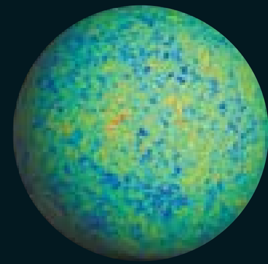
SPECIAL REPORT: FOUR KEYS TO THE COSMOS

SCIENTIFIC AMERICAN

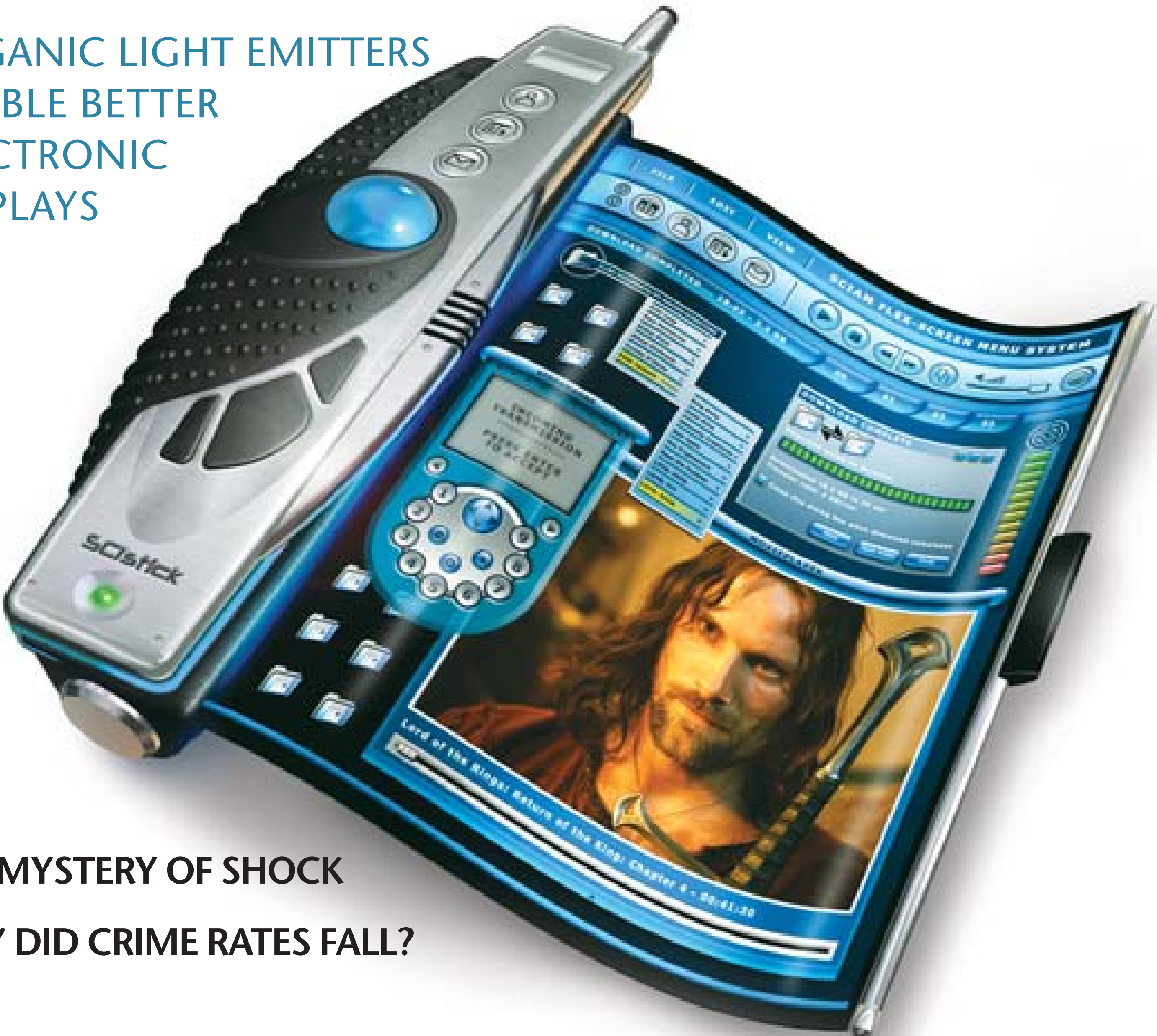
FEBRUARY 2004 $4.95 WWW.SCIAM.COM



Making Sense of Microwave Ripples, Gravity Leaks and More

THE FUTURE LOOKS **FLEXIBLE**

ORGANIC LIGHT EMITTERS ENABLE BETTER ELECTRONIC DISPLAYS



THE MYSTERY OF SHOCK

WHY DID CRIME RATES FALL?

february 2004

contents

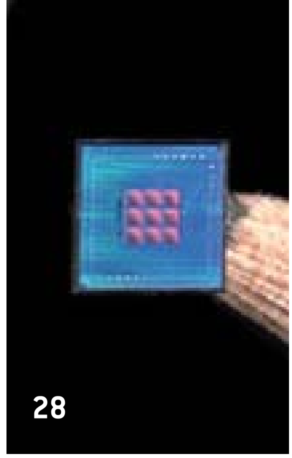
SCIENTIFIC AMERICAN Volume 290 Number 2

features

62 Type Ia supernovae shed light on changing rates of cosmic expansion

SCIENTIFIC AMERICAN Volume 290 Number 2

departments


28


92


Bonnie L. Bassler, Princeton University 34

columns

Cover image by Kenn Brown and Chris Wren, Mondolithic Studios; © New Line/courtesy of Everett Collection.

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 2004 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Periodicals postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 242764. Canadian BN No. 127387652RT; QST No. Q1015332537. Subscription rates: one year $34.97, Canada $49 USD, International $55 USD. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537. **Reprints available:** write Reprint Department, Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111; (212) 451-8877; fax: (212) 355-0408 or send e-mail to sacust@sciam.com **Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631.** Printed in U.S.A.

SA Perspectives

A Waste of Energy

America needs a new energy policy to reduce its reliance on foreign oil, but the $26-billion measure that stalled in Congress last November clearly wasn't it. The bill was bloated with $17 billion in tax breaks intended to spur production of oil, natural gas, coal and nuclear power. Although the act would have also funded efforts to reduce greenhouse-gas emissions—such as the Clean Coal Power Initiative—its strategy was wasteful and wrongheaded. The energy bill would have spent billions of taxpayer dollars on the development of unproven technologies that may never be adopted by the private sector.


COAL-FIRED power plant

Rather than resurrecting the failed 2003 bill this year, Congress should start afresh with a law focused on energy conservation. The energy saved through efficiency measures since the 1970s has been far greater than that produced by any new oil field or coal mine. As those measures came into effect between 1979 and 1986, the U.S. gross domestic product rose 20 percent while total energy use dropped 5 percent. Last year's energy bill would have set new efficiency standards for several products (traffic signals, for instance) and provided tax incentives for energy-efficient buildings and appliances, but the government can do much more.

Many economists argue that the best conservation strategy would be to establish an across-the-board energy tax. Under this approach, Congress would not dictate any efficiency standards; rather businesses and consumers would voluntarily avoid energy-guzzling appliances, heating systems and vehicles to minimize their tax bills. European countries, for example, have successfully boosted the average fuel economy of their cars by imposing high taxes on gasoline. But raising energy taxes would place a disproportionate burden on poor Americans if the new excises were not accompanied by some relief for low-income people. And the idea is a political nonstarter in Washington, D.C., anyway.

A more palatable approach would be to bolster energy conservation efforts that are already proving their worth. More than 20 states have public benefits funds that assess small charges on electricity use (typically about a tenth of a cent per kilowatt-hour) and direct the money toward efficiency upgrades. New York's Energy Smart Program, for instance, has cut annual energy bills in the state by more than $100 million since 1998, and current projects are expected to double the savings. Nationwide, however, ratepayer-financed programs lost ground in the 1990s because of utility deregulation. Congress can correct this problem by creating a federal fund that would match the state investments.

Another smart move would be to raise the Corporate Average Fuel Economy (CAFE) standards for cars and light trucks. Thanks in large part to CAFE, which was introduced in 1975, the average gas mileage of new vehicles in the U.S. reached a high of 26.2 miles per gallon in 1987. But the average has slid to 25.1 mpg since then, partly because more people are buying sport-utility vehicles, which are held to a lower standard than cars. At the very least, Congress should remove the loophole for SUVs. Automakers have the technology to improve fuel economy, and consumers will benefit in the end because their savings at the gas pump will far outweigh any markups at the car dealership.

According to the American Council for an Energy-Efficient Economy, a law that establishes a federal benefits fund and raises CAFE standards could reduce annual energy usage in the U.S. by nearly 12 percent. To put it another way, conservation would eliminate the need to build 700 new power plants. That's a lot of juice.

THE EDITORS editors@sciam.com

LESTER LEFKOWITZ *Corbis*

How to Contact Us

EDITORIAL

For Letters to the Editors:
Letters to the Editors
Scientific American
415 Madison Ave.
New York, NY 10017-1111
or
editors@sciam.com
Please include your name and mailing address, and cite the article and the issue in which it appeared. Letters may be edited for length and clarity. We regret that we cannot answer all correspondence.

For general inquiries:
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-754-0550
fax: 212-755-1976
or
editors@sciam.com

SUBSCRIPTIONS

For new subscriptions, renewals, gifts, payments, and changes of address:
U.S. and Canada
800-333-1199
Outside North America
515-247-7631
or
www.sciam.com
or
Scientific American
Box 3187
Harlan, IA 51537

REPRINTS

To order reprints of articles:
Reprint Department
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-451-8877
fax: 212-355-0408
reprints@sciam.com

PERMISSIONS

For permission to copy or reuse material from *SA*:
www.sciam.com/permissions
or
212-451-8546 for procedures
or
Permissions Department
Scientific American
415 Madison Ave.
New York, NY 10017-1111
Please allow three to six weeks for processing.

ADVERTISING

www.sciam.com has electronic contact information for sales representatives of Scientific American in all regions of the U.S. and in other countries.

New York
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-451-8893
fax: 212-754-1138

Los Angeles
310-234-2699
fax: 310-234-2670

San Francisco
415-403-9030
fax: 415-403-9033

Midwest
Derr Media Group
847-615-1921
fax: 847-735-1457

Southwest
Publicitas North America, Inc.
972-386-6186
fax: 972-233-9819

Detroit
Karen Teegarden & Associates
248-642-1773
fax: 248-642-6138

Canada
Derr Media Group
847-615-1921
fax: 847-735-1457

U.K.
The Powers Turner Group
+44-207-592-8331
fax: +44-207-630-9922

France and Switzerland
PEM-PEMA
+33-1-46-37-2117
fax: +33-1-47-38-6329

Germany
Publicitas Germany GmbH
+49-211-862-092-0
fax: +49-211-862-092-21

Sweden
Publicitas Nordic AB
+46-8-442-7050
fax: +46-8-442-7059

Belgium
Publicitas Media S.A.
+32-(0)2-639-8420
fax: +32-(0)2-639-8430

Middle East
Peter Smith Media & Marketing
+44-140-484-1321
fax: +44-140-484-1320

India
Yogesh Rao
Convergence Media
+91-22-2414-4808
fax: +91-22-2414-5594

Japan
Pacific Business, Inc.
+813-3661-6138
fax: +813-3661-6139

Korea
Biscom, Inc.
+822-739-7840
fax: +822-732-3662

Hong Kong
Hutton Media Limited
+852-2528-9135
fax: +852-2528-9281

On the Web

WWW.SCIENTIFICAMERICAN.COM

FEATURED THIS MONTH

Visit **www.sciam.com/ontheweb** to find these recent additions to the site:

Autopsies, No Scalpel Required

Television shows such as *CSI* and *Law & Order* have brought forensic science into the living room, but actual autopsies are not for the faint of heart. A recently developed technique may change that. Virtopsy, a virtual autopsy procedure, demands neither a scalpel nor a strong stomach.

Changing Climate May Leave Wintering Monarchs Out in the Cold

Every winter millions of monarch butterflies make their way from North America to Mexico in search of warmer climes. Within 50 years, however, the butterflies may find themselves with nowhere to go. Climate change—particularly an increase in wet weather in the area—may make the monarch's winter home uninhabitable.

Nearby Star May Have Planetary System Like Ours

Astronomers scanning the skies for far-flung planets have found that the region surrounding a nearby star looks very familiar. Vega, located 25 light-years away from our sun, may have an orbiting planetary system that is more similar to our own than any other yet discovered.

Ask the Experts

What kinds of patterns do scientists working on the Search for Extraterrestrial Intelligence (SETI) project look for?

Peter R. Backus, observing programs manager at the SETI Institute, explains.

Scientific American.com e-News

SIGN UP FOR FREE E-NEWS FROM SCIENTIFICAMERICAN.COM TODAY!

Get the latest science and technology news delivered straight to your mailbox with our Weekly Review, Tech Biz Alert, Best-Seller Alert and more.

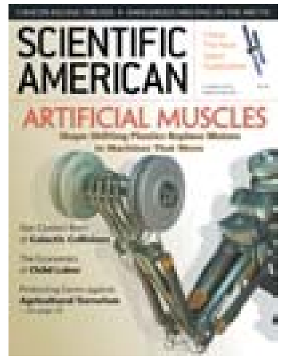
www.sciam.com/enews/

COURTESY OF RSNA AND M. THALI (*skull image*)

Letters

EDITORS@SCIAM.COM

OCTOBER IS A MONTH with a knack for drama. Soaking hurricanes attack unsuspecting crisp autumn days. Floods sweep through one state while wildfires rage in another. And although weather is not really a reliable litmus test of public mood, the unsettled conditions experienced last fall do create a fitting theatrical setting for the reader responses to the October issue. With critical probes of socioeconomic theories, nervous nods at potential new cancer therapies, faltering faith in government offices, and even some humorous commentary on our Chinese translations, the ensemble of contributions to this month's letters column truly had it all.



PARSING THE IMPASSE

In SA Perspectives ["Biotech's Clean Slate"], the editors propose the "wild thought" that good experiences with industrial biotechnology might assuage the public's fears about agricultural and medical biotechnology enough to end the "biotech impasse." Here's another wild thought: maybe the biotech impasse is not about the safety of genetically modified organisms in the first place. Maybe it stems from fears that our current risk-assessment methodology is not an adequate guide into a world where corporate entities hold ever more potent tools with ever shrinking attention to any but fiduciary responsibility.

Jim Roy
Albany, Ore.

ELABORATIONS ON CHILD LABOR

Kaushik Basu, in "The Economics of Child Labor," mentions that economic forces are not the sole determinants of child labor. But his discussion of the noneconomic determinants of child labor, including education, only scratched the surface.

The quality of the education available as an alternative to exploitative work is critical, because schools that fail to teach useful knowledge may drive children away from school learning into paid or unpaid work. For example, one often cited reason that children leave school to labor at the railway station of Bhubaneswar, Orissa, India, is that schools cannot teach usefully for students with learning disabilities such as dyslexia.

In 1995 I interviewed impoverished families in Brazil's Bahia state, in which children work long days cutting sisal alongside their parents. The parents said they wanted to send their children to school but no after-school program existed to ensure the children's safety until the parents returned home from the plantations. Economic poverty was less at issue than the poverty of state education policies and services and the failure of local communities to fill in the gaps with government or nongovernment services.

S. L. Bachman
Child Labor and the Global Village: Photography for Social Change

Basu suggests that the main factor behind child labor is poverty. I contend that the problem begins with the market economics that generate poverty and thrive on the exploitation of the weak. He points out that the outright outlawing of child labor can have negative effects on the working children themselves. This dilemma illustrates that real-life markets are never the idealized realms of free and fair exchange that some economists imagine them to be. Rather they are spaces of unequal power and other asymmetries—fields in which every available opportunity is bound to be exploited, even if that "opportunity" is the weakness of children with no options but starving or laboring at market-dictated terms.

Fidel Fajardo-Acosta
Creighton University

BASU REPLIES: Bachman is correct that child labor is caused not just by poverty but several other factors, including noneconomic ones.

She feels I downplay those causes. She is right, but I do so only to the extent that reality demands. It is true that having special provisions in schools for children with dyslexia would help keep some children in Bhubaneswar away from hard labor and that the availability of after-school care may persuade some kids in Bahia to go to school. But there are problems with viewing such an argument as a general cure for child labor. First, both examples are, at one remove, economic. How much a school can provide may not depend on household poverty but is constrained by the economic conditions of the region. Second, her examples ring true because of our implicit awareness that the households from which the children come are extremely poor. If today in the U.S. there were no special schools and no after-school care, it is still doubtful that children in these areas would get packed off for hard labor.

Thus, whereas we should be aware of the many fronts on which we need to combat child labor, it would be a failure not to recognize that the overwhelming cause of child labor is poverty.

Fajardo-Acosta is right that markets are not idealized realms of free and fair exchange. But it is also true that markets are not a detachable part of human social life. Hence, the statement that markets cause poverty is hard to comprehend because it is not clear what the negation of markets means. Government does have a role to play in distributing wealth and curbing large inequities of power, and one must not leave everything to the untrammeled forces of the market. But beyond that, it is not clear what one can do. Markets are such an integral part of human life that to lay all blame at their doorstep is like lamenting that but for gravity we would not all be down.

PATENT UPENDING

In "Kick Me, Myself and I" [Staking Claims], Gary Stix writes about a butt-kicking machine for which the U.S. patent office issued a patent. I spent a few minutes searching the Web and found two examples of butt-kicking machines that appear to predate patent 6,293,874. I also recall a few cartoons and old films in which a character submits to a self-butt-kicking machine of some sort.

If I can arrive at "prior art" with a few minutes of Web searching and a bit of retrospection, what is the patent office doing with taxpayer dollars?

Gerald A. Hanweck, Jr.
New York City

METASTASIZING CONCERNS

In "Tumor-Busting Viruses," by Dirk M. Nettelbeck and David T. Curiel, it is telling that your illustration of transductional targeting shows unmodified viruses exiting the cancer cell. If the adapter molecules that are used to tailor tumor-busting viruses to tumors are not genetically encoded within the viruses themselves, the next viral generation will be unmodified and therefore able to infect healthy tissues. Furthermore, even genetically modified viruses are not foolproof. A nonsense mutation in the genes encoding the adapter molecules could lead to infection of healthy nontarget tissues.

Finally, viruses have incredible evolutionary potential: witness the great diversity of HIV genotypes within a single infected person, as well as our inability to develop any kind of lasting immunity to common colds or influenza. Using live viruses to treat people whose systems are already weakened by cancer seems like a very risky proposition, like swallowing a spider to catch a fly.

Scott M. Ramsay
Wilfrid Laurier University
Waterloo, Ontario

NETTELBECK AND CURIEL REPLY: Despite the cues in the illustration to which Ramsay refers, adapter molecules have successfully retargeted adenoviral infection to specific cells. Adapters might provide a means to improve therapeutic efficacy by targeting the initial virus inoculum. But current research efforts focus on the direct modeling of virus-coating proteins by genetic strategies rather than on the harnessing of adapters. This genetic approach does reliably redirect the virus. A comparison between adenoviruses and HIV does not hold true for two reasons: first, adenoviruses are far less pathogenic than HIV, and second, adenoviruses possess a stable double-stranded DNA genome that does not allow the enormous mutation rate observed for the HIV RNA genome.

SHENZHOU 3 command module landed in Mongolia after one week in orbit.

MAGIC SPACE BOATS?

Your article on the Chinese space program ["China's Great Leap Upward," by James Oberg] translates the Chinese name for their spacecraft, *Shenzhou,* as "divine vessel." *Shen* does mean "divine" or "sacred" but can also translate to "magic." *Zhou* can mean "vessel," but it can also mean "boat." In view of differences in cultural perception between China and the West regarding the role of divine intervention in human affairs and bearing in mind that this is an atheist, Communist government, I wonder if "magic boat" may be a more culturally appropriate translation. After all, Chinese children know that Aladdin traveled by *Shen-tan*. Although that term could be formally translated as "sacred floor covering," "magic carpet" might be closer to the mark.

Peter C. Chen
Columbia, Md.

GETTY/AFP

Imaginary Spies ▪ Illusory Rays ▪ Erroneous Statistics

FEBRUARY 1954

RED FEAR—"The Fort Monmouth spy story fizzled out last month. Senator Joseph R. McCarthy concluded a series of public hearings, which he had said would 'show that there was espionage' in the Fort Monmouth radar laboratory. His parade of witnesses has failed to develop any testimony on spying. Of some 30 Signal Corps scientists suspended by the Army as a result of the McCarthy investigation, none was accused of espionage. The *New York Herald Tribune* writer Walter Millis reported in his column: 'This really vital and sensitive military installation has been wrecked—more thoroughly than any Soviet saboteur could have dreamed of doing it ... [through] the processes of witch-hunting, sheer bigotry, cowardice, race prejudice and sheer incompetence.'"

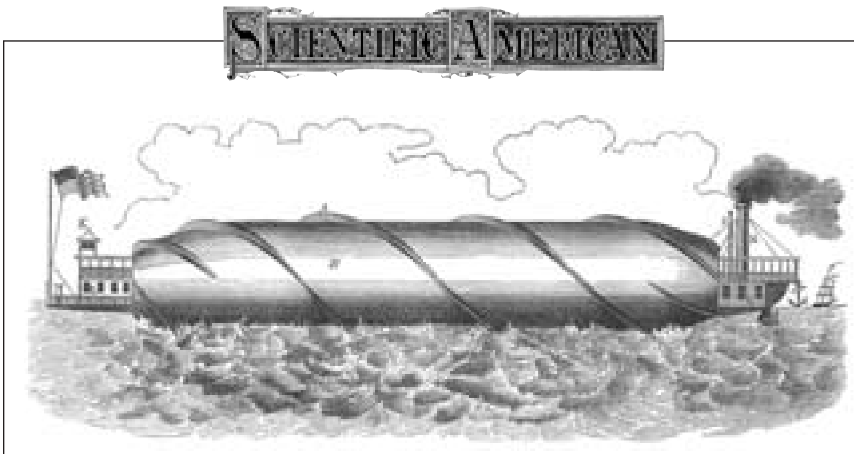
RABBIT PLAGUE—"It is with diametrically opposite feelings that different parts of the world now look upon the two-edged phenomenon which is the subject of this article—the deadly infectious disease of rabbits called myxomatosis. Introduced deliberately in Australia three years ago, it has swept rapidly over immense areas, causing great epizootics among rabbits. In Australia the disease is hailed as a measure of salvation which is ridding the continent of its major pest; in Europe, where it broke out in 1952, it is viewed as a malevolent killer which threatens to wipe out a favorite food, game, pet and laboratory animal. To check the disease in Europe, investigators are searching for a vaccine against the myxoma virus."

FEBRUARY 1904

CHIMERICAL RAYS—"M. Aug. Charpentier brings out the interesting point that the rays given out by living organisms differ from the N-rays discovered by M. René Prosper Blondlot, and he thinks they are formed of N-rays and another new form of radiation. This is especially true of the rays from the nerve centers or nerves, whose striking characteristic is that they are partially cut off by an aluminum screen. A sheet 1/50th of an inch is sufficient to cut down considerably the rays emitted by a point of the brain. On the contrary, the rays from the heart, diaphragm, and different muscles are scarcely modified by the aluminum screen. This forms a characteristic distinction between the muscular and the nerve radiations. The effect from the nerves is strongly increased by compression; that of the muscles is much less so." [*Editors' note: Both these forms of radiation were eventually disproved.*]

EIFFEL'S TOWER—"In the *Scientific American* of December 26th it was announced that the famous Eiffel Tower was about to be razed to the ground, for the reason that it displayed a marked toppling tendency. M. Eiffel denies the statement and refers to the report of M. Mascart, president of the Academy of Sciences, in which it is said that 'the tower is in a perfect state of preservation, and that no change of position has been noted either in the foundation or in the framework.' Every competent commission that has ever studied the tower has advocated the preservation of the structure, and vouched for its scientific utility."

AND NOW THE BAD GNUS—"There seems to be no doubt that the wild grotesqueness of the appearance of the gnu is a provision of nature to protect the animal. When frightened or disturbed, these remarkable antelopes go through a series of strange evolutions and extraordinary postures, in order to enhance the oddity and hideousness of their appearance, and to frighten away intruders."



MARINE LOCOMOTION before the science of fluid dynamics, 1854

FEBRUARY 1854

OCEAN PERIL—"The annexed engraving is of a Marine Locomotive, invented by Henry A. Frost of Worcester, Mass. To the outer hull are attached the screw blades. The inner cylinder being loaded at the bottom, will continually maintain the same position. It has a saloon running the full length of the vessel. The inventor is very confident that he is creating a complete revolution in ocean traveling."

STATISTICAL ERROR—"There are on the earth 1,000,000,000 inhabitants; of these 33,333,333 die every year, or one every second. These losses are about balanced by an equal number of births."

news SCAN

INFOTECH

Ballot Breakdown

FLAWS CONTINUE TO HAMPER COMPUTERIZED VOTING **BY WENDY M. GROSSMAN**

Even before the last chad was detached in the 2000 Florida election fiasco, discussions began about how to improve the voting systems in the 170,000-odd jurisdictions in the U.S. The Help America Vote Act, which passed in October 2002, allocates $3.8 billion to modernize voting systems across the nation. In large part, that modernization has led to the consideration of computerized voting. But although everyone agrees that punch cards must go, so far no one can agree on standards for the systems to replace them. The biggest bone of contention: finding a way to let voters check that their votes have been cast the way they intended. The solution, in fact, may lie with paper.

VERIFYING VOTES through a recount, as was done in the 2000 presidential election in Palm Beach County, Florida, has proved to be a stumbling block for all-electronic voting systems that have been proposed.

BRUCE WEAVER *AFP/Corbis*

To develop standards that all voting machines would meet, the Help America Vote Act turned to the Institute of Electrical and Electronics Engineers (IEEE). Project 1583 is the resulting effort and is intended, the IEEE summary says, to assure confidentiality, security, reliability, accuracy, usability and accessibility. To set standards, an IEEE working group first puts together a draft proposal, which it sends out for public comment. Then the draft must pass a vote by the members of the standards association, a subset of the IEEE's worldwide membership.

Like many standards efforts, most of the working-group members represent vendors, including Diebold Election Systems in McKinney, Tex., Election Systems and Software in Omaha, Neb., and the multinational election.com. Nonvendor members include cryptographer and digital-cash inventor David Chaum, Stanford University computer scientist David L. Dill, who also runs the Verified Voting campaign Web site, and Rebecca Mercuri, a fellow at Harvard University who wrote her dissertation on electronic voting systems.

The working group's September 2003 vote on adoption of the then current draft failed after nearly 500 people wrote to the IEEE pointing out flaws. The concerns had to do primarily with security and voter verifiability—that is, a method for polling officials to conduct a recount and for voters to ensure, before their ballots are finally cast, that they have voted the way they intended. It is unlikely that voting machines will be certified to the act's new standards before 2006.

The problem lies with all-electronic systems, known as "direct recording electronic," which are currently most likely to replace older machines. Although such systems can be tested—cast a known quantity of votes and then check that the machine has counted them correctly—there is no way to prove that the cast ballots were recorded properly or that those tallied bear any resemblance to actual votes, according to Mercuri. And there is no way to perform an independent audit. "To have a fair, democratic election, there has to be a visible, transparent way of performing recounts and confirming that ballots have been cast correctly," she explains.

A number of methods for adding voter verifiability to electronic machines have been suggested, and they all have one thing in common: paper. Chaum, for example, has demonstrated an ingenious two-part paper ballot, the top page of which is visible. Cryptographic coding ensures that while the two halves are assembled, the voter can see how the ballot was cast; once separated, the halves reveal nothing to third parties.

Simpler in conception is Mercuri's suggestion, which she has been promulgating since 1993. Electronic ballot boxes would be equipped with a glass screen and a printer. Each vote would be printed out on paper and the result dropped behind the glass screen for the voter to review before choosing to cast or void it. Such a system, she says, would reduce voter error and provide for a recount, if needed. Meanwhile the electronics could tabulate votes quickly, as our impatient society demands.

Mercuri's method is beginning to make some headway. California, for example, is considering mandating the creation of a contemporaneous paper record for each voter, and a bill in front of Congress would amend the Help America Vote Act to require a voter-verified permanent record. When it comes to votes, paper may be the wave of the future.

Wendy M. Grossman writes about information technology from London.

VOTER NONDISCLOSURE

Some observers worry that the latest attempts at computerizing voting systems could be privatizing too much power. In certain cases, vendors offering their voting machines require election officials to sign nondisclosure agreements to protect the secrets of how these machines work. The public must therefore place its trust in a handful of people who do not have to say how the software actually tabulates the votes.

ASTRONOMY

Not So Icy Stares

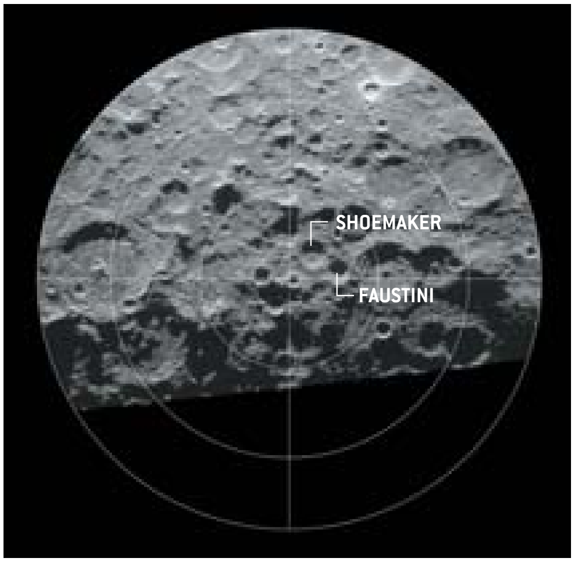
THE MOON'S WATER MAY BE DIFFICULT TO GET BY SARAH SIMPSON

For those who fantasize about a thriving human outpost on the moon, finding thick sheets of ice at the lunar poles would be like striking gold. Massive chunks could be carved out of the ground and melted for drinking, growing plants and making rocket fuel. Alas, that dreamy vision may have to be tempered. New moon scans from Arecibo Observatory in Puerto Rico suggest that future lunar colonists may have to make do with tiny ice crystals suspended within the lunar soil.

The Arecibo team, headed by Bruce A. Campbell of the Center for Earth and Planetary Studies at the Smithsonian Institution, used 70-centimeter-wavelength radar to probe up to five times as deep into the lunar surface as any of the earlier studies that found hints of ice. "We just wanted to be sure we hadn't missed anything," Campbell remarks.

Scientists have long suspected that some form of water ice survives inside deep craters near the lunar poles. There the leading edge of the sun never rises more than about two degrees above the horizon, casting long shadows that enshroud all low-lying areas in permanent darkness. When stray water molecules encounter these dark locales, they immediately solidify in the frigid temperatures, which never top –225 degrees Celsius.

Conflicting interpretations of radar measurements made in the late 1990s left open the possibility of glacierlike deposits in these so-called cold traps. Scientists still debate the meaning of the strong radar echoes they saw, which can be indicative either of thick slabs of ice or of the sloping, rugged terrain typical of crater walls. Less controversial was the 1998 announcement that the Lunar Prospector spacecraft had detected significant amounts of hydrogen—presumably from water—above the dark regions of the poles. But those early investigations probed just a meter or two into the lunar soil.



SHADOWED CRATERS at the lunar south pole—in particular, craters Shoemaker and Faustini—returned blank stares from ice-seeking radar, dashing hopes for easy access to moon water.

BRUCE A. CAMPBELL *Smithsonian Institution*

Campbell's team used the Arecibo radar to penetrate more than five meters into the floors of several small shadowed craters near the moon's north pole and into Shoemaker and Faustini craters at the south pole. When the observatory detected weak radar echoes, the researchers concluded that they must have been encountering only lunar rocks and dust, which absorb most of the beam.

Campbell admits that he was disappointed not to find thick ice deposits, but he points out that his team's results in no way imply that the moon is ice-free. Crystals as big as golf balls could be common components of the lunar soil, he explains, and they would be invisible to radar. The problem is that the crystals may well be snowflake-size or smaller, and harvesting them from other lunar material would be more challenging and costly than excavating blocks of ice.

Discovering slabs of ice below two meters would also have implied that the total supply of lunar ice was considerably greater than the best current estimate of 10 billion tons or more. That may sound like a lot, but it's really just a small lake's worth, notes Alan B. Binder, founder and director of the Lunar Research Institute in Tucson, Ariz. Lunar visitors could face tough decisions about how to conserve a very limited resource, Binder says.

Some researchers still insist that ice sheets may exist in places where no one has looked. Last year physicist Ben Bussey of Johns Hopkins University and his colleagues reported that the moon's cold traps might cover twice the area cited in previous estimates. Bussey also suggests that substantial deposits may exist in the floors of many impact craters that cannot be viewed from Earth.

No one seems to quibble that the next step is to take a closer look. NASA is already considering a mission that would probe the floor of a shrouded crater and send samples back. And with China and the Bush administration both discussing ambitious plans for manned lunar missions, astronauts may soon get a chance to see for themselves.

GETTING THE MOON ALL WET

There's no doubt that the moon gets its water from an external source, and water-laden comets often get the credit. But recent computer simulations indicate a less obvious origin: the sun. Hydrogen ions carried in the solar wind constantly pelt the lunar surface and occasionally hook up with oxygen atoms that are abundant in moon rocks. Most of the fledgling H_2O molecules break apart and escape into space, but some hop along the lunar terrain until they freeze inside sunless craters near the poles. There icebound hydrogen could constitute about 4 percent of the rocky soil—more than enough to account for the 1.5 percent hydrogen that Lunar Prospector detected in 1998.

ECOLOGY

When Blade Meets Bat

UNEXPECTED BAT KILLS THREATEN FUTURE WIND FARMS BY WENDY WILLIAMS

The interaction of bats and wind turbines is emerging as a major and unexpected problem in northern Appalachia. From mid-August through October 2003, during the fall migration period, at least 400 bats died at FPL Energy's 44-turbine Mountaineer Wind Energy Center on Backbone Mountain in West Virginia.

The bats apparently died by colliding with the wind turbines, but why so many animals were killed at this particular site remains a mystery. The public outcry over these numbers threatens to delay or halt construction of some of the additional several hundred wind turbines planned for the tristate region of West Virginia, western Maryland and south-central Pennsylvania.

Steve Stengel, a spokesperson for FPL, which is based in Juno Beach, Fla., says the company is cooperating with federal biologists to study the problem of bat kills at Mountaineer. "We don't know exactly why it happened," he states. "We're moving quickly to find out as much as we can." Some scientists believe that the migrating bats may not be using their echolocation when the collisions occur. Others speculate that the wind turbines may be emitting high-pitched sounds that draw the bats to the site. Still others suggest that the animals may be getting caught in wind shear associated with the turning turbines.

West Virginia biologists have identified the majority of the 400 bats that were recovered from the Mountaineer site—mostly common species such as red bats, eastern pipistrelles and hoary bats. "What's scary," remarks biologist Albert Manville of the U.S. Fish and Wildlife Service, "is that we may be finding only a small percentage of what's been killed." That is because bats are very small and difficult to find in the field; also, scavengers could discover the bat corpses before researchers do.

At issue is the length of time that wind-energy entrepreneurs are devoting to preconstruction wildlife studies. The Fish and Wildlife Service issued voluntary siting guidelines last summer, indicating that a census of wildlife activity should precede the building of a wind farm. Some biologists feel that such a census should last two years, although some energy companies believe this length of time to be excessive. (The guidelines are voluntary because in many cases the federal agency has little enforcement power unless an endangered or threatened animal is actually killed.)

Concerned that the endangered Indiana bat may be at risk at FPL's 20-turbine wind project in Meyersdale, Pa., wildlife advocates are threatening legal action. They allege that thorough habitat studies were not done in advance of construction at Meyersdale.

A letter last October from a bat biologist hired by the project's builders would appear to back them up. Pennsylvania State University's Michael R. Gannon spent two days last spring looking for bat caves on the future wind-farm site. He suggested that Indiana bats may use the site as a summer habitat and noted that at least a summerlong study might be appropriate. But industry biologists disagreed, Gannon says. "A two-year study should have been conducted prior to the installation of the turbines to determine the potential risk to bats," he wrote in his letter. "Unless and until these data are available, it should be assumed that this site is a flight path of the Indiana bats and that Indiana bats will be killed.... Data that are available indicate this as a very likely scenario."

FPL, which bought the project during development, still wants more information. "We are reviewing the matter," Stengel comments, "and after our review we will respond, if appropriate."

Wendy Williams, based in Mashpee, Mass., writes for Windpower Monthly, *an international news magazine.*

WHEN GREEN ENERGY meets red bats, the mammals seem to lose. Some wind farms are finding this species of bat, as well as many others, dead on their properties. Such discoveries could threaten planned wind farms and force revisions in the way turbines are sited.

LARRY L. MILLER *Photo Researchers, Inc.* (*bat*); LESTER LEFKOWITZ *Corbis* (*wind farm*)

NEED TO KNOW: BAT WATCHING

Bats have been killed at other wind turbine sites across the nation, but nothing on the scale of the recent Mountaineer incident has yet been documented. Keeping the bats away from the turbines means finding out just exactly how and why the bats are killed. One key could be infrared cameras, says James A. Simmons of Brown University, an expert on bat echolocation: "You have to watch the collisions occur. Because of the difficulty of seeing them, an infrared camera is the only way to do that." Merlin D. Tuttle of Austin, Tex.–based Bat Conservation International agrees: "There may be simple ways to solve these big problems, but first you have to take a look."

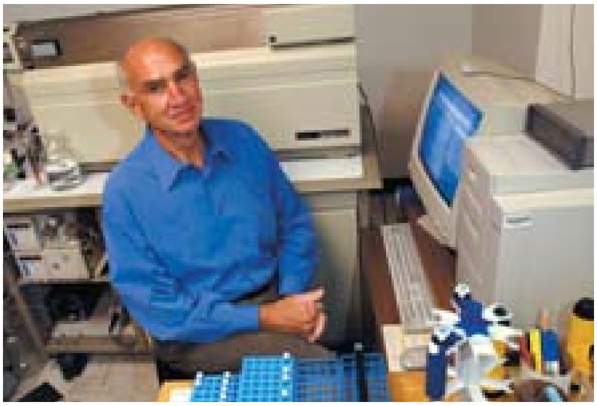
CHEMISTRY

Doping by Design

WHY NEW STEROIDS ARE EASY TO MAKE AND HARD TO DETECT

BY STEVEN ASHLEY

A **furor erupted** in the world of sports last fall when chemists announced that they had identified a new performance-enhancing synthetic steroid undetectable by standard antidoping tests. Scientists familiar with androgenic steroids and their illicit use in athletics were not at all surprised. "We've known about designer steroids for many years, but up to now we've never been able to prove that someone is actually making them," says Don H. Catlin, a molecular pharmacologist and director of the Olympic Analytical Laboratory at the University of California at Los Angeles. Catlin led the effort to isolate and analyze tetrahydrogestrinone (THG), the compound at the center of the storm. "The fact that we finally characterized one is certainly no reason to celebrate. I'm much more worried about the next THG out there that we haven't found yet."



STEROID SPOTTER: Pharmacologist Don H. Catlin feels that sports-governing bodies need to deal with underground labs proactively, before athletes start taking new drugs.

That's because it is fairly easy for organic chemists to design novel anabolic steroids that standard drug tests would not detect. (Identification depends on knowing the compound's structure beforehand; THG use was discovered only because an anonymous coach sent a spent syringe to U.S. antidoping officials.) All androgenic steroids are based on a chemical structure featuring a central complex of four hexagonal carbon rings. Small changes to the molecular groups attached to the periphery of central ring complex yield new derivatives. "Nature has made thousands of steroids, and chemists can make thousands more relatively easily," Catlin comments.

Rogue scientists start with testosterone or its commercially available analogues and then make minor structural modifications to yield similarly active derivatives. The underground chemists make no effort to test their creations for effectiveness or safety, of course. Production of a simple new steroid compound would require "lab equipment costing maybe $50,000 to $100,000," Catlin estimates. Depending on the number of chemical reactions needed for synthesis, "some of them could be made in a week or two. Others might take six months to a year."

"There are lots of good steroid chemists offshore who gained their expertise developing contraceptives and other hormone drugs decades ago," says Jean D. Wilson, an expert in androgen physiology at the University of Texas Southwestern Medical Center at Dallas. Now that birth-control pills have become a commodity product, "many of these experienced organic chemists are sitting around twiddling their thumbs," he says. "There must be thousands of people in the world who could readily synthesize designer steroids." The THG episode fuels speculation that a network of clandestine laboratories exists that develops and produces illegal steroids.

THG's chemical structure is similar to that of trenbolone and gestrinone, both synthetic anabolic steroids banned for athletic use by international sports federations. "Trenbolone is a veterinary drug used by cattle ranchers to increase the size of their stock," Wilson says. It is also popular with bodybuilders, despite toxic side effects. The structure of gestrinone, used to treat endometriosis and related illnesses, differs from that of THG by only four hydrogen atoms. In fact, once Catlin and his U.C.L.A. colleagues had deduced the structure of THG, they re-created it by hydrogenating (adding hydrogen atoms to) gestrinone, which yielded tetrahydrogestrinone.

Besides its novelty, a synthetic steroid's chemical stability under testing conditions also affects its chances of detection. THG tends to break down when prepared for analysis by standard means, which helps to explain why Catlin's team did not identify the compound in its first attempt. The U.C.L.A. chemists isolated THG's signature only after switching to a more sensitive assay process. They used liquid and gas chromatography to fractionate the sample into its molecular constituents; an electron beam then fragmented the separated molecules in a mass spectrometer to produce a

BULK UP, GET SICK

Synthetic anabolic steroids build muscle mass and endurance as well as permit faster recovery after strenuous activity, but proving those effects scientifically has been surprisingly difficult. First, athletes get a major placebo effect from doping themselves. Second, the dosages they typically use are many times as high as the levels that human-experimentation committees would approve. Focusing on the obvious benefits, users fail to recognize steroids' toxic side effects, some of which take time to manifest themselves: severe acne, heart disease, liver damage and uncontrollable rage.

ANN JOHANSSON *Corbis*

spectrum indicating the basic chemical components. The group subsequently developed a urine test for THG, which has been used to finger several well-known sports figures.

Still, the cat-and-mouse game that is athletic drug testing continues. The trouble is that the mice are fast-moving targets that never stop evolving. "We're looking forward for our next research project, and that includes looking for other designer steroids," Catlin reports. Perhaps they can pounce before the mouse disappears.

DINOSAURS

Becoming Behemoth

HOW TO PUT THE *REX* INTO *TYRANNOSAURUS* BY KATE WONG

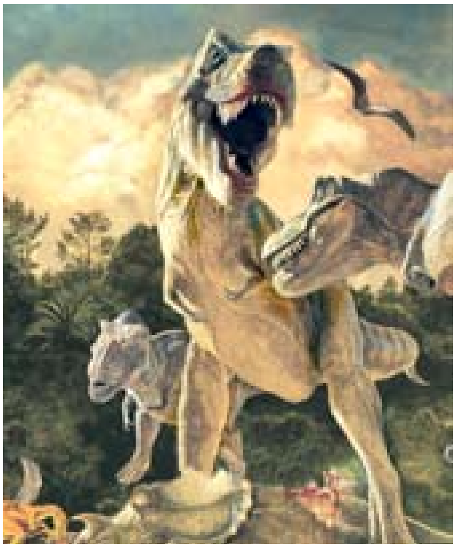
For the generation raised on *Jurassic Park* and perhaps for posterity, *Tyrannosaurus rex* will endure as a household name. It is the dinosaur people most love to loathe, a gargantuan, dagger-toothed monster every bit as fearsome as the fire-breathing dragons of fairy tales. Less widely appreciated is that the tyrant lizard king had modest roots. Indeed, before *T. rex* hit the scene, tyrannosaurs were relatively petite. Weighing one to two metric tons and standing several meters tall, these were not animals to be met in a dark alley or kept as pets. But in fact, *T. rex* broke the tyrannosaur mold, nearly tripling in body mass over its predecessors.

