SCIENTIFIC AMERICAN

NOVEMBER 1990
$3.95

Black holes, quasars and the birth of the universe.

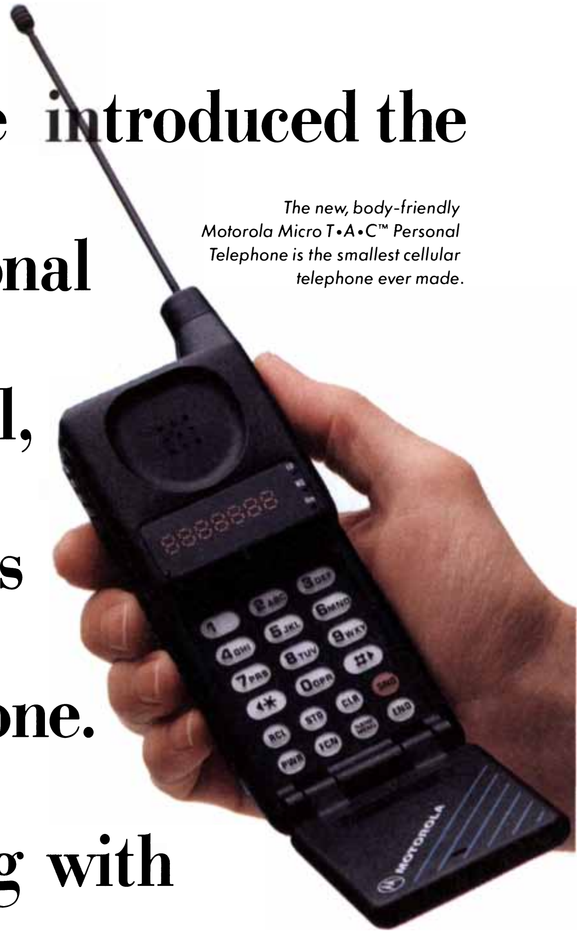
What do dreams really mean?

Treating inherited diseases with implanted genes.

This minuscule quantum dot *traps electrons in zero dimensions. It heralds a new era in designer materials.*

And, we produced the first hand-held cellular portable phone. In 1989, we introduced the Motorola Micro T·A·C™ Personal Telephone, shirt-pocket small, with a fraction of the parts of the original cellular phone.

■ These developments, along with others, are the product of an annual R&D investment more than twice the national average. Such is the measure of our belief.



The new, body-friendly Motorola Micro T•A•C™ Personal Telephone is the smallest cellular telephone ever made.

Building On Beliefs

MOTOROLA

SCIENTIFIC AMERICAN

November 1990 Volume 263 Number 5

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Copyright © 1990 by Scientific American, Inc. All rights reserved. Printed in the U.S.A. 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. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Subscription rates: one year $36, two years $64, three years $88 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. only 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537.

DEPARTMENTS

THE COVER transmission electron micrograph is an indium phosphide semiconductor structure that contains a quantum dot. Built by Henryk Temkin and his colleagues at AT&T Bell Laboratories, it measures less than a thousandth the size of a human hair at its base. When voltage is applied to the structure, electrons will be trapped in the deep purple region at the neck (made of indium gallium arsenide). When so confined, electrons exhibit unusual energy states, which may drive the next leap forward in electronics (see "Diminishing Dimensions," by Elizabeth Corcoran, page 122).

