of Christianity.) Profession notwithstanding, Renan insightfully wrote, "Science will always remain the gratification of the noblest craving of our nature; curiosity; it will always supply man with the sole means of improving his lot" – a sentiment that our scientist-essayists certainly would endorse.

On this side of the Atlantic, Charles August Kraus was one of the first scientists to anticipate the future of the scientific enterprise. A professor of chemistry at Clark University, Kraus presented "The Future of Science in America" on February 1, 1917, in honor of Founder's Day at his home insti-

tution. Like Renan, he spoke of pure science as a noble pursuit:

> The scientist, to be worthy of the name, must be possessed of an insatiable desire to extend knowledge in its most fundamental aspects. He must not count the years of preparation required to actually master his subject, nor the labor necessary to transmute a crude idea into a well polished, finished scientific product. He must never be satisfied with mediocrity, and must ever strive to increase his scope in order that he may produce results of more fundamental importance.[2]

Kraus went on to conduct supremely useful chemical research, directing the Chemical Warfare Service at Clark during World War I, after which he left for Brown University, consulted on the Manhattan Project, and helped develop, among other things, the atomic bomb, Pyrex glass, leaded gasoline, and ultraviolet lamps.

Our essayists would agree with another century-old prediction: namely, that much remains to be learned. Renowned geneticist J.B.S. Haldane addressed this issue in his 1924 essay, aptly named "Daedalus; or, Science and the Future":

> The possibility has been suggested – I do not know how seriously – that the progress of science may cease through lack of new problems for investigation. Mr. [G. K.] Chesterton in *The Napoleon of Notting Hill*, a book written fifteen years or so ago, prophesied that hansom-cabs would still be in existence a hundred years hence owing to a cessation of invention. Within six years there was a hansom-cab in a museum, and now that romantic but tardy vehicle is a memory like the trireme, the velocipede, and the 1907 Voisin biplane. I do not suggest that Mr. Chesterton be dragged – a heavier Hector – behind the last hansom cab, but I do contend that, in so far as he claims to be a prophet rather than the voice of one crying in the wilderness, he may be regarded as negligible for the purposes of our discussion. I shall try shortly to show how far from complete are any branches of science at the present time.[3]

Our twenty-first-century collection of essays similarly celebrates the power and beauty of basic research and its ever-receding frontiers. A common theme is that the boundaries of scientific knowledge are elastic, in all directions. Little-known Latinate prefixes are now common parlance; technology is nanoscale; and flyby spacecraft missions require the use of distance measures equivalent to "thirty times the Earth's distance from the sun" (that distance, 93 million miles or 150 million kilometers, is referred to as an "astronomical unit"). New vocabulary demonstrates how our knowledge is expanding: every essay contains words that would have been meaningless to Russell, Haldane, and other eminent twentieth-century authors attempting to predict the future of science. All the essayists in this volume anticipate continuing advances, with the expectation that frontiers in their respective fields will be pushed back at an accelerating pace.

The essayists writing at the turn of the last century were in agreement that predicting the utility of basic science is difficult. Haldane stated, "It is perhaps interesting to speculate on the practical consequences of Einstein's discovery." He might have been surprised by the many practical consequences of Einstein's fundamental insights, not the least of which involve nuclear fission, fusion, global positioning systems, and even semiconductors. Philosopher-mathematician Bertrand Russell, in his 1924 response to Haldane's essay, "Icarus; or, The Future of Science," specifically cited "Darwinism" when he wrote: "The effect of the biological sciences, so far, has been very small. . . . It is

probable that great effects will come from these sciences sooner or later."[4] Nearly one hundred years later, Russell's prediction has proven true. Indeed, understanding evolution is essential to comprehending much of the subject matter in this volume, including genomic science and its applicability toward curing human diseases as well as genetically modifying plants; microbial science and contributions of biofilms to microbial pathogenesis; shared developmental pathways involving limb development; and even estimating the probability of life in other solar systems.

Both Haldane and Russell were also well aware of the fact that, beyond being difficult to predict, the utility of scientific advances often depends on the political and social environments in which they are made. Russell's essay begins with a dire prediction: "Much as I would like to agree with [Haldane's] forecast, a long experience of statesmen and governments has made me somewhat skeptical. I am compelled to fear that science will be used to promote the power of dominant groups, rather than to make men happy." Today, as in the twentieth century, how (and even whether) scientific advances are used depends to an enormous extent on that substantial majority of the population who are not scientists. In this respect, I can confidently make one prediction that I believe will withstand the test of time: without a well-educated public, the future of science in the United States is bleak.