Evolutionary biologists have long pondered the factors that might have led to gigantism among theropods, the bipedal and mostly carnivorous dinosaurs. (Besides *T. rex,* four species in two other groups—the carcharodontosaurs and the spinosaurs—managed to evolve similarly supersize proportions.) In recent years, a few prime-mover hypotheses have emerged, attributing the growth spurt to such things as increased levels of atmospheric carbon dioxide (leading to elevated plant productivity, which in turn could support more of the herbivorous dinosaurs that theropods preyed on). At the annual meeting of the Society of Vertebrate Paleontology in St. Paul, Minn., last October, Scott Sampson of the University of Utah and his colleagues outlined a more nuanced model, one that takes multiple influences into account.

Piggybacking on studies of living carnivores, the team identified some likely prerequisites that permitted the rise of gigantic theropods—those that tipped the scales at three metric tons or more, as estimated from the circumference of the thigh bone. First, contrary to what many investigators have postulated, the beasts probably had to have been cold-blooded, because the costs of maintaining a constant body temperature—which is to say, being warm-blooded—at that size would have necessitated unrealistic hunting success (up to 10 times that of a lion) and a means of cooling down to avoid overheating.

The second requirement derives from the observations that meat-eating species have more extensive geographic ranges and lower population densities than vegetarians do, and big carnivores range over bigger areas than small ones do. For gigantic theropod species to succeed, the researchers argue, they would have needed continent-size landmasses to sustain populations large enough to avoid extinction. Third, titans-in-the-making had to be released from the ecological pressure of competing with other large species for food—through the extinction of rivals or the hunting of different prey, for example.

Critics find some of the conclusions hard to swallow. Kevin Padian of the University of California at Berkeley says studies of bone histology indirectly indicate that dinosaurs more closely resembled warm-blooded mammals than cold-blooded reptiles in their metabolism. Furthermore, he contends, it is impossible to determine the population sizes of extinct animals. Without living, breathing dinosaurs at hand, scientists may never know where on the metabolic spectrum they fell. But so far the fossil record upholds other predictions of the team's model. The five known



T. REX and a handful of other carnivorous dinosaurs attained gigantic sizes by being cold-blooded and having sufficiently vast landmasses free of other large competitors to call home, researchers propose.

KAZUHIKO SANO

NEED TO KNOW: LIVING LARGE

As to why some dinosaurs evolved into giants in the first place, scientists can only speculate. Perhaps intraspecies competition drove body size upward. Yet as enormous as several carnivorous dinosaurs (theropods) became, they never got much bigger than *T. rex.* In contrast, the herbivorous sauropods commonly reached 30 metric tons or more. The mechanical limitations of two-legged locomotion, as opposed to the stable quadrupedalism of sauropods, may have kept theropod gigantism in check.

heavyweights have turned up only on the vast landmasses of North America, South America and Africa. And all of them lived free from competition with other giants. (Of the two that overlapped in time and space, *Spinosaurus* and *Carcharodontosaurus*, both from Africa, the former appears to have fished for its supper, whereas the latter hunted on terra firma.)

The strongest support for this theory of theropod gigantism comes from what is known about the rise of *T. rex*, however. For more than 25 million years before the emergence of this colossus, the Western Interior Seaway divided North America, flowing from the present-day Gulf of Mexico to the Arctic. On both sides, three species of tyrannosaur, weighing in at just one to two tons, roamed in isolation. After the water retreated around 69 million years ago, thus doubling the available habitat area, only one lineage remained, and it produced the 5.5-ton *T. rex*. "There is food for thought here," Thomas R. Holtz, Jr., of the University of Maryland says of the model, "but we need some more test cases."

SENSORS

LASH Out

A BLIMP-BASED SYSTEM FOR MILITARY SURVEILLANCE **BY PHIL SCOTT**

Today's blimps are fat, happy billboards hovering above sporting events. Tomorrow's blimps may well play a much more serious role: airborne surveillance.

AIRSHIP EYES: This blimp carried aloft a sensor system called LASH, which passed its airborne tests.

U.S. Navy engineers have equipped an airship with a system known as LASH, or Littoral Airborne Sensor Hyperspectral. Basically, LASH works by detecting colors. Every object reflects light in its own unique pattern, invisible to the naked eye. LASH, developed by Science and Technology International in Honolulu, is essentially a camera that feeds the light pattern—usually in the infrared or ultraviolet range—into an onboard computer. The computer differentiates wavelengths and produces an image showing a real-time picture with enhanced color variations. "For instance, man-made camouflage is a couple of frequencies off from the surrounding natural color spectrum," says Steve Huett, director of airship advanced system development for the Office of Naval Research. "Your eyeball could never tell the difference."

But LASH can. In tests, it tracked whales swimming in a shipping lane 50 feet below the ocean's surface and detected grasshopper infestation in crops. "If you've got a little boy lost in a forest and the mother knows he was wearing a red coat, you can look for the color red," Huett explains.

Tethered blimps are already elements of police and military surveillance. But free-flying blimps with high-tech sensors offer much more flexibility. In October 2002 federal officials gave the go-ahead for a blimp outfitted with LASH and other sensors to search the Washington, D.C., area for the Beltway snipers by detecting a weapon discharge. (Lee Malvo and John Allen Muhammad, however, were caught before the blimp could take to the air.) As for its military uses, a LASH-equipped blimp could check out areas of sporadic conflicts and try to find camouflaged enemy hideouts. (Being easy targets, the blimps are not suitable for surveillance over a hot combat zone such as Iraq.) The navy or coast guard might also use the system to search American shores, looking for, say, terrorist divers or mines in harbors.

The characteristic of a blimp that makes it suitable for airborne observations—namely, the ability to loiter stably over extended periods—also has some observers worried. "I have a problem with the military looking into my backyard everyday," says Patrick Garrett, a defense analyst with GlobalSecurity.org in Alexandria, Va. "People complain about the Patriot Act already, and they worry about the military looking at them on a day-to-day basis. It would change the way we look at civil liberties dramatically. Welcome to a police state." To that, Huett responds laconically, "We don't spy on people. It's against the law."

Phil Scott writes about aviation technology from New York City.

FROM SURVEILLANCE TO DIAGNOSTICS

LASH technology could prove beneficial in hospitals. The camera can detect cancer cells by looking for their minute light variations as compared with surrounding tissue; the system is undergoing Phase II testing for use in detecting cervical cancer. It would effectively perform a painless virtual biopsy and provide results in real time, eliminating a patient's anxiety of waiting for word from the pathologist. The maker of LASH—Science and Technology International—hopes to market it in 2006.

GEORGE HALL *Corbis*

news SCAN

BY THE NUMBERS

The Great Migration

WHY AFRICAN-AMERICANS MOVED OUT OF THE SOUTH **BY RODGER DOYLE**

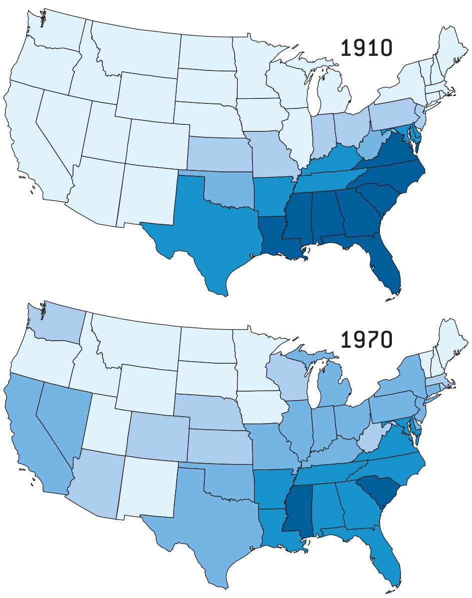
It began between 1916 and 1918, when more than 400,000 southern blacks went north; at times, entire communities and church organizations packed up and moved virtually intact to New York, Detroit, Chicago and other northern cities. During the next 50 years, net migration of blacks from the South totaled over five million, with additional millions leaving but returning after a relatively short stay. In 1900, 90 percent of black Americans lived in the South, compared with about 50 percent by 1970.

Substantial migrations had occurred before—for example, the movement to the Oklahoma Territory between 1890 and 1910—but nothing on the scale of the mass exodus of 1916–1918. Northern industry, newly deprived of immigrant labor from Europe by World War I, precipitated the migration, but conditions in the South made migration possible. At the time, southern blacks were coping with the devastating effect of the boll weevil on the cotton crop, which had thrown hundreds of thousands off the land. Industrialization in the South, much of it financed by northern interests, made obsolete many black-dominated occupations, such as blacksmithing. Meanwhile importation of cheap goods from the North had eliminated local manufacturing firms and, with them, jobs. Expanding Jim Crow laws further oppressed blacks.

Indispensable for migration was a communications system for spreading the news about the North. In part, that task fell to black newspapers, notably the *Chicago Defender*, which had many southern readers, and by agents of northern industry who came to the South to recruit blacks, at first even giving out free railroad tickets to Chicago and other cities. Soon migrants were relaying glowing messages—letters that were often read from pulpits. Nonprofit organizations, such as the Urban League, disseminated information about jobs and housing to new migrants.

During the 1930s, opportunities diminished in the North, and emigration slackened, not picking up again until World War II and the postwar boom. In the 1940s blacks began moving in large numbers up the ladder into higher-paying manufacturing positions, thanks in part to the newly mandated Fair Employment Practices Committee. Unions also helped, as their leaders realized that it was better to have blacks in the fold rather than using them as strikebreakers. But blacks lost some economic gains in the 1970s and 1980s because of deindustrialization, and they began leaving the North beginning in the 1970s, going to a now booming South where prices were low and the racial climate had improved.

It is sometimes said that the mechanical cotton picker, introduced in the 1940s, made farm hands redundant, thus putting pressure on them to migrate. But it was probably more often the case that cotton growers adopted the machines because labor had gone northward. Another notion is that southern blacks migrated to northern states for the welfare benefits, but no hard evidence for this proposition exists.



African-Americans (percent of U.S. population)
Less than 2 | 2 to 5 | 5 to 15 | 15 to 30 | 30 or more

Next month: The great migration and the formation of the black ghetto.

Rodger Doyle can be reached at rdoyle2@adelphia.net

RODGER DOYLE

FURTHER READING

The Strange Career of Jim Crow. C. Vann Woodward. Oxford University Press, 1955.

Migration and Residential Mobility in the United States. Larry Long. Russell Sage Foundation, 1988.

Farewell—We're Good and Gone: The Great Black Migration. Carole Marks. Indiana University Press, 1989.

Black Exodus: The Great Migration from the American South. Edited by Alferdteen Harrison. University of Mississippi Press, 1991.

The Great Migration in Historical Perspective: New Dimensions of Race, Class and Gender. Edited by Joe William Trotter, Jr. Indiana University Press, 1991.

DATA POINTS: DOWN BY DEGREES

The number of doctorates awarded by U.S. universities continues to drop from its 1998 peak, reaching lows not seen since 1993, according to the latest survey data. The bulk of the decline occurred in science and engineering.

Number of universities that confer research doctorates: 413

Number of degrees awarded in the academic year ending

In 1998: 42,654

In 2001: 40,790

In 2002: 39,955

Number of science and engineering doctorates awarded

In 1998: 27,283

In 2002: 24,558

In 2001: 25,525

Percent of Ph.D.s awarded to U.S. citizens and permanent residents in 2002:

In the humanities: 81

In the physical sciences: 55

In engineering: 39

Top five countries whose students earned U.S. science and engineering Ph.D.s in 2002:

China: 2,395

South Korea: 854

India: 678

Taiwan: 469

Canada: 312

SOURCES: 2002 Survey of Earned Doctorates; National Science Foundation. See also "Filling the Pipeline," by Rodger Doyle in News Scan, Scientific American, Vol. 287, No. 1; July 2002.

NANOTECH

Memorable Nanorings

Manufacturers of magnetic memory can pack chips pretty densely using lithographic techniques, but they will need another method to cheaply and reliably fashion nanoscale bits. In one possible approach, researchers have coaxed cobalt nanoparticles to assemble into rings capable of storing magnetic information at room temperature. The particles, which start by floating in an organic solvent mixed with surfactant, attract one another just like tiny magnets. They assemble into a spectrum of shapes, including rings less than 100 nanometers across; tuning the concentrations of surfactant and particles controls the fraction of rings formed. Building devices from the rings will require mechanically stabilizing them and integrating them with other types of nanofabrication, says lead investigator Alexander Wei of Purdue University. They might be combined with nanoscale wires, for example, to switch magnetic states, he adds. The work appears in the November 2003 *Angewandte Chemie*. *—JR Minkel*

BEHAVIOR

When in Doubt

If you thought the annoying tug of doubt was a distinctly human trait, think again. When shown a series of images, we tend to remember the first and last ones best; ask us about the middle images, and we're inclined to throw up our hands. Researchers gave this test to rhesus monkeys, but with a twist—the choice to decline to answer. The animals tended to decline on the middle images like human subjects did. They may not feel full-blown doubt the way humans do, comments group leader John David Smith of the University of Buffalo, but he says it makes sense for higher animals to have a mechanism for mulling options in novel situations, and he would like to begin mapping which species have the ability. (Rats and pigeons don't seem to have it; dolphins may.) A special December 2003 issue of the *Behavioral and Brain Sciences* presents the research along with numerous commentaries. *—JR Minkel*

HMM ... Nonhumans also experience doubt.

HORTICULTURE

Aroma Therapy

Perfume can attract the opposite sex, but it's expensive. A similar conundrum faces plants such as petunias and snapdragons. Once pollinated, they tend to decrease scent production in favor of other metabolic expenditures. Natalia Dudareva, Florence Negre and their colleagues at Purdue University have uncovered the molecular details whereby some flowers forfeit fragrance. A major fragrance ingredient is methylbenzoate. Dudareva's team found that pollinated petunias start making ethylene, which directly decreases the production of the enzyme responsible for a petunia's methylbenzoate. Snapdragons rely on a more complex feedback system, in which ethylene and the ratio of two other compounds determine levels of methylbenzoate. Understanding such natural mechanisms for decreased scent could be a first step toward restoring redolence in commercial plants, which are bred for long life at the expense of sweet smell. *—Steve Mirsky*

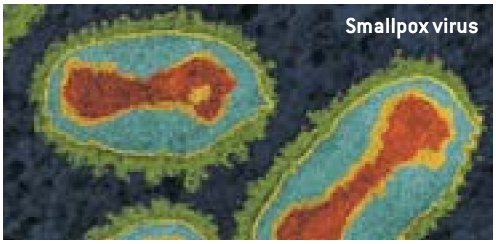
UNPOWERED FLOWERS: Petunias lose their scent after pollination.

GEORGE D. LEPP *Corbis* (*left*); SHAI GINOTT *Corbis* (*right*); ILLUSTRATION BY MATT COLLINS

IMMUNOLOGY

AIDS Resistance Thanks to Smallpox?

Those people who can resist HIV have a mutation that prevents the body from generating a cell protein onto which the virus latches. This mutation is limited mostly to about 10 percent of Europeans, so scientists think it arose as recently as 700 years ago. Some researchers suggest that this mutation could have guarded against the bubonic plague that once ravaged Europe. Population geneticists at the University of California at Berkeley argue that the mutation protected against smallpox instead. Smallpox's geographic distribution fits well with the mutation frequencies in Europe. And smallpox better meshes with the idea that a mutation is more likely to remain in a gene pool if it provides an advantage: bubonic plague stopped being a major cause of death worldwide 250 years ago; smallpox persisted until 1977. Moreover, plague results from a bacterium, whereas smallpox hinges on an RNA-based virus the way HIV does. The study appears in the December 9, 2003, *Proceedings of the National Academy of Sciences.* —*Charles Choi*

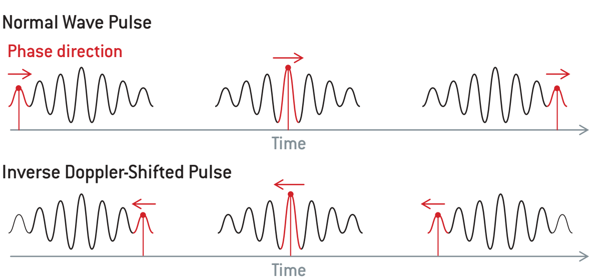


RESISTANCE TO HIV may stem from a mutation that emerged 700 years ago to protect against smallpox.

PHYSICS

Inverting the Doppler Shift

The Doppler effect makes a train whistle seem higher in pitch when the train approaches (as sound waves bunch up) and lower when it recedes (as sound waves stretch out). Scientists at BAE Systems Advanced Technology Center in Bristol, England, have effectively inverted the Doppler shift. The researchers bounced a wave pulse off a receding (electromagnetic) barrier in a specially designed circuit. The nature of the circuit meant that in the receding pulse, every tiny crest and trough making it up—specifically, the phase of the pulse—could travel in the opposite direction of the wave pulse as a whole, not unlike salmon swimming upstream. The reflected pulse rose in frequency as the barrier receded—opposite to the Doppler effect. (Technically speaking, the frequency of a wave pulse depends on the phase velocity, not on the movement of the pulse as a whole, or group velocity.) This phenomenon could soon allow new control over electromagnetic waves, for use in applications such as medicine and telecommunications. The findings appear in the November 28, 2003, *Science.* —*Charles Choi*



INVERTING DOPPLER: A circuit bounces a bundle of waves (a pulse) off a receding barrier (*not shown*). The phase of that pulse (represented by the position of the red dot) moves in the direction opposite to that of a normal wave pulse. As a result, the phase is shortened (the times when the red dot reaches a peak are bunched together), yielding a higher frequency indicative of an inverted Doppler shift.

EYE OF SCIENCE *Photo Researchers, Inc.* (*top*); NADIA STRASSER (*bottom*)

BRIEF POINTS

- **The oldest marsupial fossil yet found has turned up in northern China. The age of the unusually complete, mouse-size skeletal remains—125 million years—predates other evidence of marsupials by 15 million years.**

  Science, December 12, 2003; www.sciam.com/news_directory.cfm

- **Researchers have discovered that sections of midocean ridges are moving apart annually just 0.15 to 0.8 inch, as compared with the more typical one to seven inches. The ultraslow spreading could change the thinking about the formation of crustal plates.**

  Nature, November 27, 2003

- **Besides the presence of chemicals, simple mechanical cues can determine what stem cells turn into. Given a lot of space, certain human stem cells become bone cells; in narrow quarters, they differentiate into fat cells instead.**

  Presentation at the American Society for Cell Biology meeting, December 2003

- **Physicists have demonstrated imaging with a flat lens at microwave frequencies. The technique, based on negative refraction, could produce sharper images than ordinary bending by curved lenses.**

  Nature, November 27, 2003

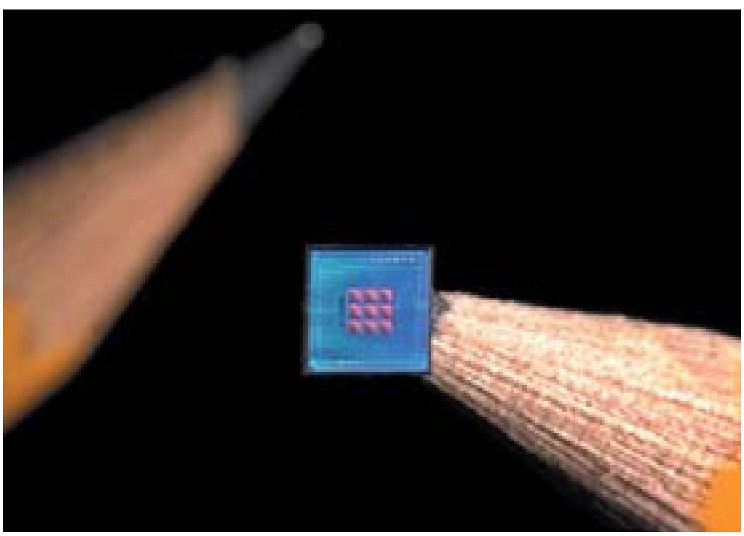
Micro(mechanical)phones

Integrating microphones and speakers on a chip could be a big deal for MEMS By GARY STIX

In 1987 a camera attached to a microscope snapped images of tiny gears, each of which had a diameter approaching that of the proverbial human hair. The black-and-white photographs that graced the pages of national magazines at the time evoked prospects of a true-to-life *Fantastic Voyage*. One of the creators of the microgears was a researcher at Bell Labs, just a few years removed from a graduate degree in electrical engineering at the Massachusetts Institute of Technology.

In the ensuing 17 years, the gears took a quick beeline to nowhere. The minuscule elements, producing a lot less torque than a mechanical watch part, will never power a submarine through the blood vessels on a trip to the islets of Langerhans. But the researcher, Kaigham (Ken) J. Gabriel, did go on to become a major figure in shaping the still emerging field of microelectromechanical systems, or MEMS.

Fabricating a gear that could serve as a bracelet on the leg of an ant did, in fact, accomplish more than just selling magazines (see the cover of the November 1992 *Scientific American*). It taught a number of lessons about what not to do to advance the nascent technology. At Bell Labs, Gabriel learned from making gears, miniature tongs and other micronovelties that, to become a commercial reality, any device would need to have a well-defined purpose—and it would have to be readily manufacturable using conventional semiconductor processes. He took those lessons with him when he became the head of the first MEMS program at the Defense Advanced Research Projects Agency (DARPA).



SOUND CHIP: Small membranes on the chip's surface (the array of pink squares) allow for superior detection of faint sounds over a wide range of frequencies.

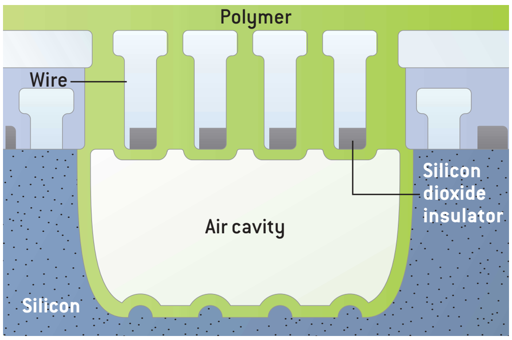
AKUSTICA

When Gabriel arrived at the agency in 1992, MEMS was suffering from gizmo creep: ideas and prototypes for machines and sensors ranging from micromotors to mite-size robots populated laboratory benchtops. He viewed his job as propelling MEMS beyond the graduate project stage. "I asked, 'What are the things that people would care about and pay for?' " he says. Under his watch, DARPA funded, among other projects, a specialized fabrication facility for MEMS devices.

By 1997, however, Gabriel had done what he set out to do and was tiring of reading the 100th-plus application for yet another approach to making acceleration sensors. He felt it was time to put into practice the principles he had espoused during his five-year tenure at DARPA. After accepting a faculty position at Carnegie Mellon University, he directed his graduate students to a relatively neglected area, that of acoustic MEMS. The two components that Alexander Graham Bell would recognize in a modern-day cellular phone are the microphone and the speaker, Gabriel was always quick to note: "That had to change. We needed to bring sound into the 21st century."

The idea for a sound chip built on Gabriel's interest in acoustics, something he had neglected ever since writing a doctoral thesis on binaural hearing—how humans use two ears to process acoustic signals and suppress background noise. Both microphones and speakers are essentially vibrating membranes that, in a MEMS system, extend in size up to half a millimeter square (big for MEMS devices). Membrane oscillations are turned into an electrical signal that corresponds to the pitch and loudness of a dog barking or 50 Cent rapping.

Concocting the membranes was the biggest technical challenge faced by Gabriel's laboratory at Carnegie Mellon. Like every other step in chipmaking, membrane fabrication takes place by building up thin layers of either conducting or insulating materials (typically metals or silicon dioxide) on a silicon substrate and then etching them away to create structures such as transistors and the wires that connect them. The lithographic and etching process for a membrane creates a mesh of metal wires with silicon dioxide filling the space between them. After the rest of the components are built up layer by layer, the mesh is excavated from the strata, and once it is exposed, a gas plasma etches the silicon substrate under it, creating a cavity that allows the now released structure to vibrate freely. Finally, the entire mesh is encased in a polymer to form the completed membrane [*see illustration above*].



LITTLE, TINY MEMBRANE: A suspended polymer studded with wires vibrates in response to sound waves, setting up an electrical signal in the wires that then gets channeled through the chip's signal-processing electronics.

A MEMS-based microphone chip should help address the poor sound quality produced by many of today's cell phones. The conventional microphone in run-of-the-mill cell phones has only one membrane. The ordinary phone represents a trade-off between a large enough membrane to detect faint sounds and a small enough one to be able to pick up higher frequencies. A MEMS chip does not have to strike this compromise. The first-generation MEMS sound chip will have more than five membranes, each optimized for either sensitivity to soft sounds or detection of a high C. The output from separate membranes can be integrated by the signal-processing electronics on the chip. The same design flexibility can also offer superior sound quality in MEMS-based speakers.

Two years ago Gabriel realized that he could not go any further in the development of the technology in his university lab. He could supply a demonstration of the capabilities of a sound chip, but he had no idea about the relevant specifications for the type of microphone or speaker demanded in the commercial sector. Taking a leave of absence from Carnegie Mellon, Gabriel joined with entrepreneur James H. Rock to launch Akustica to commercialize the technology. The 18-employee company has so far raised more than $12 million in two major rounds of financing, has reached agreements with both a major cell-phone and a hearing-aid maker, and expects to deliver chip samples this summer that can be designed into next-generation products.

Akustica will be able to function as a virtual company, one that does not need its own manufacturing operations. The entire process of building a sound chip is fully compatible with the standard industry process for semiconductor manufacturing, called complementary metal oxide semiconductor, or CMOS.

Thus, the crafting of the membrane, the chip's mechanical element, occurs in tandem with the manufacture of the signal-processing electronics. Its virtual status lets Akustica support only a small staff that designs the chips and subsequently taps into existing manufacturing capacity, wherever it exists throughout the world. Taking advantage of already available resources becomes particularly important if demand for microphone chips that would go into cell phones or video cameras balloons to many millions of units. (One final step, the excavation of the membrane, must currently be outsourced, although that, too, will eventually be handled at a single CMOS foundry.)

The prospect for sound chips may help counter recent setbacks that the MEMS field has experienced. Thought to be the next big wave, MEMS-based optical switching devices fizzled when the telecommunications market imploded. And much of the buzz about micro devices has shifted to nanotechnology, an endeavor that makes MEMS researchers look like hard-nosed pragmatists. In some cases, before even providing laboratory proof of the capabilities of their molecular-size creations, nanotechnologists have started to muse about whether infinitesimal robots will run amok and consume the planet.

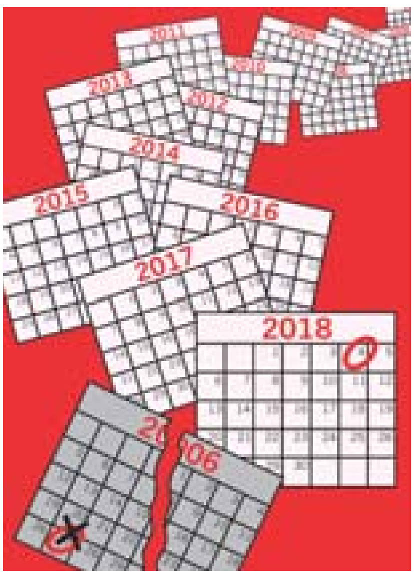
The excitement over micro(mechanical)phones has a far better grounding in the real world, whether big or little. The technology promises to provide the benefits—in sizing and cost—that heretofore have been the monopoly of digital devices. "Existing microphone and speaker technologies have reached their limits," Gabriel says. "You can't shrink them without compromising their performance." With sound chips, technology designers will be now be able to apply the tenet of smaller and cheaper to the analog world in which we speak and listen. SA

SARA CHEN; SOURCE: AKUSTICA

Working the System

A duo of antibody makers tries to prolong ownership of a key technology   By GARY STIX

A patent is supposed to last for a predetermined period so that the know-how contained therein eventually passes into the public domain. But if it is important enough, companies will try to get around the inconvenience of an expiration date. A dispute currently in the courts illustrates just how far some will go.



The case involves a patent that one litigant calls the "fundamental technology" needed for the artificial synthesis of antibodies. Last spring MedImmune, a maker of monoclonal antibodies, sued biotechnology giant Genentech, the City of Hope National Medical Center (a research partner of Genentech) and the British company Celltech. The suit levels antitrust charges, claiming that Genentech and Celltech colluded illegally to extend a monopoly over monoclonal antibody technology for more than a decade beyond a patent's 2006 expiration date. MedImmune has also asked that a patent that prolongs the rights to the technology be invalidated and stipulates that the agreement between Genentech and Celltech has "profoundly and fundamentally altered the competitive landscape in the biotechnology industry." The making of antibody drugs is one of the most dynamic industry sectors. In 2002, for example, Genentech and its marketing partners, took in total revenues of more than $1 billion for Rituxan, an antibody-based lymphoma drug.

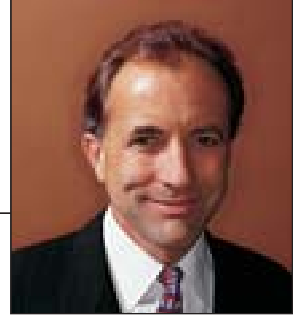
The wrangling began after Celltech received a very broad patent in 1989 for making monoclonal antibodies. Genentech then initiated a proceeding at the U.S. Patent and Trademark Office, claiming that it had invented the technology first and thus should retain patent rights.

The PTO held for Celltech. Subsequently, the two companies entered into a confidential settlement that resulted in a federal district court ordering on March 16, 2001, that the PTO should revoke Celltech's patent, slated for expiration in 2006, and issue a new one to Genentech, with an expiration of 2018. "The same claims that were about to expire in 2006 have a new lease on life in this important field in which recombinant antibody products are just coming to market," says Duncan Greenhalgh, a Boston attorney who coauthored an article on the MedImmune suit. Other drug firms using this critical technology now have to license it from Genentech until 2018, and the company could also simply refuse a license to a competitor. A press release issued by Celltech indicated that Genentech had agreed to compensate Celltech for royalties it would have received through the original 2006 expiration date. The agreement also gives the British firm a "preferential license" to the technology for the term of the new patent, according to court papers.

The effect of issuing a new patent to Genentech essentially results in a patent term of 29 years—from 1989 to 2018—allowing Genentech to reap unreasonable gains from licensing fees during the patent extension, in MedImmune's view. At the time that Celltech and Genentech filed for patents, a patent term was nominally 17 years from the date of issuance.

As of late 2003, the case was still snaking its way through the courts. Genentech had filed a motion for a summary judgment asking that MedImmune's antitrust complaint be thrown out; a hearing on that plea was scheduled for mid-December. Whatever happens, the interminable legal maneuvering serves as a demonstration that if a patent is important enough, the companies involved will stop at nothing. SA

JENNIFER KANE



A Bounty of Science

A new book reexamines the mutiny on the *Bounty,* but science offers a deeper account of its cause By MICHAEL SHERMER

The most common explanation for the *Bounty* mutiny pits a humane Fletcher Christian against an oppressive William Bligh. In her 2003 revisionist book, *The Bounty,* Caroline Alexander recasts Bligh as hero and Christian as coward. After 400 pages of gripping narrative, Alexander hints that the mutiny might have involved "the seductions of Tahiti" and "Bligh's harsh tongue" but concludes that it was "a night of drinking and a proud man's pride, a low moment on one gray dawn, a momentary and fatal slip in a gentleman's code of discipline."

A skeptic's explanation may seem less romantic, but it is more intellectually satisfying because it is extrapolated from scientific evidence and reasoning. There are, in fact, two levels of causality to consider: proximate (immediate historical events) and ultimate (deeper evolutionary motives). Both played a role in the *Bounty* debacle.

Evolutionarily adaptive emotions expressed nonadaptively caused mutiny.

A count of every lash British sailors received from 1765 through 1793 while serving on 15 naval vessels in the Pacific shows that Bligh was not overly abusive compared with contemporaries who did not suffer mutiny. Greg Dening's *Mr. Bligh's Bad Language* computed the average percentage of sailors flogged from information in ships' logs at 21.5. Bligh's was 19 percent, lower than James Cook's 20, 26 and 37 percent, respectively, on his three voyages, and less than half that of George Vancouver's 45 percent. Vancouver averaged 21 lashes per man, compared with the overall mean of five and Bligh's 1.5.

If unusually harsh punishment didn't cause the mutiny, what did? Although Bligh preceded Charles Darwin by nearly a century, the ship commander comes closest to capturing the ultimate cause: "I can only conjecture that they have Idealy assured themselves of a more happy life among the Otaheitians than they could possibly have in England, which joined to some Female connections has most likely been the leading cause of the whole business."

Indeed, crews consisted of young men in the prime of sexual life, shaped by evolution to bond in serial monogamy with women of reproductive age. Of the crews who sailed into the Pacific from 1765 through 1793, 82.1 percent were between the ages of 12 and 30, and another 14.3 percent were between 30 and 40. When the men arrived in the South Pacific, the results, from an evolutionary point of view, were not surprising. Of the 1,556 sailors, 437 (28 percent) got the "venereals." The *Bounty*'s infection rate was among the highest, at 39 percent.

After 10 months at sea, Bligh was not surprised by the reaction to the natives: "The Women are handsome … and have sufficient delicacy to make them admired and beloved—The chiefs have taken such a liking to our People that they have rather encouraged their stay among them than otherwise, and even made promises of large possessions. Under these and many other attendant circumstances equally desirable it is therefore now not to be Wondered at … that a Set of Sailors led by Officers and void of connections … should be governed by such powerfull inducement … to fix themselves in the most of plenty in the finest Island in the World where they need not labour, and where the alurements of disipation are more than equal to anything that can be conceived."

Neuroscience shows that the attachment bonds between men and women, especially in the early stages of a relationship, are chemical in nature and stimulate the pleasure centers of the brain in a manner resembling addictive drugs. In her book *The Oxytocin Factor,* for example, Kerstin Uvnäs-Moberg shows that oxytocin is secreted into the blood by the pituitary gland during sex, particularly orgasm, and plays a role in pair bonding, an evolutionary adaptation for long-term care of infants.

Ten months at sea weakened home attachments of the *Bounty*'s crew. New and powerful bonds made through sexual liaisons in Tahiti (that in some cases led to cohabitation and pregnancy) culminated in mutiny 22 days after departure, as the men grew restless to renew those fresh attachments; Christian, in fact, had been plotting for days to escape the *Bounty* on a raft.

Proximate causes of mutiny may have been alcohol and anger, but the ultimate reason was evolutionarily adaptive emotions expressed nonadaptively, with irreversible consequences. SA

Michael Shermer is publisher of Skeptic *magazine (www.skeptic.com) and author of* The Science of Good and Evil.

BRAD HINES

Talking Bacteria

Microbes seem to talk, listen and collaborate with one another—fodder for the truly paranoid. Bonnie L. Bassler has been eavesdropping and translating By MARGUERITE HOLLOWAY

It is far too early in the morning, and Bonnie L. Bassler is charging across the Princeton University campus, incandescent purple coat flying, brown curls bouncing, big laugh booming. She has come directly from the aerobics class she teaches every morning at 6:15—"I get up at exactly 5:42, not a minute earlier, not a minute later," she says emphatically. She says most things with similar energy, and when the conversation turns to her work, she becomes, impossibly, even more dynamic. "I am not meant to be stopped in time," she laughs. "I am supposed to be a blur."

ABIGAIL POPE

BONNIE L. BASSLER: EARS TO BACTERIA

- Discoverer of potential "bacterial Esperanto": the compound AI-2, which may present a target for a new class of antibacterial drugs.
- Hoped to be a veterinarian but became queasy during dissections.
- Nonbacterial passions: drama, song and dance, with occasional performances. Married her dance instructor, Todd Reichart.
- On achieving scientific success: "I am just really a person for whom nothing is ever good enough. I always think everything is wrong. And my little gift to my students is that they think everything is wrong, too."

The 41-year-old Bassler—a professor of molecular biology, winner of a 2002 MacArthur Foundation genius award, and occasional actress, dancer and singer—studies bacteria and how they communicate among their own kind and with other species. Quorum sensing, as this phenomenon is called, is a young science. Until recently, no one thought bacteria talked to one another, let alone in ways that changed their behavior, and Bassler has been instrumental in the field's rapid ascension. She has figured out some of the dialects—the genetic and molecular mechanisms different species use—but is best known for identifying what might be a universal language all species share, something she has jokingly referred to as "bacterial Esperanto."

As its moniker suggests, quorum sensing describes the ways in which bacteria determine how many of them there are in the vicinity. If enough are present (a quorum), they can get down to business or up to mischief. For instance, millions of bioluminescent bacteria might decide to emit light simultaneously so that their host, a squid, can glow—perhaps to distract predators and escape. Or salmonella bacteria might wait until their hordes have amassed before releasing a toxin to sicken their host; if the bacteria had acted as independent assassins rather than as an army, the immune system most likely would have wiped them out. Researchers have shown that bacteria also use quorum sensing to form the slimy biofilms that cover your teeth and eat through ship hulls and to regulate reproduction and the formation of spores.

If it all holds up, the implications are enormous. Quorum sensing offers a way to think about evolution.

Perhaps early bacteria communicated, then organized themselves according to different functions and, ultimately, into complex organisms. More practically, quorum sensing provides a strategy for medicine: muck up the communication system of dangerous bacteria, such as antibiotic-resistant enterococcus, and perhaps the bugs can't so effectively orchestrate their assault. As Bassler puts it, "You can either make them deaf or you can make them mute."

The study of quorum sensing has its roots in the late 1960s. Two scientists—J. Woodland Hastings and Kenneth H. Nealson—discovered that a marine bacterium, *Vibrio fischeri*, produced light when its population reached a critical size. When fewer were present, the bacteria didn't bioluminesce. The two researchers speculated that the bacteria released a signal—something they called an autoinducer—that cried out, like Horton the elephant's dust speck in the Dr. Seuss book, "We are here! We are here! We are here! We are here!" When the cacophony became loud enough, the assemblage glowed. In 1983 Michael R. Silverman, then at the Agouron Institute in La Jolla, Calif., and a colleague identified the genes for *V. fischeri*'s autoinducer and its receptor.

COLONIES of the glowing bacterium *Vibrio harveyi* provided clues about quorum sensing to bacteriologist Bonnie Bassler.

Bassler came to work with Silverman in 1990, after finishing her doctorate at Johns Hopkins University. She decided to focus on another glowing marine bacterium, *V. harveyi*, to determine whether its signaling system was similar. She got to work making mutant bacteria—disabling a gene here, a gene there, to see if she could impair the one that triggered the bug to bioluminesce when it was in like company. "You turn off the lights in the room and just look for the ones that are dark when they should be bright or bright when they should be dark. It is genetics for morons," she quips. Bassler found the genes for *V. harveyi*'s autoinducer and its receptor.

She also discovered something surprising. If she knocked out those two genes and put the altered *V. harveyi* in mixed company—that is, around masses of different species of bacteria—it glowed. "So I knew there was a second system," Bassler remarks. Bacteria "don't have enough room in their genome to be stupid, so there had to be a separate purpose for this system." The foreign bacteria were emitting something that *V. harveyi* responded to. Bassler called that something autoinducer two (AI-2). In 1994, as the field of quorum sensing was coming alive, Bassler moved to Princeton. Over time, she and others showed that quorum sensing initiates the release of toxins by bacteria such as *V. cholerae*. And they found that every bacterium they tested has its own personal autoinducer, the one it uses to communicate with its own kind. Gram-negative bacteria such as *Pseudomonas aeruginosa* use different versions of AHL molecules (acylated homoserine lactones); gram-positive bacteria such as *Staphylococcus aureus* use peptides.

But most bacteria Bassler looked at also used AI-2. By 1997 "we could see that all these bacteria made this molecule and that it was not just weird, crazy bacteria from the ocean," Bassler recalls. "So we got the idea that the bacteria must have a way of knowing self from other." For Bassler, the idea that different bacteria chat makes perfect sense. "There are 600 species of bacteria on your teeth every morning, and they are in exactly the same structure every single time: this guy is next to that one, is next to that one," she says. "It just seemed to us that you can't do that if the only thing you can detect is yourself. You have to know 'other.'"

Bassler and her students set out to purify and characterize AI-2. Finally, through the efforts of postdoctoral student Stephan Schauder and the crystallography of Frederick M. Hughson and Xin Chen, they got it. AI-2 is an unusual package—a sugar with a boron sitting in the middle of it. "What is amazing about that molecule is that it is the first *ever* to have a biological function for boron. *Ever!*" Bassler exclaims.

Now Bassler and her colleagues are trying to determine whether AI-2 is, indeed, one molecule that works alone as a signal and does not combine with other molecules to give rise to slightly different "languages." If it is the latter, no more Esperanto. "Her work has been truly superb," comments microbiologist Richard P. Novick of New York University. "But there is argument about where [AI-2] comes from and why. And what role it plays in different systems is unclear."

Some scientists are also concerned that aspects of quorum sensing—but not Bassler's findings—have been slightly overinterpreted. "Do bacteria want to communicate with each other, or is it just by accident?" asks Stephen C. Winans, a microbiologist at Cornell University. "This idea has taken hold that these bacteria want to communicate with each other. It may be just too good to be true."

Bassler's drive—her friend and former mentor Silverman describes her as "intensely motivated," "on a quest" and "just fierce"—suggests that she will hear bacteria's every last word. For the time being, she remains focused on understanding AI-2. "I want it all to be one thing, so I am sure that is wrong," she says. "I want it to be one thing because that is better if you want to make a drug, right?" Bassler is one of several quorum-sensing researchers working with companies to develop drugs. In 1999 she formed a company called Quorex with a former colleague from Agouron. Although her involvement is limited at the moment, she is hopeful that the start-up will find new antibacterials. "This was really considered fringe science," Bassler says. "Now it is this amazing field that didn't even exist 10 years ago." SA

MEDICAL PERSONNEL care for a patient who went into shock after losing a substantial amount of blood.

INSIGHTS INTO SHOCK

Still a last step before death for thousands of people, shock is shedding some of its medical mystery and becoming more treatable

By Donald W. Landry and Juan A. Oliver

Whatever the cause—a heart attack, a car accident, a serious bacterial infection—the glassy-eyed catatonia of a person in shock often portends death. Every year in the U.S. alone, about 500,000 people go into sudden shock, and half die from it. For millions more, it is the final stage of terminal illness. Doctors know a good deal about what causes the condition: very low blood pressure that results in dangerously reduced delivery of blood to tissues. And they know that it kills when the lack of oxygen irreparably damages the brain and other vital organs. They also have a few tools for reversing shock before it goes too far, at least in some people. But all too often treatment is ineffective, especially when a runaway infection is the trigger.

Because shock is so devastating, many investigators are aggressively trying to develop better treatments. Yet despite their initial promise, several seemingly helpful drug candidates have failed in recent years. To our great satisfaction, though, a chance discovery we made not long ago has led to a successful therapy. This agent does not cure the conditions that caused the shock, but it is already helping to treat thousands of shock victims. In addition, during the course of our research into this compound, we learned new information about the underlying mechanisms of shock. With luck, our insights and those of others may lead to further advances in treatment.

RICH PRESS

Under Pressure

TO UNDERSTAND SHOCK and its recondite nature, it helps to know a bit about the circulatory system. Early life in the earth's ancient seas used a simple principle to obtain oxygen and nutrients and to dispose of carbon dioxide and waste: diffusion. In this process, molecules move naturally from areas of high concentration to those of low concentration. But such life-sustaining diffusion proved efficient across only millimeters. Bigger creatures needed a more robust mechanism. Nature's solution was the circulatory system, in which blood carries molecules over long distances. For most organisms, the heart provides the driving force for this flow, which, in turn, delivers crucial gases and nutrients to every part of the body.

Oxygen-rich blood makes contact with tissues through intimate arcades of branching vessels that culminate in small, highly permeable vessels called capillaries. Arteries carry the blood from the heart to the narrowest arteries, or arterioles, which then lead into the capillaries. For blood to circulate, the heart must generate enough force to overcome the resistance it meets as the passageways become smaller and smaller. Blood pressure is a measurement of the force applied to blood as it is pumped.

In humans, the heart pumps five liters or so of blood through 10 miles of blood vessels about 1,000 times a day. A mere six-second cessation in blood flow can render an individual unconscious. Even a modest drop in blood pressure can deprive the brain of oxygen and leave a patient limp and dazed. In minutes other organs can become impaired. Shock has set in. If it persists and organs are irreversibly damaged, shock will lead to death.