THE ILLUSTRATIONS

Cover micrograph by Henryk Temkin, AT&T Bell Laboratories

Page	Source
43	Granger Collection
44	Photo Researchers, Inc.
48	© Mary Evans Picture Library, Photo Researchers, Inc.
50	Bettman Hulton
53	Granger Collection
54	General Electric Company
57	George Retseck
58	Rick Perley, National Radio Astronomy Observatory (*top*), George Retseck (*bottom*)
59	Palomar Observatory, California Institute of Technology (*top*), Steve Kent, Harvard-Smithsonian Center for Astrophysics (*second from top*), George Retseck (*bottom*)
60-64	George Retseck
66	Fred Lo, University of Illinois at Urbana-Champaign
69	Gamma-Liaison Network
70-72	Dana Burns-Pizer
81	James M. Wilson, University of Michigan at Ann Arbor, and J. Roy-Chowdhury, Albert Einstein College of Mecidine
82	Dana Burns-Pizer
87	Marc Chagall, Scala/Art Resource
88	Carol Donner
90	Patricia J. Wynne
92	Carol Donner (*top*), Gabor Kiss (*bottom*)
94	Gabor Kiss
96	Carol Donner (*top*), Gabor Kiss (*bottom*)
99–103	Michael Goodman
107	A. Trevor Hodge
108–111	Tom Prentiss
115	Stephen Dalton, Natural History Photographic Agency, Sussex, England
116	David Askew, University of Exeter (*top*), Patricia J. Wynne (*bottom*)
117	Patricia J. Wynne, David J. S. Newman, University of Exeter (*inset*)
118–120	Patricia J. Wynne
122–123	Edward M. Clausen, Jr., and James P. Harbison, Bell Communications Research
124	John R. Deecken
125	Spectra Diode Laboratories, (*both photographs*), John R. Deecken (*inset*)
126	Kim Y. Lee, IBM Corporation (*left*), David M. Hwang, Bell Communications Research (*right*), John R. Deecken (*inset*)
127	John McGrail
128	Henryk Temkin, AT&T Bell Laboratories (*left*), Kim Y. Lee, IBM Corporation (*right*), John R. Deecken (*inset*)
129	John N. Randall, Texas Instruments (*left*), Galen D. Stucky, University of California, Santa Barbara (*right*), John R. Deecken (*inset*)
130	Edward Bell
131	Russell F. Smith, Sandia National Laboratories
142–144	Edward Bell

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Alan Hall, *Executive Editor;* Michelle Press, *Managing Editor;* Timothy M. Beardsley; Elizabeth Corcoran; Deborah Erickson; Marguerite Holloway; John Horgan; Philip Morrison, *Book Editor;* Corey S. Powell; John Rennie; Philip E. Ross; Ricki L. Rusting; Russell Ruthen; Gary Stix; Paul Wallich; Philip M. Yam

ART: Samuel L. Howard, *Art Director;* Joan Starwood, *Design Director;* Edward Bell, *Associate Art Director;* Johnny Johnson

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Jonathan Goodman

PRODUCTION: Richard Sasso, *Vice President Production and Distribution; Managers:* Carol Albert, *Prepress;* Carol Eisler, *Manufacturing and Distribution;* Carol Hansen, *Electronic Composition;* Madelyn Keyes, *Systems;* Leo J. Petruzzi, *Manufacturing and Makeup;* Carl Cherebin; William Sherman

CIRCULATION: Bob Bruno, *Circulation Director;* Lorraine Terlecki, *Business Manager;* Cary Zel, *Promotion Manager*

ADVERTISING: Robert F. Gregory, *Advertising Director.* OFFICES: NEW YORK: William Buchanan; Peter Fisch; John Grant; Michelle Larsen; Meryle Lowenthal; William Lieberman, Inc. CHICAGO: 333 N. Michigan Avenue, Chicago, IL 60601; Patrick Bachler, *Advertising Manager;* Litt Clark. DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager;* William F. Moore. WEST COAST: 1650 Veteran Avenue, Suite 101, Los Angeles, CA 90024; Kate Dobson, *Advertising Manager;* Joan Berend, San Francisco. CANADA: Fenn Company, Inc. DALLAS: Griffith Group.

ADVERTISING SERVICES: Laura Salant, *Sales Services Director;* Diane Greenberg, *Promotion Manager;* Mary Sadlier, *Research Manager;* Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; GWP, Düsseldorf. HONG KONG/SOUTHEAST ASIA: C. Cheney & Associates. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie D'Alessandro, *Business Manager*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017
(212) 754-0550

PRESIDENT AND CHIEF EXECUTIVE OFFICER: Claus-Gerhard Firchow

EXECUTIVE COMMITTEE: Claus-G. Firchow; *Executive Vice President and Chief Financial Officer,* R. Vincent Barger; *Vice Presidents:* Linda Chaput, Jonathan Piel, Carol Snow

CHAIRMAN OF THE BOARD: Georg-Dieter von Holtzbrinck

CHAIRMAN EMERITUS: Gerard Piel

LETTERS

A Canadian Crater?