The U.S. scientific enterprise, more so than in most nations, depends on a public not only supportive of federal funding for basic research but also capable of crafting and adopting policies based on solid science. So with knowledge expanding at such a rapid pace, can the vast majority of Americans who are not professional scientists keep up? There are disturbing indicators that a gap is widening between scientists and the general public. According to a 2006 study, less than 40 percent of Americans understand the concept of evolution by common descent – a proportion lower than that of any other industrialized nation.[5] A lack of understanding of such a well-supported, widely accepted fundamental scientific principle could undermine the future productivity of science writ large. Moreover, there are indications that the problem in the American public is more than not understanding; tolerance (even encouragement) of willful rejection of well-established science is on the rise. One-third of Americans polled averred that evolution is "absolutely false," a significantly higher proportion than in any other Western country.

Our competitiveness as a nation in the twenty-first century depends, much as it did in the twentieth century, on the ability of scientific advances to improve the human condition; and the viability of the scientific enterprise depends on an informed and supportive electorate. Our peculiar outlier status among the scientifically advanced nations of the world with regard to evolutionary theory is symptomatic of a larger problem: that is, fewer than one in five Americans possesses the background skills required to read an article about science in a newspaper, follow a television program on a scientific subject, or understand a popular science book.[6]

Perhaps even more important than knowing scientific facts is understanding the scientific process – that theories have predictive power yet are provisional, that there are many ways to test hypotheses, and that conclusions must be based on strong and repeatable evidence. In a recent report based on the 2012 *Science and Engineering Indicators*, the National Science Foundation revealed that a significant number of Americans tested on their

ability to understand experimental design, in addition to knowledge of factual material, lacked this basic understanding. On questions "measuring the concepts of scientific experiment and controlling variables," the number of correct responses ranged from 29 percent to 57 percent, with only 12 percent of respondents answering all these questions correctly and 20 percent answering none correctly.[7] Understanding the scientific process is arguably more important for scientific literacy than knowing specific scientific facts, as this understanding makes even the most arcane of science disciplines graspable.

Fortunately, there is reason for optimism. The same survey that revealed problems with American understanding of scientific methods has for three decades shown that Americans understand why basic science is important. Over the years, the *Science and Engineering Indicators* surveys have asked Americans whether "even if it brings no immediate benefits, scientific research that advances the frontiers of knowledge is necessary and should be supported by the Federal Government"; remarkably, the proportion who agree has continued to rise, reaching 84 percent in 2008. Moreover, the 11 percent of respondents in 2006 and 2008 who felt that the federal government invested too much in research represented "the lowest levels registered since 1981."

Haldane wrote in 1924 that "we must educate our poets and artists in science." To that sentiment, which we heartily endorse, we add that we must educate everyone in science; this mission includes educating our scientists beyond the boundaries of their own disciplines. We hope that our contributors, in presenting their vision of science in the twenty-first century, have succeeded in stirring that "noblest craving of our nature" – curiosity – our best investment for a happy, healthy, and well-informed future.

***

Acknowledgments: I am very grateful to Dr. Richard W. Burkhardt, a superbly knowledgeable historian of science who generously helped me examine and understand the history of the future of science.

ENDNOTES

1 Ernest Renan, "The Future of Science: Ideas of 1848," translated from the French (London: Chapman and Hall, 1891). Biographical information on Ernest Renan is from "Notes & Obituary Notes," *Popular Science Monthly* 42 (December 1892).

2 Charles August Kraus, "The Future of Science in America: An Address Delivered on Founder's Day February 1, 1917," Publications of the Clark University Library, vol. 5, no. 3 (Worcester, Mass.: Clark University Press, 1917), http://www.archive.org/details/futureofscience100kraurich.

3 J.B.S. Haldane, "Daedalus; or, Science and the Future," a paper read to the Heretics, Cambridge, on February 4, 1923 (London: K. Paul, Trench, Trubner, 1925).