Shock can be triggered in several ways and is often classified by its triggers. One of the most common causes, leading to what is called hypovolemic shock, is a rapid decrease in the volume of blood—as can occur when a trauma or a stomach ulcer causes extensive bleeding or when severe diarrhea drains fluid from the body. The heart pumps too little blood with each beat, and although it tries to compensate by increasing its pumping rate, it cannot do enough: blood pressure falls, and nourishment does not reach the tissues. In cases of hypovolemic shock, of which there are thousands every year, physicians try to stanch bleeding or other fluid loss and administer blood or saltwater, or do both, to replenish what was lost. And researchers are investigating new ways to stop bleeding—applying a paste to enhance coagulation, for example—as well as using substitutes for blood in cases where enough is not available.

Another form of shock, termed cardiogenic, arises when the heart stops pumping properly. If, say, a blood clot were blocking a coronary artery, preventing oxygen from reaching the heart muscle fed by that artery, a heart attack would occur: part of the muscle would become starved for oxygen and die, often leaving the heart unable to function normally. Alternatively, arrhythmia—too fast, too slow or nonsynchronized beating—or the failure of a heart valve to seal can also lead to cardiogenic shock. In the roughly 280,000 cases of cardiogenic shock that occur annually in the U.S., physicians often try to perform one of various interventions. They administer medicines to increase the heart muscle's ability to contract, they undertake valve replacements (using a mechanical or a pig valve), or they implant a defibrillator, a device that delivers an electrical charge to the heart, keeping the heart muscle pumping at the right rate. If all else fails, they then try to find a heart for transplant.

Even a modest drop in blood pressure can DEPRIVE the brain of OXYGEN.

The third and most common type of shock—the vasodilatory form—can result from cardiogenic or hypovolemic shock that has lasted for several days. In such cases, the heart may have been repaired or blood transfused, yet shock has persisted. But vasodilatory shock most frequently results from sepsis, a severe infection in which bacteria or fungi run rampant in the blood, setting in motion an inflammatory response. White blood cells and other immune system agents disrupt the function of tissues throughout the body in a deranged attempt to fight infection. Sepsis affects 500,000 people in the U.S. every year; about half of them develop septic shock, and 125,000 die from it. In this condition, the heart is blameless: the organ is pumping a high flow of blood, and the patient's skin feels warm to the touch. Instead the problem lies far away in the arterioles.

Researchers have long suspected that an understanding of what goes wrong in the arterioles could lead to improved therapy for vasodilatory shock. Indeed, efforts to tease out the source of arteriole malfunction led us to our unexpected discovery six years ago.

The story of why the arterioles behave abnormally begins well before shock sets in. The body's first reaction to falling blood pressure is compensatory—an effort to forestall shock—and this response centers in the arterioles. These hollow tubes are ringed by muscle cells that contract or relax, varying the

Overview/*Treating Shock*

- Every year in the U.S. about 500,000 people go into shock from myriad causes, including massive bacterial infections called sepsis. The condition is fatal for about half of them.
- Finding better treatments has been exceptionally difficult, as several lines of research and possible drugs have hit dead ends.
- Fortuitously, an existing drug is proving highly effective. Traditionally used to prevent bleeding in the esophagus, the hormone vasopressin has unexpectedly turned out to be powerful as a treatment for shock.

TYPES OF SHOCK

IN A NUTSHELL, shock is dangerously low blood pressure. When the pressure remains low for more than several minutes, blood does not adequately nourish key organs, which may then fail. Shock is often classified by its causes.

TYPE	EXAMPLES OF CAUSES	HEART FUNCTION	ARTERIOLE FUNCTION	TREATMENTS
1 HYPOVOLEMIC, arising when something causes excessive bleeding or fluid loss	Trauma (for instance, a gunshot or car accident); bleeding stomach ulcer; severe diarrhea	Heart works normally but does not have enough blood to pump—it may push only an average of three liters per minute as opposed to the five needed	Arterioles (key regulators of blood distribution) constrict in arms and legs, making those limbs cold and clammy to the touch, as blood is redirected to critical organs	Stop bleeding; give fluids, such as blood and saltwater; administer blood substitutes. New clotting factors are under study
2 CARDIOGENIC, arising when the heart has a problem	Heart attack (which damages the heart muscle); heart valve damage (which leads to obstruction or leakage); arrhythmia (heartbeat is too fast or too slow)	Heart cannot pump normally, even though there is enough blood—so, for instance, only three liters may be pumped out	Arterioles constrict in arms and legs, again in an effort to redirect blood to critical organs	Give medications that can help the heart muscle function more effectively; replace damaged valve; implant defibrillator; conduct a heart transplant in cases of severe damage
3 VASODILATORY, arising when small blood vessels called arterioles fail to constrict properly	Prolonged hypovolemic or cardiogenic shock (which can sometimes persist even when the primary problem has been fixed); sepsis (raging bacterial or fungal infection)	Heart operates properly and the circulatory system has enough blood, but arterioles malfunction	Arterioles dilate in arms and legs (keeping them warm to the touch), which prevents blood from being shunted to critical organs	In addition to any needed treatment above: administer steroids to decrease inflammation; give vasopressin

width of the tube. The normal orchestration of the arterioles is highly complex and entails the input of myriad compounds—including norepinephrine, vasopressin, angiotensin II, dopamine and nitric oxide. As blood pressure falls, some of these actors become involved. Both norepinephrine and angiotensin II, which constrict the arteriole muscles, are secreted into the bloodstream; at the same time, the body halts the secretion of atrial natriuretic peptide, a protein that causes arteriole muscles to relax and the arterioles to dilate. If successful, these maneuvers cause the arterioles in places such as the skin and certain nonessential muscles to constrict, increasing their resistance to the incoming blood; meeting this resistance allows the blood to flow to critical organs such as the brain. To visualize this, imagine a garden hose that branches in two; if one branch constricts, the pressure in and flow through the other branch increases. It is the same with arterioles.

Falling Resistance

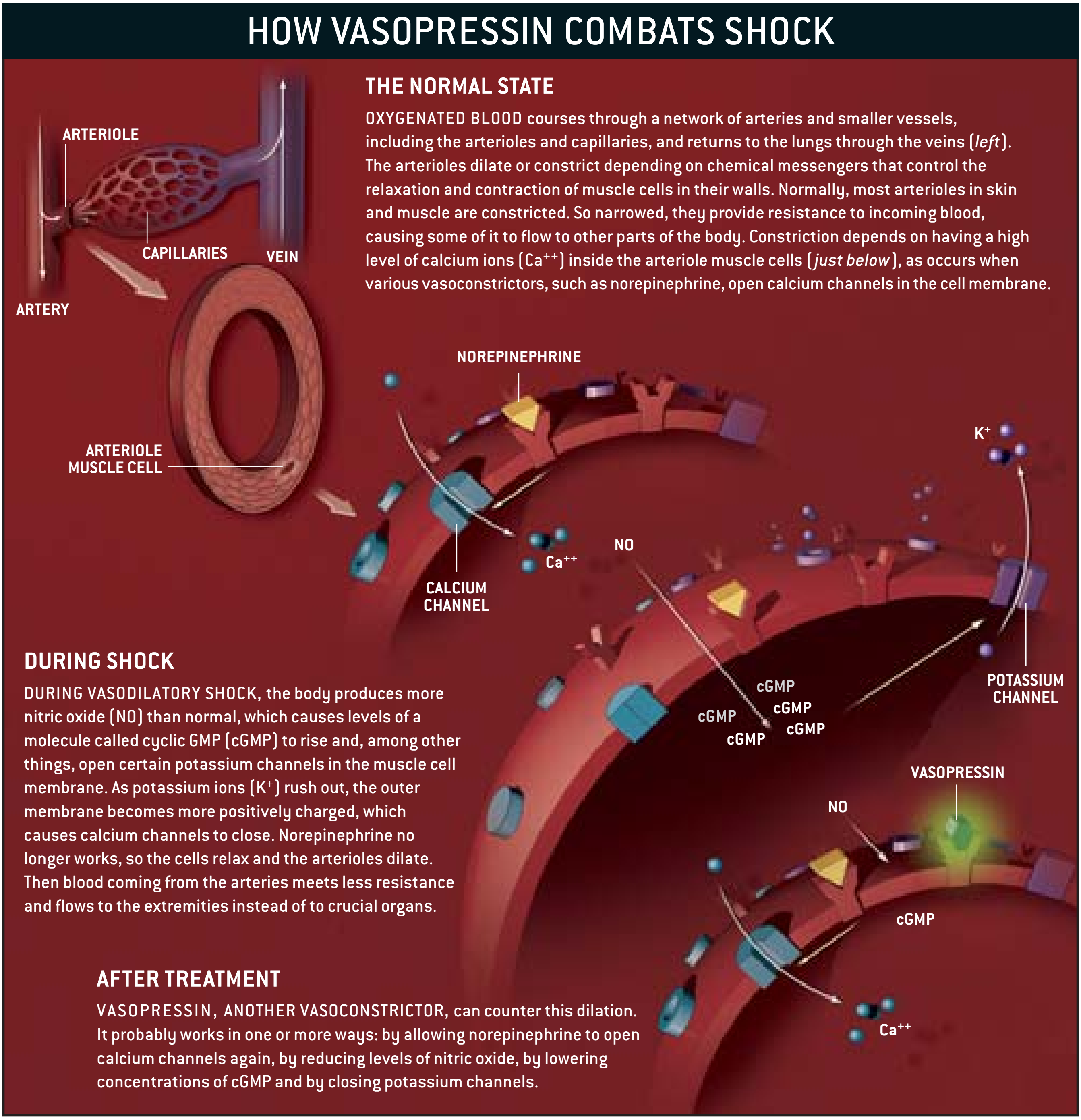
BUT IF SOMETHING goes wrong and certain arterioles fail to constrict, the blood does not encounter the resistance necessary to direct it on toward vital regions. Strangely, patients experiencing vasodilatory shock have high blood levels of both norepinephrine and angiotensin II. This fact suggests that the absence of constricting signals is not the problem. Clinical experience also supports this observation: when shock patients are given these two compounds, relatively little happens. Because of this puzzling result, many experts came to the conclusion long ago that something in the muscle cells of the arterioles was not functioning: the cells were not responding to their normal cues.

In the mid-1980s, however, researchers discovered that one root of the problem was not an error on the part of arteriole muscle cells; it was instead the action of a dilating agent. The body's most prominent dilator is nitric oxide, a simple molecule with wide-ranging effects [see "Biological Roles of Nitric Oxide," by Solomon H. Snyder and David S. Bredt; Scientific American, May 1992]. It became clear that the very infections that cause sepsis—such as pneumonia or meningitis—cause cells to increase their synthesis of nitric oxide. This news was greeted with excitement, and investigators designed a clinical trial to test a nitric oxide inhibitor—the idea being that once the dilator was taken off the scene, the constrictors (norepinephrine and angiotensin II) would succeed at their jobs. Tragically, the new treatment caused higher than expected rates of death and complications. Nitric oxide has so many diverse and poorly understood roles in the human body that inhibiting it led to serious and unanticipated problems.

Then, in 1992, we discovered an alternative way to constrict the arterioles during vasodilatory shock. Our insight came from brainstorming about how cell membranes work. It has long

THE AUTHORS

DONALD W. LANDRY and *JUAN A. OLIVER* work together at Columbia University's College of Physicians and Surgeons. Landry, who is an associate professor of medicine, directs the division of nephrology and the division of experimental therapeutics, where he makes artificial enzymes. Oliver, a native of Catalonia, Spain, earned his medical degree at the University of Barcelona. After completing research fellowships at Harvard, he joined Columbia as an associate professor of clinical medicine.



TAMI J. TOLPA

been known that every cell membrane has an electrical potential across it—in other words, the inside and the outside are differently charged. This happens for the most part because positively charged potassium ions reside inside cells, along with all kinds of negatively charged entities, but they also tend to leak outside, causing the outer part of the membrane to be sometimes slightly more positively charged than the inside [see "Patch Clamp Technique," by Erwin Neher and Bert Sakmann; Scientific American, March 1992].

In the case of arteriole muscle cells, this electrical potential is used to regulate the influx of calcium ions through calcium channels, which play a role in constriction. If the membrane's polarization is slightly more negatively charged on the outside, calcium channels open in response to norepinephrine or angiotension II, and calcium rushes into the cell. The cell then constricts. If the outer membrane becomes more positively charged, the calcium channels close, despite the urgings of the vasoconstricting hormones, and as calcium levels inside the cell fall, the muscle cell dilates. Thus, the electrical potential determines the responsiveness of calcium channels to the hormones that brings about constriction.

Simply put, the behavior of the arteriole muscles is orchestrated by calcium channels. But the passage of calcium ions depends on potassium-carrying channels to control membrane polarization properly. The channels are, in turn, regulated by a variety of compounds, including adenosine triphosphate

(ATP), a form of cellular energy created by the oxygen-based metabolism of nutrients. When ATP levels drop, certain potassium channels open, allowing potassium to flood (rather than merely leak) out of the cell, which causes the outer membrane to become more positively charged than normal, the calcium channels to close and the cell to relax.

We wondered whether the low-oxygen conditions of shock could reduce levels of ATP, leading to the relaxation of the muscle cells and a consequent decline in blood pressure. So we administered a compound called glibenclamide, which blocked the activity of ATP-sensitive potassium channels. And, indeed, the move boosted blood pressure. This mechanism explains why doctors giving norepinephrine or angiotensin II had little success increasing constriction: those compounds do not work well when the potassium channels are open.

Yet, as with the nitric oxide inhibitor described earlier, there were problems with the drug. Its effect was short-lived, and it led to low blood sugar when given in the high levels needed to reverse shock. (At lower levels, glibenclamide works to increase the pancreas's release of insulin and is used to treat diabetes.) It was a frustrating time. We knew that potassium channels regulated by ATP were important, and we knew that nitric oxide was important. But we could not figure out how to regulate them without causing harm elsewhere.

A New Approach

IN 1997 A SERENDIPITOUS observation changed the entire direction of our work. We had a patient who was bleeding in the esophagus and who later developed a serious infection. On admission, he had been placed on a hormone that would constrict his esophageal blood vessels and stop the bleeding there. This hormone, called vasopressin, was well known for its role in constricting arterioles—it acted throughout the body when released by the pituitary gland in response to low blood pressure. But previous clinical studies showed that when administered as therapy, it worked its magic solely on esophageal vessels. So we were not expecting to see any effect on our patient's blood pressure. To our surprise, we found that his blood pressure dropped when we stopped giving him the vasopressin. When we started administering it again, his blood pressure rebounded. Perhaps, we thought, the infection had somehow made our patient more sensitive to the hormone.

We had to determine whether this was a fluke. We needed to find a patient with septic shock, and we had to be very careful about the dosage and adhere to the heartfelt dictum of physicians, *primum non nocere* ("first do no harm"). So we gave a shock patient one tenth of the amount we had given the patient with esophageal bleeding, expecting to see no effect until we slowly increased it. To our amazement, his blood pressure rose dramatically. Further studies revealed that vasopressin levels in this and other septic shock patients were very low, even though logic dictated that the body would produce vasopressin to try to get blood pressure up.

We began to wonder why vasopressin deficiency developed in patients with vasodilatory shock. Our subsequent studies showed that at the outset of shock—no matter its origin—levels of vasopressin are exceedingly high. But after a few hours, vasopressin declines. The body's stores are released when shock starts, the compound then degrades in the bloodstream, and replacement vasopressin takes a long time to synthesize. We subsequently found two reports (neglected because everyone had concluded that vasopressin did not raise blood pressure) that vasopressin reduces nitric oxide's dilating effects on arterioles and blocks ATP-sensitive potassium channels, allowing the calcium channels to open and the cell to contract.

Since these early discoveries, vasopressin has been examined in 10 small studies around the world, and it has been found to reliably restore blood pressure—with no significant side effects. Anecdotally as well, we hear from doctor after doctor about cases of restored blood pressure saving patients from the scythe of shock, and many large medical centers around the world are now using it. An extensive multicenter trial is under way in sepsis patients to determine more definitively whether restoring blood pressure will reduce shock-related symptoms and deaths. Luckily, vasopressin is not patented, which means it can be produced without high cost.

The ongoing work on vasopressin is not the only front researchers and clinicians are investigating to thwart shock. In recent years, for example, scientists have pinpointed elements of the inflammatory cascade triggered by sepsis, which ultimately leads to shock. They are attempting to design antibodies—such as INNO 202—and other compounds that interfere with certain of the actors in the inflammatory response. They are also looking at the role steroids play in curbing the inflammatory response in some patients. It is hoped that these lines of research will result in a set of lifesaving therapies for sepsis and shock.

It has been very exciting for us to see the different threads of knowledge about the cellular and molecular mechanisms of constriction, dilation and shock come together through a chance observation. To have it translate so quickly into clinical practice in so many places has been most gratifying of all. SA

A serendipitous observation changed THE ENTIRE DIRECTION of our work.

MORE TO EXPLORE

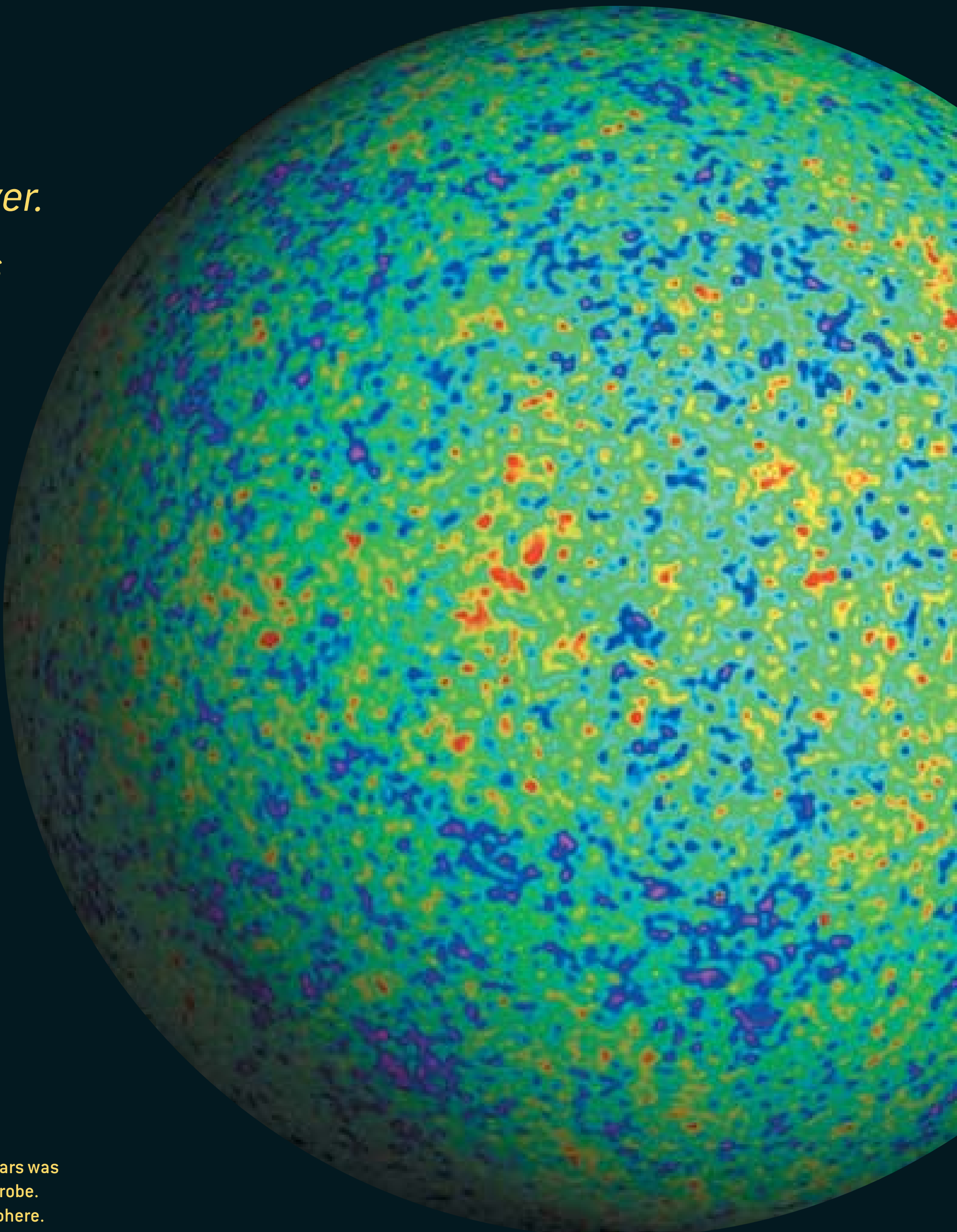
Vasopressin Deficiency Contributes to the Vasodilation of Septic Shock. Donald W. Landry et al. in *Circulation,* Vol. 95, No. 5, pages 1122–1125; March 4, 1997.

The Pathogenesis of Vasodilatory Shock. Donald W. Landry and Juan A. Oliver in *New England Journal of Medicine,* Vol. 345, No. 8, pages 588–595; August 23, 2001.

SPECIAL REPORT

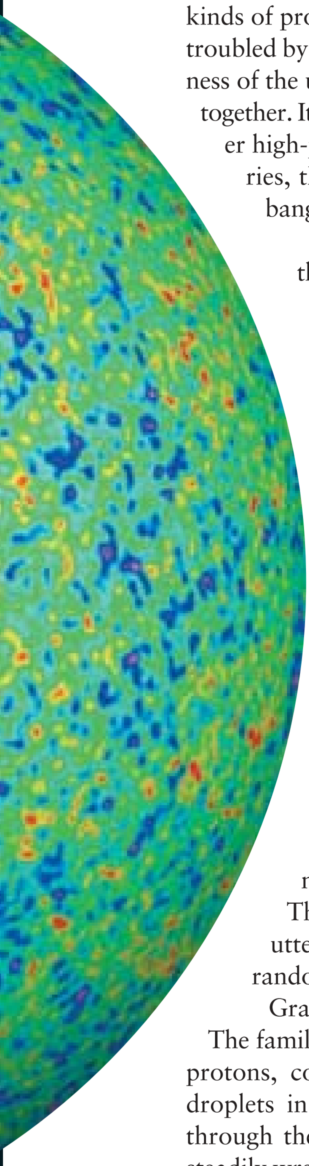
FOUR KEYS TO COSMOLOGY

The big bang theory works better than ever. If only cosmologists could figure out that mysterious acceleration. . . .



SNAPSHOT OF THE UNIVERSE at age 380,000 years was taken by the Wilkinson Microwave Anisotropy Probe. Earth is located at the center of this celestial sphere. Red corresponds to warm regions, blue to cooler ones.

FOR COSMOLOGISTS, it was like stepping on the brakes and feeling the car speed up, an exhilarating but disconcerting sensation that something wasn't working quite as it should.



NASA/WMAP SCIENCE TEAM

In what is widely regarded as the most important scientific discovery of 1998, researchers turned their telescopes to measure the rate at which cosmic expansion was decelerating and instead saw that it was accelerating. They have been gripping the steering wheel very tightly ever since.

As deeply mysterious as acceleration is, if you just accept it without trying to fathom its cause, it solves all kinds of problems. Before 1998, cosmologists had been troubled by discrepancies in the age, density and clumpiness of the universe. Acceleration made everything click together. It is one of the conceptual keys, along with other high-precision observations and innovative theories, that have unlocked the next level of the big bang theory.

The big bang is often described as an event that occurred long ago, a great explosion that created the universe. In actuality, the theory says nothing about the moment of creation, which is a job for quantum physics (or metaphysics). It simply states that as far back as we can extrapolate, the cosmos has been expanding, thinning out and cooling down. The big bang is best thought of not as a singular event but as an ongoing process, a gradual molding of order out of chaos. The recent observations have given this picture a coherence it never had before.

From the perspective of life on Earth, cosmic history started with inflation—a celestial reboot that wiped out whatever came before and left the cosmos a featureless place. The universe was without form, and void. Inflation then filled it with an almost completely uniform brew of radiation. The radiation varied from place to place in an utterly random way; mathematically, it was as random as random could be.

Gradually the universe imposed order on itself. The familiar particles of matter, such as electrons and protons, condensed out of the radiation like water droplets in a cloud of steam. Sound waves coursed through the amorphous mix, giving it shape. Matter steadily wrested control of the cosmos away from radiation. Several hundred thousand years after inflation, matter declared final victory and cut itself loose from radiation. This era and its dramatic coda have now been probed by high-precision observations of the fossil radiation [see "The Cosmic Symphony," on page 44].

Over the ensuing eons, matter organized itself into bodies of increasingly large size: subgalactic scraps, majestic galaxies, galactic clusters, great walls of galaxies. The universe we know—a set of distinct bodies separated by vast expanses of essentially empty space—is a fairly recent development, cosmologically speaking. This arrangement has now been systematically mapped [see "Reading the Blueprints of Creation," on page 54]. Starting several billion years ago, matter has been losing control to cosmic acceleration. Evidently the big bang has gotten a second wind, which is good for it but will be bad for us. The ever faster expansion has already arrested the formation of large structures and, if it continues, could rip apart galaxies and even our planet [see "From Slowdown to Speedup," on page 62].

In developing a cohesive and experimentally successful account of cosmic history, cosmologists have settled the disputes that once animated their field, such as the old debates between the big bang theory and the steady state theory and between inflation and its alternatives. Nothing in science is absolutely certain, but researchers now feel that their time is best spent on deeper questions, beginning with the cause of the cosmic acceleration.

Although the discovery of acceleration was revolutionary, cosmologists' initial response was fairly conservative. They dusted off an idea of Einstein's, the so-called cosmological constant, which represents a new type of energy—an example of what is more generally known as dark energy. But many physicists are thinking that a revolutionary discovery calls for a revolutionary response. Maybe the law of gravity works differently on gigantic scales than it does on humble, everyday ones [see "Out of the Darkness," on page 68].

Just as a nuclear missile cannot be fired unless two keys are turned simultaneously, the explosive progress in cosmology has depended on multiple observational and theoretical keys being turned at once. Will the rush of new ideas lead to chaos? Will order reemerge? Must the cosmos be "preposterous," as one of the authors of this special report once put it? Or will it start to make sense again? *—George Musser, staff editor*

THE COSMIC SYMPHONY

By Wayne Hu and Martin White

New observations of the cosmic microwave background radiation show that the early universe resounded with harmonious oscillations

In the beginning, there was light. Under the intense conditions of the early universe, ionized matter gave off radiation that was trapped within it like light in a dense fog. But as the universe expanded and cooled, electrons and protons came together to form neutral atoms, and matter lost its ability to ensnare light. Today, some 14 billion years later, the photons from that great release of radiation form the cosmic microwave background (CMB).

Tune a television set between channels, and about 1 percent of the static you see on the screen is from the CMB. When astronomers scan the sky for these microwaves, they find that the signal looks almost identical in every direction. The ubiquity and constancy of the CMB is a sign that it comes from a simpler past, long before structures such as planets, stars and galaxies formed. Because of this simplicity, we can predict the properties of the CMB to exquisite accuracy. And in the past few years, cosmologists have been able to compare these predictions with increasingly precise observations from microwave telescopes carried by balloons and spacecraft. This research has brought us closer to answering some age-old questions: What is the universe made of? How old is it? And where did objects in the universe, including our planetary home, come from?

Arno Penzias and Robert Wilson of AT&T Bell Laboratories detected the CMB radiation in 1965 while trying to find the source of a mysterious background noise in their radio antenna. The discovery firmly established the big bang theory, which states that the early universe was a hot, dense plasma of charged particles and photons. Since that time, the CMB has been cooled by the expansion of the universe, and it is extremely cold today—comparable to the radiation released by a body at a temperature of 2.7 kelvins (that is, 2.7 degrees Celsius above absolute zero). But when the CMB was released, its temperature was nearly 3,000 kelvins (or about 2,727 degrees C).

In 1990 a satellite called COBE (for Cosmic Background Explorer) measured the spectrum of the CMB radiation, showing it to have exactly the expected form. Overshadowing this impressive achievement, however, was COBE's detection of slight variations—at the level of one part in 100,000—in the temperature of the CMB from place to place in the sky. Observers had been diligently searching for these variations for more than two decades because they hold the key to understanding the origin of structure in the universe: how the primordial plasma evolved into galaxies, stars and planets.

Since then, scientists have employed ever more sophisticated instruments to map the temperature variations of the CMB. The culmination of these efforts was the launch in 2001 of the Wilkinson Microwave Anisotropy Probe (WMAP), which travels around the sun in an orbit 1.5 million kilometers beyond Earth's. The results from WMAP reveal that the CMB temperature variations follow a distinctive pattern predicted by cosmological theory: the hot and cold spots in the radiation fall into characteristic sizes. What is more, researchers have been able to use these data to precisely estimate the age, composition and geometry of the universe. The process is analogous to determining the construction of a musical instrument by carefully listening to its notes. But the cosmic symphony is produced by some very strange players and is accompanied

SOUND WAVES in the early universe—represented by tuning forks in this photograph—reveal the age, composition and geometry of the cosmos.

DAVID EMMITE PHOTOGRAPHY

by even stranger coincidences that cry out for explanation.

Our basic understanding of the physics behind these observations dates back to the late 1960s, when P. James E. Peebles of Princeton University and graduate student Jer Yu realized that the early universe would have contained sound waves. (At almost the same time, Yakov B. Zel'dovich and Rashid A. Sunyaev of the Moscow Institute of Applied Mathematics were coming to very similar conclusions.) When radiation was still trapped by matter, the tightly coupled system of photons, electrons and protons behaved as a single gas, with photons scattering off electrons like ricocheting bullets. As in the air, a small disturbance in gas density would have propagated as a sound wave, a train of slight compressions and rarefactions. The compressions heated the gas and the rarefactions cooled it, so any disturbance in the early universe resulted in a shifting pattern of temperature fluctuations.

Sounding Out Origins

WHEN DISTANCES in the universe grew to one thousandth of their current size—about 380,000 years after the big bang—the temperature of the gas decreased enough for the protons to capture the electrons and become atoms. This transition, called recombination, changed the situation dramatically. The photons were no longer scattered by collisions with charged particles, so for the first time they traveled largely unimpeded through space. Photons released from hotter, denser areas were more energetic than photons emitted from rarefied regions, so the pattern of hot and cold spots induced by the sound waves was frozen into the CMB. At the same time, matter was freed of the radiation pressure that had resisted the contraction of dense clumps. Under the attractive influence of gravity, the denser areas coalesced into stars and galaxies. In fact, the one-in-100,000 variations observed in the CMB are of exactly the right amplitude to form the large-scale structures we see today [see "Reading the Blueprints of Creation," by Michael A. Strauss, on page 54].

Yet what was the prime mover, the source of the initial disturbances that triggered the sound waves? The question is troubling. Imagine yourself as an observer witnessing the big bang and the subsequent expansion. At any given point you will see only a finite region of the universe that encompasses the distance light has traveled since the big bang. Cosmologists call the edge of this region the horizon, the place beyond which you cannot see. This region continuously grows until it reaches the radius of the observable universe today. Because information cannot be conveyed faster than light, the horizon defines the sphere of influence of any physical mechanism. As we go backward in time to search for the origin of structures of a particular physical size, the horizon eventually becomes smaller than the structure [*see illustration on opposite page*]. Therefore, no physical process that obeys causality can explain the structure's origin. In cosmology, this dilemma is known as the horizon problem.

Fortunately, the theory of inflation solves the horizon problem and also provides a physical mechanism for triggering the primordial sound waves and the seeds of all structure in the universe. The theory posits a new form of energy, carried by a field dubbed the "inflaton," which caused an accelerated expansion of the universe in the very first moments after the big bang. As a result, the observable universe we see today is only a small fraction of the observable universe before inflation. Furthermore, quantum fluctuations in the inflaton field, magnified by the rapid expansion, provide initial disturbances that are approximately equal on all scales—that is, the disturbances to small regions have the same magnitude as those affecting large regions. These disturbances become fluctuations in the energy density from place to place in the primordial plasma.

Evidence supporting the theory of inflation has now been found in the detailed pattern of sound waves in the CMB. Because inflation produced the density disturbances all at once in essentially the first moment of creation, the phases of all the sound waves were synchronized. The result was a sound spectrum with overtones much like a musical instrument's. Consider blowing into a pipe that is open at both ends. The fundamental frequency of the sound corresponds to a wave (also called a mode of vibration) with maximum air displacement at either end and minimum displacement in the middle [*see top illustration in box on page 48*]. The wavelength of the fundamental mode is twice the length of the pipe. But the sound also has a series of overtones corresponding to wavelengths that are integer fractions of the fundamental wavelength: one half, one third, one fourth and so on. To put it another way, the frequencies of the overtones are two, three, four or more times as high as the fundamental frequency. Overtones are what distinguish a Stradivarius from an ordinary violin; they add richness to the sound.

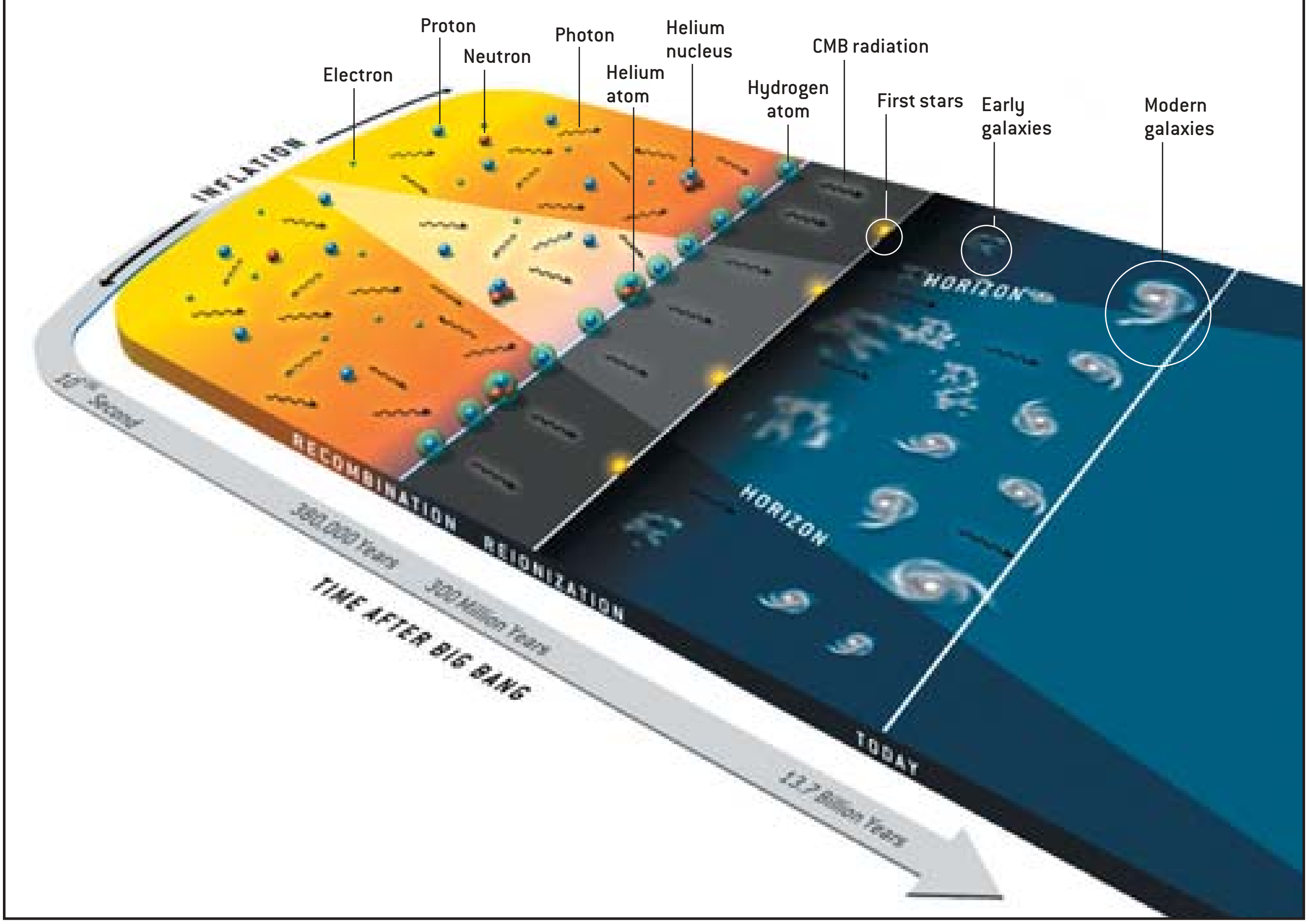
The sound waves in the early universe are similar, except now we must imagine the waves oscillating in time instead of space [*see bottom illustration in box on page 48*]. In this analogy, the length of the pipe represents the finite duration when sound waves traveled through the primordial plasma; the waves start at inflation and end at recombination about 380,000 years later. Assume that a certain region of space has a maximum positive displacement—that is, maximum temperature—at inflation. As the sound waves propagate, the density of the region will begin to oscillate, first heading toward average tempera-

Overview/*Cosmic Acoustics*

- Inflation, the rapid expansion of the universe in the first moments after the big bang, triggered sound waves that alternately compressed and rarefied regions of the primordial plasma.
- After the universe had cooled enough to allow the formation of neutral atoms, the pattern of density variations caused by the sound waves was frozen into the cosmic microwave background (CMB) radiation.
- By studying the acoustic signals in the CMB, cosmologists have estimated the age, composition and geometry of the universe. But the results suggest that the biggest component of the modern cosmos is a mysterious entity called dark energy.

TIMELINE OF THE UNIVERSE

AS INFLATION EXPANDED the universe, the plasma of photons and charged particles grew far beyond the horizon (the edge of the region that a hypothetical viewer after inflation would see as the universe expands). During the recombination period about 380,000 years later, the first atoms formed and the cosmic microwave background (CMB) radiation was emitted. After another 300 million years, radiation from the first stars reionized most of the hydrogen and helium.



ture (minimum displacement) and then toward minimum temperature (maximum negative displacement). The wave that causes the region to reach maximum negative displacement exactly at recombination is the fundamental wave of the early universe. The overtones have wavelengths that are integer fractions of the fundamental wavelength. Oscillating two, three or more times as quickly as the fundamental wave, these overtones cause smaller regions of space to reach maximum displacement, either positive or negative, at recombination.

How do cosmologists deduce this pattern from the CMB? They plot the magnitude of the temperature variations against the sizes of the hot and cold spots in a graph called a power spectrum [*see box on page 51*]. The results show that the regions with the greatest variations subtend about one degree across the sky, or nearly twice the size of the full moon. (At the time of recombination, these regions had diameters of about one million light-years, but because of the 1,000-fold expansion of the universe since then, each region now stretches nearly one billion light-years across.) This first and highest peak in the power spectrum is evidence of the fundamental wave, which compressed and rarefied the regions of plasma to the maximum extent at the time of recombination. The subsequent peaks in the power spectrum represent the temperature variations caused by the overtones. The series of peaks strongly supports the theory that inflation triggered all the sound waves at the same time. If the perturbations had been continuously generated over time, the power spectrum would not be so harmoniously ordered. To return to our pipe analogy, consider the cacophony that would result from blowing into a pipe that has holes drilled randomly along its length.

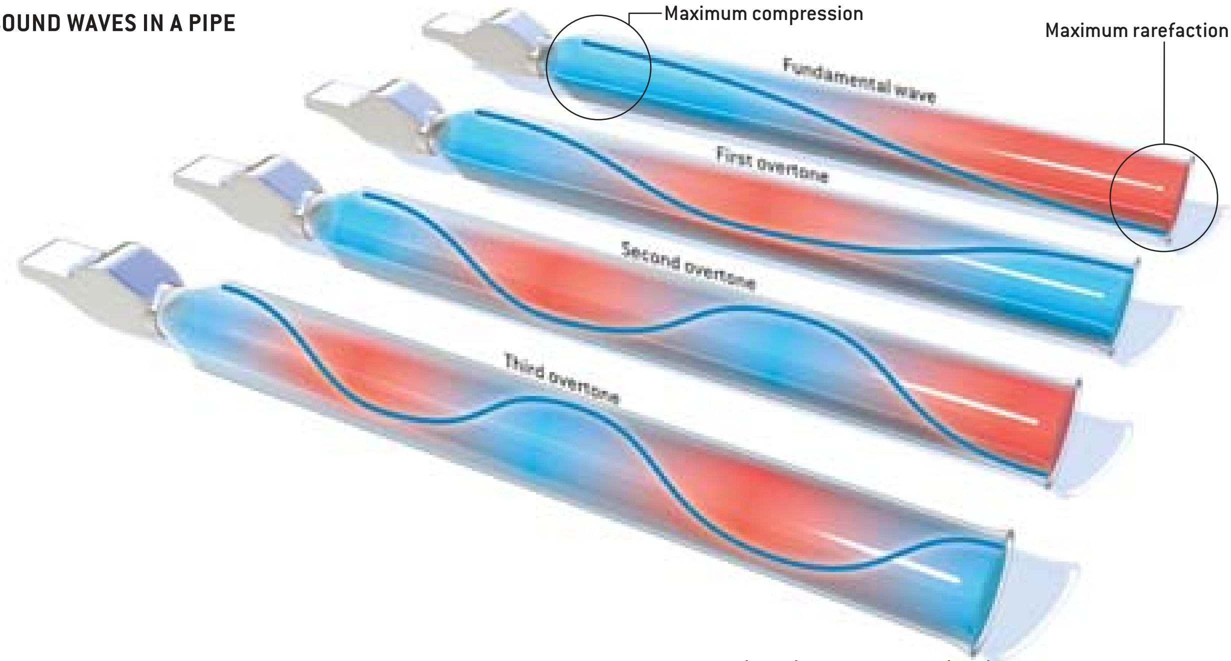
The theory of inflation also predicts that the sound waves should have nearly the same amplitude on all scales. The power spectrum, however, shows a sharp drop-off in the magnitude of temperature variations after the third peak. This discrepancy can be explained by the fact that sound waves with short wavelengths dissipate. Because sound is carried by the collisions

BRYAN CHRISTIE DESIGN

COSMIC HARMONICS

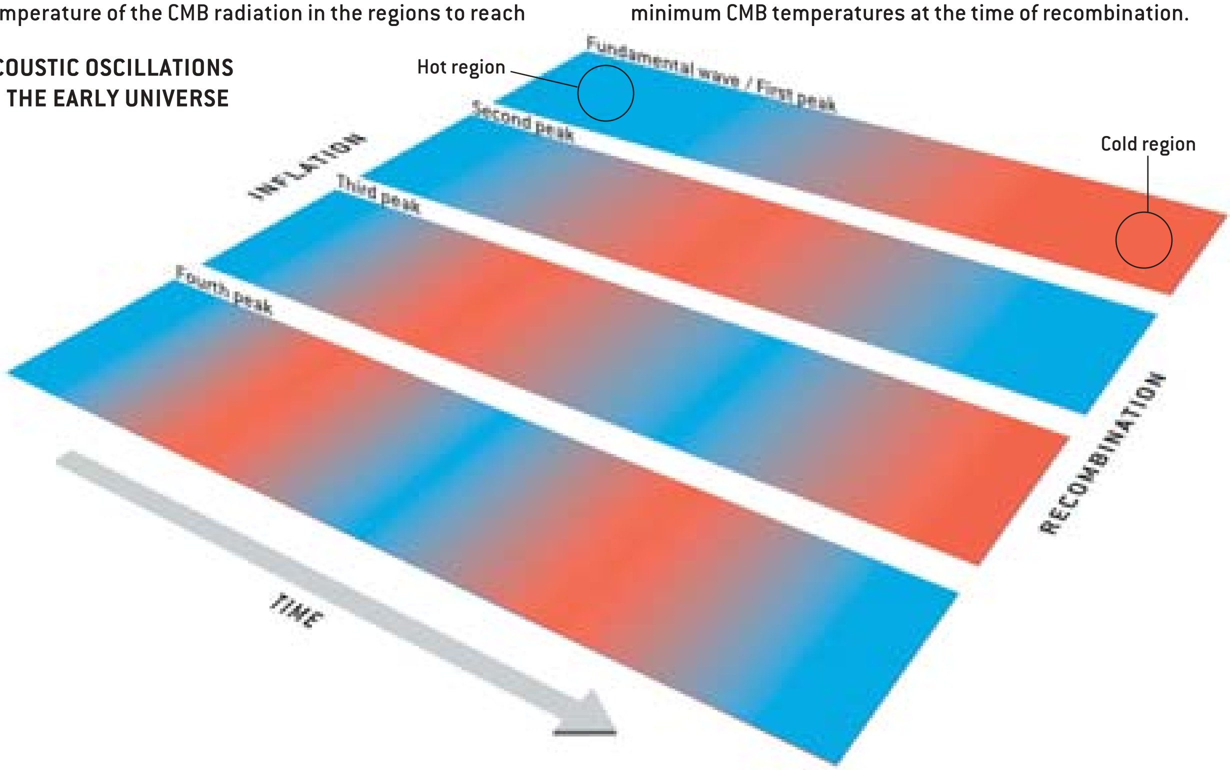
THE SOUND SPECTRUM of the early universe had overtones much like a musical instrument's. If you blow into a pipe, the sound corresponds to a wave with maximum air compression (*blue*) at the mouthpiece and maximum rarefaction (*red*) at the end piece. But the sound also has a series of overtones with shorter wavelengths that are integer fractions of the fundamental wavelength. (The wavelengths of the first, second and third overtones are one half, one third and one fourth as long.)