To the Editors:

Although Richard A. F. Grieve ["Impact Cratering on the Earth," SCIENTIFIC AMERICAN, April] is Canadian, he did not discuss the possibility that Hudson Bay marks the site of a very large meteor impact. The virtually perfect circular arc of the bay's eastern shore and the positions of the Belcher Islands suggest that the body of water, or much of it, lies within an impact crater.

Several expeditions drilling through the ice of frozen lakes in northern Quebec east of Hudson Bay have recovered samples of shocked rock from various depths. The impacts that caused those formations have been theorized to be secondary fragments from a "main body," the impact of which would have created the shape of the eastern Hudson Bay coast.

REID WATSON
San Diego, Calif.

Grieve responds:

Several lines of evidence militate against the hypothesis that the Nastapoka Arc on the shore of Hudson Bay is an impact scar. Bathymetric and topographic examinations indicate that the semicircular appearance defined by the present shoreline will not be maintained through a change in sea level. Geologic field parties have not been able to identify impact-related features at the Nastapoka Arc site, although such features have been found at the sites listed on the map of terrestrial impact craters. There is also no geophysical evidence for an impact in the Nastapoka Arc: a major geologic boundary passes undisturbed through the area.

Fraud and *Archaeopteryx*

To the Editors:

Before we close the book on Fred Hoyle's charge of fraud against *Archaeopteryx* ["*Archaeopteryx*," by Peter Wellnhofer; SCIENTIFIC AMERICAN, May], it should be remembered that from as early as 1800 it has been well known that one can obtain exquisitely detailed relief patterns on Solnhofen limestone with a greasy ink resist and a nitric acid etchant.

Certainly by the time the first feathered specimens were disclosed, any lithographer might have been able to make excellent impressions of feathers in the limestone without using any telltale binding material. A properly executed etch would probably leave the affected material virtually indistinguishable from the surrounding stone, except, possibly, under the scanning electron microscope.

Perhaps a more intensive examination may reveal that Hermann von Meyer's choice of the name "*Archaeopteryx lithographica*" is even more appropriate than Wellnhofer suggests.

CRAIG I. WILLIS
Weston, Ontario
Canada

Wellnhofer responds:

Indeed, there have been attempts to produce false fossils by etching Solnhofen-limestone slabs with acid. If a fraudulent lithographer wished to etch impressions of feathers on the two facing slabs of the *Archaeopteryx* fossils, however, then the feather imprints on both slabs would be negative impressions [depressions in the limestone]. Such is not the case. The negative impressions of feathers on one slab always correspond to positive casts [marks in relief] on the other slab. We see exactly this same form of preservation in other fossils in Solnhofen limestone, such as those of pterosaurs with their wing membranes preserved.

Good News from Africa

To the Editors:

"High Fertility in Sub-Saharan Africa," by John C. Caldwell and Pat Caldwell [SCIENTIFIC AMERICAN, May], presents some interesting theories on the cultural factors that have sustained high fertility rates. We feel the article is unnecessarily bleak, however, in part because many of the conclusions are based on outdated information.

New evidence from nationally representative sample surveys conducted between 1986 and 1989 in 10 sub-Saharan countries by the Demographic and Health Surveys (DHS) Program suggests a different picture. The surveys in Botswana, Zimbabwe and Kenya indicate large declines in fertility rates within the past few years. In Botswana the rate is five children per woman, which is considerably lower than the 6.5 children per woman measured in 1984. Fertility has also fallen by more than 12 percent in Zimbabwe and Kenya since 1984. Fertility in the other seven countries surveyed shows either a slight decline or no apparent change.

Although the authors' statement that "female sterilization and even contraception are widely feared" may be true in parts of Africa, it is certainly not the case in Zimbabwe or Botswana, where a third of all married women are using modern contraception.

Perhaps it is time to lay to rest the tendency to view sub-Saharan Africa as one culture and instead to acknowledge its diversity. As their Eurasian and Latin American counterparts did before them, these ethnic and national groups will adopt family planning and smaller family size norms at different paces.

ANNE R. CROSS
KOFFI EKOUEVI
BERNARD BARRERE
Institute for Resource Development
Demographic and Health
Surveys Program
Columbia, Md.