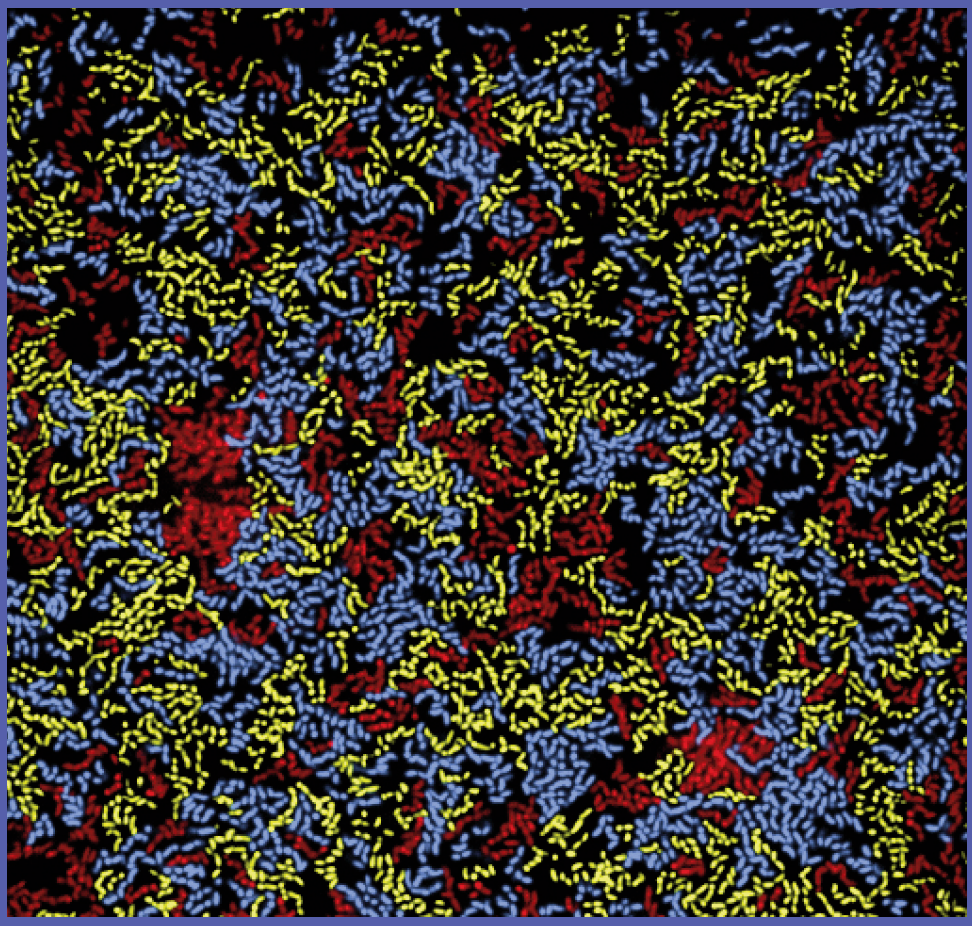
4 Bertrand Russell, "Icarus; or, The Future of Science" (London: K. Paul, Trench, Trubner, 1924).

5 Jon D. Miller, Eugenie C. Scott, and Shinji Okamoto, "Public Acceptance of Evolution," policy forum, *Science* 313 (5788) (August 2006): 765 – 766.

6 Jon D. Miller, "Civic Scientific Literacy: A Necessity in the 21st Century," *FAS Public Interest Report: The Journal of the Federation of American Scientists* 55 (2002): 3 – 6.

7 National Science Foundation Science Board, *Science and Engineering Indicators*, 2012, http://www.nsf.gov/statistics/seind12/.

Inside back cover: This image, a scanning confocal micrograph, shows a bacterial community in which three bacterial strains were grown together. The strains are identical except for the color that they express (red, blue, or yellow); the image illustrates how cell lineages that are initially well-mixed grow into pockets that contain only close relatives. This spatial effect is important for bacterial social interactions, including how they communicate with each other. Image courtesy of Dr. Carey Nadell and Dr. Bonnie Bassler, Department of Molecular Biology, Princeton University.



coming up in Dædalus:

Public Opinion	Lee Epstein, James Druckman & Thomas Leeper, Robert Erikson, Linda Greenhouse, Diana Mutz, Kevin Quinn & D. James Greiner, Gary Segura, James Stimson, James Gibson, and others
The Alternative Energy Future, vol. 2	Robert Fri, Stephen Ansolabehere, Jon Krosnick, Naomi Oreskes, Kelly Sims Gallagher, Thomas Dietz, Paul Stern & Elke Weber, Roger Kasperson & Bonnie Ram, Michael Dworkin, Pamela Matson & Rosina Bierbaum, Dallas Burtraw, Ann Carlson, Robert Keohane & David Victor, and others
The Common Good	Norman Ornstein, William Galston, Amy Gutmann & Dennis Thompson, Mickey Edwards, Thomas Mann, Deborah Tannen, Howard Gardner, Geoffrey Stone, and others

plus Immigration & the Future of America, New American Music &c.

AMERICAN ACADEMY
OF ARTS & SCIENCES
Cherishing Knowledge · Shaping the Future

U.S. $13; www.amacad.org