SOUND WAVES IN A PIPE



SOUND WAVES also oscillated in the plasma of the early universe. After inflation, the fundamental wave compressed some regions of plasma and rarefied others, causing the temperature of the CMB radiation in the regions to reach maximum (*blue*) and minimum (*red*) values by the time of recombination. The overtones oscillated two, three or more times as quickly, causing smaller regions to reach maximum and minimum CMB temperatures at the time of recombination.

ACOUSTIC OSCILLATIONS IN THE EARLY UNIVERSE



BRYAN CHRISTIE DESIGN

of particles in gas or plasma, a wave cannot propagate if its wavelength is shorter than the typical distance traveled by particles between collisions. In air, this distance is a negligible 10^{-5} centimeter. But in the primordial plasma just before recombination, a particle would typically travel some 10,000 light-years before striking another. (The universe at this stage was dense only in comparison with the modern universe, which is about a billion times as rarefied.) As measured today, after its 1,000-fold expansion, that scale is about 10 million light-years. Therefore, the amplitudes of the peaks in the power spectrum are damped below about 10 times this scale.

Just as musicians can distinguish a world-class violin from an ordinary one by the richness of its overtones, cosmologists can elucidate the shape and composition of the universe by examining the fundamental frequency of the primordial sound waves and the strength of the overtones. The CMB reveals the angular size of the most intense temperature variations—how large these hot and cold spots appear across the sky—which in turn tells us the frequency of the fundamental sound wave. Cosmologists can precisely estimate the actual size of this wave at the time of recombination because they know how quickly sound propagated in the primordial plasma. Likewise, researchers can determine the distance CMB photons have traveled before reaching Earth—about 45 billion light-years. (Although the photons have traveled for only about 14 billion years, the expansion of the universe has elongated their route.)

So cosmologists have complete information about the triangle formed by the wave and can check whether its angles add up to 180 degrees—the classic test of spatial curvature. They do so to high precision, showing that aside from the overall expansion, the universe obeys the laws of Euclidean geometry and must be very close to spatially flat. And because the geometry of the universe depends on its energy density, this finding implies that the average energy density is close to the so-called critical density—about 10^{-29} gram per cubic centimeter.

The next thing cosmologists would like to know is the exact breakdown of the universe's matter and energy. The amplitudes of the overtones provide the key. Whereas ordinary sound waves are driven solely by gas pressure, the sound waves in the early universe were modified by the force of gravity. Gravity compresses the gas in denser regions and, depending on the phase of the sound wave, can alternately enhance or counteract sonic compression and rarefaction. Analyzing the modulation of the waves reveals the strength of gravity, which in turn indicates the matter-energy composition of the medium.

As in today's universe, matter in the early universe fell into two main categories: baryons (protons and neutrons), which make up the bulk of so-called ordinary matter, and cold dark matter, which exerts gravity but has never been directly observed because it does not interact with ordinary matter or light in any noticeable way. Both ordinary matter and dark matter supply mass to the primordial gas and enhance the gravitational pull, but only ordinary matter undergoes the sonic compressions and rarefactions. At recombination, the fundamental wave is frozen in a phase where gravity enhances its compression of the denser regions of gas [*see box on page 52*]. But the first overtone, which has half the fundamental wavelength, is caught in the opposite phase—gravity is attempting to compress the plasma while gas pressure is trying to expand it. As a result, the temperature variations caused by this overtone will be less pronounced than those caused by the fundamental wave.

The cosmic symphony is produced by very strange players and is accompanied by even stranger coincidences.

This effect explains why the second peak in the power spectrum is lower than the first. And by comparing the heights of the two peaks, cosmologists can gauge the relative strengths of gravity and radiation pressure in the early universe. This measurement indicates that baryons had about the same energy density as photons at the time of recombination and hence constitute about 5 percent of the critical density today. The result is in spectacular agreement with the number derived from studies of light-element synthesis by nuclear reactions in the infant universe.

The general theory of relativity, however, tells us that matter and energy gravitate alike. So did the gravity of the photons in the early universe also enhance the temperature variations? It did, in fact, but another effect counterbalanced it. After recombination, the CMB photons from denser regions lost more energy than photons from less dense areas, because they were climbing out of deeper gravitational-potential wells. This process, called the Sachs-Wolfe effect, reduced the amplitude of the temperature variations in the CMB, exactly negating the enhancement caused by the gravity of the photons. For regions of the early universe that were too big to undergo acoustic oscillations—that is, regions stretching more than one degree across the sky—temperature variations are solely the result of the

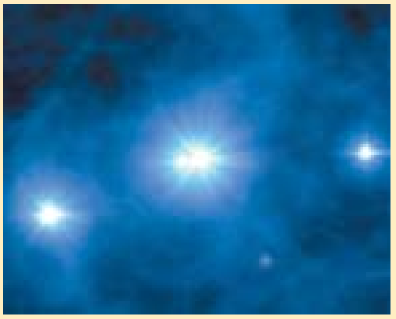
THE AUTHORS

WAYNE HU and *MARTIN WHITE* are trying to unveil the history of the universe. Hu is associate professor of astronomy and astrophysics at the University of Chicago. He received his Ph.D. in physics from the University of California, Berkeley, in 1995. His research pursuits include the investigation of dark energy, dark matter and the formation of cosmological structure. White, professor of astronomy and physics at Berkeley, earned his Ph.D. in physics from Yale University in 1992. In addition to exploring how structure in the universe came to be, he is interested in the connections between astrophysics and fundamental physics.

NOTES OF DISCORD

AFTER THE EMISSION of the cosmic microwave background (CMB) radiation, about 380,000 years after the big bang, most of the photons traveled across the observable universe without scattering. But some photons did scatter off charged particles, polarizing the radiation across wide swaths of the sky. Observations of this large-angle polarization by the WMAP spacecraft imply that about 17 percent of the CMB photons were scattered by a thin fog of ionized gas a few hundred million years after the big bang.

This relatively large fraction is perhaps the biggest surprise from the WMAP data. Cosmologists had previously theorized that most of the universe's hydrogen and helium would have been ionized by the radiation from the first stars, which were extremely massive and bright. (This process is called reionization because it returned the gases to the plasma state that existed before the emission of the CMB.) But the theorists estimated that this event occurred nearly a billion years after the big bang, and therefore only about 5 percent of the CMB photons would have been scattered. WMAP's evidence of a higher fraction indicates a much earlier reionization and presents a challenge for the modeling of the first rounds of star formation. The discovery may even challenge the theory of inflation's prediction that the initial density fluctuations in the primordial universe were nearly the same at all scales. The first stars might have formed sooner if the small-scale fluctuations had higher amplitudes.



FIRST STARS reionized the surrounding gas.

The WMAP data also contain another hint of deviation from scale invariance that was first observed by the COBE satellite. On the biggest scales, corresponding to regions stretching more than 60 degrees across the sky, both WMAP and COBE found a curious lack of temperature variations in the CMB. This deficit may well be a statistical fluke: because the sky is only 360 degrees around, it may not contain enough large-scale regions to make an adequate sample for measuring temperature variations. But some theorists have speculated that the deviation may indicate inadequacies in the models of inflation, dark energy or the topology of the universe.

—W.H. and M.W.

Sachs-Wolfe effect. At these scales, paradoxically, hot spots in the CMB represent less dense regions of the universe.

Finally, cosmologists can use the CMB to measure the proportion of dark matter in the universe. The gravity from baryons alone could not have modulated the temperature variations much beyond the first peak in the power spectrum. An abundance of cold dark matter was needed to keep the gravitational-potential wells sufficiently deep. By measuring the ratios of the heights of the first three peaks, researchers have determined that the density of cold dark matter must be roughly five times the baryon density. Therefore, dark matter constitutes about 25 percent of the critical density today.

Remarkable Concord

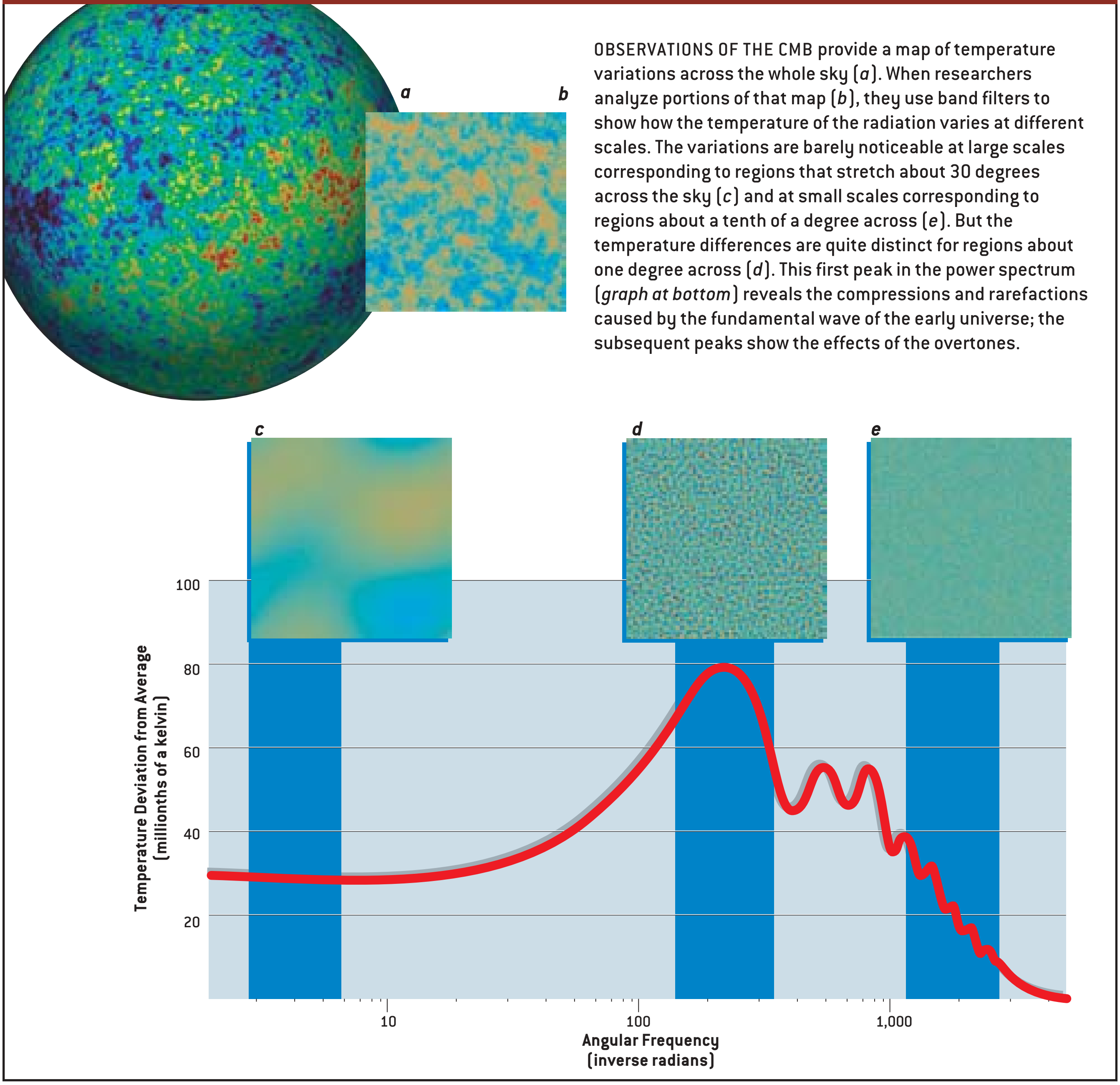
UNFORTUNATELY, these calculations of the modern universe's matter and energy leave about 70 percent of the critical density unspecified. To make up the difference, theorists have posited a mysterious component called dark energy, whose relative influence has grown as the universe has expanded [see "Out of the Darkness," by Georgi Dvali, on page 68]. We are thus led by degrees to an improbable conclusion: most of the universe today is composed of invisible dark matter and dark energy. Worse yet, dark matter and dark energy seem to be coincidentally comparable in energy density today, even though the former vastly outweighed the latter at recombination. Physicists dislike coincidences; they prefer to explain the world in terms of cause and effect rather than dumb luck. What is more, another mysterious component, the inflaton, dominated the very early universe and seeded cosmic structure. Why should we believe a cosmological model that is based on the seemingly fanciful introduction of three enigmatic entities?

One reason is that these three entities explain a wealth of previously known facts. Dark matter was first postulated in the 1930s to explain measurements of the local mass density in galaxy clusters. Albert Einstein introduced the concept of dark energy in 1917 when he included the so-called cosmological constant in his equations to counteract the influence of gravity. He later disavowed the constant, but it was resurrected in the 1990s, when observations of distant supernovae showed that the expansion of the universe is accelerating [see "From Slowdown to Speedup," by Adam G. Riess and Michael S. Turner, on page 62]. The energy densities of dark matter and dark energy, as measured from the CMB, are in striking accord with these astronomical observations.

Second, the standard cosmological model has predictive power. In 1968 Joseph Silk (now at the University of Oxford) predicted that the small-scale acoustic peaks in the CMB should be damped in a specific, calculable way. As a result, the corresponding radiation should gain a small but precisely known polarization. (Polarized radiation is oriented in a particular direction.) One might assume that the CMB would be unpolarized because the scattering of the photons in the primordial

NASA/WMAP SCIENCE TEAM

THE POWER SPECTRUM



OBSERVATIONS OF THE CMB provide a map of temperature variations across the whole sky (*a*). When researchers analyze portions of that map (*b*), they use band filters to show how the temperature of the radiation varies at different scales. The variations are barely noticeable at large scales corresponding to regions that stretch about 30 degrees across the sky (*c*) and at small scales corresponding to regions about a tenth of a degree across (*e*). But the temperature differences are quite distinct for regions about one degree across (*d*). This first peak in the power spectrum (*graph at bottom*) reveals the compressions and rarefactions caused by the fundamental wave of the early universe; the subsequent peaks show the effects of the overtones.

NASA/WMAP SCIENCE TEAM (*a–e*); BRYAN CHRISTIE DESIGN (*graph*)

plasma would have randomized their direction. But on the small scales where damping occurs, photons can travel with relatively few scatterings, so they retain directional information that is imprinted as a polarization of the CMB. This acoustic polarization was measured by the Degree Angular Scale Interferometer (an instrument operated at the Amundsen-Scott South Pole Station in Antarctica) and later by WMAP; the value was in beautiful agreement with predictions. WMAP also detected polarization on larger scales that was caused by scattering of CMB photons after recombination [*see box on opposite page*].

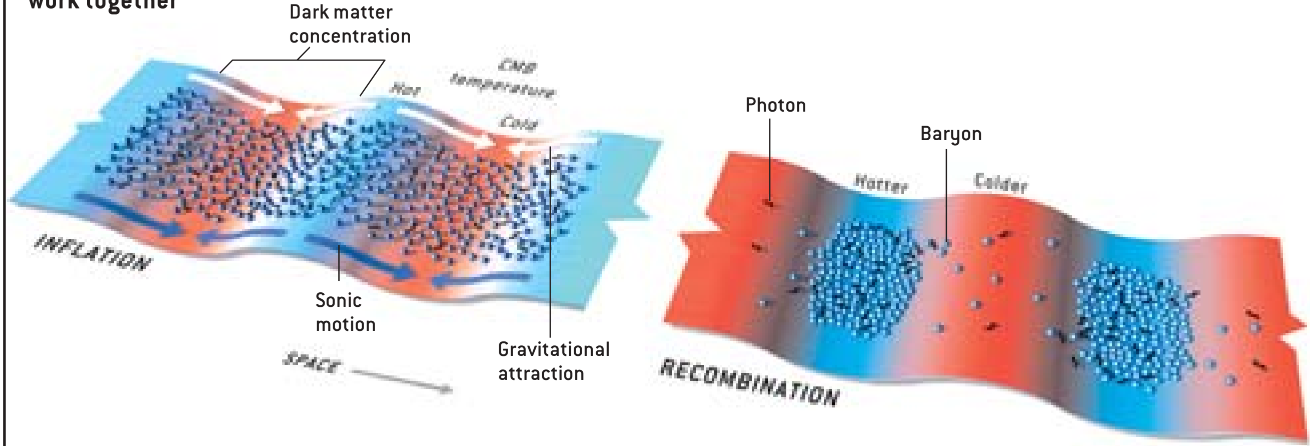
Furthermore, the existence of dark energy predicts additional phenomena in the CMB that are beginning to be observed. Because dark energy accelerates the expansion of the universe, it weakens the gravitational-potential wells associated with the clustering of galaxies. A photon traveling through such a region gets a boost in energy as it falls into the potential well, but because the well is shallower by the time the photon climbs back out, it loses less energy than it previously gained. This phenomenon, called the integrated Sachs-Wolfe effect, causes large-scale temperature variations in the CMB. Observers have recently seen hints of this correlation by comparing large structures in galaxy surveys with the WMAP data. The amount of dark energy needed to produce the large-scale temperature variations is consistent

GRAVITATIONAL MODULATION

INFLUENCE OF DARK MATTER modulates the acoustic signals in the CMB. After inflation, denser regions of dark matter that have the same scale as the fundamental wave (*represented as troughs in this potential-energy diagram*) pull in baryons and photons by gravitational attraction. (The troughs are shown in red because gravity also reduces the temperature of any escaping photons.) By the time of recombination, about 380,000 years later, gravity and sonic motion have worked together to raise the radiation temperature in the troughs (*blue*) and lower the temperature at the peaks (*red*).

FIRST PEAK

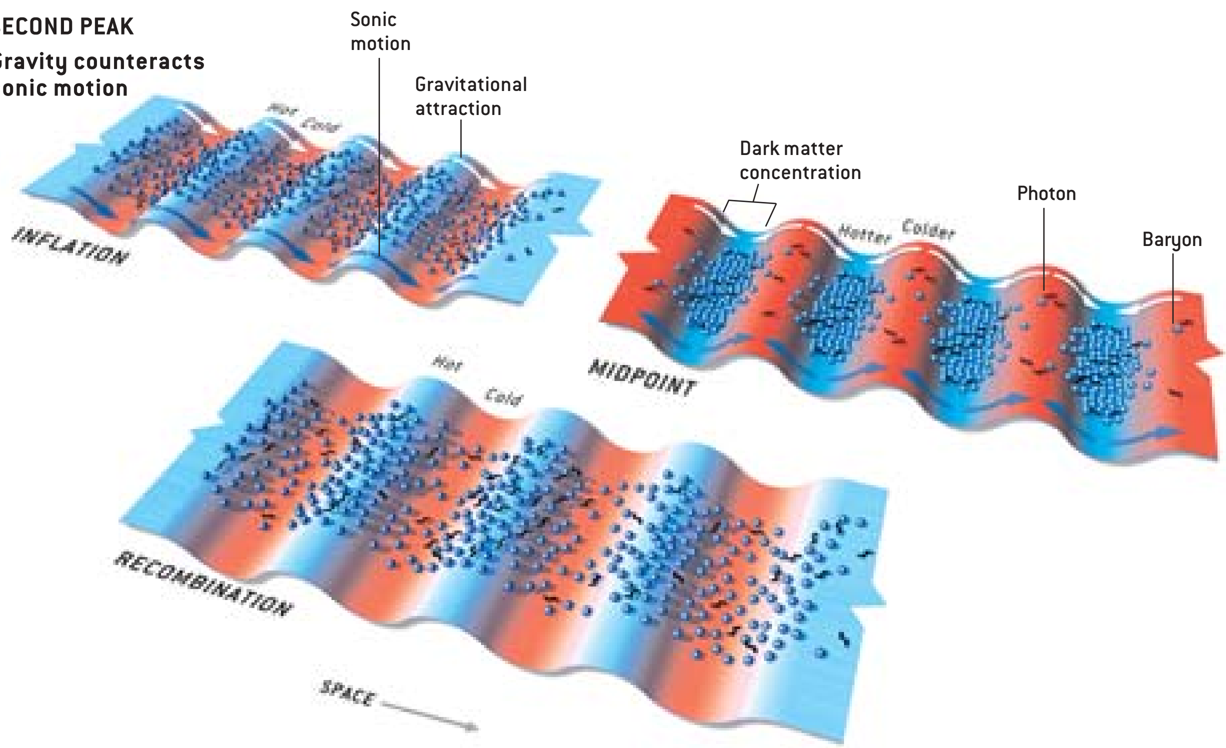
Gravity and sonic motion work together



AT SMALLER SCALES, gravity and acoustic pressure sometimes end up at odds. Dark matter clumps corresponding to a second-peak wave maximize radiation temperature in the troughs long before recombination. After this midpoint, gas pressure pushes baryons and photons out of the troughs (*blue arrows*) while gravity tries to pull them back in (*white arrows*). This tug-of-war decreases the temperature differences, which explains why the second peak in the power spectrum is lower than the first.

SECOND PEAK

Gravity counteracts sonic motion



BRYAN CHRISTIE DESIGN

with the amount inferred from the acoustic peaks and the distant supernovae. As the data from the galaxy surveys improve and other tracers of the large-scale structure of the universe become available, the integrated Sachs-Wolfe effect could become an important source of information about dark energy.

No Requiem Yet

THE CMB MAY ALSO provide crucial new evidence that could explain what happened during the very first moments after the big bang. Few aspects of cosmology are more bizarre than the period of inflation. Did the universe really inflate, and, if so, what was the nature of the inflaton, the theoretical field that caused the rapid expansion? Current measurements of the CMB have dramatically strengthened the case for the simplest models of inflation, which assume that the amplitudes of the initial density fluctuations were the same at all scales. But if more detailed observations of the CMB reveal that the amplitudes varied at different scales, the simple inflation models would be in trouble. More baroque alternatives would need to be invoked or altogether different paradigms adopted.

Another exciting possibility is that we could learn about the physics of inflation by determining the energy scale at which it took place. For example, physicists believe that the weak nuclear force and the electromagnetic force were different aspects of a single electroweak force when the universe was hotter than 10^{15} kelvins. If researchers determine that inflation occurred at this energy scale, it would strongly imply that the inflaton had something to do with electroweak unification. Alternatively, inflation could have occurred at the much higher temperatures at which the electroweak force merges with the strong nuclear force. In this case, inflation would most likely be associated with the grand unification of the fundamental forces.

A distinctive signature in the CMB could allow researchers to settle this issue. In addition to spawning density perturbations, inflation created fluctuations in the fabric of spacetime itself. These fluctuations are gravitational waves whose wavelengths can stretch across the observable universe. The amplitude of these gravitational waves is proportional to the square of the energy scale at which inflation took place. If inflation occurred at the high energies associated with grand unification, the effects might be visible in the polarization of the CMB.

Last, further observations of the CMB could shed some light on the physical nature of dark energy. This entity might be a form of vacuum energy, as Einstein had hypothesized, but its value would have to be at least 60 and perhaps as much as 120 orders of magnitude as small as that predicted from particle physics. And why is dark energy comparable to dark matter in density now and apparently only now? To answer these questions, researchers can take advantage of the fact that CMB photons illuminate structures across the entire observable universe. By showing the amplitude of density fluctuations at different points in cosmic history, the CMB can reveal the tug-of-war between matter and dark energy.

We are led by degrees to an improbable conclusion: most of the universe today is composed of invisible dark matter and dark energy.

Measurements of two CMB phenomena could be particularly useful. The first, called the Sunyaev-Zel'dovich effect, occurs when CMB photons are scattered by the hot ionized gas in galaxy clusters. This effect allows galaxy clusters to be identified during the crucial period, about five billion years ago, when dark energy began to accelerate the expansion of the universe. The number of galaxy clusters, in turn, indicates the amplitude of density fluctuations during this time. The second phenomenon, gravitational lensing, happens when CMB photons pass by a particularly massive structure that bends their trajectories and hence distorts the pattern of temperature and polarization variations. The degree of lensing reveals the amplitude of the mass density fluctuations associated with these structures.

To conduct these investigations of inflation and dark energy, however, researchers will need a new generation of CMB telescopes that can observe the radiation with even greater sensitivity and resolution. In 2007 the European Space Agency plans to launch the Planck spacecraft, a microwave observatory that will be placed in the same orbit as WMAP. Planck will be able to measure CMB temperature differences as small as five millionths of a kelvin and detect hot and cold spots that subtend less than a tenth of a degree across the sky. Such measurements will enable scientists to glimpse the full range of acoustic oscillations in the CMB and thus sharpen their picture of the inflationary spectrum. A multitude of ground-based experiments are also under way to study CMB effects associated with structure in the current epoch of accelerated expansion.

Although the standard cosmological model appears to work remarkably well as a phenomenological description of the universe, a deeper understanding of its mysteries awaits the findings of these experiments. It seems clear that the cosmic symphony will continue to enchant its listeners for some time to come. SA

MORE TO EXPLORE

Wrinkles in Time. George Smoot and Keay Davidson. William Morrow, 1994.

3K: The Cosmic Microwave Background Radiation. R. B. Partridge. Cambridge University Press, 1995.

The Inflationary Universe: The Quest for a New Theory of Cosmic Origins. Alan H. Guth and Alan P. Lightman. Perseus, 1998.

More information about WMAP and the cosmic microwave background can be found at **map.gsfc.nasa.gov** and **background.uchicago.edu**

READING THE BLUEPRINTS of CREATION

By Michael A. Strauss

The latest maps of the cosmos have surveyed hundreds of thousands of galaxies, whose clustering has grown from primordial fluctuations

As late as the 1970s, cosmology, the study of the universe as a whole, was a field filled with much speculation but few hard facts. New observations and theoretical work over the past two decades have changed that dramatically. Cosmology has become a rigorous, quantitative branch of astrophysics with a strong theoretical foundation backed by abundant data. The big bang model, which states that almost 14 billion years ago the universe started expanding from a state of extremely high density and temperature, is able to explain galaxy motions, the abundance of hydrogen and helium, and the properties of the cosmic microwave background (CMB), the remnant heat from the expanding and cooling gas.

Cosmologists can now go to the next level and claim an understanding of the formation of structures in the universe. Measurements of the large-scale distribution of galaxies, as mapped by cartography projects such as the ongoing Sloan Digital Sky Survey (SDSS), are in beautiful agreement with theoretical predictions. We currently have a coherent model that tracks the growth of subtle density fluctuations laid down in the early universe to the present richness of the night sky.

The universe around us exhibits structure on all scales. Stars are not scattered uniformly through space; they are grouped into galaxies. Our sun is one of several hundred billion stars in the Milky Way galaxy, a highly flattened disk 100,000 light-years across. The Milky Way, in turn, is one of tens of billions of galaxies in the observable universe. Our nearest large galactic neighbor is about two million light-years away. But galaxies are not randomly sprinkled like raisins in a muffin. Between 5 and 10 percent are grouped into clusters containing up to 1,000 galaxies in a volume a few million light-years across.

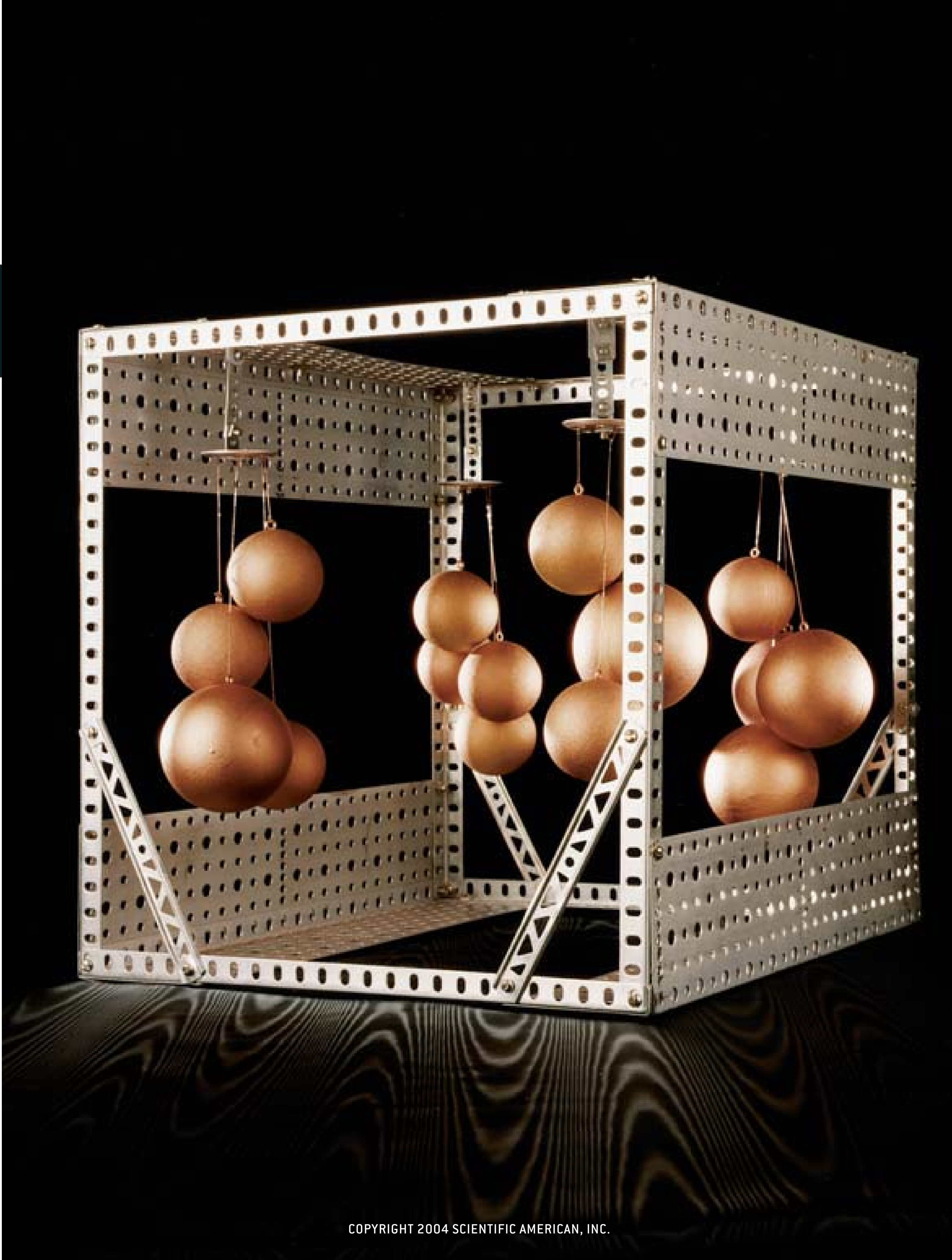
Most astronomers used to think that galaxy clusters were the largest coherent structures in existence. Whereas stars belong to galaxies and many galaxies belong to clusters, the clusters did not seem to be clumped into larger bodies. This picture fit neatly with theorists' understanding of the big bang. When Einstein first applied his general theory of relativity to the universe, he made a dramatic simplifying assumption: the universe, on average, was homogeneous (it had no big lumps) and isotropic (it looked the same in all directions). He called this assumption the cosmological principle, and it underlies all modern scientific models of the universe.

Becoming Aware of Large Structures

TESTING WHETHER the principle applies beyond galaxy clusters requires some depth perception. When you point a telescope at the night sky, the eyepiece reveals stars, planets and galaxies. But without further information, you will not know which objects are small and nearby or large and far away. Fortunately, the telescope can provide that information. For galaxies, the key is that we live in an expanding universe. Galaxies are receding from one another, and the more distant a galaxy is, the faster it is moving away from us. This motion manifests itself as a redshift in the spectrum of the galaxy. The energy of its photons decreases (shifts in wavelength from blue to red) by an amount that depends on its distance. Having established this relation for objects of known distance, researchers use it to study galaxies of unknown distance. They obtain their spectra,

THE UNIVERSE IS HIGHLY STRUCTURED on scales up to a billion light-years or so. Matter is not randomly scattered; gravity has organized it.

DAVID EMMITE PHOTOGRAPHY



determine their redshifts and infer how far away they are.

By the late 1970s, advances in telescope and detector technology made it feasible to carry out extensive redshift surveys of galaxies to create three-dimensional maps of the local cosmos. As a junior in college, I read a *Scientific American* article by Stephen A. Gregory and Laird A. Thompson ["Superclusters and Voids in the Distribution of Galaxies," March 1982] that detailed some of these first 3-D maps. The authors described hints that Einstein's cosmological principle might be wrong: the discovery of coherent structures that were much larger than single clusters, and great voids many tens of millions of light-years across. I was fascinated. This exercise in cosmography, in discovering entirely new structures in the universe, struck me as one of the most exciting things happening in science, and it led me to my present career.

In 1986 Valérie de Lapparent, Margaret J. Geller and John P. Huchra of the Harvard-Smithsonian Center for Astrophysics (CfA) published a map of the distribution of 1,100 galaxies, out of what eventually would be a survey containing 18,000 galaxies. This survey confirmed the richness and ubiquity of large structures. It revealed an unmistakably frothy appearance to the galaxy distribution; galaxies were located along filaments, leaving enormous voids. Among the map's most notable features was a structure dubbed the Great Wall, which stretched 700 million light-years from one edge of the surveyed region to the other. Given that the map did not reveal its end, the full extent of the wall was unknown.

The presence of the Great Wall and the uncertainty about its extent added to the suspicion that the cosmological principle, and therefore our basic theoretical underpinning of the expanding universe, might be incorrect. Was Einstein wrong? Was the universe not homogeneous on average? It was clear that we needed to survey larger volumes to find out.

The big bang paradigm holds that the structure we see in galaxy distribution today grew out of variations present in the almost perfectly smooth early universe. These initial fluctuations were subtle; the density typically varied from one region to another by only one part in 100,000, as measured in the temperature of the cosmic microwave background [see "The Cosmic Symphony," by Wayne Hu and Martin White, on page 44]. If a region of space had a density higher than the average, it had a greater gravitational pull, and thus matter in its vicinity was drawn into it. Similarly, a region of space that was slightly less dense than average lost mass with time. Through this process of gravitational instability, the denser regions eventually became the huge galactic superclusters we see today; the less dense regions became the vast empty voids.

Overview/*Cosmic Structure*

- Astronomers, acting as cosmic cartographers, are creating ever more detailed three-dimensional maps of the locations of galaxies and galaxy clusters. The largest of these projects, the Sloan Digital Sky Survey, is plotting a total of one million galaxies out to a distance of two billion light-years.
- The maps show that galaxies are organized into huge structures, stretching hundreds of millions of light-years. These surveys have quantified the degree of clustering to high precision. The results agree with the clustering expected from extrapolating the cosmic microwave background (CMB) fluctuations to the present. The agreement indicates that astronomers finally have a consistent account of 14 billion years of cosmic evolution.

Running Hot and Cold

ABOUT THE TIME the early redshift surveys were completed, astronomers realized that the story had a twist: the stars and gas we see in galaxies represent just a small fraction (about 2 percent) of the total matter in the universe. The rest of the matter is revealed indirectly through its gravitational effects. Astronomers proposed a variety of models to describe this dark matter. These fell into two broad categories, cold and hot, and the difference is crucial to the evolution of structure.

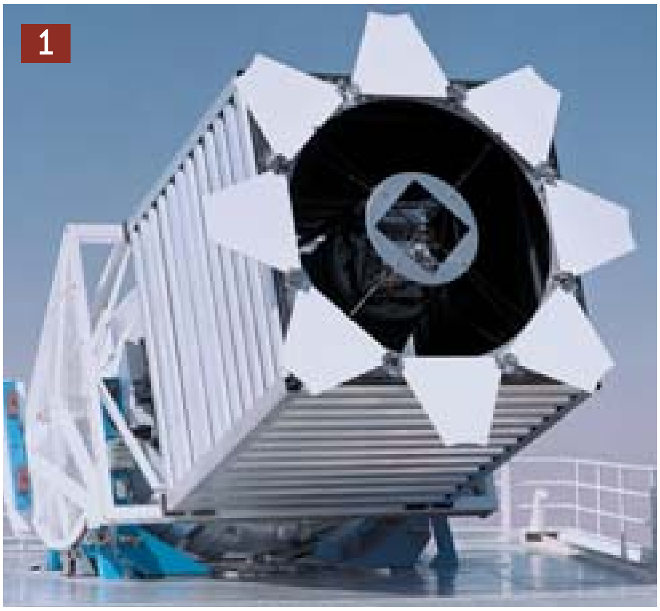
In the cold dark matter scenario, suggested by P. James E. Peebles of Princeton University and others, the first structures to form were relatively small objects such as galaxies and pieces of galaxies. As time went on, gravity brought these together in ever larger structures. In this model, the Great Wall formed relatively recently. In the hot dark matter scenario, posited by Yakov B. Zel'dovich and his colleagues at Moscow State University, dark matter moved sufficiently quickly in the early universe to smooth out any clustering on small scales. The first things to form were large sheets and filaments tens or hundreds of millions of light-years in extent, which only later fragmented to form galaxies. In other words, the Great Wall is ancient.

Thus, the next generation of surveys would not just test Einstein's cosmological principle and identify the largest structures in the universe; it would also probe the nature of dark matter. One such survey was conducted from 1988 to 1994 by Stephen A. Shectman of the Carnegie Institution of Washington and his collaborators using the Las Campanas 2.5-meter telescope in Chile [see "Mapping the Universe," by Stephen D. Landy; SCIENTIFIC AMERICAN, June 1999]. The survey contained 26,418 galaxy redshifts and covered an appreciably larger volume than the original CfA survey. As team member Robert P. Kirshner of the CfA phrased it, the Las Campanas survey found "the end of greatness." It revealed a galaxy distribution similar to that of the CfA survey but saw no structures much larger than the Great Wall. Einstein's cosmological principle appeared to hold: the cosmos is homogeneous and isotropic over vast distances.

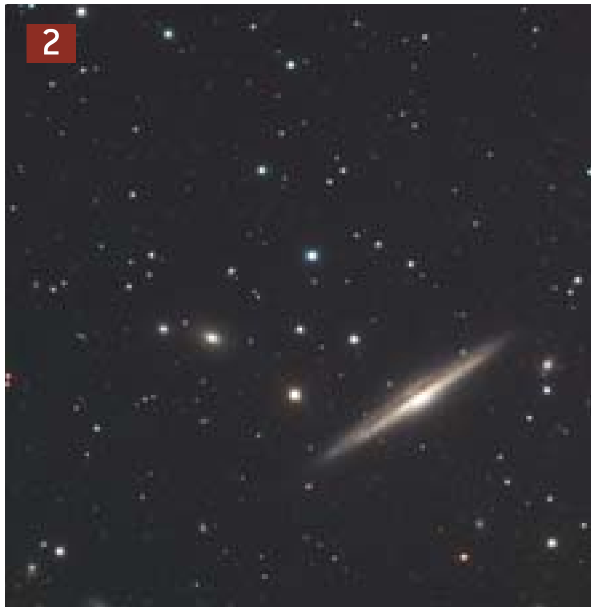
The Las Campanas survey was still not large enough to be definitive, however. It said nothing about what was happening in regions of space one to two billion light-years across. It is on these largest scales that clustering is the easiest to interpret theoretically yet the hardest to measure observationally. The variations in galaxy numbers over such a volume are subtle, and it is easy to introduce errors into the sample; artifacts of the selection procedure might masquerade as clustering.

HOW TO SURVEY THE COSMOS IN FOUR NOT-SO-EASY STEPS

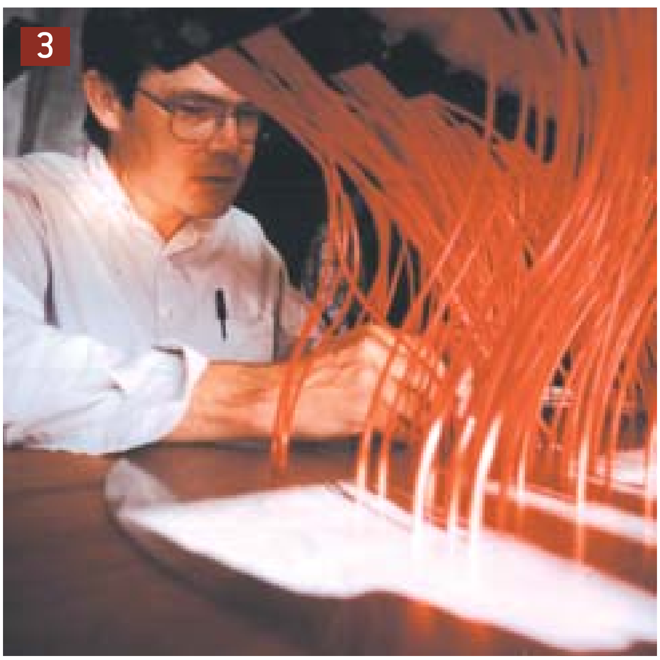
THE SLOAN DIGITAL SKY SURVEY, the most advanced of the current generation of astronomical surveys, is compiling an atlas of a quarter of the sky. It will take five years, using a dedicated 2.5-meter telescope atop Apache Point in New Mexico.



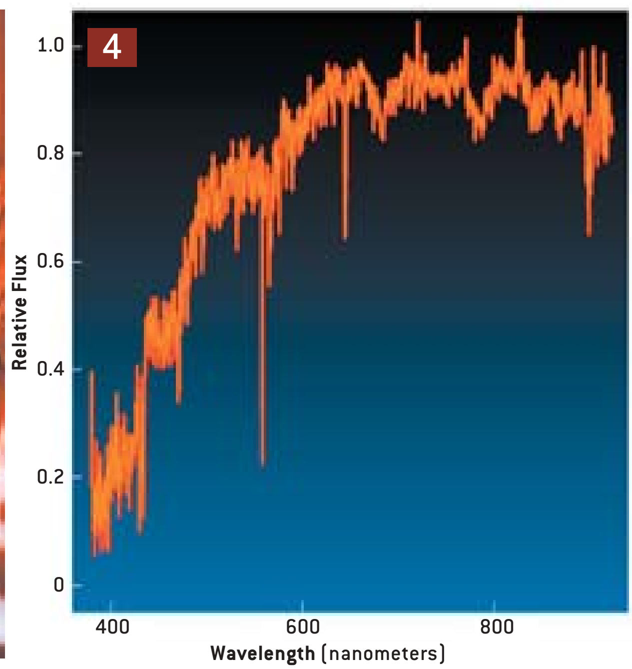
THE TELESCOPE operates in camera mode on clear nights, taking pictures through five color filters at the rate of 20 square degrees an hour—netting millions of celestial bodies per night.



GALAXIES and other objects are identified by software and selected for follow-up spectroscopy. The object shown here is the spiral galaxy UGC 03214 in the constellation Orion.



OPTICAL FIBERS are plugged into a metal plate with 640 holes. Each fiber channels light from a celestial body to a spectrograph, which operates when the sky is not so clear.