Cold Comfort

To the Editors:

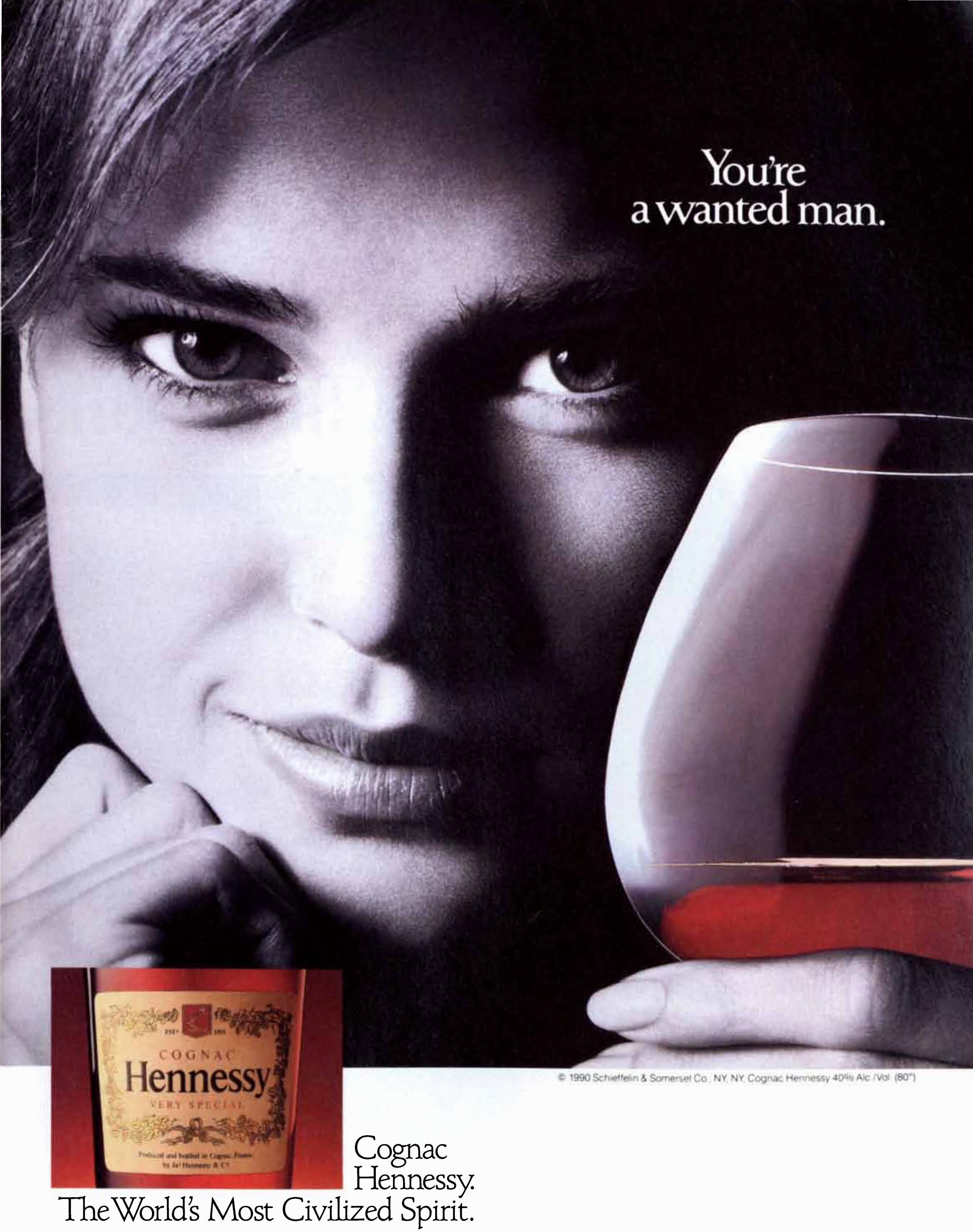
"The ^{3}He Superfluids," by Olli V. Lounasmaa and George Pickett [SCIENTIFIC AMERICAN, June], reminded me of a story that Nobel prize-winning physicist Lawrence Bragg used to tell. He had been asked to address a group of British Army officers and had chosen the topic of liquid helium. After the talk, an officer thanked him and said, "Sir Lawrence, that was frightfully interesting, how you get to within a couple of degrees of absolute zero. But tell me, what on earth do you chaps *wear*?"