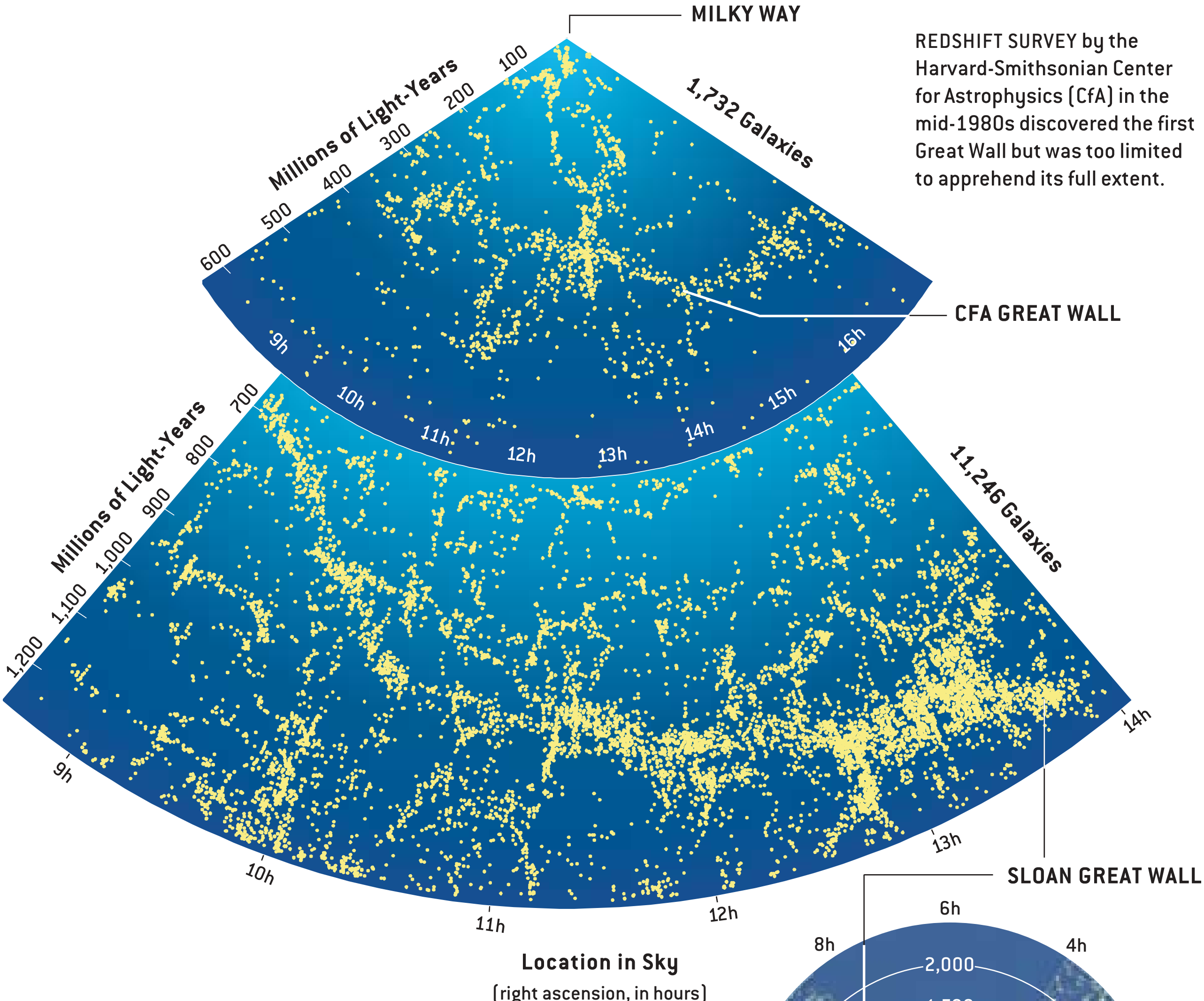


THE SPECTRA THAT RESULT provide a precise way to classify objects. From them, astronomers also determine the redshifts, hence distances, of the objects.

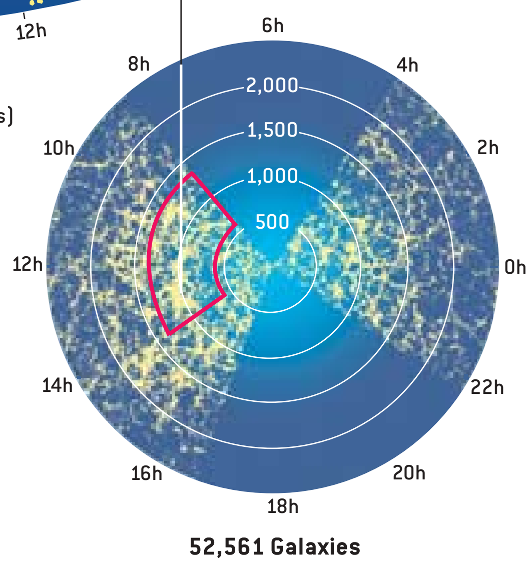
SDSS COLLABORATION

COSMIC MAPS

THESE WEDGE-SHAPED FIGURES show the distribution of galaxies (*dots*) in two volumes of space. The third dimension, which spans an angle of several degrees, has been flattened. The figures reveal two dramatic "Great Walls" containing thousands of galaxies each, as well as filaments and voids at all scales.



REDSHIFT SURVEY by the Harvard-Smithsonian Center for Astrophysics (CfA) in the mid-1980s discovered the first Great Wall but was too limited to apprehend its full extent.

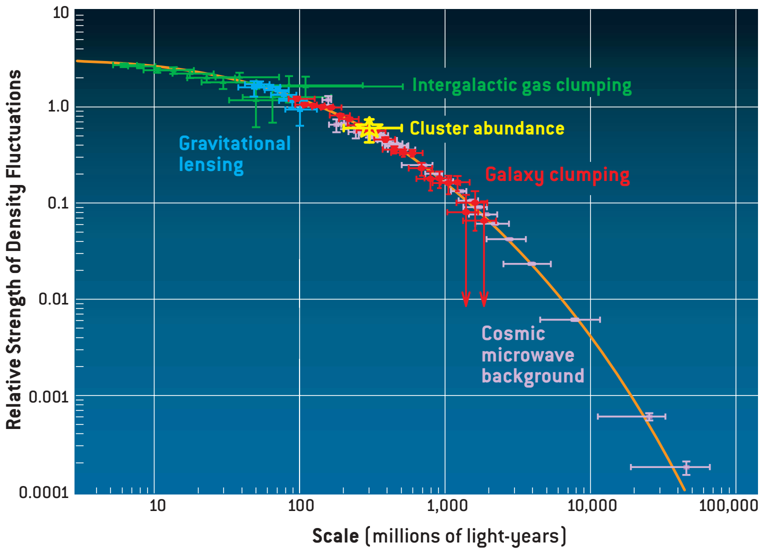


SLOAN DIGITAL SKY SURVEY, which is now under way, covers a much larger volume. It found another Great Wall more than one billion light-years long. The region shown above is about 1 percent of the full Sloan volume. The bull's-eye diagram (*right*) covers a volume six times as large; the outline indicates the location of the above wedge. The concentric circles indicate distance in millions of light-years.

BRYAN CHRISTIE DESIGN; SOURCES: J. RICHARD GOTT AND MARIO JURIĆ *Princeton University* (*wedge*); SDSS COLLABORATION (*bull's-eye*)

OVERALL DISTRIBUTION OF COSMIC STRUCTURE

THE MAPS produced by galaxy surveys can be boiled down to a power spectrum, which shows the fractional density variation (*vertical axis*) from one position to another in regions of different sizes (*horizontal axis*). Other data—cosmic microwave background, gravitational lensing, galaxy-cluster surveys, hydrogen gas clouds—can be plotted in the same way. They follow the same universal curve (*solid line*). The relative fluctuations approach zero, substantiating Einstein's cosmological principle. The arrows represent upper limits.



Astronomers, for example, typically choose all galaxies brighter than a certain value to be included in a given redshift survey. If they overestimate galaxy brightnesses in one part of the sky, the sample will have too many galaxies in that region, yielding a false measurement of clustering. Thus, a definitive redshift survey must not only cover a huge volume, it must be exquisitely calibrated.

On a Clear Night...

IN THE LATE 1980S James E. Gunn of Princeton, Richard G. Kron and Donald G. York of the University of Chicago and others began a collaboration to do the problem right. That is, they sought to measure the distribution of galaxies in the largest volume to date, with careful control of calibration. About a decade later the Sloan Digital Sky Survey, an $80-million, 200-astronomer collaboration, started operation. The SDSS features a special-purpose telescope with a 2.5-meter-wide primary mirror. The telescope operates in two modes. On the most pristine nights, it uses a wide-field camera to take carefully calibrated pictures of the night sky in five broad wavebands. The camera uses CCDs, highly sensitive electronic detectors whose response can be calibrated with an accuracy of 1 percent.

On nights with moonshine or mild cloud cover, the telescope instead uses a pair of spectrographs to obtain spectra, and therefore redshifts, of 608 objects at a time. For reference, the device also takes spectra of 32 blank patches of sky. Unlike traditional telescopes, for which nights are parceled out among many scientific programs, this telescope is devoted solely to the survey, every night for five years. The project is now approaching the halfway point in its goal of measuring one million galaxy and quasar redshifts. As a midterm report, my colleagues and I recently completed an analysis of the first 200,000 galaxies with redshifts.

In a parallel effort, a team of Australian and British astronomers built a spectrograph for the 3.9-meter Anglo-Australian Telescope, capable of measuring the spectra of 400 objects at a time over a field of view two degrees on a side (thus earning the name "Two Degree Field," or 2dF). The 2dF team worked from galaxy catalogues drawn from carefully calibrated and electronically scanned photographic atlases that were already available. Now complete, the survey measured the redshifts of 221,414 galaxies over a period of five years.

Our surveys describe the distribution of galaxies. They do not see dark matter, which constitutes the bulk of the mass of the universe. Researchers have no reason to assume that the distribution of galaxies is the same as the distribution of dark matter. For example, galaxies might tend to form only in regions that contain an above-average density of dark matter—a scenario astronomers refer to as biasing.

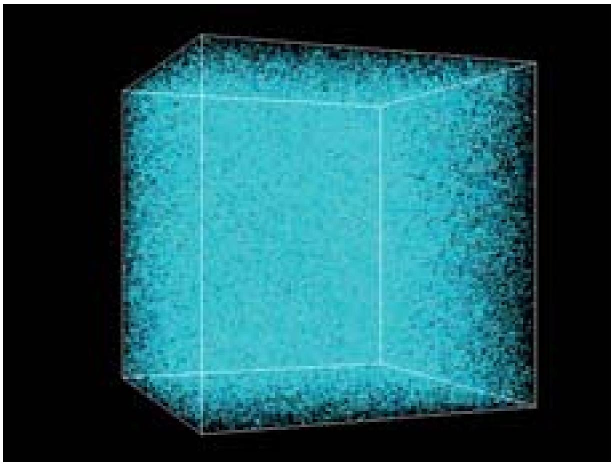
By analyzing previous generations of redshift surveys, my colleagues and I had shown that the galaxy and dark matter distributions were closely related, but we were unable to distinguish between simple models of bias and the unbiased case. More recently Licia Verde of the University of Pennsylvania and her colleagues used the 2dF galaxy redshift survey to measure triplets of galaxies. It turns out that the number of these trios

THE AUTHOR

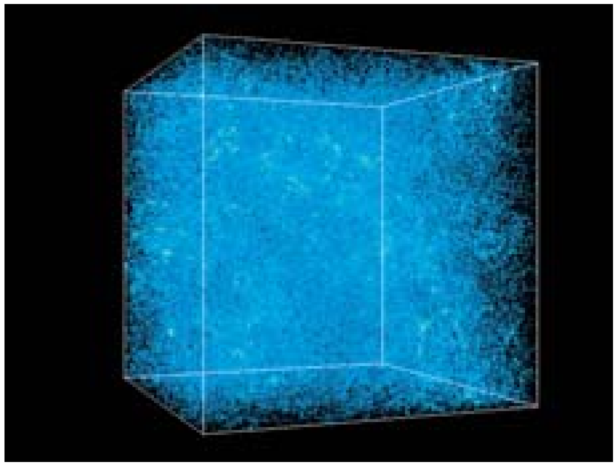
MICHAEL A. STRAUSS is deputy project scientist and project spokesperson for the Sloan Digital Sky Survey, an effort to make a complete map of a quarter of the sky. He received a Ph.D. in physics from the University of California, Berkeley, did postdoctoral work at the California Institute of Technology and at the Institute for Advanced Study in Princeton, N.J., and now holds a faculty position at Princeton University. He thanks his Sloan project colleagues for a fabulous data set. Strauss was featured in the September 2002 issue of *New Jersey Monthly* as having one of the best jobs in the state of New Jersey.

BUILDING A UNIVERSE

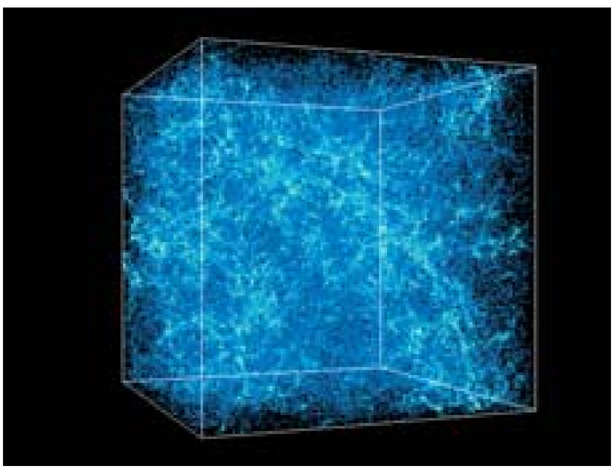
TO CONNECT THE CLUMPING of matter today (revealed by Sloan and other mapping projects) with the clumping of matter in the early universe (revealed by observations of the cosmic microwave background radiation, or CMB), cosmologists run computer simulations. Each frame is a snapshot at a time after the onset of the big bang expansion. Because the universe is expanding, the frames are not to scale: the first one is about five million light-years across, the last one about 140 million light-years across. Dots represent matter. The simulation was performed at the National Center for Supercomputer Applications (the full movie is available at cfcp.uchicago.edu/lss/filaments.html).



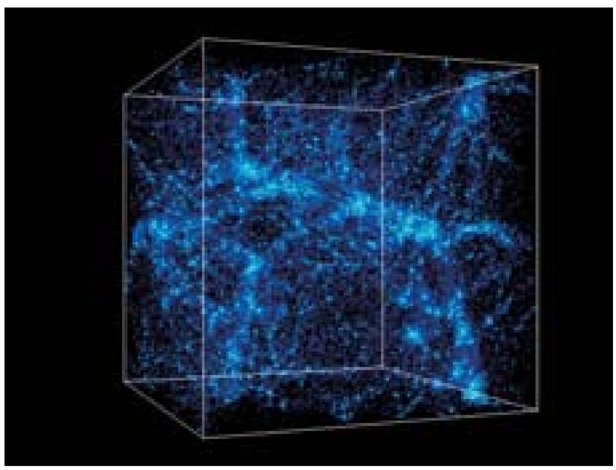
120 MILLION YEARS: Early on, matter was spread out in a nearly uniform sea with subtle undulations.



490 MILLION YEARS: Dense regions gained material at the expense of less dense ones. The first galaxies formed.



1.2 BILLION YEARS: Over time, gravity pulled matter into vast filaments and emptied the intervening voids.



13.7 BILLION YEARS (TODAY): The growth of large structures has ceased because cosmic acceleration counteracts clumping.

ANDREY KRAVTSOV *University of Chicago* AND ANATOLY KLYPIN *New Mexico State University*

depends on the total mass, dark matter included. The researchers found that the galaxy distribution is essentially unbiased: the density field of galaxies is the same as that of the dark matter, which means that the galaxy surveys accurately reflect the overall arrangement of matter in the cosmos.

The Power of the Power Spectrum

WITH THIS CAVEAT ADDRESSED, cosmologists can interpret the galaxy maps. Among the most useful statistical tools to describe galaxy clustering is the power spectrum. Imagine placing a series of spheres of a given radius (say, 40 million light-years) at random in the universe and counting the number of galaxies in each one. Because galaxies are clustered, that number will vary substantially from one sphere to another. The variation in the number of galaxies is a measure of the lumpiness of the galaxy distribution on a scale, in this case, of 40 million light-years. Cosmologists repeat the exercise with spheres of various radii to measure this lumpiness at different scales.

An analogy is to express a complex sound in terms of the contributions from sound waves of different wavelengths. A graphic equalizer on a home audio system can perform this function: it shows how loud the deep bass notes (of very long wavelength) are, how loud the treble (of shorter wavelength) is, and so on. In a live concert, a person with a musically trained ear can easily pick out the piccolo from the bassoon. Cosmologists do the same thing with the distribution of galaxies. The relative amount of structure on large and small scales is a powerful cosmological probe.

The power spectrum has been measured by both the 2dF and Sloan teams, with consistent results. The first thing to note is that the fluctuations are weaker as one proceeds to larger scales [*see illustration on page 59*]. Weak fluctuations mean that the galaxy distribution is very close to homogeneous, exactly as Einstein's cosmological principle requires.

Second, the power spectrum, when plotted on a logarithmic scale, does not follow a straight line. The deviation from straightness is confirmation that the dynamics of the universe have changed with time. From other observations, astronomers have concluded that the energy density of the universe is dominated by matter and a mysterious component known as dark energy. Photons, their energy sapped by cosmic expansion, are negligible. Extrapolating backward in time, however, photons dominated when the universe was less than 75,000 years old. When photons ruled, gravity did not cause fluctuations to grow with time the same way they do today. That, in turn, caused the power spectrum to behave differently on the largest scales (more than about 1.2 billion light-years).

The exact scale of this deviation provides a measure of the total density of matter in the universe, and the result—roughly 2.5×10^{-27} kilogram per cubic meter of space—agrees with the value from other measurements. Finally, the combination of these results strongly suggests that the dark matter is all of the cold variety. Hot dark matter would smooth out the fluctuations in the galaxy distribution on smaller scales, and that is not seen.

The fluctuations we observed in the galaxy distribution on large scales should simply be an amplified version of those of the early universe. These early fluctuations are apparent directly in the CMB, so we can directly compare the CMB and galaxy power spectra. Amazingly, we get consistent answers from these two approaches. On scales approaching one billion light-years, the galaxy density fluctuates by about one part in 10. The CMB reveals fluctuations of one part in 100,000, which, when extrapolated to the present, are in beautiful agreement. This gives us confidence that our cosmological picture—big bang, gravitational instability and all—is actually correct.

The Future of Large-Scale Structure Studies

THE MAIN SDSS GALAXY SURVEY probes the structure of the cosmos on scales from 100 million to more than one billion light-years. To probe yet larger scales, SDSS has a second, auxiliary sample of extremely luminous galaxies that extends more than five billion light-years away. For smaller scales, a third sample looks at absorption lines in the spectra of distant quasars, whose light passes through a dense network of clouds of hydrogen gas not yet formed into galaxies.

With these data, cosmologists are working to make an even tighter connection between cosmic structures (seen today and the not so distant past) and the CMB (which probes cosmic structures in the very early universe). In particular, the power spectrum of the microwave background shows a series of distinctive bumps, which reflect the relative amounts of dark and ordinary matter. Researchers hope to find the equivalent bumps in the present-day power spectrum. If they do, it will be further confirmation that the fluctuations observed today evolved directly from those seen in the early universe.

Another way to trace the development of structures over time is to probe the distribution of more distant galaxies—looking to great distances is looking back in time. The dark matter at those early times should be weakly clustered because gravitational instability had not yet had as much time to operate. But surveys carried out with the European Southern Observatory's Very Large Telescope in Chile and the Keck Observatory in Hawaii show that tremendously distant galaxies are just as clustered as today and are arranged in the same filamentary, bubbly structures that nearby galaxies are. This is odd. Unlike today's galaxies, which follow the dark matter, these early galaxies must be much more strongly clustered than the underlying dark matter is. This pattern is an important clue to how galaxies formed.

Researchers are close to a complete understanding of the development of the structure of the cosmos, from undulations in the primordial plasma to the bright galaxy clusters of the modern universe. That said, their work is cut out for them in the coming years. What exactly is the mechanism that gave rise to the initial fluctuations in the microwave background? How exactly did the galaxies form? Why do they have the properties that they do? And could it have been any other way—could one imagine a universe with fluctuations that started out with much higher or lower amplitudes? These are among the big questions that perhaps a high school or college student reading this article will be inspired to tackle. SA

MORE TO EXPLORE

Cosmological Physics. John A. Peacock. Cambridge University Press, 1998.

Cosmology: The Science of the Universe. Second edition. Edward Harrison. Cambridge University Press, 2000.

The 2dF Galaxy Redshift Survey: The Power Spectrum and the Matter Content of the Universe. Will J. Percival et al. in *Monthly Notices of the Royal Astronomical Society,* Vol. 327, No. 4, pages 1297–1306; November 2001. Available online at **arXiv.org/abs/astro-ph/0105252**

From Here to Eternity. Valerie Jamieson in *New Scientist,* Vol. 180, No. 2422, pages 36–39; November 22, 2003. The centerfold map is also available at **www.astro.princeton.edu/~mjuric/universe/**

The Three-Dimensional Power Spectrum of Galaxies from the Sloan Digital Sky Survey. Max Tegmark et al. in *Astrophysical Journal* (in press). **arXiv.org/abs/astro-ph/0310725**

The official Web site of the Sloan Digital Sky Survey is **www.sdss.org**
The official Web site of the 2dF Galaxy Redshift Survey is **msowww.anu.edu.au/2dFGRS**

SPECIAL REPORT

FROM SLOWDOWN to SPEEDUP

By Adam G. Riess and Michael S. Turner

Distant supernovae are revealing the crucial time when the expansion of the universe changed from decelerating to accelerating

From the time of Isaac Newton to the late 1990s, the defining feature of gravity was its attractive nature. Gravity keeps us grounded. It slows the ascent of baseballs and holds the moon in orbit around the earth. Gravity prevents our solar system from flying apart and binds together enormous clusters of galaxies. Although Einstein's general theory of relativity allows for gravity to push as well as pull, most physicists regarded this as a purely theoretical possibility, irrelevant to the universe today. Until recently, astronomers fully expected to see gravity slowing down the expansion of the cosmos.

In 1998, however, researchers discovered the repulsive side of gravity. By carefully observing distant supernovae—stellar explosions that for a brief time shine as brightly as 10 billion suns—astronomers found that they were fainter than expected. The most plausible explanation for the discrepancy is that the light from the supernovae, which exploded billions of years ago, traveled a greater distance than theorists had predicted. And this explanation, in turn, led to the conclusion that the expansion of the universe is actually speeding up, not slowing down. This was such a radical finding that some cosmologists suggested that the falloff in supernova brightness was the result of other effects, such as intergalactic dust dimming the light. In the past few years, though, astronomers have solidified the case for cosmic acceleration by studying ever more remote supernovae.

But has the cosmic expansion been speeding up throughout the lifetime of the universe, or is it a relatively recent development—that is, occurring within the past five billion years or so? The answer has profound implications. If scientists find that the expansion of the universe has always been accelerating, they will have to completely revise their understanding of cosmic evolution. But if, as cosmologists expect, the acceleration turns out to be a recent phenomenon, researchers may be able to determine its cause—and perhaps answer the larger question of the destiny of the universe—by learning when and how the expansion began picking up speed.

Battle of Titans

ALMOST 75 YEARS AGO astronomer Edwin Hubble discovered the expansion of the universe by observing that other galaxies are moving away from ours. He noted that the more distant galaxies were receding faster than nearby ones, in accordance with what is now known as Hubble's law (relative velocity equals distance multiplied by Hubble's constant). Viewed in the context of Einstein's general theory of relativity, Hubble's law arises because of the uniform expansion of space, which is merely a scaling up of the size of the universe [*see top illustration in box on page 65*].

In Einstein's theory, the notion of gravity as an attractive force still holds for all known forms of matter and energy, even on the cosmic scale. Therefore, general relativity predicts that the expansion of the universe should slow down at a rate determined by the density of matter and energy within it. But general relativity also allows for the possibility of forms of energy with strange properties that produce repulsive gravity [*see box on page 66*]. The discovery of accelerating rather than decelerating expansion has apparently revealed the presence of such an energy form, referred to as dark energy.

TO MEASURE DISTANCES across the universe, astronomers rely on type Ia supernovae, which are represented in this photograph by lightbulbs.

DAVID EMMITE PHOTOGRAPHY


Ia
Ia
Ia
Ia
Ia
Ia
Ia
Ia

Whether or not the expansion is slowing down or speeding up depends on a battle between two titans: the attractive gravitational pull of matter and the repulsive gravitational push of dark energy. What counts in this contest is the density of each. The density of matter decreases as the universe expands because the volume of space increases. (Only a small fraction of matter is in the form of luminous stars; the bulk is believed to be dark matter, which does not interact in a noticeable way with ordinary matter or light but has attractive gravity.) Although little is known about dark energy, its density is expected to change slowly or not at all as the universe expands. Currently the density of dark energy is higher than that of matter, but in the distant past the density of matter should have been greater, so the expansion should have been slowing down then [*see right illustration in box on page* 67].

Cosmologists have other reasons to expect that the expansion of the universe has not always been speeding up. If it had been, scientists would be at a loss to explain the existence of the cosmic structures observed in the universe today. According to cosmological theory, galaxies, galaxy clusters and larger structures evolved from small inhomogeneities in the matter density of the early universe, which are revealed by variations in the temperature of the cosmic microwave background (CMB). The stronger attractive gravity of the overdense regions of matter stopped their expansion, allowing them to form gravitationally bound objects—from galaxies such as our own to great clusters of galaxies. But if the expansion of the universe had always been accelerating, it would have pulled apart the structures before they could be assembled. Furthermore, if the expansion had been accelerating, two key aspects of the early universe—the pattern of CMB variations and the abundances of light elements produced seconds after the big bang—would not agree with current observations.

Nevertheless, it is important to look for direct evidence of an earlier, slowing phase of expansion. Such evidence would help confirm the standard cosmological model and give scientists a clue to the underlying cause of the present period of cosmic acceleration. Because telescopes look back in time as they gather light from far-off stars and galaxies, astronomers can explore the expansion history of the universe by focusing on distant objects. That history is encoded in the relation between the distances and recession velocities of galaxies. If the expansion is slowing down, the velocity of a distant galaxy would be relatively greater than the velocity predicted by Hubble's law. If the expansion is speeding up, the distant galaxy's velocity would fall below the predicted value. Or, to put it another way, a galaxy with a given recession velocity will be farther away than expected—and hence fainter—if the universe is accelerating [*see bottom illustration on opposite page*].

Overview/*Cosmic Expansion*

- In 1998 observations of distant supernovae indicated that the expansion of the universe is speeding up. Since then, astronomers have solidified the case for cosmic acceleration.
- By studying ever more remote supernovae, researchers have found evidence that the expansion slowed down before it sped up—just as cosmologists had predicted.
- Determining the time when the expansion switched from slowdown to speedup may reveal the nature of dark energy and the ultimate fate of the universe.

Supernova Hunting

TO TAKE ADVANTAGE of this simple fact requires finding astronomical objects that have a known intrinsic luminosity—the amount of radiation per second produced by the object—and that can be seen across the universe. A particular class of supernovae known as type Ia are well suited to the task. These stellar explosions are so bright that ground telescopes can see them halfway across the visible universe, and the Hubble Space Telescope can view them from even farther away. Over the past decade, researchers have carefully calibrated the intrinsic luminosity of type Ia supernovae, so the distance to one of these explosions can be determined from its apparent brightness.

Astronomers can deduce the recession velocity of a supernova by measuring the redshift of the light from the galaxy in which it lies. Radiation from receding objects is shifted to longer wavelengths; for example, light emitted when the universe was half its present size will double in wavelength and become redder. By gauging the redshift and apparent brightness of a large number of supernovae located at a variety of distances, researchers can create a record of the universe's expansion.

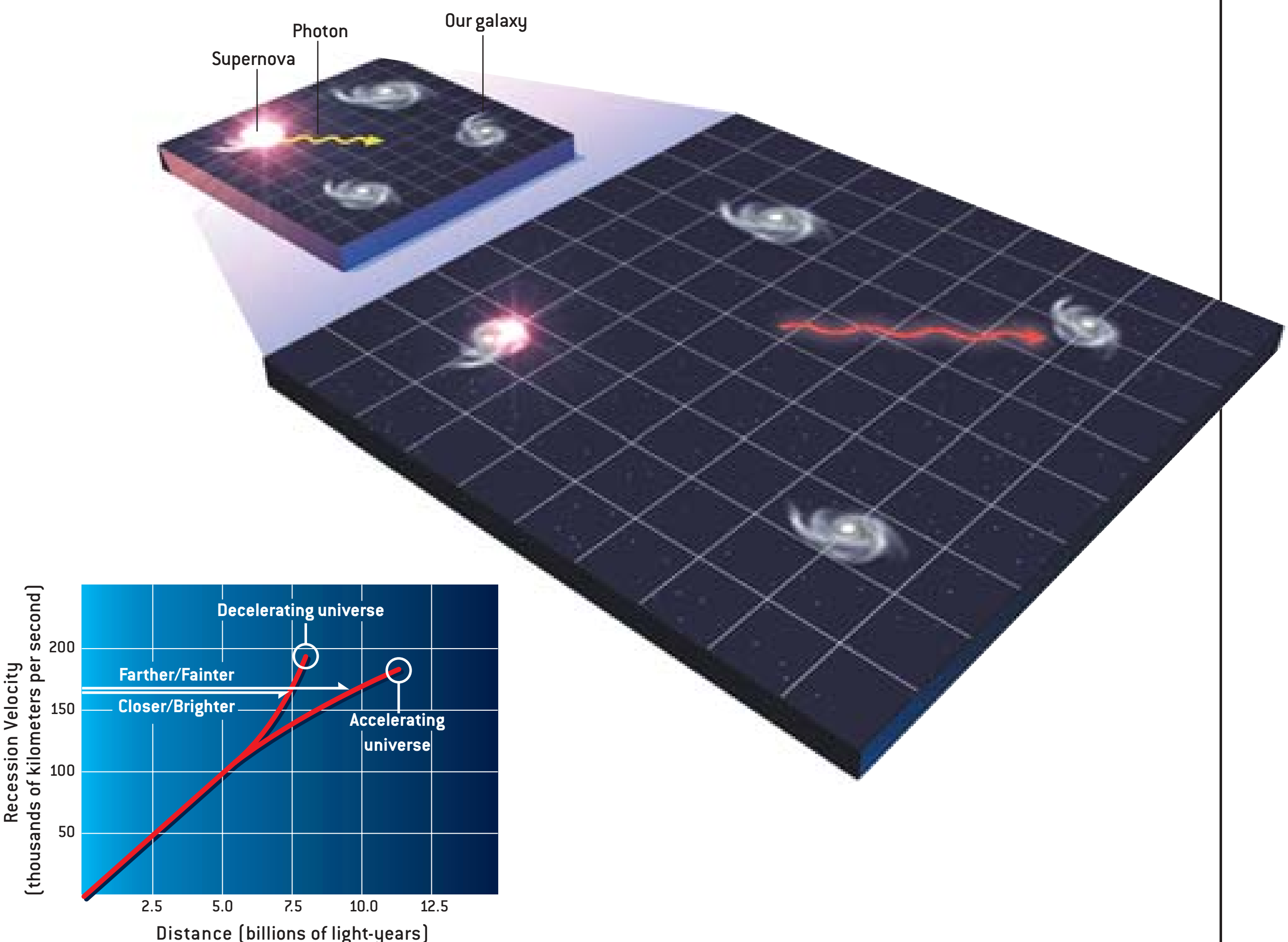
Unfortunately, type Ia supernovae are rare, occurring in a galaxy like the Milky Way only once every few centuries on average. The technique used by supernova hunters is to repeatedly observe a patch of sky containing thousands of galaxies and then compare the images. A transient point of light that appears in one image but not in a previous one could be a supernova. The 1998 results showing evidence of cosmic acceleration were based on the observations of two teams that looked at supernovae that exploded when the universe was about two thirds of its present size, about five billion years ago.

Some scientists wondered, though, whether the teams had correctly interpreted the data from the supernovae. Was it possible that another effect besides cosmic acceleration could have caused the supernovae to appear fainter than expected? Dust filling intergalactic space could also make the supernovae appear dim. Or perhaps ancient supernovae were just born dimmer because the chemical composition of the universe was different from what it is today, with a smaller abundance of the heavy elements produced by nuclear reactions in stars.

Luckily, a good test of the competing hypotheses is available. If supernovae appear fainter than expected because of an astrophysical cause, such as a pervasive screen of dust, or because past supernovae were born dimmer, the putative dimming effects should increase with the objects' redshift. But if the dimming is the result of a recent cosmic speedup that followed an earlier era of deceleration, supernovae from the slowdown period would

EXPANDING SPACE

IMAGINE THAT A SUPERNOVA exploded in a distant galaxy when the universe was half its present size (*left*). By the time the radiation from the explosion reached our galaxy, its wavelength would have doubled, shifting the light toward the red part of the spectrum (*right*). (Note that the galaxies are not drawn to scale; the distance between them would actually be much greater than shown.) If the expansion of the universe were decelerating, the supernova would be closer and brighter than expected; if the expansion were accelerating, the supernova would be farther away and dimmer (*graph at bottom*).



appear relatively brighter. Therefore, observations of supernovae that exploded when the universe was less than two thirds of its present size could provide the evidence to show which of the hypotheses is correct. (It is possible, of course, that an unknown astrophysical phenomenon could precisely match the effects of both the speedup and slowdown, but scientists generally disfavor such artificially tuned explanations.)

Finding such ancient and far-off supernovae is difficult, however. A type Ia supernova that exploded when the universe was half its present size is about one ten-billionth as bright as Sirius, the brightest star in the sky. Ground-based telescopes cannot reliably detect the objects, but the Hubble Space Telescope can. In 2001 one of us (Riess) announced that the space telescope had serendipitously imaged an extremely distant type Ia supernova (dubbed SN 1997ff) in repeated observations. Given the redshift of the light from this stellar explosion—which occurred about 10 billion years ago, when the universe was one third its current size—the object appeared much brighter than it would have been if the dusty universe hypothesis were true. This result was the first direct evidence of the decelerating epoch. The two of us proposed that observations of more high-redshift supernovae could provide definitive proof and pin down the transition from slowdown to speedup.

The Advanced Camera for Surveys, a new imaging instrument installed on the space telescope in 2002, enabled scientists to turn Hubble into a supernova-hunting machine. Riess led an

BRYAN CHRISTIE DESIGN

How Can Gravity Be Repulsive?

IN NEWTON'S THEORY, gravity is always attractive and its strength depends on the mass of the attracting object. The twist in Einstein's theory is that the strength of the gravitational pull exerted by an object also depends on its composition. Physicists characterize the composition of a substance by its internal pressure. An object's gravity is proportional to its energy density plus three times the pressure. Our sun, for example, is a hot sphere of gas with positive (outward) pressure; because gas pressure rises with temperature, the sun's gravitational pull is slightly greater than that of a cold ball of matter of equivalent mass. On the other hand, a gas of photons has a pressure that is equal to one third its energy density, so its gravitational pull should be twice that of an equivalent mass of cold matter.

Dark energy is characterized by negative pressure. (Elastic objects—for instance, a rubber sheet—also have negative, or inward, pressure.) If the pressure falls below $-1/3$ times the energy density, then the combination of energy plus three times the pressure is negative and the gravitational force is repulsive. The quantum vacuum has a pressure that is -1 times its energy density, so the gravity of a vacuum is very repulsive. Other hypothetical forms of dark energy have a pressure that is between $-1/3$ and -1 times its energy density. Some of these types of energy have been invoked to explain the inflationary epoch, a very early period of cosmic acceleration. Other types are candidates for the dark energy powering the acceleration observed today. *—A.G.R. and M.S.T.*

effort to discover the needed sample of very distant type Ia supernovae by piggybacking on the Great Observatories Origins Deep Survey. The team found six supernovae that exploded when the universe was less than half its present size (more than seven billion years ago); together with SN 1997ff, these are the most distant type Ia supernovae ever discovered. The observations confirmed the existence of an early slowdown period and placed the transitional "coasting point" between slowdown and speedup at about five billion years ago [*see left illustration in box on opposite page*]. This finding is consistent with theoretical expectations and thus is reassuring to cosmologists. Cosmic acceleration was a surprise and a new puzzle to solve, but it is not so surprising as to make us rethink much of what we thought we understood about the universe.

THE AUTHORS

ADAM G. RIESS and *MICHAEL S. TURNER* have led the way in exploring the history of the universe's expansion. Riess is an associate astronomer at the Space Telescope Science Institute (the science headquarters for the Hubble Space Telescope) and adjunct associate professor of physics and astronomy at Johns Hopkins University. In 1998 he was lead author of the study published by the High-z Supernova Team announcing the discovery of an accelerating universe. Turner, the Rauner Distinguished Service Professor at the University of Chicago, is now serving as the assistant director for mathematical and physical sciences at the National Science Foundation. His 1995 paper with Lawrence M. Krauss predicted cosmic acceleration, and he coined the term "dark energy."

Our Cosmic Destiny

THE ANCIENT SUPERNOVAE also provided new clues about dark energy, the underlying cause of the cosmic speedup. The leading candidate to explain dark energy's effects is vacuum energy, which is mathematically equivalent to the cosmological constant that Einstein invented in 1917. Because Einstein thought he needed to model a static universe, he introduced his "cosmological fudge factor" to balance the attractive gravity of matter. In this recipe, the constant's density was half that of matter. But to produce the observed acceleration of the universe, the constant's density would have to be twice that of matter.

Where could this energy density come from? The uncertainty principle of quantum mechanics requires that the vacuum be filled with particles living on borrowed time and energy, popping in and out of existence. But when theorists try to compute the energy density associated with the quantum vacuum, they come up with values that are at least 55 orders of magnitude too large. If the vacuum energy density were really that high, all matter in the universe would instantly fly apart and galaxies would never have formed.

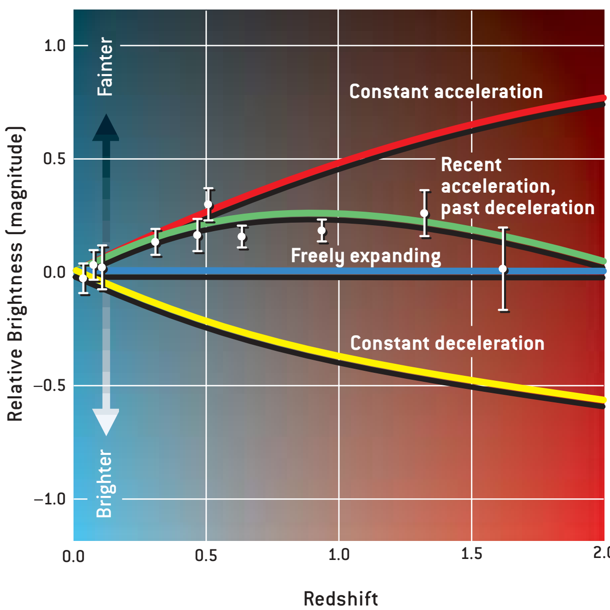
This discrepancy has been called the worst embarrassment in all of theoretical physics, but it may actually be the sign of a great opportunity. Although it is possible that new attempts to estimate the vacuum energy density may yield just the right amount to explain cosmic acceleration, many theorists believe that a correct calculation, incorporating a new symmetry principle, will lead to the conclusion that the energy associated with the quantum vacuum is zero. (Even quantum nothingness weighs nothing!) If this is true, something else must be causing the expansion of the universe to speed up.

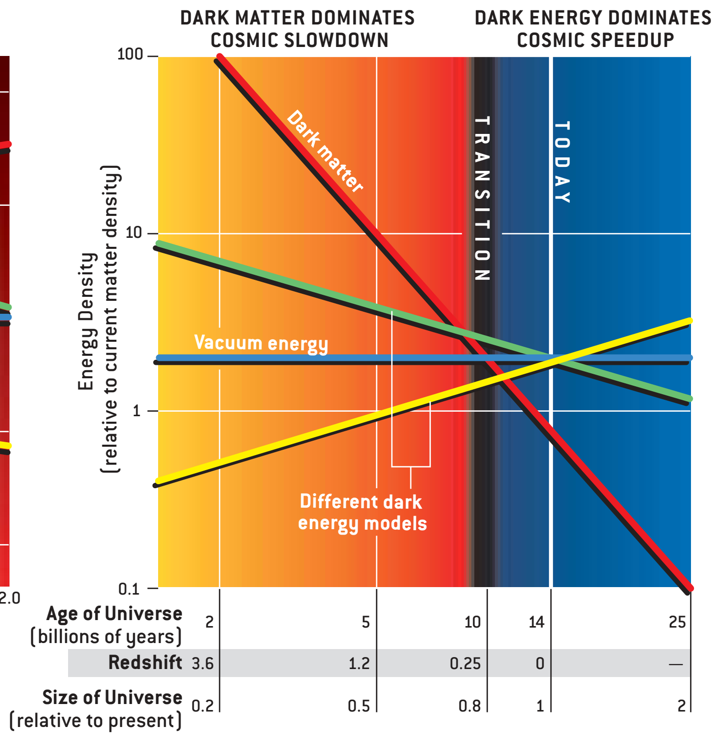
Theorists have proposed a variety of ideas, ranging from the influence of extra, hidden dimensions to the energy associated with a new field of nature, sometimes called quintessence [see "Out of the Darkness," by Georgi Dvali, on page 68]. In general, these hypotheses posit a dark energy density that is not constant and that usually decreases as the universe expands. (But the suggestion that dark energy density is actually increasing as the universe expands has also been put forth.) Perhaps the most radical idea is that there is no dark energy at all but rather that Einstein's theory of gravity must be modified.

Because the way the dark energy density varies is dependent on the theoretical model, each theory predicts a different time for the transition point when the expansion of the universe switched from slowdown to speedup. If the dark energy density decreases as the universe expands, then the switch-over point occurs earlier in time than it would for a model assuming constant dark energy density. Even theoretical models where gravity is modified lead to a discernible signature in the switch-over time. The latest supernova results are consistent with theories positing a constant dark energy density, but they also agree with

THE TRANSITION POINT

RECENT OBSERVATIONS of remote supernovae indicate that the expansion of the universe was decelerating before it began accelerating (*graph at left*). Astronomers found that type Ia supernovae with redshifts greater than 0.6 were brighter than what would be expected if the universe had always been accelerating or if intergalactic dust were dimming their light. (Each plot point is an average of supernovae with nearly the same redshift.) The results show that the transition point between slowdown and speedup occurred about five billion years ago. If astronomers can determine this transition time more precisely, they may learn how the energy density of dark energy has evolved over time and perhaps discover its nature (*right*).





most of the models that assume a varying dark energy density. Only theories stipulating large variations in dark energy density have been ruled out.

To narrow the range of theoretical possibilities, the Hubble Space Telescope is continuing to gather supernova data that could pin down the details of the transition phase. Although the space telescope remains the only means to probe the early history of cosmic expansion, more than half a dozen ground-based programs are trying to improve the precision of the measurement of recent cosmic speedup enough to reveal the physics of dark energy. The most ambitious project is the Joint Dark Energy Mission (JDEM) proposed by the U.S. Department of Energy and NASA. JDEM is a two-meter, wide-field space telescope dedicated to discovering and accurately measuring thousands of type Ia supernovae. Supernova hunters hope to see JDEM launched at the start of the next decade; until then, they will have to rely on the Hubble telescope to detect the most distant stellar explosions.

Solving the mystery of cosmic acceleration will reveal the destiny of our universe. If the dark energy density is constant or increasing with time, in 100 billion years or so all but a few hundred galaxies will be far too redshifted to be seen. But if the dark energy density decreases and matter becomes dominant again, our cosmic horizon will grow, revealing more of the universe. Even more extreme (and lethal) futures are possible. If dark energy density rises rather than falls, the universe will eventually undergo a "hyper speedup" that would tear apart galaxies, solar systems, planets and atomic nuclei, in that order. Or the universe might even recollapse if dark energy density falls to a negative value. The only way to forecast our cosmic future is to figure out the nature of dark energy. SA

MORE TO EXPLORE

Do Type Ia Supernovae Provide Direct Evidence for Past Deceleration in the Universe? Michael S. Turner and Adam G. Riess in *Astrophysical Journal,* Vol. 569, Part 1, pages 18–22; April 10, 2002. Available online at **arXiv.org/abs/astro-ph/0106051**

The Extravagant Universe: Exploding Stars, Dark Energy and the Accelerating Cosmos. Robert P. Kirshner. Princeton University Press, 2002.

Connecting Quarks with the Cosmos. Committee on the Physics of the Universe, National Research Council. National Academies Press, 2003.

Is Cosmic Speed-Up Due to New Gravitational Physics? Sean M. Carroll, Vikram Duvvuri, Mark Trodden and Michael S. Turner in *Physical Review Letters* (in press). **arXiv.org/abs/astro-ph/0306438**

SPECIAL REPORT

OUT OF THE DARKNESS

By Georgi Dvali

Maybe cosmic acceleration isn't caused by dark energy after all but by an inexorable leakage of gravity out of our world

Cosmologists and particle physicists have seldom felt so confused. Although our standard model of cosmology has been confirmed by recent observations, it still has a gaping hole: nobody knows why the expansion of the universe is accelerating. If you throw a stone straight up, the pull of Earth's gravity will cause it to slow down; it will not accelerate away from the planet. Similarly, distant galaxies, thrown apart by the big bang expansion, should pull on one another and slow down. Yet they are accelerating apart. Researchers commonly attribute the acceleration to some mysterious entity called dark energy, but there is little physics to back up these fine words. The only thing that is becoming clear is that at the largest observable distances, gravity behaves in a rather strange way, turning into a repulsive force.

The laws of physics say that gravity is generated by matter and energy, so they attribute a strange sort of gravity to a strange sort of matter or energy. That is the rationale for dark energy. But maybe the laws themselves need to be changed. Physicists have a precedent for such a change: the law of gravity that Newton formulated in the 17th century, which had various conceptual and experimental limitations, gave way to Einstein's general theory of relativity in 1915. Relativity, too, has limitations; in particular, it runs into trouble when applied to extremely short distances, which are the domain of quantum mechanics. Much as relativity subsumed Newtonian physics, a quantum theory of gravity will ultimately subsume relativity.

Over the years, physicists have come up with a few plausible approaches to quantum gravity, the most prominent being string theory. When gravity operates over microscopic distances—for instance, at the center of a black hole, where a huge mass is packed into a subatomic volume—the bizarre quantum properties of matter come into play, and string theory describes how the law of gravity changes.

Over greater distances, string theorists have generally assumed that quantum effects are unimportant. Yet the cosmological discoveries of the past several years have encouraged researchers to reconsider. Four years ago my colleagues and I asked whether string theory would change the law of gravity not just on the smallest scales but also on the largest ones. The feature of string theory that could bring about this revision is its extra dimensions—additional directions in which particles can roam. The theory adds six or seven dimensions to the usual three.

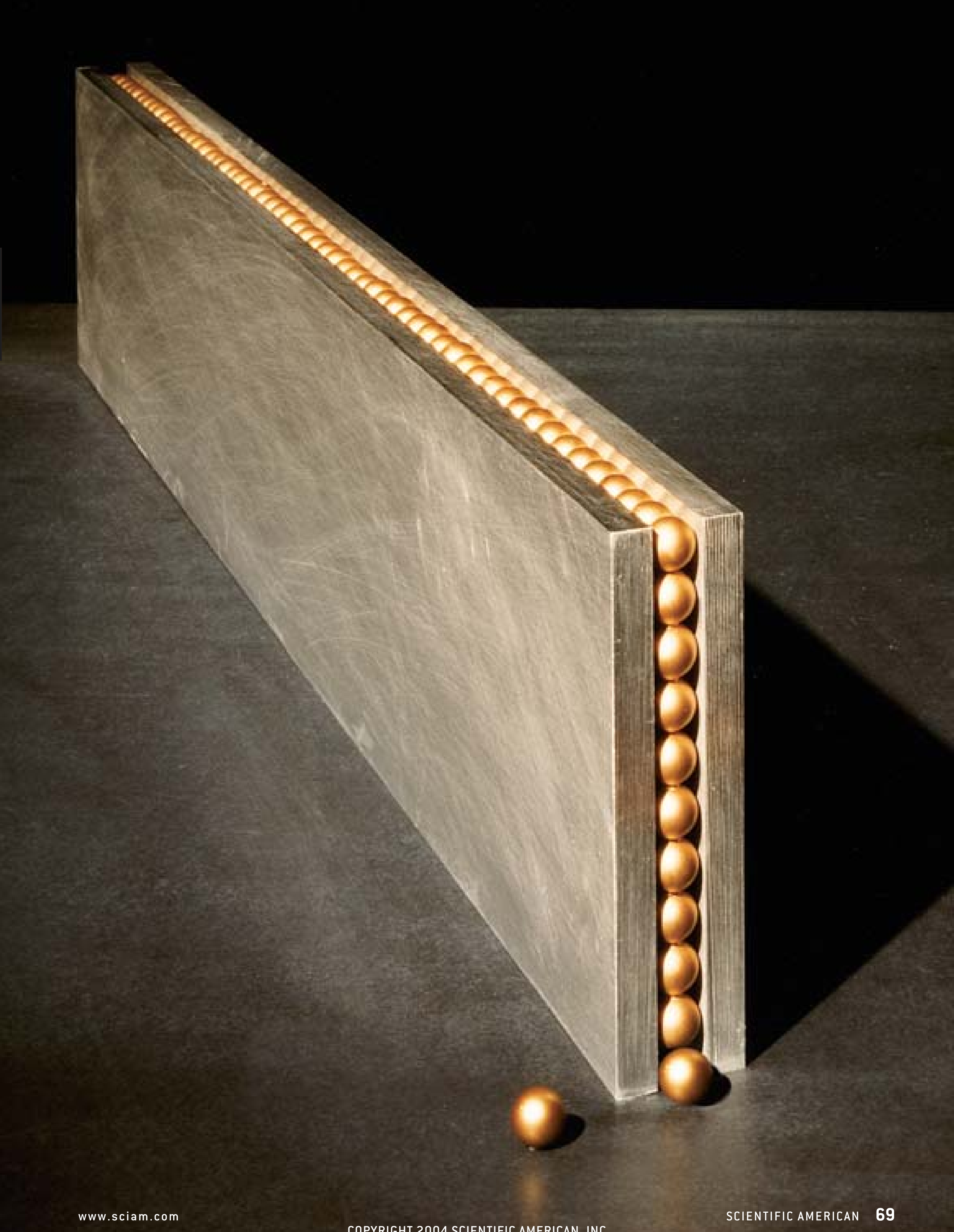
In the past, string theorists have argued that the extra dimensions are too small for us to see or move in. But recent progress reveals that some or all of the new dimensions could actually be infinite in size. They are hidden from view not because they are small but because the particles that make up our bodies are trapped in three dimensions. The one particle that eludes confinement is the particle that transmits the force of gravity, and as a result, the law of gravity changes.

Quintessence Even from Nothingness

WHEN ASTRONOMERS ENCOUNTERED the cosmic acceleration, their first reaction was to attribute it to the so-called cosmological constant. Notoriously introduced and then retracted by Einstein, the constant represents the energy inherent in space

LEAKING OUT of our universe, particles of gravity could explore a higher-dimensional space. The leakage becomes apparent only on cosmic scales.

DAVID EMMITE PHOTOGRAPHY



itself. A completely empty volume of space, devoid of all matter, would still contain this energy—equivalent to roughly 10^{-26} kilogram per cubic meter. Although the cosmological constant is consistent with all the existing data so far, many physicists find it unsatisfying. The problem is its inexplicable smallness, so small that it had little effect for most of cosmic history, including the formative early period of the universe. Worse, it is much smaller than the energy scales of the physical processes that would create it [see "From Slowdown to Speedup," by Adam G. Riess and Michael S. Turner, on page 62].

To get around this problem, a number of physicists have proposed that the acceleration is caused not by space itself but by an energy field that suffuses space like a thin fog. The potential energy of certain spatially uniform fields can act much like a cosmological constant. One such field, known as the inflaton, is thought to have driven a period of accelerated expansion, or inflation, in the early universe. Perhaps another such field has reared its head, driving the universe into another inflationary period. This second field goes under the name of quintessence. Like the cosmological constant, it must have a bizarrely small value, but proponents argue that it should be easier for a dynamic entity to settle into such a value than for a static constant to do so [see "The Quintessential Universe," by Jeremiah P. Ostriker and Paul J. Steinhardt; Scientific American, January 2001].

Both the cosmological constant and quintessence fall into the general category of dark energy. So far a compelling explanation for either one remains absent, which is why physicists are thinking seriously about higher-dimensional theories. The appeal of additional dimensions is that they would automatically alter how gravity behaves. When gravity operates according to the rules of either Newton's theory or general relativity, its strength falls off with the square of the distance between objects. The reason is simple geometry: according to a principle formulated by 19th-century physicist Carl Friedrich Gauss, the strength of gravity is determined by the density of lines of gravitational force, and as the distance increases these lines are spread out over an ever larger boundary. In three-dimensional space, the boundary is a two-dimensional surface—that is, an area, the size of which increases as the square of the distance.

But if space were four-dimensional, the boundary would be three-dimensional—a volume, whose size increases as the cube of the distance. In this case, the density of lines of force would decrease with the cube of the distance. Gravity would thus be weaker than in a three-dimensional world. On cosmological scales, the weakening of gravity can lead to cosmic acceleration, for reasons I will discuss later.

If gravity is free to move into the extra space, why have we not noticed it before? Why does the standard three-dimensional inverse-square law explain the motions of baseballs, rockets and planets so precisely? The traditional answer in string theory is that the additional dimensions are compact—curled up into finite, minuscule circles. For a long time, the size of these circles was assumed to be the so-called Planck length, about 10^{-35} meter, but recent theoretical and experimental work shows they could be as big as 0.2 millimeter ["The Universe's Unseen Dimensions," by Nima Arkani-Hamed, Savas Dimopoulos and Georgi Dvali; Scientific American, August 2000]. If the dimensions are curled up, they interfere with the workings of gravity only over short distances—comparable to or smaller than the radius of the compact dimensions. Over larger distances, the standard law of gravity holds.

Overview/*Gravitational Leaks*

- Astronomers typically ascribe the accelerated expansion of the universe to a shadowy dark energy. It might, however, be a sign that the standard laws of physics break down on the largest scales.
- A new law of gravity emerges from string theory, one of the leading efforts to prepare an ultimate unified theory of nature. String theory is usually considered a theory of the very small, but it can also have macroscopic consequences.
- In particular, the theory predicts that the universe has extra dimensions into which gravity, unlike ordinary matter, may be able to escape. This leakage would warp the spacetime continuum and cause cosmic expansion to accelerate. It might even have a minute but observable effect on planetary motion.

Prison Life

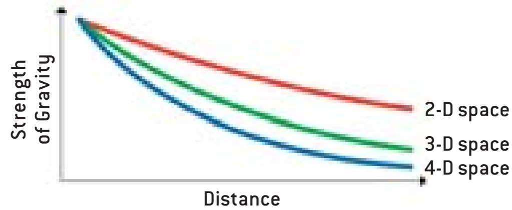
The idea of compact dimensions has its difficulties, however. One could ask, for example, why some dimensions (the extra ones) are tightly knotted, whereas others (the familiar ones) go on forever? To put it a different way, under the influence of the matter and energy in the universe, the curled-up dimensions should uncurl, unless something stabilizes them. One interesting possibility is that magneticlike fields predicted by string theory prevent the dimensions from either shrinking or expanding. Another potential solution emerged in 1999. Maybe all the dimensions, even the extra ones, are infinite in size. The observable universe is a three-dimensional surface, or membrane ("brane" for short), in a higher-dimensional world. Ordinary matter is confined to the brane, but some forces, such as gravity, can escape.

Gravity has this Houdini-like ability because it is fundamentally unlike other forces. According to quantum field theory, the force of gravity is carried by a special particle called the graviton. Gravitational attraction results from a flow of gravitons between two bodies, much as the force of electricity or magnetism results from a flow of photons between two charged particles. When gravity is static, these gravitons are "virtual"—although their effects can be measured, they cannot be observed as independent particles. The sun holds Earth in orbit because it emits virtual gravitons that our planet absorbs. "Real," or directly observable, gravitons correspond to the gravitational waves that are given off under certain circumstances [see "Ripples in Spacetime," by W. Wayt Gibbs; Scientific American, April 2002].

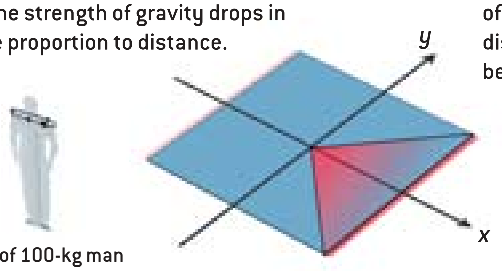
As conceived by string theory, gravitons, like all particles,

FROM FLATLAND TO FOUR DIMENSIONS

A FAMOUS POSTER by artist Gerry Mooney proclaims: "Gravity: It isn't just a good idea. It's the law." Yet the law is actually rather flexible. For instance, it depends on the number of spatial dimensions. The key is that gravity weakens with distance because, as it propagates, it gets spread out over an ever larger boundary (*red in diagrams below*).



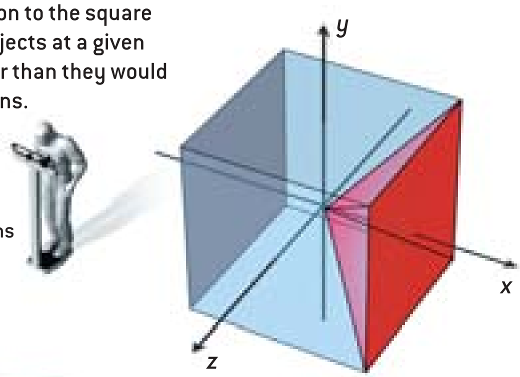
TWO DIMENSIONS: The boundary is one-dimensional (a line) and grows in direct proportion to the propagation distance. Thus, the strength of gravity drops in inverse proportion to distance.



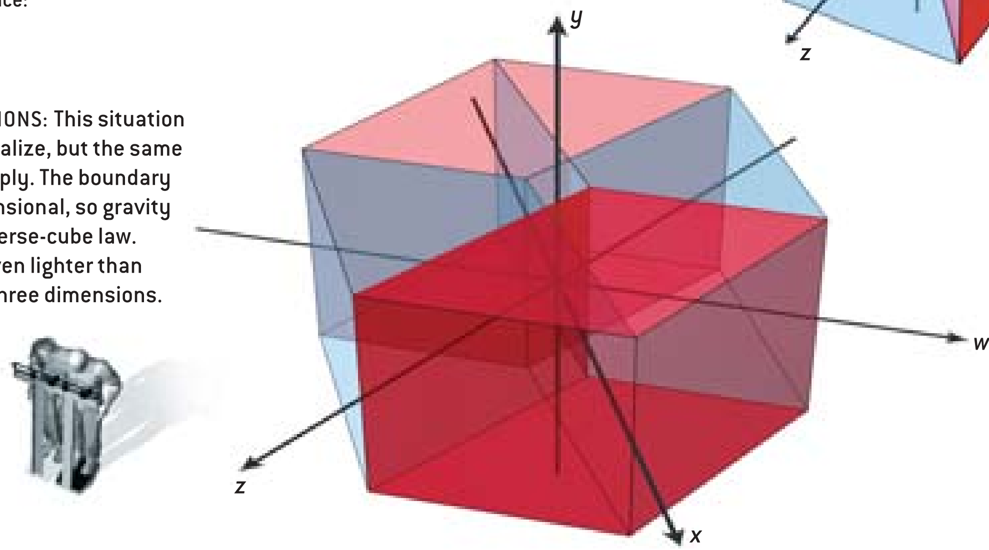
Weight of 100-kg man on Earth's surface: 10^{45} newtons

THREE DIMENSIONS: The boundary is two-dimensional, so gravity attenuates in inverse proportion to the square of the distance. Objects at a given distance are lighter than they would be in two dimensions.

Weight: 10^3 newtons



FOUR DIMENSIONS: This situation is hard to visualize, but the same basic rules apply. The boundary is three-dimensional, so gravity follows an inverse-cube law. Objects are even lighter than they were in three dimensions.



Weight: 10^{-39} newton

BRYAN CHRISTIE DESIGN; SOURCE: JEFFREY WEEKS (*4-D cube*)

are ultimately the vibrations of tiny strings. But whereas the electron, proton and photon are vibrations of open-ended strings, like violin strings, the graviton is the vibration of a closed loop, like a rubber band. Joseph Polchinski of the Kavli Institute for Theoretical Physics in Santa Barbara has shown that the ends of open strings cannot flap around; they must be tied down to a brane. If you tried to pull an open string out of a brane, it would get longer, like an elastic cord, but remain attached to the brane. In contrast, closed strings such as gravitons cannot get stuck. They are free to explore the full 10-dimensional space.

To be sure, gravitons cannot have absolute freedom. If they did, the standard law of gravity would fail conspicuously. The authors of the infinite-dimensions hypothesis, Lisa Randall of Harvard University and Raman Sundrum of Johns Hopkins University, suggested that gravitons are hindered because the extra dimensions, unlike our familiar three, are very strongly curved—creating a steep-walled valley that is hard to leave.

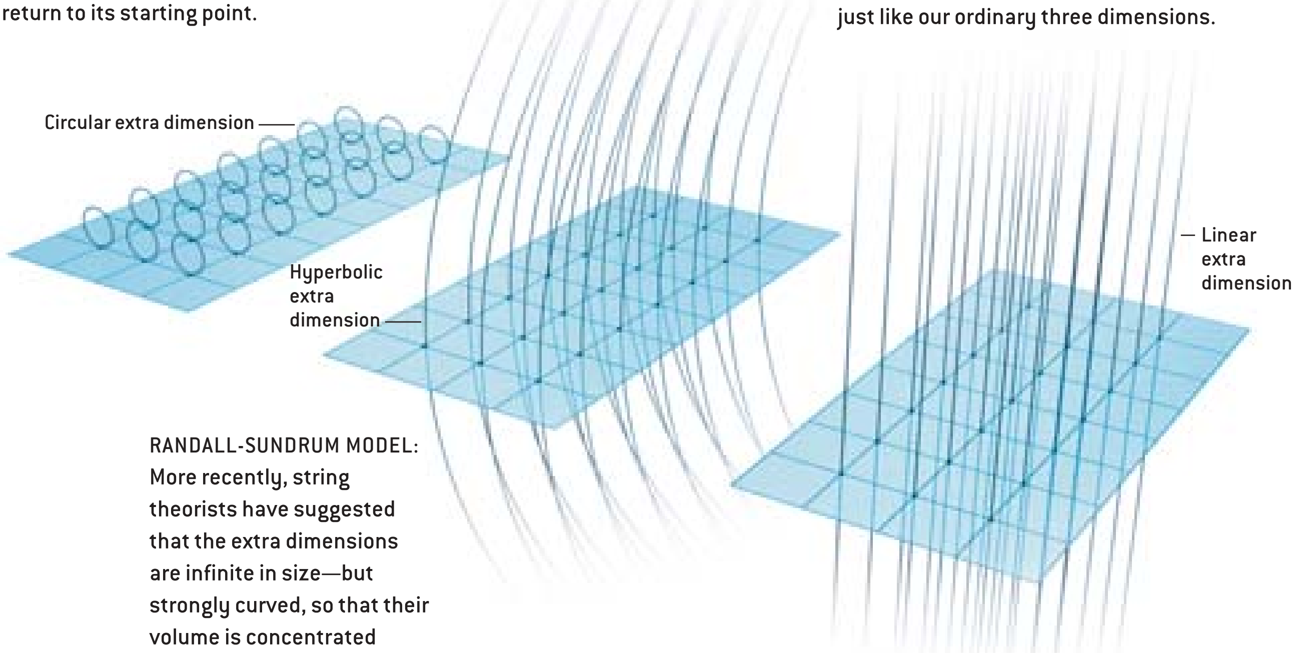
The trick is, because the extra dimensions are so strongly curved, their volume is effectively finite, even though they are infinite in extent. How can an infinite space have a finite volume? Imagine pouring gin into a bottomless martini glass whose radius shrinks in inverse proportion to its depth. To fill the glass, a finite amount of gin would suffice. Because of the curvature of the glass, its volume is concentrated near the top. This is very similar to what happens in the Randall-Sundrum scenario. The volume of the extra space is concentrated around our brane. Consequently, a graviton is forced to spend most of its time near the brane. The probability of detecting the graviton quickly diminishes as a function of distance. In quantum jargon, the wave function of the graviton is peaked

THREE WAYS TO ADD A DIMENSION

ALBERT EINSTEIN and other scientists of his generation, notably Theodor Kaluza and Oskar Klein, were enamored of the idea that space has hidden dimensions. This hypothesis lives on in string theory. For the sake of clarity, think of our three-dimensional universe as a flat grid. At each grid point is a line that represents one of the extra dimensions.

TRADITIONAL STRING THEORY: String theorists long assumed that the extra dimensions were finite in size—small circles of sub-subatomic size. Moving in this dimension, a tiny creature would eventually return to its starting point.

INFINITE-VOLUME MODEL: The author and his colleagues have proposed that the extra dimensions are infinite in size and uncurved, just like our ordinary three dimensions.



RANDALL-SUNDRUM MODEL: More recently, string theorists have suggested that the extra dimensions are infinite in size—but strongly curved, so that their volume is concentrated around our universe.

at the brane—an effect referred to as localization of gravity.

Though conceptually different from the idea of compact dimensions, the Randall-Sundrum scenario has much the same outcome. Both models modify the law of gravity on short distances but not on large distances, so neither bears on the problem of cosmic acceleration.

Physics on the Brane

A THIRD APPROACH, though, does predict the breakdown of the standard laws of gravity on cosmological scales and explains acceleration without having to invoke dark energy. In 2000 Gregory Gabadadze and Massimo Porrati, both now at New York University, and I proposed that the extra dimensions are exactly like the three dimensions that we see around us. They are neither compact nor strongly curved.

Even so, gravitons are not completely free to go where they like. Emitted by stars and other objects located on the brane, they can escape into the extra dimensions, but only if they travel a certain critical distance. The gravitons behave like sound in a metal sheet. Hitting the sheet with a hammer creates a sound wave that travels along its surface. But the sound propagation is not exactly two-dimensional; part of the energy is lost into the surrounding air. Near the location of the hammer blow, this energy loss is negligible. Farther away, however, it becomes noticeable.

This leakage has a profound effect on the gravitational force between objects separated by more than the critical distance. Virtual gravitons exploit every possible route between the objects, and the leakage opens up a huge number of multidimensional detours, which bring about a change in the law of gravity. Real gravitons that leak away are simply lost forever, and for those of us stuck on the brane, it looks as though they have disappeared into thin air.

The extra dimensions also reveal themselves on very small scales, just as in the compact and Randall-Sundrum scenarios. Over intermediate distances—larger than the size of the strings but smaller than the leakage distance—gravitons are three-dimensional and closely obey the conventional law of gravity.

BRYAN CHRISTIE DESIGN

The key to this scenario is the brane itself. It is a material object in its own right, and gravity spreads through it differently than through the surrounding space. The reason is that ordinary particles such as electrons and protons can exist on the brane and only on the brane. Even a seemingly empty brane contains a seething mass of virtual electrons, protons and other particles, continuously created and destroyed by quantum fluctuations. Those particles both generate and respond to gravity. The surrounding space, in contrast, is truly empty. Gravitons can flutter through it but have nothing to act on except one another.

An analogy is a dielectric material, such as plastic, ceramic or pure water. The material, unlike a vacuum, contains electrically charged particles and can respond to an electric field. Although charged particles cannot flow through a dielectric (as they can through an electrical conductor), they can still redistribute themselves within it. If you apply an electric field, the material becomes electrically polarized. In water, for example, the molecules rotate so that their positive ends (the two hydrogen atoms) point in one direction and their negative ends (the oxygen atom) point in the opposite direction. In sodium chloride, the positive sodium ions and negative chloride ions move slightly apart.

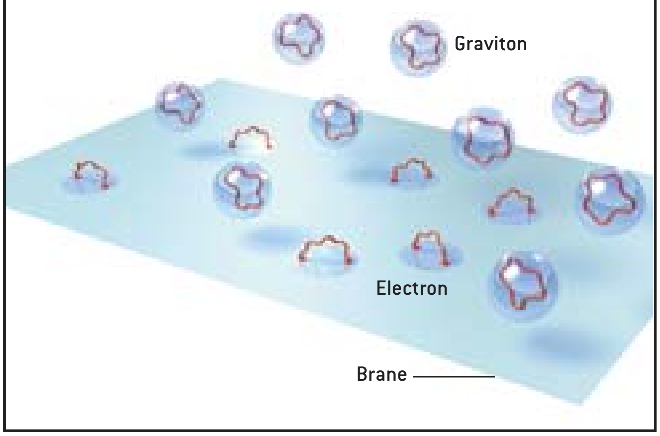
The redistributed charges set up an electric field of their own, which partially cancels the external field. A dielectric can thus affect the propagation of photons, which are nothing more than oscillating electric and magnetic fields. Photons penetrating into a dielectric polarize it and, in turn, are partially canceled out. To bring about this effect, a photon must have a wavelength in a certain range: long-wavelength (low-momentum) photons are too weak to polarize a dielectric, and short-wavelength (high-momentum) photons oscillate too quickly for the charged particles to respond. For this reason, water is transparent to radio waves (which have a long wavelength) and to visible light (short wavelength) but opaque to microwaves (intermediate wavelength). Microwave ovens rely on this effect.

Similarly, quantum fluctuations convert the brane into the gravitational equivalent of a dielectric. It is as if the brane is populated by positive-energy and negative-energy virtual particles. If you apply an external gravitational field, the brane becomes gravitationally polarized. Positive-energy particles move slightly away from negative-energy ones. A graviton, which embodies an oscillating gravitational field, can polarize the brane and get canceled out if its wavelength falls into the right range—which, we calculate, lies between 0.1 millimeter (or smaller, depending on the number of extra dimensions) and approximately 10 billion light-years.

This cancellation affects only gravitons traveling into or out of the brane. Gravitons, like photons, are transverse waves: they oscillate perpendicular to their direction of propagation. A graviton entering or exiting the brane tends to push particles along the brane, a direction in which the particles are able to move. Thus, these gravitons can polarize the brane and, in turn, get cancelled out. But gravitons moving along the brane try to push particles out of the brane, a direction in which they cannot go. Therefore these gravitons do not polarize the brane.

THE SURLY BONDS OF BRANES

SADLY, even if extra dimensions exist, humans will never be able to venture into them. Particles in our bodies—electrons, protons, neutrons—are thought to be the vibrational patterns of open-ended strings. By their very nature, they are tied down to the membrane, or brane, that constitutes our universe. Gravitons, the particles of gravitational force, elude these bonds because they have no ends to tie down.



They move without encountering resistance. In practice, most gravitons fall between these two extremes. They zip through space at an oblique angle to the brane and may cover billions of light-years before getting canceled out.

Brane (and Brain) Bending

IN THIS WAY, THE BRANE shields itself from the extra dimensions. If an intermediate-wavelength graviton attempts to escape from or penetrate into the brane, particles within the brane redistribute themselves and block it. The gravitons must instead move along the brane, so gravity follows an inverse-square law. Long-wavelength gravitons, however, are free to pass through the extra dimensions. These gravitons are insignificant over short distances but dominate on distances comparable to their wavelength, and they undermine the brane's ability to isolate itself from the extra dimensions. The law of gravity approaches an inverse-cube law (if only one of the extra dimensions is infinite), an inverse-fourth-power law (if two are infinite) or an even steeper law. In all these cases, gravity is weakened.

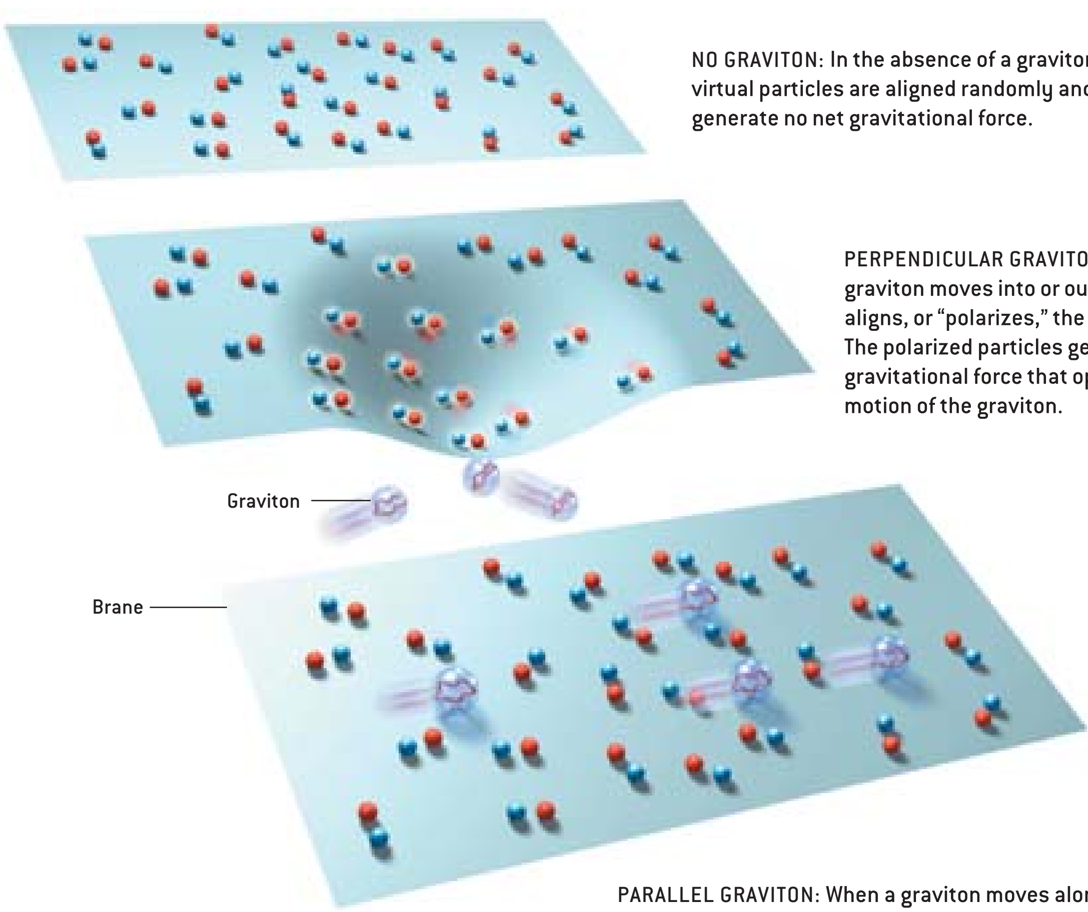
Cédric Deffayet, now at the Paris Institute of Astrophysics, Gabadadze and I have found that the extra dimensions not only sap the strength of gravity but also force cosmic expansion to

THE AUTHOR

GEORGI DVALI grew up in the former Soviet republic of Georgia and received his Ph.D. from the Andronikashvili Institute of Physics in Tbilisi. After working at the University of Pisa in Italy, at CERN near Geneva and at the International Center for Theoretical Physics in Trieste, he joined the physics faculty of New York University. He enjoys overcoming gravity by mountain hiking, as well as taking advantage of this mysterious force by downhill skiing.

THE POLARIZED BRANE

GRAVITONS do not have untrammeled freedom to traipse through extra dimensions. Our three-dimensional universe, or brane (shown here as a flat sheet), is filled with "virtual" particles that bubble in and out of existence. One way to model their effect on gravitons is to think of them as coming in pairs. One particle in each pair is endowed with positive energy (*blue*), the other with negative energy (*red*). Such pairs can block gravitons from entering or exiting the brane.



NO GRAVITON: In the absence of a graviton, the virtual particles are aligned randomly and generate no net gravitational force.

PERPENDICULAR GRAVITON: When a graviton moves into or out of the brane, it aligns, or "polarizes," the virtual particles. The polarized particles generate a gravitational force that opposes the motion of the graviton.

PARALLEL GRAVITON: When a graviton moves along the brane, it has no effect on the virtual particles, because the forces it exerts act at right angles to the brane. The virtual particles, in turn, do not impede the graviton.

accelerate without any need to stipulate the existence of dark energy. It is tempting to say that by weakening the gravitational glue that retards expansion, graviton leakage reduces the deceleration, so much so that the deceleration becomes negative—that is, an acceleration. But the effect is more subtle. It has to do with how leakage alters general relativity.

The central idea of Einstein's theory is that gravitation is a consequence of the curvature of spacetime, which is related to the density of matter and energy within it. The sun attracts Earth by warping the spacetime around it. No matter and no energy mean no warping and no gravity. In the higher-dimensional theory, however, the relation between curvature and density changes. The extra dimensions introduce a correction term into the equations, which ensures that the curvature of an empty brane is not zero. In effect, graviton leakage puts tension on the brane, giving it an irreducible warp that does not depend on the density of matter and energy within it.

Over time, as matter and energy get diluted, the curvature that they cause decreases, and so the irreducible warp becomes increasingly important. The curvature of the universe approaches a constant value. The same effect would be brought about if the universe were filled with a substance that did not get diluted over time. Such a substance is none other than a cosmological constant. Therefore, the irreducible warp of the brane acts like a cosmological constant, which speeds up cosmic expansion.

BRYAN CHRISTIE DESIGN

Scofflaws

OUR THEORY IS NOT the only one that postulates the breakdown of the standard gravitational law at large distances. In 2002 Thibault Damour and Antonios Papazoglou of the Institute for Higher Scientific Studies in France and Ian Kogan of the University of Oxford suggested that gravitons come in an additional variety—one that, unlike normal gravitons, possesses a small mass. As physicists have long known, if gravitons have mass, gravity does not obey an inverse-square law. They are unstable and gradually decay, with much the same effect as graviton leakage: gravitons traveling for long distances vanish, gravity gets weaker, and cosmic expansion accelerates. Sean Carroll, Vikram Duvvuri and Michael Turner of the University of Chicago and Mark Trodden of Syracuse University have modified Einstein's theory in three dimensions by introducing tiny terms that are inversely proportional to spacetime curvature. Such terms would be negligible in the early universe but would speed expansion later on. Other research teams have also suggested modifying the law of gravity, but their proposals do not eliminate the need for dark energy to cause acceleration.

Observations will be the final arbiter of all these models. Supernova surveys provide one direct test. The transition from deceleration to acceleration is very different in a leakage scenario than in other dark-energy scenarios. Further improvements in the precision of these surveys could differentiate among the theories.

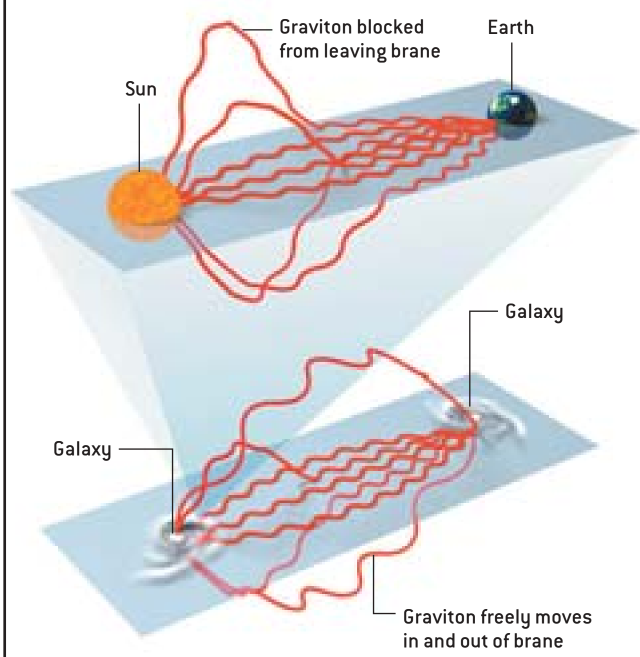
Planetary motion offers another empirical test. A gravitational wave, just like an ordinary electromagnetic wave, can have a preferred direction of oscillation. General relativity permits two such directions, but alternative theories of gravity allow for more. These additional possibilities modify the gravitational force in a small but nonnegligible way, yielding potentially observable corrections to planetary motion. Andrei Gruzinov and Matias Zaldarriaga of New York University and I have calculated that graviton leakage would cause the moon's orbit to precess slowly. Every time the moon completed one orbit, its closest approach to Earth would shift by about a trillionth of a degree, or about half a millimeter. This motion is almost large enough to be seen by lunar-ranging experiments, which monitor the moon's orbit by bouncing laser beams off mirrors left on the lunar surface by the Apollo astronauts. Current ranging measurements have a precision of one centimeter, and Eric Adelberger and his colleagues at the University of Washington propose using more powerful lasers to improve the sensitivity 10-fold. Spacecraft tracking could look for a similar precession of Mars's orbit.

The mere fact that observers are talking about probing string theory is exciting. For years, the theory was assumed to be a theory of the very small—so small that no experiment could ever prove or disprove it. Cosmic acceleration may be a rear window of opportunity, a gift from nature, that lets us peer into the extra dimensions that are otherwise invisible to us. It may be a bridge between the very small and the ultralarge. The fate of the universe may be hanging from a string. SA

GRAVITY NEAR AND FAR

PARTICLES IN OUR UNIVERSE tend to block out gravitons, but only if the gravitons have enough momentum to provoke a response. Low-momentum gravitons (which have a long wavelength) enter or exit the brane at will.

The sun exerts a force on Earth by emitting virtual gravitons. These gravitons have a relatively short wavelength (high momentum), so they are blocked from leaving the brane. They behave as if the extra dimensions did not exist.



Two distant galaxies emit gravitons with a long wavelength (low momentum). These gravitons are not blocked from escaping into the extra dimensions. The law of gravity changes, weakening the force between the galaxies.

BRYAN CHRISTIE DESIGN

MORE TO EXPLORE

The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory. Brian Greene. W. W. Norton, 2003.

An Alternative to Compactification. Lisa Randall and Raman Sundrum in *Physical Review Letters,* Vol. 83, No. 23, page 4690–4693; December 6, 1999. Available online at **arXiv.org/abs/hep-th/9906064**

Accelerated Universe from Gravity Leaking to Extra Dimensions. Cédric Deffayet, Gia Dvali and Gregory Gabadadze in *Physical Review D,* Vol. 65, paper number 044023; 2002. **arXiv.org/abs/astro-ph/0105068**

The Accelerated Universe and the Moon. Gia Dvali, Andrei Gruzinov and Matias Zaldarriaga in *Physical Review D,* Vol. 68, paper number 024012; 2003. **arXiv.org/abs/hep-ph/0212069**

Tests of the Gravitational Inverse-Square Law. E. G. Adelberger, B. R. Heckel and A. E. Nelson in *Annual Review of Nuclear and Particle Science,* Vol. 53, pages 77–121; December 2003. **arXiv.org/abs/hep-ph/0307284**

An introduction to string theory can be found at **superstringtheory.com**

Light-emitting organic materials offer brighter and more efficient displays than LEDs. And you'll be able to unroll them across a tabletop

By Webster E. Howard

BETTER DISPLAYS with ORGANIC FILMS

THE FIRST active-matrix OLED display on the market provides a 2.2-inch screen for the Kodak EasyShare LS633 digital camera.

EASTMAN KODAK COMPANY

Before the videocassette recorder there was the movie projector and screen. Perhaps you remember your fifth-grade teacher pulling down a screen—or Dad hanging a sheet on the wall, ready to show visiting friends the enthralling account of your summer vacation at the shore. Just as the film got started, the projector bulb often blew out.

Those days did have one advantage, though: the screen was light, paper-thin and could be rolled into a portable tube. Compare that with bulky television and computer screens, and the projector screen invokes more than just nostalgia. Could yesterday's convenience be married to today's technology?

The answer is yes, thanks to organic light-emitting materials that promise to make electronic viewing more convenient and ubiquitous. Used in displays, the organic materials are brighter, consume less energy and are easier to manufacture (thus potentially cheaper) than current options based on liquid crystals. Because organic light-emitting diodes (OLEDs) emit light, they consume significantly less power, especially in small sizes, than common liquid-crystal displays (LCDs), which require backlighting. OLEDs also offer several exciting advantages over common LEDs: the materials do not need to be crystalline (that is, composed of a precisely repeating pattern of planes of atoms), so they are easier to make; they are applied in thin layers for a slimmer profile; and different materials (for different colors) can be patterned on a given substrate to make high-resolution images. The substrates may be inexpensive glass or flexible plastic or even metal foil.

In the coming years, large-screen televisions and computer monitors could roll up for storage. A soldier might unfurl a sheet of plastic showing a real-time situation map. Smaller displays could be wrapped around a person's forearm or

FLEXIBLE VIDEO SCREEN prototype from Universal Display Corporation gives new meaning to the words "motion picture."

UNIVERSAL DISPLAY CORPORATION

incorporated into clothing. Used in lighting fixtures, the panels could curl around an architectural column or lie almost wallpaperlike against a wall or ceiling.

LEDs currently have longer lifetimes than organic emitters, and it will be tough to beat the widespread LED for use in indicator lamps. But OLEDs are already demonstrating their potential for displays. Their screens put out more than 100 candelas per square meter (about the luminance of a notebook screen) and last tens of thousands of hours (several years of regular use) before they dim to half their original radiance.

Close to 100 companies are developing applications for the technology, focusing on small, low-power displays [*see box on page 80*]. Initial products include a nonflexible 2.2-inch (diagonal) display for digital cameras and cellular phones made jointly by Kodak and Sanyo, introduced in 2002, and a 15-inch prototype computer monitor produced by the same collaborative venture. The global market for organic display devices was about $219 million in 2003 and is projected to jump to $3.1 billion by 2009, according to Kimberly Allen of iSuppli/Stanford Resources, a market-research firm specializing in displays.

What LED to OLED

CRYSTALLINE semiconductors—the forerunners of OLEDs—trace their roots back to the development of the transistor in 1947, and visible-light LEDs were invented in 1962 by Nick Holonyak, Jr. They were first used commercially as tiny sources of red light in calculators and watches and soon after also appeared as durable indicator lights of red, green or yellow. (When suitably constructed, LEDs form lasers, which have spawned the optical-fiber revolution, as well as optical data storage on compact discs and digital video discs.) Since the advent of the blue LED in the 1990s [see "Blue Chip," by Glenn Zorpette; SCIENTIFIC AMERICAN, August 2000], full-color, large-screen television displays made from hundreds of thousands of LED chips have appeared in spectacular fashion on skyscrapers and in arenas [see "In Pursuit of the Ultimate Lamp," by M. George Crawford, Nick Holonyak, Jr., and Frederick A. Kish, Jr.; SCIENTIFIC AMERICAN, February 2001]. Yet the smaller sizes used in devices such as PDAs (personal digital assistants) and laptops are not as practical.

LEDs and OLEDs are made from layers of semiconductors—materials whose electrical performance is midway between an excellent conductor such as copper and an insulator such as rubber. Semiconducting materials, such as silicon, have a small energy gap between electrons that are bound and those that are free to move around and conduct electricity. Given sufficient energy in the form of an applied voltage, electrons can "jump" the gap and begin moving, constituting an electrical charge. A semiconductor can be made conductive by doping it; if the atoms added to a layer have a smaller number of electrons than the atoms they replace, electrons have effectively been removed, leaving positively charged "holes" and making the material "*p*-type." Alternatively, a layer that is doped so that it has an excess of negatively charged electrons becomes "*n*-type" [*see box on opposite page*]. When an electron is added to a *p*-type material, it may encounter a hole and drop into the lower band, giving up an amount of energy (equal to the energy gap) as a photon of light. The wavelength depends on the energy gap of the emitting material.

For the production of visible light, organic materials should have an energy gap between their lower and higher conduction bands in a relatively small range, about two to three electron volts. (One electron volt is defined as the kinetic energy gained by an electron when it is accelerated by a potential difference of one volt. A photon with one electron volt of energy corresponds to the infrared wavelength of 1,240 nanometers, and a photon of two electron volts has a wavelength half as much—620 nanometers—a reddish color.)