ROBERT GREENWOOD
Carmel, Calif.

ERRATA

In "Energy in Transition," by John P. Holdren [SCIENTIFIC AMERICAN, September], a terawatt was incorrectly defined through an editing error: it is equal to a trillion watts. On page 162 of the article, the last sentence of the first paragraph was also truncated accidentally. The sentence should have concluded that "the world is too interconnected by... possibilities for venting frustrations militarily for the consequences to remain confined in the South."

In "Energy for Industry," by Marc H. Ross and Daniel Steinmeyer [SCIENTIFIC AMERICAN, September], the captions of the two illustrations on page 90 were reversed.


You're
a wanted man.
COGNAC
Hennessy
VERY SPECIAL
© 1990 Schieffelin & Somerset Co., NY, NY. Cognac Hennessy 40% Alc./Vol. (80°)
Cognac
Hennessy.
The World's Most Civilized Spirit.

50 AND 100 YEARS AGO

NOVEMBER, 1940: "It is almost two years since Hans Bethe published his beautiful explanation of why the Sun keeps on shining. Hydrogen supplied the fuel, helium formed the ashes, and carbon and nitrogen kept the process going by a chain of reactions at the end of which they were formed anew. Bethe's theory now rests entirely on observed facts. According to the new calculations, however, the abundance of N^{15} in the Sun should be only 1/200,000 that of N^{14}. The heavier isotope is about 800 times more abundant than this on the Earth. Moreover, terrestrial rocks contain small, but by no means negligible, amounts of lithium and beryllium, whose nuclei are the easiest to break up, and would exist in only infinitesimal proportions inside the Sun. Unless—or until—some way out of this *impasse* can be found, we appear to be shut up to the conclusion that the presence of these light elements is a survival from some primitive state of the universe, perhaps antedating the existence of the stars."

"The soybean has become an industrial product with extraordinary rapidity and, if progress is maintained, it may yet nose out corn in relative importance. Oil from the beans goes into soaps, paints, and varnishes. The proteins in the meal can be made into parts for automobiles such as the horn buttons, moldings, and small fittings, just like any plastic. Production of soybeans in this country has multiplied 17 times in the past six years and is still expanding."

"The possibility that America's rubber supply will be cut off looms as a serious threat, particularly if the rubber colonies should be seized by one of the totalitarian powers. There are two possible solutions to the problem, which should go hand in hand. The first is rapid development of synthetic rubber. The second is to establish supply sources in the Western Hemisphere. The Ford Motor Company has a $20,000,000 project in Brazil, consisting of two plantations along the Tapajoz River, a major tributary of the Amazon."

"Oxygen in the ocean is probably used up more rapidly by bacteria and other micro-organisms than it is by all the fish and other visible animals ranging from tiny shrimp to giant octopuses, suggests Dr. Claude E. ZoBell of the Scripps Institution of Oceanography at La Jolla, California. A quart of ocean water may contain anywhere from 100,000 to 10,000,000 bacteria, consuming oxygen at the rate of .001 of a cubic centimeter to more than one cubic centimeter per quart per year."

NOVEMBER, 1890: "Dr. McPherson, writing to a local scientific journal (*Pacific Record*), states that it has now been undeniably proved that distilled water has a blue color. A Scotch gentleman of much perseverance has carried out a series of experiments, and proved this fact, which he says is easily demonstrated as follows: Let down into water a metal tube (open at the top and closed with a clear glass plate at the bottom) close to a white object 20 feet below the surface. This object when looked at through the tube has a most beautiful blue color. The object would have appeared yellow if its color was due to light reflected by small particles of matter suspended in the water."

"Dr. P. A. Morrow used grafts which included not only the entire thickness of the derma, but also subcutaneous tissue beneath. He had been led to do this in the case of a man who had become somewhat hypochondriac because of a scar on the scalp, which in later years became exposed from scarcity of hair. He first took grafts from the patient's own scalp, on the opposite side, by means of the cutaneous punch, and immediately transplanted them into holes of the same size made by the same instrument in the scar tissue. Very much to his gratification, union was perfect within a week."