A Surprising Glow

ORGANIC semiconductors are formed as aggregates of molecules that are, in the technologies being pursued, amorphous—a solid material, but one that is noncrystalline and without a definite order. There are two general types of organic light emitters, distinguished by "small" and "large" molecule sizes. The first practical *p-n*-type organic LED, based on small molecules, was invented in 1987 by Ching W. Tang and Steven A. Van Slyke of Eastman Kodak, after Tang noticed a surprising green glow coming from an organic solar cell he was working on. The duo recognized that by using two organic materials, one a good conductor of holes and the other a good conductor of electrons, they could ensure that photon emission would take place near the contact area, or junction, of the two materials, as in a crystalline LED. They also needed a material that held its electrons tightly, meaning that it would be easy to inject holes. For the light to escape, one of the contacts must be transparent, and the scientists benefited from the fortunate fact that the most widely used transparent conducting material, indium tin oxide, bound its electrons suitably for *p*-type contact material.

WEARABLE ELECTRONICS could be fashioned with flexible OLED displays, such as shown in this prototype by Universal Display Corporation.

The structure they came up with has not changed much over the years and is often called "Kodak-type," because Kodak had the basic patent [*see box on opposite page*]. Beginning with a glass substrate, different materials are deposited layer by layer. This process is accomplished by evaporating the constituent materials and letting them condense on the substrate. The total thickness of the organic layers is only 100 to 150 nanometers, much thinner than that of a conventional LED (which is at least microns in thickness) and less than 1 percent of the thickness of a human hair. Because the molecules of the materials used are relatively lightweight—even lighter than a small protein—the Kodak-type OLEDs are referred to as "small molecule" OLEDs.

After their initial insight, Tang and Van Slyke tinkered with the design to increase efficiency. They added a small amount of the fluorescent dye coumarin to the emitter material tris (8-hydroxyquinoline) aluminum. The energy released by the recombination of holes and electrons was transferred to the dye, which emitted light with greatly increased efficiency. Deposition of additional thin layers of indium tin oxide and other compounds next to the electrodes altered the interaction of the thicker layers and also improved the efficiency of the injection of holes and electrons, thereby further upping the overall power efficiency of the fluorescent OLED.

Organic LEDs of this small-molecule type are used to make red, green and blue light, with green light having the highest efficiency. Such green-emitting OLEDs can exhibit luminous efficiencies of 10 to 15 candelas per ampere—about as efficient as commercial LEDs today—and seven to 10 lumens per watt, values that are comparable to those for common incandescent lamps.

Spin the Puddle

THE SECOND TYPE of organic light emitter is the large-molecule light-emitting polymer, also known as the polymer light-emitting diode (PLED). Reported in 1990 by Jeremy Burroughes and his colleagues at the University of Cambridge, this device incorporates polymers—made from unions, often chains, of smaller organic molecules—to form the layers. Polymer LEDs are formed by spin coating: applying a thin layer of polymer to

Overview/Organic Light Emitters

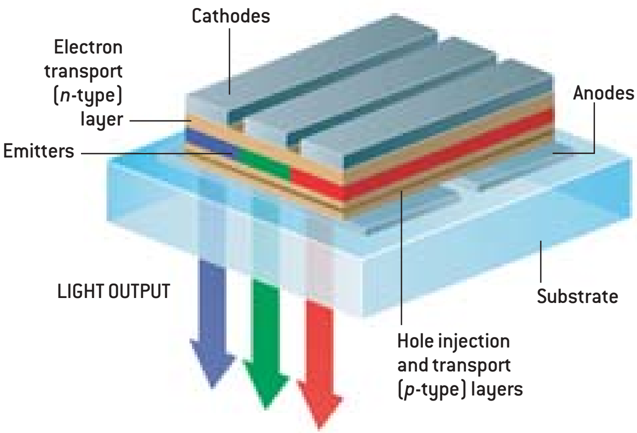
- The familiar LED is common in devices from calculators and watches to gigantic displays at public arenas. Yet because the technology is too costly and difficult to manufacture, it is not very practical for the small, high-resolution displays used in laptops, PDAs, digital cameras and home TVs.
- Enter the OLED—"O" is for "organic." OLEDs are easier to produce than LEDs, more energy-efficient and can be laid down in thin films atop a variety of inexpensive substrates—including flexible plastic or metal foil.
- The technology is on the market in some digital cameras and cellular phones. But the future holds applications that include large TVs and computer monitors that could be rolled up like projection screens, portable battlefield displays for real-time updates, and plastic laminates that could replace lighting fixtures.

UNIVERSAL DISPLAY CORPORATION

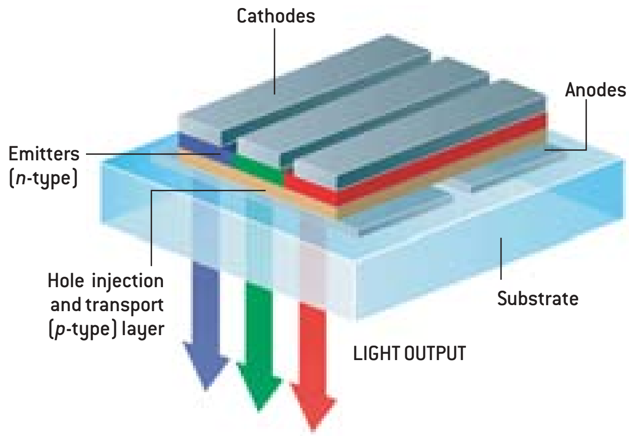
A LOOK INSIDE

GLOW OF COLOR emanates from a sandwich of materials when voltage is applied to an OLED (*left*) or PLED (*right*) display. Holes (positive charges) from *p*-type layers and electrons from *n*-type (negative) layers give off photons of specific wavelengths when they meet in the active, emitter layers. Hues depend on the compounds used in the emitter layers.

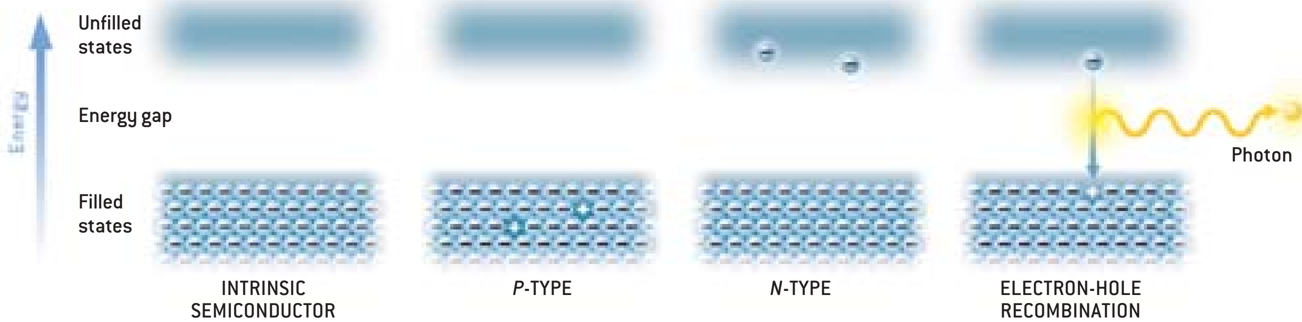
OLED



PLED



HOW ORGANIC EMITTERS PRODUCE LIGHT

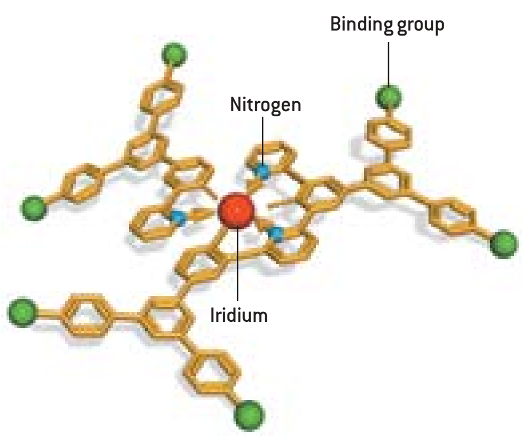


1 A semiconducting material such as silicon has an energy gap between its lower, filled electron states (called the valence band) and its upper, unfilled electron states (the conduction band).

2 Removing some electrons from the lower state leaves behind positively charged "holes," and the material is labeled "*p*-type."

3 Adding electrons to the unfilled upper states makes the material "*n*-type."

4 As electrons drop to the lower state and occupy holes, photons of visible light are emitted.



BUILDING BLOCKS

PHOSPHORESCENT DENDRIMER (*right*) serves as the core around which branching rings are bonded, forming a large molecular ball. The use of dendrimers in organic emitters combines the key benefit of OLEDs—which today produce more colorful displays—with that of PLEDs—which have a more economical manufacturing method. Compounds attached to binding groups at the ends of the branching arms determine solubility for ease of manufacturing.

SAMUEL VELASCO; SLIM FILMS (*dendrimer*)

Organic Displays Coming to Market

CLOSE TO 100 manufacturers are at work developing applications for organic light emitters. Here are some examples.

APPLICATION	USES	PRODUCTS (DISPLAY MAKERS)	STATUS
Small displays	Personal electronic equipment	Digital camera (Kodak/Sanyo); cellular phones (Pioneer, RiTdisplay); car audio components (Pioneer, TDK); electric razor (Philips)	On the market
Large displays	TVs; computers; billboards; vehicle windshields	15.5-inch OLED (Samsung SDI); 17-inch PLED (Toshiba); 20-inch OLED (ChiMei/IBM); 24-inch multipanel screen (Sony)	Prototype; two to four years from market
Bendable displays	Clothing; portable devices	Wearable computer (Pioneer); rollable display (Universal Display Corporation)	Prototype; several years from market
General illumination	"Wallpaper" for walls or ceilings; safety lights for vehicles or clothing	None yet defined	Under development; five to 10 years from market

a flat substrate and then spinning the substrate at a high speed (typically 1,200 to 1,500 revolutions per minute) to spread the polymer puddle by centrifugal force. The spin coating is followed by a baking step to remove the solvent and, in some cases, to complete the polymerization. This film formation process is generally more economical than the thermal evaporation method discussed above. Polymers have had an edge over Kodak-type small-molecule devices in power efficiencies because the greater electronic conductivity of the polymer layers allows lower operating voltages.

The original PLED consisted of a single active layer of a polymer called polyphenylene vinylene (PPV) between dissimilar metal contacts such as indium tin oxide and calcium, as in an OLED, to provide injection of both holes and electrons [*see box on preceding page*]. Indium tin oxide is a metal that tends to inject holes, and calcium is a metal that tends to inject electrons. Current PLEDs use a second polymer layer for hole injection and transport. The polymer PPV produces yellow light, with an excellent efficiency and lifetime; at computer-monitor luminance levels, such a PLED can last more than 10,000 hours—about 10 years of regular use. (Full color has been demonstrated, but the only commercial product to date uses yellow.) Other polymers and mixed polymers (two different polymers in solution) have been developed, notably by Dow Chemical, based on the organic molecule polyfluorene. These configurations can be modified to produce a full range of colors, from red to green, by varying the lengths of the segments of the co-polymers. Unfortunately, the display lifetimes of these colors have not been comparable to that of PPV, and blue is not yet ready for prime time.

Overcoming Limitations

ALTHOUGH TODAY'S fluorescent OLEDs and PLEDs are fairly energy-conserving, there is room for improvement. One limiting factor of the first-generation devices is electron spin, an intrinsic quantum property that determines how the particles react to a magnetic field. When an electron and a hole come together, they form an entity known as an exciton. The quantum-mechanical rules of interaction of the electron and hole spins dictate that for most materials only one in four excitons formed will be able to give up its energy as a photon, whereas the others will lose their energy as heat.

This problem was overcome in 1998 by a team led by Stephen R. Forrest of Princeton University and Mark E. Thompson of the University of Southern

THE AUTHOR

WEBSTER E. HOWARD received a Ph.D. in physics from Harvard University in 1962 and spent his early years in semiconductor physics research. Since 1974 he has worked on various display technologies at IBM Research, ATT/Lucent and eMagin Corporation. His major efforts were in thin-film transistors and liquid-crystal displays. Most recently, as chief technology officer of eMagin, he led the development of OLED-based microdisplays for use in helmets and headsets. In 2003 he was awarded the Jan Rajchman Prize of the Society for Information Display for his contributions to flat-panel display technology. He is currently enjoying retirement, with more time for family, woodworking and golf.

REPORTING BY CHRIS JOZEFOWICZ

California. They developed OLED devices in which the emitting material contained a heavy metal such as platinum or iridium. In a heavy metal atom the outer electrons are, on average, far from the nucleus, and consequently have a large angular momentum of rotation. This momentum interacts with the spin of other electrons, essentially creating conditions where all the excitons can emit light rather than heat, raising the theoretical efficiency to nearly 100 percent. These new emitters are referred to as phosphorescent OLEDs, to distinguish them from the more common fluorescent OLEDs.

The phosphorescent small-molecule OLEDs generate impressive efficiency gains without any significant loss of lifetime, except for the case of blue emitters. No good blue-emitting phosphorescent material yet exists, although major efforts are under way at numerous laboratories to remedy this situation.

Is there a way to pair the many colors produced by small-molecule OLEDs with the economic benefits of the spin solution coating used in PLEDs? One recent exciting development builds on the successful application of these phosphorescent materials to potentially do just that. Workers at the University of Oxford, the University of St. Andrews and OpSys in Oxford, England, have synthesized molecules known as dendrimers that may prove to combine the desirable aspects of small molecules and polymers. In a dendrimer, a phosphorescent molecule serves as the core around which layers of branching ring structures are bonded, forming a large molecular ball [*see box on page 79*]. If the branching elements are suitably chosen, these molecules can be dissolved, and films can be formed by spin coating and baking, as with polymer materials. Already dendrimer devices have yielded very high luminous efficiencies (more than 50 candelas per ampere and 40 lumens per watt).

Organic light emitters achieve the same image quality as bulky cathode-ray tubes (CRTs) and LCDs, although the manufacturing process differs. For good perceived image quality, the red, green and blue pixel-producing groups, called triads, must be more densely spaced than 100 per inch (or roughly 40 per centimeter) so that the eye can merge the colors without noticing the underlying image structure. Both CRTs and LCDs achieve such precision with photolithography, in which a projected light image creates a pattern for laying out a material design. Each phosphor or color-filter material is suspended in a photosensitive resin matrix and spread over the entire display area. Wherever light hits the material, the matrix is rendered insoluble and the rest of the material is then dissolved, or developed, away. Three successive applications, one for each color, complete the process.

In contrast, small-molecule color OLEDs are patterned by evaporating the emitter materials through a shadow mask. The mask is displaced slightly for the successive evaporations of the red-, green- and blue-producing materials and can be manipulated with a 10-micron accuracy, for 250-micron triad dimensions. The preferred application method for full-color PLEDs is ink-jet printing, a technique widely used in PC color printers. High-precision versions of printers with polymer inks can pattern the red, green and blue polymer emitters with a precision of five to 10 microns.

Products Today and Tomorrow

AT PRESENT, small molecule–based prototype displays look better—they have more colors—than polymer versions, but PLED proponents are confident about closing the gap.

Color OLEDs and PLEDs can also be made by using a white light emitter and tiny color filters, as is done with LEDs. This approach has been used by eMagin Corporation (the author's former organization) in Hopewell Junction, N.Y., to provide 800-by-600-pixel resolution in a 0.6-inch (diagonal) color microdisplay, built on a silicon microchip active matrix. Such displays are used, with magnifying optics, in helmets for pilots, soldiers and firefighters, as well as in headsets for applications such as three-dimensional PC games—an example of an innovative product made possible by the new organic emitter technology.


LARGE-SCREEN OLED from Samsung SDI Company measures 15.5 inches diagonally—but is only 1.8 millimeters thick.

SAMSUNG SDI COMPANY

The military and others are also investigating technologies to make flexible organic light emitters. The U.S. Army Research Laboratory recently awarded $2 million to Universal Display Corporation in Ewing, N.J., to develop such an OLED display, which could be unrolled from a pen-size communications device. The Defense Advanced Research Projects Agency has also funded work on the flexible devices for battlefield displays.

Less dramatically, one can expect consistent and rapid progress in organic emitter and transport materials, as the number of interested companies and dedicated scientists grows. Already it has become clear that organic emitters offer almost unlimited opportunity for molecular engineering and synthesis to improve energy efficiency, product life and manufacturing cost. SA

MORE TO EXPLORE

The Dawn of Organic Electronics. Stephen Forrest, Paul Burrows and Mark Thompson in *IEEE Spectrum,* Vol. 37, No. 8, pages 29–34; August 2000.

A Bright Future for Displays. Bob Johnstone in *Technology Review,* Vol. 104, No. 3, pages 80–85; April 2001.

OLED-Display Technology. Kathleen M. Vaeth in *Information Display,* Vol. 19, No. 6, pages 12–17; June 2003.

Organic Light-Emitting Devices. Edited by Joseph Shinar. Springer Verlag, 2003.

the case of the Unsolved Crime

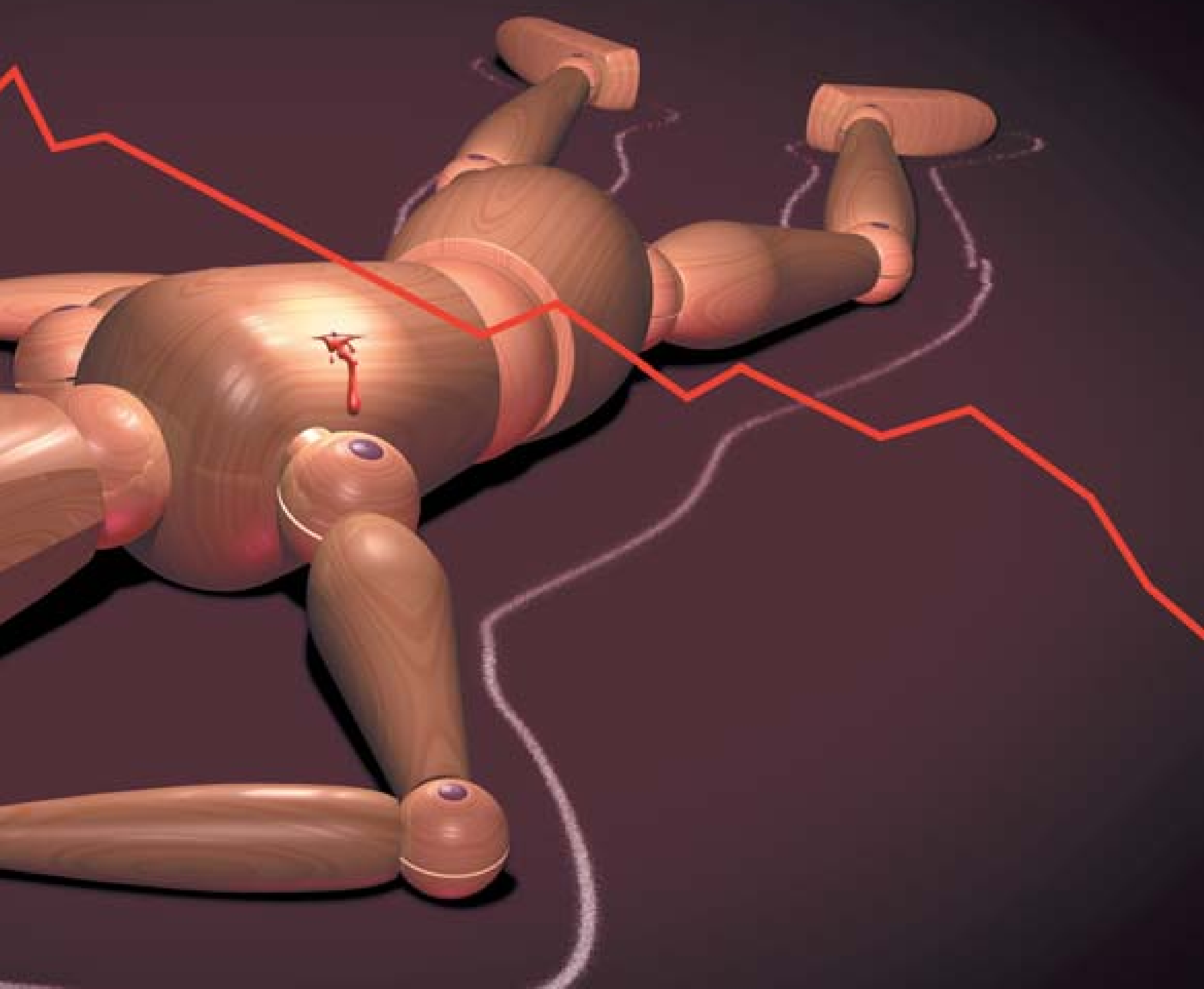
By Richard Rosenfeld

For a short period during the closing decade of the last century, U.S. crime rates dropped precipitously. Homicide, burglary and robbery rates fell more than 40 percent, to levels not seen since the 1960s. The reduction in serious felonies per capita stunned criminologists, who have struggled to provide a satisfying explanation for such an unexpected and complex phenomenon. The research community has reached a consensus on the basic contours of the 1990s crime decline—the who, what, when and where—but still argues about the why.

Today, as crime rates are again creeping upward, it seems appropriate to examine the evidence associated with the 1990s drop and the theories put forth to account for it. Such an analysis could help society to better understand the causes underlying shifts in national crime statistics and may even be used to forestall future increases in serious offenses. In this article, I will weigh the relative merits of the leading explanations and present some suggestions for policies and experiments that could help prevent the next rise in criminal activity.

Decline

Criminologists have not yet cracked the case, but they now know more about why U.S. crime rates plummeted in the 1990s—and how to help keep them down



HOMICIDE RATES per 100,000 Americans fell more than 40 percent during the 1990s.

Exhibit A: The Facts

TO BETTER EVALUATE the various theories, it helps to take a closer look at the available data. The Federal Bureau of Investigation compiles and confirms cases of serious violent and property crimes reported to local police departments and then converts the tallies into averaged rates expressed as victims per 100,000 people. Of the crime categories the FBI tracks through its Uniform Crime Reporting program, homicide statistics are the most reliable because nearly all the cases are known to the police. The graph of crime levels from 1982 through 2001 [*see box on opposite page*] shows that the national homicide rate peaked at a high of 9.8 per 100,000 in 1991 and then fell to 5.5 by 2000—a 44 percent decline. The slump in the murder rate was accompanied by similar decreases in every major FBI crime category. The incidence of burglary (unlawful entry into a structure to commit a felony or theft) fell by 42 percent between 1991 and 2000; robbery (theft accompanied by force or the threat of force) dipped 47 percent.

The evidence depicting a drop in serious felonies in the 1990s is not limited to crimes compiled in the FBI reports, which may be affected by victims' willingness to notify police. Large contractions in burglary and robbery rates were also noted by the country's other major crime gauge, the National Crime Victimization Survey, which conducts annual polls of crime victims. This survey includes incidents that were not reported to police. Homicide victims cannot participate in polls, of course, but data collected by the National Center for Health Statistics from death certificates and coroner's reports match the homicide trends identified by the FBI.

Exhibit B: The Perpetrators

THE TRENDS are not uniform, however. According to demographic studies, the U.S. actually experienced two crime drops during the last two decades of the 20th century: one among adolescents and young adults (those under the age of 25) and the other among adults. Although the decrease in crime rates among youths has attracted more attention, the adult decline started sooner and lasted longer. Rates of homicide committed by adults have slid steadily since 1980. In contrast, youth homicide levels did not begin to fall until 1993 or 1994 and followed a dramatic increase that had originated about 10 years earlier. Robbery rates among adolescents and young adults traced the same trajectory, rising precipitously from the mid-1980s until about 1994 and descending just as sharply thereafter.

The so-called youth violence epidemic, during both its escalation in the 1980s and its subsidence during the 1990s, was itself highly concentrated in another population subset: young black men [*see box on opposite page*]. Changes in the crime rate for young women and whites were much less pronounced. Between 1984 and 1993, homicide offenses grew nearly fivefold among black male adolescents and more than doubled among black male young adults. The rates for both categories fell rapidly afterward. Meanwhile the offense rates among black male adults dropped by more than half; those for whites of any age have shown comparatively little change over the past 20 years.

Because people of Hispanic origin may be of any race, the FBI data for Hispanics and non-Hispanic whites and blacks cannot be broken out separately. The National Center for Health Statistics series on the causes of mortality, however, has permitted such a partitioning for homicide victims (not offenders) since 1990. Those data portray roughly the same pattern among young Hispanic males—a rise in homicide incidence until the early 1990s, followed by a decline through the end of the decade; however, the Hispanic rates were lower and their increase and subsequent decrease less severe than among young black men.

The nationally aggregated crime statistics obscure important differences in the timing and location of the youth violence epidemic as it arose in various areas around the country—valuable clues regarding the dimensions of the crime drop. The rise in youth homicide and robbery rates started in the largest cities during the early 1980s and then spread to smaller ones a few years later; violent crime rates also peaked and ebbed earlier in the biggest cities.

Exhibit C: The Weapons

A FINAL FACTOR to note was the proliferation of firearms possession among young minority males during the past two decades, reflected in the rising proportion of violent crimes committed with guns. Most homicides and a large proportion of robberies in the U.S. are perpetrated with firearms, whereas few burglars use guns. FBI statistics indicate that nearly all the growth in youth homicide rates (the crime for which weapon use is best documented) during the 1980s and early 1990s involved firearms—usually handguns. Killings in which other or no weapons were implicated actually decreased during the escalation phase of the youth homicide epidemic and continued to drop after 1993, when firearm-related homicide rates also started to fall.

Although it is clear that guns played a prominent role in the rise in homicides beginning in the mid-1980s, this does not necessarily imply that firearms "caused"

Overview/*The 1990s Crime Drop*

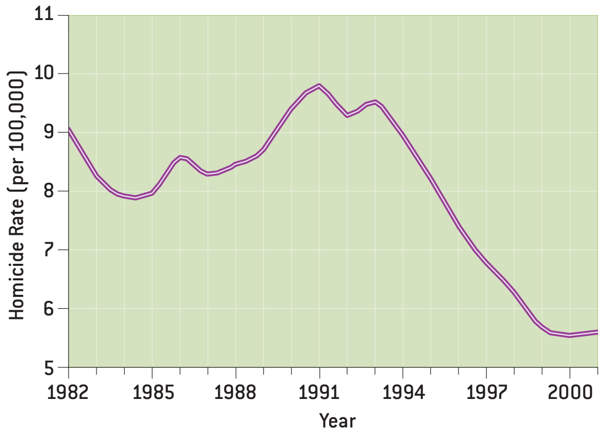
- Serious violent and property crime rates—for homicide, burglary and robbery—in the U.S. decreased substantially during the 1990s. The rates for these serious offenses tumbled by more than 40 percent.
- Analysts have attributed the 1990s crime drop to various causes. These include changes in demographics, law-enforcement practices, economic conditions, incarceration rates, domestic violence and firearm policies, and the use of guns by young crack cocaine dealers.
- Society can draw three lessons from research on the crime decline that may help anticipate and even head off the next rise: divide crime trends into their component parts, look for unintended policy effects and engage in research-based policy experiments before new programs are implemented.

SLIM FILMS (*preceding pages*)

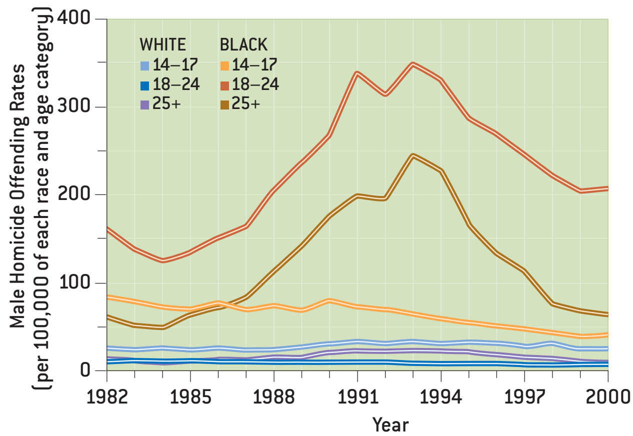
JUST THE FACTS

FOLLOWING A GENERAL DECLINE, the trend lines for homicide, burglary and robbery began to level off by the end of the 1990s and then rose slightly from 2000 to 2001. (Note that the national homicide tally for 2001 excludes the deaths resulting from the terrorist attacks of September 11.) Preliminary data for 2002 from the Federal Bureau of Investigation's Uniform Crime Reporting (UCR) program indicates that crime continued to drop in some cities, notably New York, while increasing in others, including Chicago and Los Angeles. A spot check of crime statistics for 10 large U.S. cities reveals similarly mixed results through the first quarter of 2003.

U.S. HOMICIDE RATE 1982 to 2001



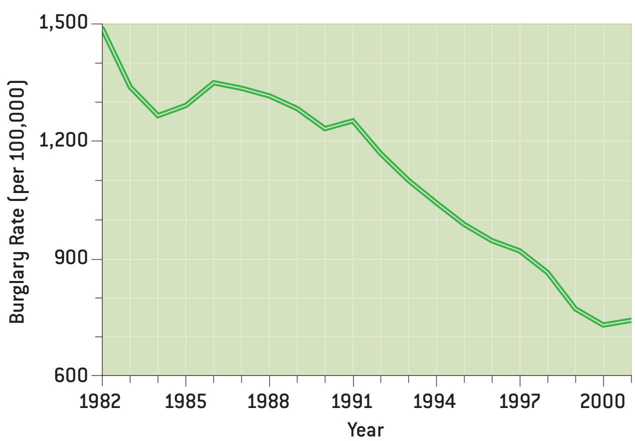
U.S. MALE HOMICIDE Offending Rates by Race and Age, 1982 to 2000



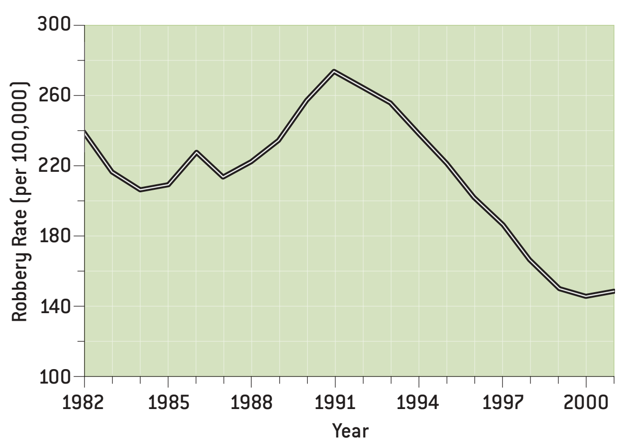
THE HOMICIDE RATE IN THE U.S. rose to a high of 9.8 per 100,000 people in 1991 and then dropped to 5.5 nine years later—a 44 percent decrease, according to the FBI's UCR program.

The so-called youth violence epidemic—the outbreak of serious crime incidents in urban areas that arose in the 1980s and then moderated during the 1990s—was perpetrated mainly by young black men (*right graph*). Young women and whites were much less likely to participate in this phenomenon. These patterns are clearly documented by data from the FBI's Supplementary Homicide Reports, an extension of the UCR data. Between 1984 and 1993, homicide offenses grew nearly fivefold among black male adolescents (ages 14 to 17) and more than doubled among black male young adults (ages 18 to 24). (People of races other than white and black make up only about 1 percent of homicide offenders and are not shown.)

U.S. BURGLARY RATE 1982 to 2001



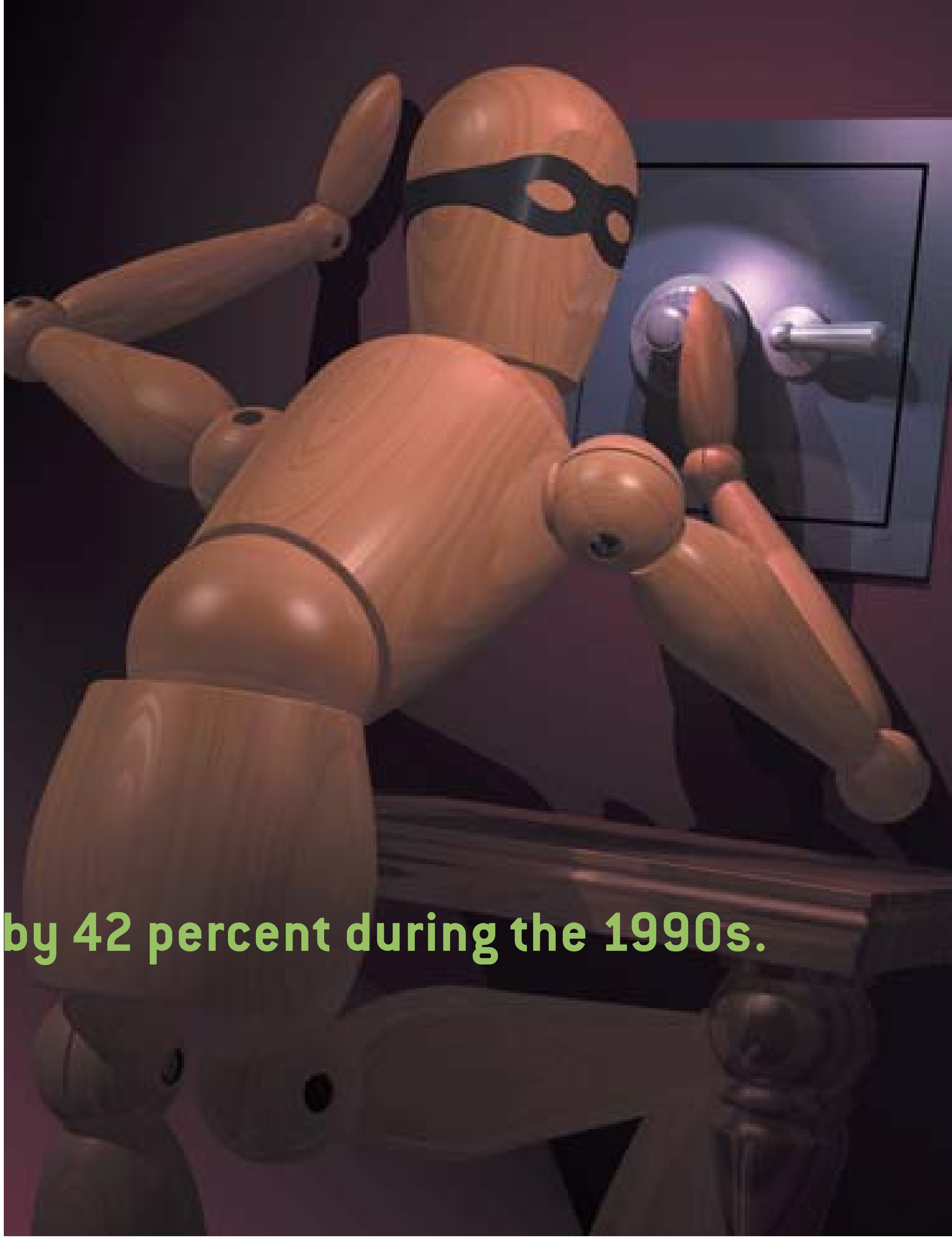
U.S. ROBBERY RATE 1982 to 2001



THE BURGLARY RATE IN THE U.S. declined 42 percent in the 1990s. Unlike homicide and robbery rates, the lower incidence of burglaries was not limited to the closing decade of the 20th century. Burglary rates began to slip in the early 1980s and fell by more than half during the next 20 years.

The falls in the murder and burglary rates during the 1990s were accompanied by a similar decrease in the occurrence of robbery (a 47 percent drop), according to statistics compiled by the FBI from police reports.

NINA FINKEL



SLIM FILMS

the youth violence epidemic. The relation between guns and the decline in homicide is less clear, because both firearm- and nonfirearm-related killings abated during the 1990s. Two researchers who study violent acts, Philip J. Cook of Duke University and John H. Laub of the University of Maryland, have argued that the factors associated with the "way in" to the violence epidemic differ in some respects from those contributing to the "way out." The decrease was steeper and broader than the increase, affecting crimes committed with and without guns. Keep that point in mind during the forthcoming evaluation of firearms' role in explaining the decrease in crime during the 1990s.

The Arguments

THE FACTS HAVE THUS led us to a set of unresolved questions that any credible explanation of the 1990s crime drop must accommodate: Why did the decline in youth violence arrive on the heels of a dramatic increase in youth violence? Why was the youth violence epidemic concentrated among young minority males? Why was the rise in youth violence—but not the fall—restricted to crimes involving firearms? Why did it start and end first in the big cities? And, not least, why has adult violence diminished over the past 20 years?

Burglary rates fell by 42 percent during the 1990s.

No theory accounts for all these facts, but certain hypotheses do a better job than others at explaining the many aspects of the crime decline. And some popularly held explications are plainly wrong.

Demographics. A few explanations of the crime drop can be eliminated simply because their timing is off. One such theory is based on the premise that crime rates will rise or fall in step with corresponding changes in the size of the age cohorts that are disproportionately responsible for crime. A greater proportion of adolescents and young adults take part in homicide and other crimes than adults. When the younger segment of the population shrinks, some researchers hypothesize, crime rates should as well, all else being equal. Unfortunately, all else is rarely equal. Other conditions affecting crime rates tend to change more rapidly, if less predictably, than the size of the age group mainly involved in crime. Moreover, the relative size of the demographic group at highest risk, 14- to 24-year-old black males, changed little from 1993 (the peak of the youth-violence epidemic) to 2000 (the trough)—and actually decreased during the first stages of the epidemic.

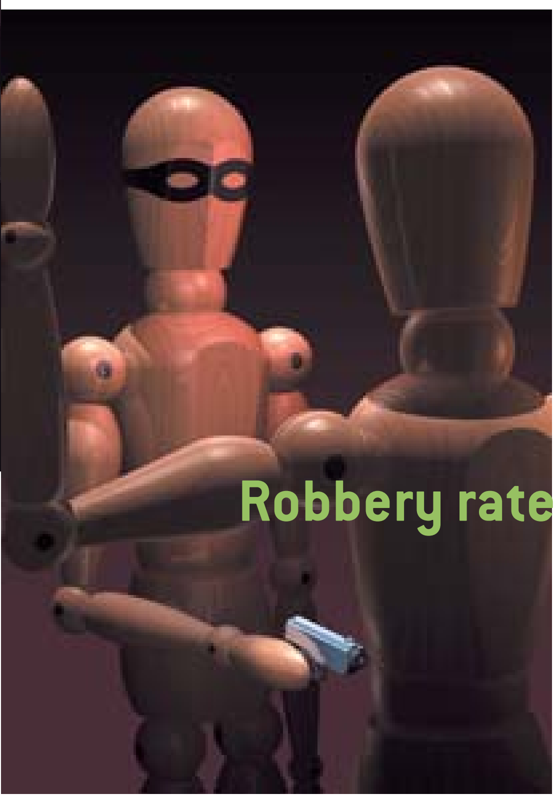
Economists John J. Donohue III of Stanford University and Steven D. Levitt of the University of Chicago offer an intriguing alternative to conventional demographic explanations of the crime drop. They attribute as much as half of the crime decline during the 1990s to the legalization of abortion in the 1970s. This change resulted in fewer births of unwanted children to low-income women, thereby, they claim, preventing the crimes those disadvantaged children would have committed some 15 to 20 years later.

THE AUTHOR

RICHARD ROSENFELD is professor in and chair of the department of criminology and criminal justice at the University of Missouri–St. Louis. He is co-author with Steven F. Messner of *Crime and the American Dream,* third edition (Wadsworth, 2001), and has published widely on the social sources of violent crime. Rosenfeld is a member of the National Academy of Sciences's Committee on Crime, Law and Justice and is on the steering committee of the National Consortium on Violence Research.

Though not implausible, the analysis implies that youth homicide trends should have slackened earlier than they did.

Law enforcement. Policing efforts in a number of cities underwent revision during the 1990s, from the introduction of "community policing" strategies to blanket crackdowns on minor infractions. But many of these changes did not go into effect until well after the fall in crime had begun. It is true that local evidence supports the effectiveness of targeted "gun patrols" (in which police saturate areas that have a high incidence of firearm use) and of ensuring swift punishment of gang members who carry guns. Local evidence also backs computerized accountability programs, such as New York City's COMSTAT, that hold police commanders accountable for the crimes that occur in their districts.



Robbery rates dropped by 47 percent in the 1990s.

Yet sizable crime declines occurred in cities without such programs, too. No one can say whether changes in policing policies contributed to the decrease in *national* crime rates.

Kids, crack and guns. A more promising explanation for the decrease in crime has been offered by Carnegie Mellon University criminologist Alfred Blumstein, who links the phenomenon to shrinking demand for crack cocaine in the early 1990s, which presumably resulted in less violence related to the drug's sale. Unlike other theories, this one jibes both with the timing of the crime reductions and with that of the youth violence epidemic. The popularity of cheap crack cocaine created a boom in illicit drug trade during the 1980s. To meet growing demand, dealers recruited inner-city youths to sell crack and armed them with firearms for protection against thieves, unscrupulous buyers and rival sellers. The arms race soon moved beyond the neighborhood drug markets into surrounding communities, as gun violence engendered retaliation in kind.

For reasons that remain poorly understood, crack has shown itself to be a single-generation drug. As the original addicts grew older and either stopped taking it or died, younger drug users, who preferred marijuana, did not replace them. Thus, the demand for crack began to subside. Following a lag of a year or two, rates of gun violence also began to fall off, first in the largest cities where crack took hold earlier, then afterward in smaller cities where both crack and the youth violence epidemic arrived later.

Blumstein's crack/firearms diffusion explanation squares with most of the facts of the 1990s crime drop. It accounts for why the violence epidemic was ignited by males and minority youths who sold crack and why rates of nonfirearm-related violence were not affected. (In general, drug sellers do not settle disputes with fists, clubs or knives.) Blumstein's theory is consistent with survey evidence showing that inner-city youths, including gang members, acquire guns mainly for protection. Additionally, it coincides with studies of the violence common to crack markets and the comparative lack thereof in the marijuana trade.

Economic expansion. Unfortunately, the crack/firearms diffusion hypothesis does not account for the length and breadth of the decrease in youth violence, and it says nothing about the drop in property crime rates or the long-term decline in violence among adults. It is tempting to invoke the economic boom of the 1990s, especially the steep declines in unemployment rates, to explain why both adults and youth might turn away from crime and toward legal and safer sources of income. Yet property crime and adult violence rates also dipped during the less favorable economic climate of the 1980s. The relation between unemployment and crime is in fact decidedly more complex than it first appears. The so-called opportunity effects of more and bigger paychecks, which make potential crime victims more attractive targets, may cancel out the crime-cutting effects of falling unemployment levels. On the other hand, unemployed people spend more time at home, resulting in fewer home burglaries.

In addition, job and income growth may have differing consequences regarding the occurrence of crime, depending on the availability of chances for illicit income. Although dealing drugs is dangerous and uncertain work, thousands of inner-city teenagers and young adults engage in it. Legitimate employment may be more attractive when those illegal moneymaking opportunities disappear, such as during the crash in the crack markets in the early 1990s. This situation implies that the growth of legal job availability has a greater effect on crime rates when and where drug markets have dried up because of lower demand or effective law-enforcement efforts. To date, no published studies have examined this hypothesis.

Prison expansion. If the long-term decline in adult crime cannot be explained by improved economic conditions, perhaps it is associated with the corresponding escalation in incarceration rates during the past two decades. The U.S. incarcerates a larger proportion of its citizens than any other nation, and the size of the American prison population quadrupled between 1980 and 2000. It would be surprising if mass incarceration had no impact on crime, but like economic conditions, the ramifications of imprisonment are complex. Growth in imprisonment undoubtedly cuts crime in the short run; one study estimates that roughly a quarter of the drop in crime during the 1990s can be attributed to the sharp escalation in incarceration rates. Large-scale confinement, however, may boost crime rates in the long term by breaking up families, driving up unemployment rates, and otherwise depleting the social capital of those communities hardest hit by both crime and imprisonment.

The incarceration boom also may have contributed, albeit indirectly, to the youth homicide epidemic. As the demand for crack climbed during the 1980s and adult vendors increasingly sat behind bars, drug dealers turned to younger sellers. If such a labor shortage was created and all else held constant, sharper rises in youth homicide should have been observed in those areas with the greatest increases in adult imprisonment for drug crimes. This hypothesis merits further research.