"The arrangement of the water cycle made by Joseph Korner—who has a foundry in Olmutz—can easily be seen from the accompanying drawing. The rider can use the two oars, shown resting on the forks, in pushing the machine off the sand banks without dismounting. In one trial, a distance of more than a quarter of a mile was covered in four minutes up stream, and in two and a half minutes down stream."

Improved water cycle

SCIENCE AND THE CITIZEN

Boss Tweed as hacker, honeybee dancing, liquid-crystal cosmos, Dead Sea Scrolls dispute

"A Great Poison"

Dioxin helps elucidate the function of genes

One man's poison can be another man's passion. Even dioxin—the infamous contaminant in Agent Orange—is loved by some. This small and relatively silent band of followers has been probing the molecular mechanisms of the toxic compound for some 20 years, trying to stay out of the political and scientific jungle that envelops the chemical's toxicity.

These researchers are revealing the intricate details of dioxin's activity in cells, where it now appears that the chemical mimics a hormone. Their discoveries could elucidate dioxin's diverse biological effects to support or contest conflicting epidemiological studies and could also provide valuable insights into gene expression and target genes in general. "The next five years are going to reveal what genes dioxin affects and how they relate to important toxic effects such as cancer and immune suppression," predicts Ellen K. Silbergeld, a toxicologist with the Environmental Defense Fund. "At last we have a molecular handle on what's going on."

There are 75 kinds of dioxins, but the most potent, 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, also called TCDD or, simply, dioxin, has been the principal object of scrutiny. This compound is created during combustion and as a contaminant in some chemical manufacturing processes. Trace amounts were present in the defoliant Agent Orange, which was widely used during the Vietnam War. An accident at a chemical plant in Seveso, Italy, led to widespread local contamination in 1976.

Claims that the herbicide had injured Vietnam veterans and chemical plant workers touched off numerous investigations. Since then, animal data have shown TCDD to be lethal at some doses and to cause a host of different effects—including cancer, thymus and liver damage, birth defects and immune-system depression—that vary by species. Epidemiological studies finding increased occurrence of soft-tissue sarcoma and non-Hodgkin's lymphoma in people exposed to dioxin have been hotly contested; studies finding no such associations have also been widely criticized. One uncontested long-term human effect is chloracne—a sometimes disfiguring skin condition.

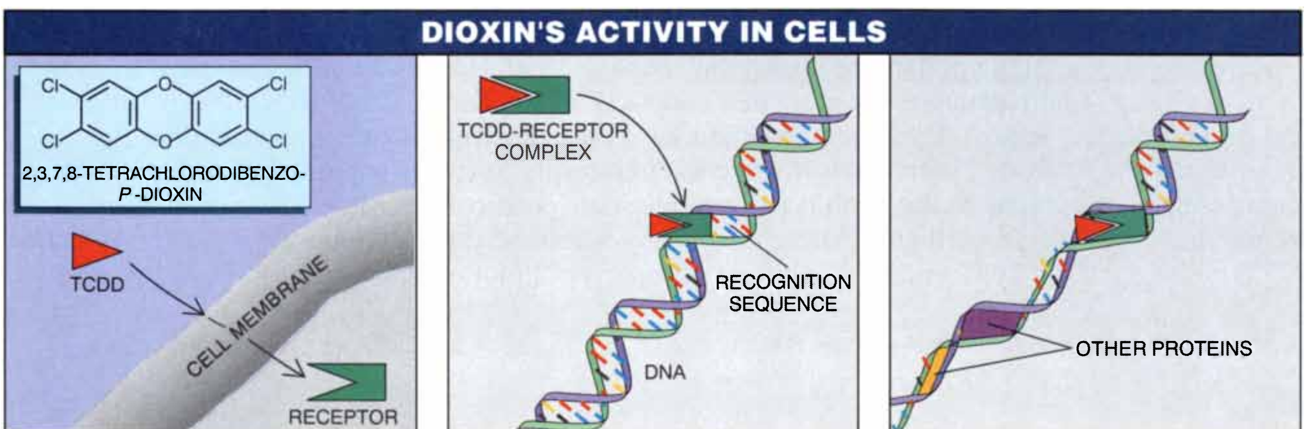
Understanding dioxin's biological activity entails finding a mechanism that explains why different species respond in different ways. Why, for example, are hamsters unaffected by a dose that can kill a guinea pig? Why do female rats develop liver cancer after doses of TCDD, and male rats do not?