Domestic violence and firearms policies. Two additional arguments have been offered for the decline in adult crime. The first links the drop in domestic homicide to the expansion of hotlines, shelters, judicial protection orders, and other domestic violence resources during the 1980s and 1990s. The second attributes lower incidence of adult victimization to the expansion of "concealed carry" laws, which permit adults to bear concealed weapons. Laura Dugan of the University of Maryland, Daniel Nagin of Carnegie Mellon and I conducted research that showed that domestic homicide rates fell more rapidly in cities with the greatest growth in legal advocacy and other services for victims of domestic abuse. We found, though, that other responses to

Pretested crime-control policies may yield better success.

domestic violence, such as a policy of mandatory arrest for offenders, may actually increase the likelihood of homicide under some conditions, presumably because offenders are angered by the legal intervention or because the resulting sanctions are not sufficient to protect victims from further violence.

Economist John R. Lott, Jr., of the American Enterprise Institute has proposed that laws permitting the carrying of concealed weapons have reduced violent crime rates by making would-be offenders aware that potential victims could be armed. His research indicates that the rates of serious crime are lower in places with concealed-carry laws than elsewhere, controlling for other conditions affecting crime. Other scholars using similar data and methods, however, have not been able to reproduce Lott's results. For now, the case for "more guns, less crime" remains unproved.

Clearly, a single theory encompassing all the facts of the crime drop does not exist. The closest approximation of such an explanation is the link between the rise and fall in youth violence and corresponding shifts in the crack cocaine trade. The longer declines in adult violence and property crime are probably associated with the explosive growth in imprisonment, the adoption of domestic violence policies and programs, and the economic boom of the 1990s—but mass incarceration, tougher arrest policies, more

SLIM FILMS

Growing incarceration rates affected the crime drop.

jobs and larger incomes may increase as well as reduce the occurrence of crime.

The Verdict

CONSIDERING the complexities of crime, what lessons can society draw to help anticipate and even head off another rise in criminal activity? Three guidelines seem appropriate.

Take apart the trends. As we have seen, the crime drop seems to have resulted from the confluence of two separate crime trends: one for adults and one for youths. Distinct explanations and policies apply to each of them. Adults are subject to incarceration in state and federal prisons, are legally entitled to possess firearms and are directly affected by efforts to stem domestic violence. Youths have a low risk of imprisonment and cannot legally possess a handgun if they are younger than 21 (or own a long gun until they are 18). Domestic and family violence policies do affect children and adolescents, but only indirectly, through their parents—and the consequences of such policies for youth crime may not show up for a number of years.

Young people did not participate equally in the violence epidemic of the late 1980s and early 1990s, which was concentrated primarily among inner-city black and Hispanic males. Small-town and suburban white youths were involved in several highly publicized school shootings, but fortunately these events were not numerous enough to reverse the decline in youth firearm-related homicide. Policies and programs to prevent school shootings or reduce the already comparatively low level of violent crime among more affluent youths are likely to differ from those intended to suppress an arms race among inner-city youths. A wise strategy would be to tailor crime-control policies to the particular circumstances of different groups of victims and offenders.

Watch for accidental policy effects. Stiffer sentences for adult drug offenders may facilitate the criminal careers and shorten the lives of the youthful drug sellers who take their place. Mass incarceration reduces crime in the short run—but at great monetary and social costs—and may contribute to the chronically high levels of crime in those distressed communities from which prisoners are disproportionately drawn and to which they return. Some domestic violence policies may result in more, not less, violence and abuse. That is not to say society should close prisons or stop arrests. Instead communities should stay alert for the unexpected aftereffects that large-scale social interventions inevitably produce. Policymakers must understand the trade-offs between intended and unintended consequences and change policies if they are doing more harm than good. This lesson should be applied first to sentencing policies such as "three strikes and you're out" and mandatory minimum sentences, which have made the U.S. the world leader in incarceration.

Conduct experiments in crime-control policy. Public safety would benefit from a reversal of the standard sequence of implementing a crime-control policy or program first and then determining if it works. The National Institute of Justice, the research arm of the U.S. Department of Justice, has funded many field studies of domestic violence interventions, innovative policing practices, alternatives to prison for drug offenders, and greater penalties for gang members and other youths who carry guns. The term "experiment" must be used advisedly when referring to such research: for example, batterers cannot be assigned randomly to households, and legal barriers prevent the alteration of sentences for the most serious law violators. But threats to the validity of crime-control research can be lessened through the careful selection of comparison groups and the introduction of rigorous statistical controls. The point is to perform policy experimentation before implementing policies. Research findings alone will not prevent future increases in crime, but control policies that have been pretested stand a much better chance of succeeding. SA

MORE TO EXPLORE

The Crime Drop in America. Edited by Alfred Blumstein and Joel Wallman. Cambridge University Press, 2000.

After the Epidemic: Recent Trends in Youth Violence in the United States. Philip J. Cook and John H. Laub in *Crime and Justice: A Review of Research,* Vol. 29. Edited by Michael Tonry. University of Chicago Press, 2002.

The Crime Decline in Context. Richard Rosenfeld in *Contexts,* Vol. 1, No. 1, pages 25–34; Summer 2002.

Exposure Reduction or Retaliation? The Effects of Domestic Violence Resources on Intimate-Partner Homicide. Laura Dugan, Daniel Nagin and Richard Rosenfeld in *Law & Society Review,* Vol. 37, No. 1, pages 169–198; March 2003.

Evaluating Gun Policy: Effects on Crime and Violence. Jens Ludwig and Philip J. Cook. Brookings Institution, 2003.

Bureau of Justice Statistics, National Crime Victimization Survey: **www.ojp.usdoj.gov/bjs/cvict.htm#ncvs**

Federal Bureau of Investigation, Uniform Crime Reports: **www.fbi.gov/ucr/ucr.htm**

National Center for Health Statistics, National Vital Statistics: **www.cdc.gov/nchc/nvss.htm**

WORKINGKNOWLEDGE

AUTOMOBILE BLACK BOX

Data Driven

Cars and trucks have a small, rugged box of electronics that can reveal how the vehicle was operating before a crash. Few drivers were aware that they were sitting on such hardware until last August, when police sought the box from U.S. Representative Bill Janklow's Cadillac after he hit and killed a motorcyclist in South Dakota.

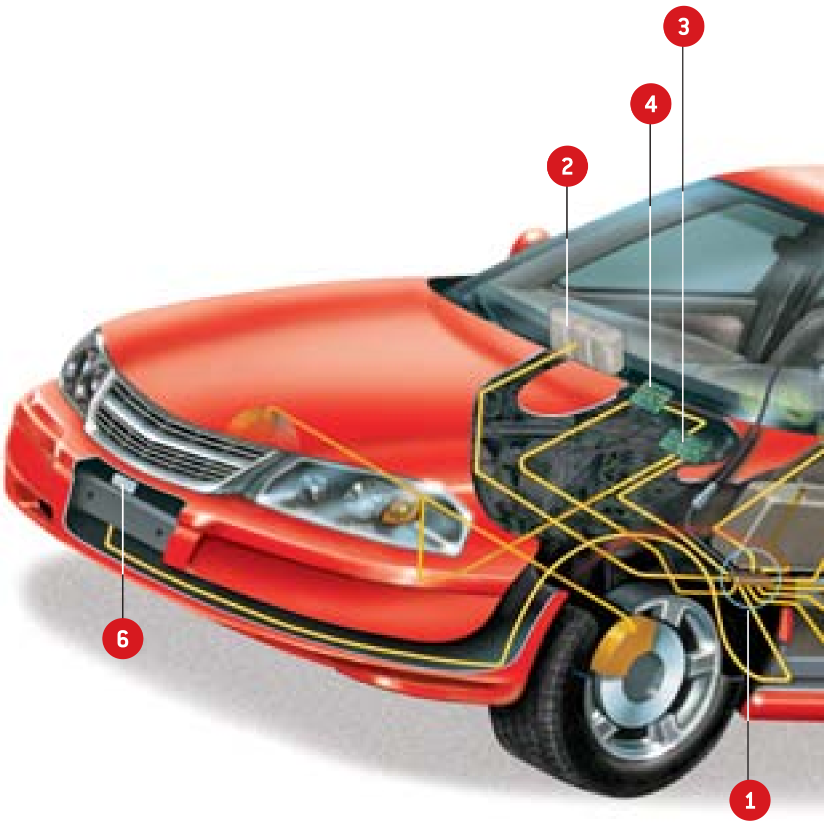
Event data recorders began to be installed in U.S. passenger vehicles in the mid-1970s to deploy air bags when they detect extreme changes in velocity. Automakers gather data from boxes in random accidents to analyze which velocity deltas, stored by the boxes, caused deployment—useful for improving designs. But since 1999 the units, typically installed under the front seat carpet, now record vehicle speed, engine rpm, degree of throttle (accelerator) and brake deployment, and seat-belt engagement. Coupled with crash injury reports, the data are used to enhance performance or to help in recalls of faulty systems.

The boxes store only the most recent five seconds of information, but police and prosecuting attorneys have found that such vital statistics for the moments preceding a crash can strongly influence whether a driver bears guilt: Was he speeding? Did he fail to stop at a stop sign? Did he accelerate to run a red light?

This application, never intended by automakers, "has caught everyone a bit by surprise," says James Kerr of Vetronix Corporation, which makes software to decode black boxes. That surprise includes little protection over what civil libertarians are now calling private information. General Motors states that it will obtain a vehicle owner's permission before deciphering the box. But most other manufacturers are keeping mum, and, except for California, there are no laws preventing anyone from obtaining readouts without a court order or the owner's okay.

Media attention has added "an amazing amount of misinformation," says GM's Jim Schell. The boxes cannot reveal where a car has been or how fast it was going beyond five seconds before a crash. But giving the box a heftier memory chip and linking it to an onboard Global Positioning System transceiver could provide such details. That prospect suddenly seems to be worrying drivers. *—Mark Fischetti*

BLACK BOX 1 tells front air bags 2 to inflate if an accelerometer in the box detects a rapid enough drop in velocity, indicating the first milliseconds of strong impact. Meanwhile the box monitors the vehicle's power-train control module 3 and braking module 4, and records the vehicle's speed, engine rpm, and degree of throttle and brake deployment for five seconds before impact. Buckle sensors 5 tell the box if front seat belts are fastened. The box may also read (but not yet record) accelerometers in pole crash sensors 6, side air bags 7 and rollover prevention monitors 8 in vehicles so equipped.



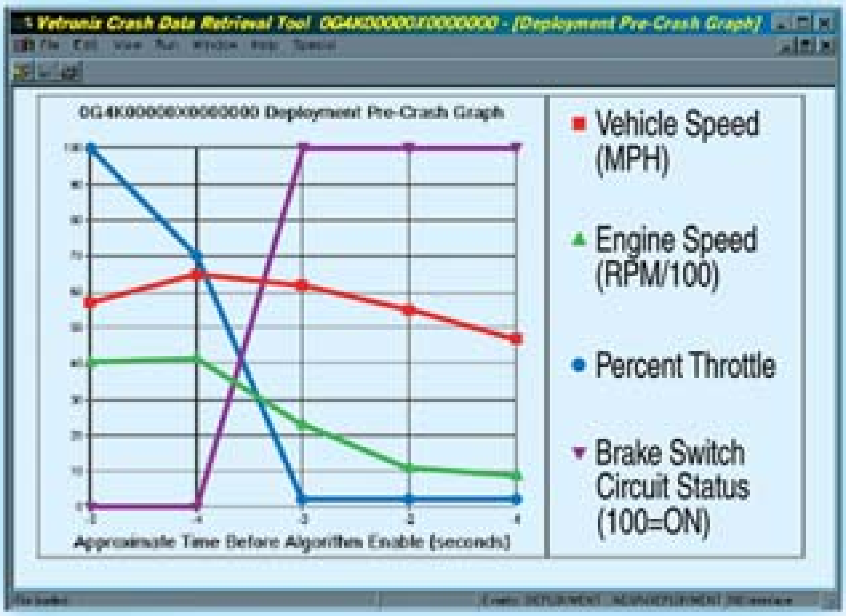
KENT SNODGRASS *Precision Graphics*

DID YOU KNOW . . .

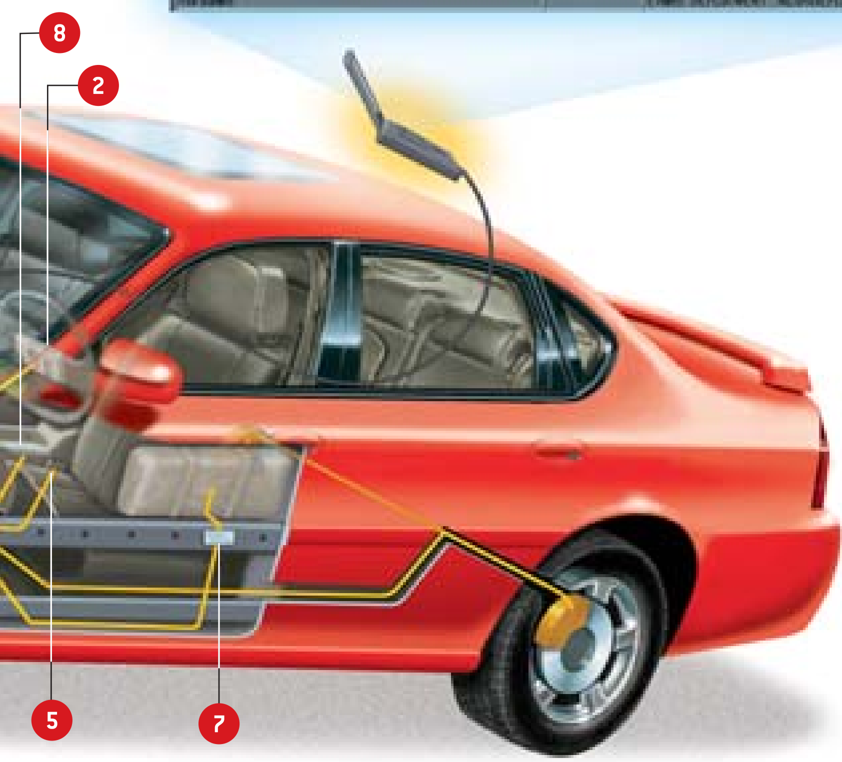
➤ WHO DOES IT: Carmakers use proprietary computers to analyze black box data. Only GM, Ford and Isuzu allow an outside vendor, Vetronix, to make laptop software that can decipher their recorders. That package costs $2,500, but the car companies hope police departments and law firms will buy it to speed analysis, saving the manufacturers time and money spent decoding boxes for investigations. No standards yet exist for data recording or output, but efforts are under way at the Institute of Electrical and Electronics Engineers and the Society of Automotive Engineers.

➤ DOWN THE PIKE: Future data recorders may track accelerometers that sense a car's violent lateral motion—the trigger for deploying side air bags—and sudden steering changes that would spark rollover prevention actuators. Insurers and litigators would most likely seek these data to help determine, for example, if a driver swerved or not. The car companies might not need to store such information to control safety systems, but they might anyway, because they know it is increasingly being requested to investigate accidents.

➤ GOOD AND BAD: In a recent GM survey, consumers were equally divided in their opinion of black boxes. Spokesperson Jim Schell surmises that people who think of themselves as good drivers welcome a processor that can verify their innocence in accidents; bad drivers don't want it tattling on their transgressions.



GRAPH rendered from black box data shows that five seconds before air bags deployed, the driver quickly let up on the gas pedal (throttle). One second later he hit the brakes, cutting engine speed but only reducing vehicle speed to 48 miles per hour one second before a collision triggered the air bags.



FLIGHT DATA RECORDER (in tail) far exceeds an automotive black box. It stores the last 25 hours of aircraft altitude, direction of flight, speed, acceleration and thrust from each engine. A cockpit voice recorder (in nose) captures 30 to 120 minutes of crew conversation and ambient sounds. The units can withstand impact at 268 knots, fire of 1,100 degrees Celsius and pressure of 20,000 feet undersea for 30 days.

Send topic ideas to workingknowledge@sciam.com

A Walk in the Woods

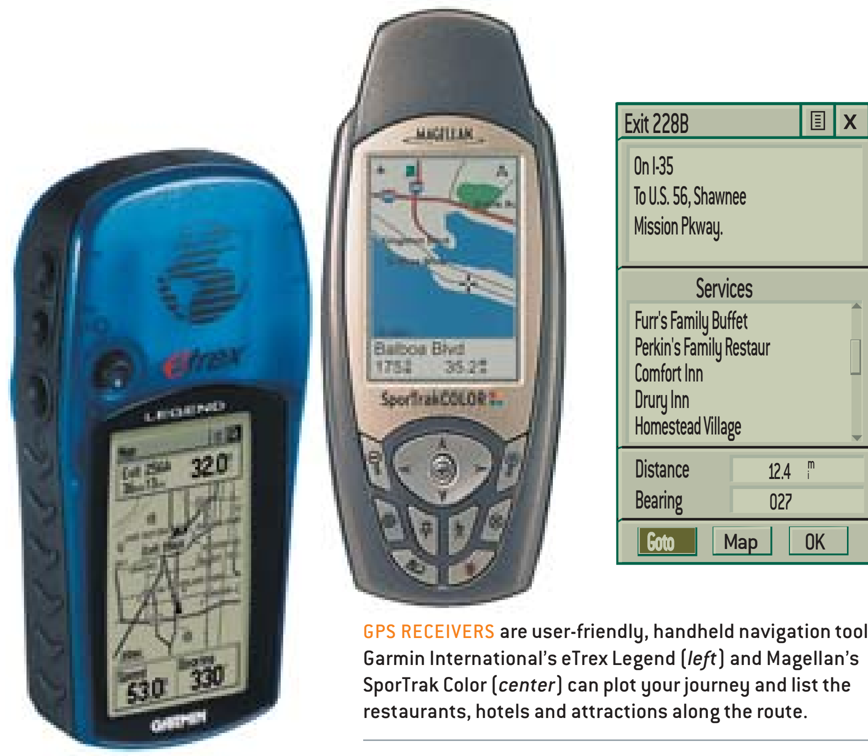
SATELLITES SHOW THE WAY IN THE NEW SPORT OF GEOCACHING BY MARK CLEMENS

Congratulations! *You've found it! Intentionally or not! What is this hidden container sitting here for? What the heck is this thing doing here with all these things in it?*

Standing on a hillside in Lake George, N.Y., I pored over the letter left in a Tupperware container that I'd found just moments before on a rock shelf. The note explained the somewhat mischievous rules of a new high-tech sport called geocaching (pronounced "geocashing"). Flush with excitement, I read on:

This container is part of a worldwide game dedicated to GPS (Global Positioning System) users. The game basically involves a GPS user hiding "treasure" (this container and its contents) and publishing the exact coordinates so other GPS users can come on a "treasure hunt" to find it. Geocaching is a relatively new phenomenon. Therefore, the rules are very simple:

1. Take something from the cache.
2. Leave something in the cache.
3. Write about it in the logbook.



GPS RECEIVERS are user-friendly, handheld navigation tools. Garmin International's eTrex Legend (*left*) and Magellan's SporTrak Color (*center*) can plot your journey and list the restaurants, hotels and attractions along the route.

COURTESY OF GARMIN LIMITED (*eTrex GPS*); COURTESY OF THALES NAVIGATION (*Magellan GPS*)

Before my first geocaching experience, I thought that handheld GPS units were used only by wilderness hikers and the military, but in the past few years the devices have become more accessible. The GPS system consists of 24 satellites and five ground stations. Orbiting about 12,000 miles above the earth, the satellites transmit low-power radio signals on frequencies in the UHF band. By measuring the travel time of the signals from at least four of the satellites, a handheld GPS unit can (at least in theory) triangulate its position anywhere on the planet.

Until recently, the U.S. government scrambled GPS signals so that only the military could receive the most accurate readings. But in May 2000 the Clinton administration stopped the signal degradation, allowing civilian GPS units to determine their location to within 20 to 40 feet. Soon afterward several companies introduced affordable, simple-to-operate receivers with built-in maps and screens that work much like a Web page. At about the same time, the Federal Aviation Administration implemented the Wide Area Augmentation System (WAAS), a network of ground stations that improves GPS precision by correcting for timing errors and variations in the satellites' orbits. WAAS-enabled devices can determine a user's position to within nine feet, assuming that the user has a clear view of the sky and that the weather is not too cloudy.

I've enjoyed orienteering in the woods since my days in the Boy Scouts, so when a college friend brought up the idea of geocaching, I jumped at the opportunity. On my first foray I relied on the GPS III Plus, a now discontinued handheld unit made by Garmin International. Guided by the device's readings, we climbed an uphill trail, hopped across streams, marched through a meadow of waisthigh grass and descended into a boulder-strewn ravine. A word to the wise: Don't look down at the GPS screen while you're walking through the forest. I paid the price for my inattention, getting whacked in the head by a low-hanging fir branch.

After we had trekked for 1.2 miles the device told us that we'd reached the cache's latitudinal and longitudinal coordinates (called waypoints in GPS lingo). But because of the dense tree cover, the potential error was as much as 45 feet. We assiduously searched every rock, tree and crack in the area for 20 minutes but found nothing. Then came a holler of triumph from Dave Taylor, my geocaching mentor, who pointed to a rock shelf. Under a pile of sticks was our prize, a four-inch-by-12-inch plastic container with "geocaching.com" written in black marker on the lid.

We carefully lifted the container and opened the lid. Inside we found a collection of trinkets—a whistle, a disposable camera, a couple of key chains, a miniature flashlight, a fishhook and so on—along with a logbook and pen. After reading the entries scribbled by other geocachers, I signed my name in the logbook and removed the fishhook from the container, knowing as I did so that I'd been hooked by the game. We deposited our own memento—a rubber brain bearing the words "Scientific American"—and took a group photo of ourselves using the disposable camera. Then we returned the container to its hiding spot and covered it with sticks. In a victorious mood we hiked down the mountain, led by the GPS unit to the previously entered coordinates of our parked car.

On subsequent geocaching expeditions, I tried four different GPS devices to rate their accuracy and ease of use. I chose two beginner's units, Garmin International's Geko 201 ($149) and Magellan's SporTrak Map ($229), along with two more advanced units, Garmin's eTrex Venture ($194) and Magellan's SporTrak Color ($499). Each unit displays a map, a digital compass and a satellite tracker that lets you know how many GPS signals you're receiving. Entering the waypoints of your destination is just as easy as writing a text message on a cell phone. And to get position information from the GPS satellites, all you need to do is hold the unit face-up toward the sky. Interpreting the maps and compass readings becomes easier with experience.

Logging onto www.geocaching.com, I did a search for caches near my home in suburban north New Jersey. The Web site listed half a dozen within four miles. Each listing is posted by the person who hid the cache and gives the coordinates of the hiding place. I recruited one of *Scientific American*'s editors, and we set off to find the nearest cache, carrying all four handheld receivers. After entering the cache's waypoints into each unit, we outlined a route from my house to the hiding place on the map display. Then we followed the route by walking in the direction specified by the compass.

As we walked through a nearby park, the GPS units constantly updated our distance from the cache. Although the Garmin devices provided less frequent updates than the Magellan receivers, they appeared to gauge the distance more accurately. To get the most precise readings of one's location, it is best to stand still in an area with an unobstructed view of the sky for 10 to 15 seconds. On this excursion, the potential error was only 15 feet, so when we reached the coordinates of the cache we needed to search for only a few minutes to find the hidden treasure.

We soon learned, however, that the devices don't work so well in urban settings. GPS signals can pass through clouds, glass and plastic, but they are completely stymied by skyscrapers. The best accuracy we could get in midtown Manhattan was 150 feet. Nevertheless, we were able to find a National Geodetic Survey disk on the side of the New York Public Library's main building at 42nd Street. (The three-inch copper disks, which are scattered all over the U.S., enable precise surveying for civil engineering and cartography.)

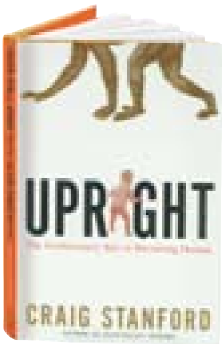
After two months of exploring, I have become a geocaching veteran. So far I have located 96 geocaches and 27 survey disks in four states (New York, New Jersey, Pennsylvania and Connecticut) and three countries (the U.S., Canada and England). I prefer the advanced GPS units over the beginner's models because of extra features such as the ability to store more than 100 waypoints. I particularly like the most expensive of the devices I tested, Magellan's SporTrak Color, which displays its map in color. Because roads and rivers are shown in different colors on this unit's display, I didn't have to worry about driving my car into a creek.

Despite my frantic activity, it will take some doing to find all the treasures scattered across the globe. According to the geocaching Web site, there are currently 76,477 caches in 190 countries and nearly 10,000 people trying to uncover them. What's more, an average of 50 new caches are being hidden every week. So I have my work cut out for me. SA

Mark Clemens is an assistant art director at Scientific American.

Good with Their Feet

BECOMING BIPEDAL, THE AUTHOR ARGUES, MADE US HUMAN **BY BLAKE EDGAR**



COURTESY OF HOUGHTON MIFFLIN

UPRIGHT: THE EVOLUTIONARY KEY TO BECOMING HUMAN
by Craig Stanford
Houghton Mifflin, 2003 ($24)

"Anyone who thinks that six million years of human evolution has created an optimally designed, anatomically perfect human has never looked closely at his own body," asserts biological anthropologist Craig Stanford in *Upright,* his brief, breezy tour through the most vexing problem (aside from language) in human origins: why we, alone among primates and all mammals, took to full-time, two-footed locomotion. Fossil evidence unearthed during the past 30 years has confirmed that bipedalism emerged long before culture, big brains, stone tools and other attributes we have upheld as uniquely ours. Bipedalism still guarantees inclusion as a hominid, a member of the human family, although that is one of the bones that Stanford has to pick in his fifth book.

Rather than harp on why bipedal humans are so unusual, Stanford, a professor at the University of Southern California, places people into a broad ecological and evolutionary perspective. He points to another group of captivating fossil creatures—dinosaurs—as nature's first experiment with erect posture. Dinosaurs came in all shapes and sizes and employed different kinds of bipedalism for different circumstances.

So did early hominids. Six million years ago several bipeds and protobipeds were running—or at least ambling awkwardly—in Africa. Some we are just discovering through fossils such as *Sahelanthropus* and *Ardipithecus,* whereas others, Stanford believes, remain to be found. Some were true hominids, but others will turn out to be apes. One of Stanford's take-home messages, repeated frequently, is the fallacy of reading any linear progression into the story of why certain bipeds, our ancestors, became proficient walkers. Today we can walk a straight line, but that ability stems from a meandering evolutionary journey.

STANDING CHIMPANZEE bends its knees and—because it does not have properly positioned gluteal muscles, which form the human buttocks—leans forward for balance.

How did bipedalism overhaul standard ape anatomy? To cite a few changes, our pelvis became much broader, shorter and saddle-shaped. An obscure thigh muscle ballooned into buttocks, the largest human muscles, which shifted function from propulsion to stabilizing the hip in midstep. Our skull and spine were realigned, and our center of gravity shifted up and forward. The list goes on.

Standing upright entailed a Faustian bargain of evolutionary trade-offs. We gained stability at the expense of propulsive power. We increased energy efficiency for forward motion but lost some climbing capability. We spread around the planet but inherited a legacy of lower back pain, broken hips and difficult childbirth. The initial and subsequent benefits of bipedalism, though, must have outweighed the costs. Writes Stanford, "As our way of moving about changed, so did our niche in the world, our perspective, and our prospects."

Plenty of explanations have been proposed for why our ancestors stood up. Stanford touches on most of them: bipeds make more energy-efficient foragers; upright males could carry high-caloric foods to mates saddled with offspring; standing displays might diffuse potentially lethal scuffles. Some he dismisses with too little description and discussion, and the idea that standing upright provided better thermoregulation on a sunbaked savanna receives scant mention. In general, Stanford summarizes others' research concisely and evenly while portraying the rancorous debates that often arise in paleoanthropology. Unfortunately, he misidentifies some historic anthropologists and extinct species.

In his analysis of explanations, Stanford rejects ideas that essentially work backward from the human form of walking. Whatever benefits accrued to early bipeds, he reasons, must have helped a creature that looked and acted much like an ape, especially a chimpanzee, live another day.

This brings *Upright* to the author's expertise, the foraging behavior of wild chimpanzees. Stanford treats the reader to anecdotes from the field. He writes of a mother chimp feeding in a treetop: "With her baby clinging precariously to her chest, the mother reached above her head while standing on three limbs on a huge branch. Suddenly, the mother stood upright and used her long arms to reach several feet above her head to grab the figs. For thirty seconds she stood on two legs, her muscular feet anchoring her on the branch and her hands steadying her above. Her baby rode along, bobbing up and down as her mom plucked handfuls of figs and shoved them into her mouth."

Modest movements like this, Stanford argues, may have been the starting point for the rest of human evolution. Once on the ground, bipeds could expand their diets to include more meat, eventually fashioning stone tools to access this important resource, making marathon migrations, and putting their primate brains and hands to good use.

Upright rambles a bit near the end, rehashing the discovery and disappearance of the Peking man fossils and the controversial Eve hypothesis and evoking science-fiction aliens. But until then, Stanford provides an engaging and intriguing introduction to our most enduring feature. SA

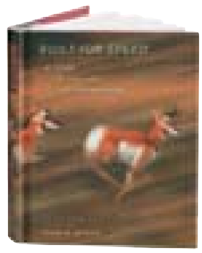
Blake Edgar is the co-author of three books on human origins, including The Dawn of Human Culture *(John Wiley & Sons, 2003).*

THE EDITORS **RECOMMEND**

BUILT FOR SPEED: A YEAR IN THE LIFE OF PRONGHORN

by John A. Byers. Harvard University Press, Cambridge, Mass., 2003 ($24.95)

"Pronghorn are unquestionably the fastest mammal in North America. They accelerate explosively from a standing start to quickly reach a top speed close to 60 miles per hour. Pronghorn also can cruise at 45 miles per hour for several miles." Byers, professor of zoology at the University of Idaho, has spent 20 years closely observing pronghorn on the National Bison Range in Montana. His account of the animal's ways is thorough and fascinating. He tells of racing a male pronghorn with his truck; after two miles at 45 mph, the pronghorn "wasn't even breathing hard." Byers also recounts the difficulty of finding and catching fawns, which he does in order to tag them for long-term study. They hide in the grass until they are about three weeks old. "If I'm in reasonable physical condition, I can usually run down a 5-day-old fawn. A contest against a 7-day-old fawn is a toss-up, and a 10-day-old fawn can in effect thumb its nose at me with impunity."



DRAW THE LIGHTNING DOWN: BENJAMIN FRANKLIN AND ELECTRICAL TECHNOLOGY IN THE AGE OF ENLIGHTENMENT

by Michael Brian Schiffer. University of California Press, Berkeley, Calif., 2003 ($34.95)

BOLT OF FATE: BENJAMIN FRANKLIN AND HIS ELECTRIC KITE HOAX

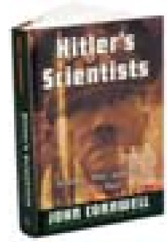
by Tom Tucker. PublicAffairs, New York, 2003 ($39)

Did Franklin really fly that kite? Schiffer says yes, Tucker says no. Nobody knows. Schiffer, professor of anthropology at the University of Arizona, accepts the tradition of the kite experiment, although he says it is "a long and inconclusive story." Tucker, a science writer, offers two reasons in particular for rejecting the kite story. One is that in describing the experiment in the newspaper, the *Pennsylvania Gazette,* Franklin does not say that he did it. The other is that the experiment as Franklin described it would be unlikely to succeed because of the design of the kite and the difficulty of flying it under the conditions outlined by Franklin. Whatever the truth, the kite experiment is a footnote to the rapid blossoming of electrical technology in the 18th century. Both authors tell that story well.

HITLER'S SCIENTISTS: SCIENCE, WAR AND THE DEVIL'S PACT

by John Cornwell. Viking Penguin, New York, 2003 ($29.95)

What you see in the title of this book is less than what you get. Cornwell, who is in the department of history and philosophy of science at the University of Cambridge, does write about Hitler's scientists and their attitudes toward the warped science of the Nazi regime. But he also looks at German science, good and bad, throughout the 20th century. And he explores an issue of profound importance: the relationship between science and the good society. It would be "naive," he warns, to assume that the relationship is morally sound in a democracy. In the face of such challenges as secrecy, the militarization of science, and abuse of the environment, scientists must not "relax their moral and political vigilance."



All the books reviewed are available for purchase through www.sciam.com

All or Nothing

BY DENNIS E. SHASHA

Any message can be represented as a number. For example, the word "meet" in "Let's meet at the corner of Constitution and Lake" would be represented in most computers by 109 101 101 116 in decimal, using the encoding known as ASCII.

Suppose you want to send a secret directive—perhaps a rendezvous time and location—via five couriers, but you fear that one or two will be caught. You therefore want to spread the message among the couriers such that any three of them can together reconstruct it but two or fewer cannot.

Because messages are encoded numbers, you could think of this problem as having the five couriers share the secret of a number. Intuitively, you might think it wise to give each of them part of the number, but that is not the most secure approach. Instead you want to establish an information "cliff": two couriers give you no useful information, but three give you the entire message. To reach that goal, you have to think of a cleverer plan.

To warm up, suppose I think of a point in a plane—say (13,6)—and ask three friends to identify that spot. I want any two of the three, but no single friend, to find it. As a clue, I give Martha the line $x = 13$, Jaime the line $y = 6$, and Valeria the line $y = x - 7$ [*see illustration below*]. How could my pals use that information? Can you see that two are necessary and sufficient? Similar reasoning will give you a solution to the five-courier problem. SA

Dennis E. Shasha is professor of computer science at the Courant Institute of New York University.

Answer to Last Month's Puzzle

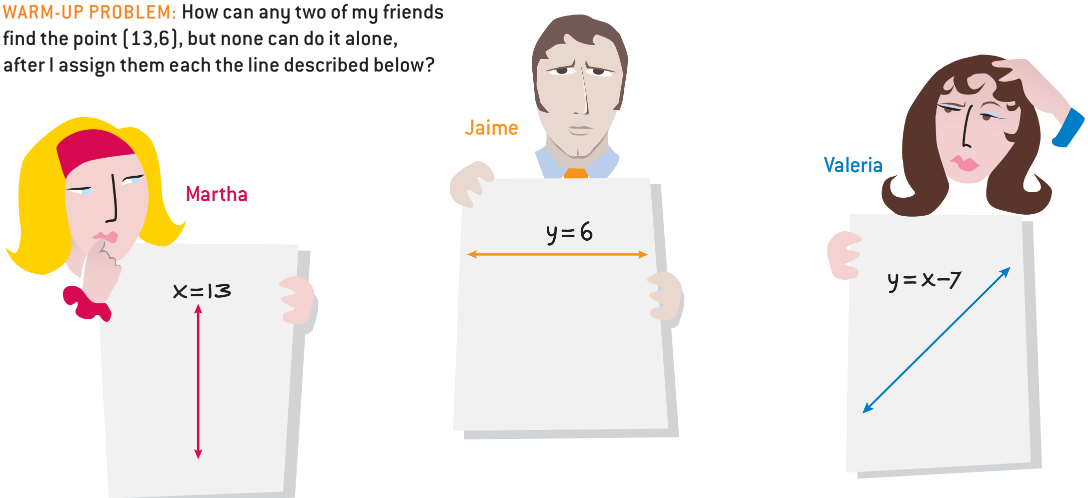
Two tests are needed to verify the four-element circuit. In the first test, put 0 in inputs A, B and C and 1 in input D; output E should be 1. In the second test, put 1 in inputs A and C and 0 in inputs B and D; output E should be 0. If the circuit has four AND gates, three tests would be required, with the following inputs: 0111 (meaning input A is 0 and the others are 1), 1011 and 1110.

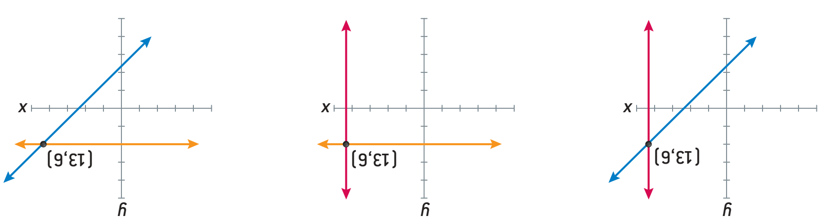
Only two circuits in the three-element configuration cannot be verified with one test: the circuit in which element 2 is an AND gate and the others are OR gates, and the circuit in which element 2 is an OR gate and the others are AND gates.

Web Solution

For a peek at the answer to this month's problem, visit www.sciam.com

WARM-UP PROBLEM: How can any two of my friends find the point (13,6), but none can do it alone, after I assign them each the line described below?





ANSWER: With just one line, no friend has useful information: my point could be any one of the infinite points in the assigned line. But any two of my friends can find the point (13,6) instantly by determining where their lines intersect on a standard *x,y* grid. My pals go from facing infinite uncertainty to having precise knowledge.

CORNELIA BLIK

It Is High, It Is Far

BUT IT'S NOT GONE, BECAUSE THERE ARE SOME LAWS YOU JUST CAN'T BREAK **BY STEVE MIRSKY**

As has been said often in these pages, there's a clear and present need for better math and science education in this nation. One obvious place for improvement in our math and science skills can be found among the hosts of and callers to the country's many sports talk radio programs.

Recently I was listening to a show on which the host contended that the Boston Red Sox's payroll had swelled to the point where Sox fans couldn't complain that the New York Yankees' even larger payroll gave the Yankees any advantage. A Boston caller disagreed, saying, "The Red Sox payroll is only $120 million, and the Yankees is $180 million. You know what percentage $120 million is of $180 million? Seventy-five percent." The host did not dissuade the caller. This display came from two grown men who spend an inordinate amount of their time calculating batting averages.

That exchange was about simple arithmetic. But the next morning I was treated to a lively discussion of Newtonian physics. The two morning hosts had left the subject of sports for a moment to discuss national news headlines. One story involved a Ku Klux Klan initiation ceremony at which a Klansman had fired a gun into the air. By the way, part of the radio hosts' conversation went something like this:

"The Klu Klux Klan."

"It's not Klu. It's *Ku*. It's not Klu Klux Klan, it's Ku Klux Klan."

"I didn't say Klu Klux Klan, I said *Klu* Klux Klan."

"You said it again, you said Klu."

"I did not say Klu Klux Klan, I said *Klu* Klux Klan."

"You said it again. You said Klu."

I was grateful I *didn't* have a gun. Fortunately, they eventually departed from the science of linguistics and returned to classical mechanics. As noted, some Klan member was firing a gun straight up in the air. You've probably guessed by now that

a bullet came back down, which they tend to do. Well, all the bullets came back down. But one in particular found its way to the ground only after going clean through the top and then out the bottom of the skull of one of the celebrants, critically injuring him. (A British newspaper headlined this story "Ku Klux Klan Man Shot as Initiation Goes Wrong." Which raises the question: What Klan initiation has ever gone right?)

Anyway, the talk show hosts were incredulous that a bullet could have come down hard enough to do that kind of head damage "just from gravity," as one put it.

Now, I wouldn't necessarily expect Klansmen to go in for the kind of book learning that would reveal that a bullet returning to earth after being shot straight up could return fast enough to cause serious injury. What did surprise me was that two men who basically watch the trajectory of projectiles for a living—batted baseballs, for example—would be incredulous at the speed at which some objects return to earth. Hadn't they ever noticed that when a catcher fields a major league pop-up, where the baseball has gone almost straight up and down, the ball smashes into that catcher's mitt pretty darn hard?

In summary, as a public service for guys waving guns or news copy: stuff that goes up fast comes down fast. In a vacuum, where air resistance is not a factor, an object sent on a flight has a final downward speed that is, amazingly enough, equal to its initial upward speed. I've seen the equations—it's true! Closer to home, air resistance does indeed slow down a bullet or baseball, but both still gallop back to earth at quite a clip. And a bullet is pointy.

Clearly, the gunman, who has been charged with aggravated assault and reckless endangerment, did not intend to hit his victim. Why, he shot in the completely opposite direction! But ignorance of the laws of motion is no excuse. Nevertheless, the bet here is that the shooter gets off and that Isaac Newton gets charged with being an accessory after the fact. Or, actually, three centuries before. SA

ILLUSTRATION BY MATT COLLINS; PHOTOGRAPH BY FRANK VERONSKY

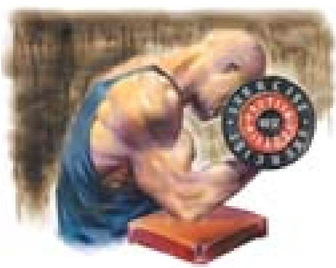
How does exercise make your muscles stronger?

—B. Thrall, St. Louis

Mark A. W. Andrews, professor of physiology and director of the independent study program at the Lake Erie College of Osteopathic Medicine, explains:

The exact mechanism by which exercise augments strength remains unclear, but its basic principles are understood. Two processes appear to be involved: hypertrophy, or the enlargement of cells, and neural adaptations that enhance nerve-muscle interaction.

Muscle cells subjected to regular bouts of exercise, followed by periods of rest that include a sufficient intake of dietary protein, undergo hypertrophy. (This should not be confused with short-term swelling resulting from water uptake into cells.) Improved muscle protein synthesis and incorporation of these proteins into cells cause the muscle-building effect. When a muscle cell is activated by its nerve cell, the interaction of the proteins responsible for muscle contraction—actin and myosin—generates force via changes in protein structure called power strokes. The total force generated depends on the sum of all the power strokes occurring simultaneously within all the cells of a muscle. Because more potential power strokes accompany an increased presence of actin and myosin, the muscle can exhibit greater strength of contraction. In addition, hypertrophy is aided by certain hormones and has a strong genetic component.



MATT COLLINS

The neural basis of muscle strength enhancement primarily involves the ability to recruit more muscle cells—and thus more power strokes—simultaneously. This process, called synchronous activation, is in contrast to the firing pattern seen in untrained muscle, where the cells take turns firing in a more asynchronous manner. Training also decreases inhibitory neural feedback, a natural protective response of the central nervous system to feedback arising from the muscle. Such inhibition keeps the muscle from overworking and possibly ripping itself apart as it creates a level of force to which it is not accustomed. This neural adaptation generates significant strength gains with minimal hypertrophy and is responsible for much of the strength gains seen in women and adolescents who exercise. It also utilizes nerve and muscle cells already present and accounts for most of the strength growth recorded in the initial stages of all strength training; hypertrophy is a much slower process, depending as it does on the creation of new muscle proteins. Thus, overall, the stress of repeated bouts of exercise yields neural as well as muscular changes to add to muscle strength.

What causes a mirage?

Edwin Meyer, professor of physics at Baldwin-Wallace College, provides this answer:

Mirages are caused by photons (particles of light) taking the path of minimum time as they travel through air of differing temperatures and densities.

Ideal conditions for a mirage are a hot, sunny day, when the air is still over a flat surface baked by the sun. The air closest to the surface is hottest, and thus the air density gradually increases with height. Incoming photons take a curved path from the sky to the viewer's eye. The standard freshman physics explanation for this phenomenon is that cooler air has a higher index of refraction than warmer air does. Accordingly, photons travel through hot, less dense air faster than they do through cool air. The quantum electrodynamics rationale is that photons always take the path of minimum time when traveling from one point to another. To get from one point to another in a minimum time, photons take shortcuts, even though the length of the path is curved and it covers a longer distance than the direct route.

Why does your brain see the mirage of water? The mirage occurs because past experience has taught you that when you look at a surface ahead and see the sky, you are *usually* looking at a reflection off of water. So when the conditions are right for a mirage, the brain assumes that water must also be present. SA

For a complete text of these and other answers from scientists in diverse fields, visit www.sciam.com/askexpert


CLUES POINTING TO UNIVERSE'S ORIGINS FOUND!
Pamphlet
SO YOU WANT TO START A UNIVERSE!
Measuring cup and spoons, mixing spoon, and bowl
Empty tin and opener
COSMOS-IN-A-CAN
A bunch of Polaroids
Coffee grounds
Cassette tape
Music to Create Everything By
Overdue library book
1,001 TIPS FOR MAKING THE WHOLE KIT AND CABOODLE FROM SCRATCH
BEFORE
AFTER
Receipts
R. Chast