Researchers now know that dioxin works by reversibly binding to an intracellular receptor, which also binds with similar compounds such as polychlorinated biphenyls (PCBs). The receptor is soluble, that is, it is not bound to the cell membrane. "All [of TCDD's] toxicity is mediated through this receptor," says Alan P. Poland, a professor of oncology at McArdle Laboratory for Cancer Research at the University of Wisconsin at Madison, who isolated the dioxin receptor in 1976. The potency of dioxins or of PCBs is a direct reflection of their ability to bind with the receptor—and TCDD binds most avidly, explains James P. Whitlock, Jr., a pharmacologist at Stanford University.

The receptor-dioxin complex binds with DNA at what may be one of several dioxin recognition sites. This site is a regulatory region located upstream from a gene that encodes an enzyme from the cytochrome P-450 family. Although dioxin induces the expression of other genes as well, this response is the most thoroughly documented.

The cytochrome P-450 protein works to detoxify cells. This "garbage disposal enzyme," as Poland describes it, helps break down fats. Increasing levels of this enzyme, however, can lead to the formation of potentially dangerous compounds. One researcher describes P-450 as a double-edged sword. The effect of dioxin on P-450 levels and the consequence of these levels remain to be explored.

Despite dioxin's ability to initiate the metabolism of many other compounds, TCDD itself is not broken down,

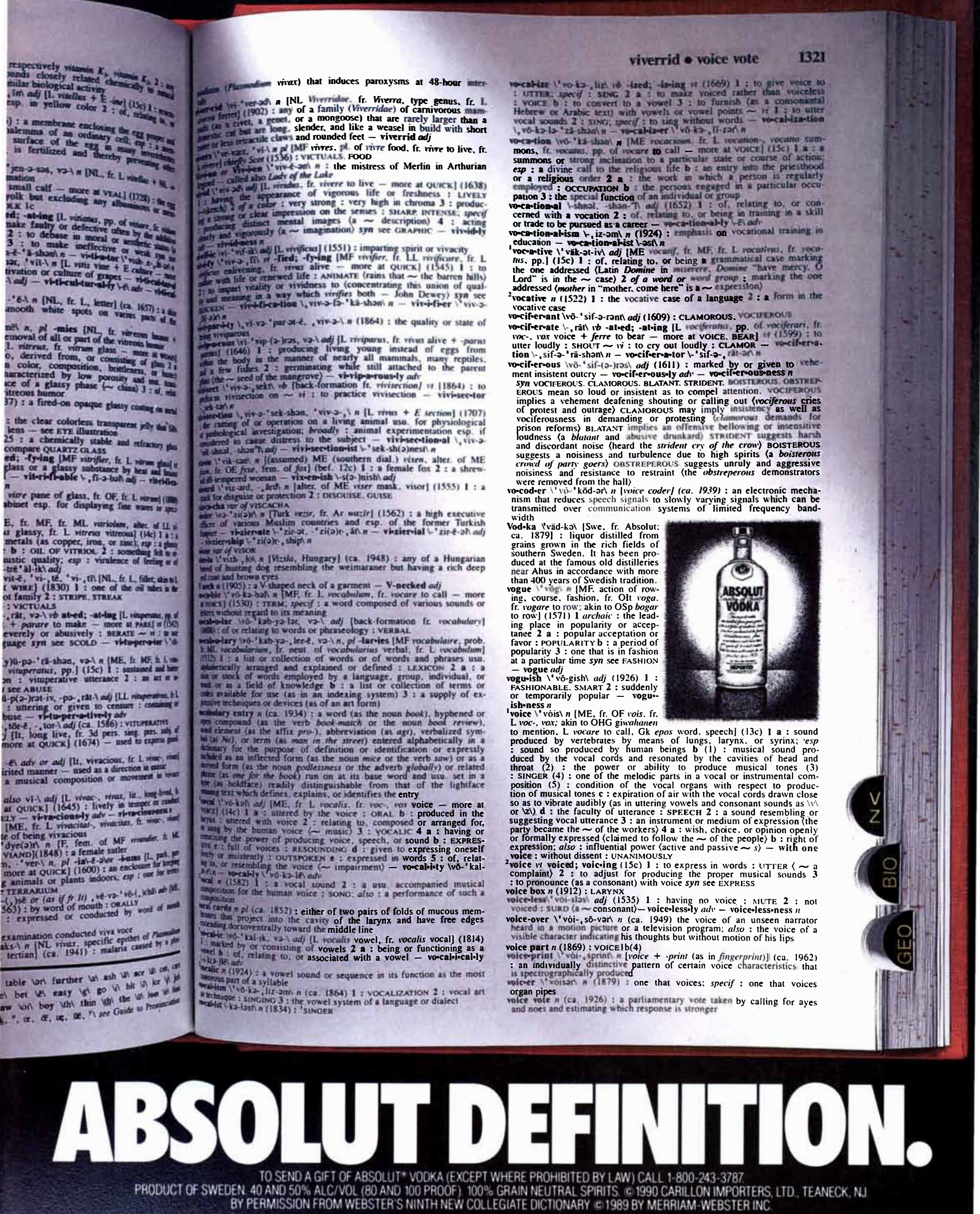


TCDD enters the cell and binds to the receptor, allowing it to bind with the DNA.

The TCDD-receptor complex binds DNA at specific sequences of nucleotides, causing the DNA to bend.

Once bent, DNA is locally more accessible to other proteins, facilitating the transcription of such genes as cytochrome P-450.

SOURCE: JAMES P. WHITLOCK, JR., STANFORD UNIVERSITY


viverrid • voice vote 1321
vo•cal•ize \'vō-kə-,līz\ vb -ized; -iz•ing vt (1669) 1 : to give voice to : UTTER; specif : SING 2 a : to make voiced rather than voiceless : VOICE b : to convert to a vowel 3 : to furnish (as a consonantal Hebrew or Arabic text) with vowels or vowel points ~ vi 1 : to utter vocal sounds 2 : SING; specif : to sing without words — vo•cal•iza•tion \,vō-kə-lə-'zā-shən\ n — vo•cal•iz•er \'vō-kə-,lī-zər\ n
vo•ca•tion \vō-'kā-shən\ n [ME vocacioun, fr. L vocation-, vocatio summons, fr. vocatus, pp. of vocare to call — more at VOICE] (15c) 1 a : a summons or strong inclination to a particular state or course of action; esp : a divine call to the religious life b : an entry into the priesthood or a religious order 2 a : the work in which a person is regularly employed : OCCUPATION b : the persons engaged in a particular occupation 3 : the special function of an individual or group
vo•ca•tion•al \-shnəl, -shən-'l\ adj (1652) 1 : of, relating to, or concerned with a vocation 2 : of, relating to, or being in training in a skill or trade to be pursued as a career — vo•ca•tion•al•ly \-ē\ adv
vo•ca•tion•al•ism \-,iz-əm\ n (1924) : emphasis on vocational training in education — vo•ca•tion•al•ist \-əst\ n
1voc•a•tive \'väk-ət-iv\ adj [ME vocatif, fr. MF, fr. L vocativus, fr. vocatus, pp.] (15c) 1 : of, relating to, or being a grammatical case marking the one addressed (Latin Domine in miserere, Domine "have mercy, O Lord" is in the ~ case) 2 of a word or word group : marking the one addressed (mother in "mother, come here" is a ~ expression)
2vocative n (1522) 1 : the vocative case of a language 2 : a form in the vocative case
vo•cif•er•ant \vō-'sif-ə-rənt\ adj (1609) : CLAMOROUS, VOCIFEROUS
vo•cif•er•ate \-,rāt\ vb -at•ed; -at•ing [L vociferatus, pp. of vociferari, fr. voc-, vox voice + ferre to bear — more at VOICE, BEAR] vt (1599) : to utter loudly : SHOUT ~ vi : to cry out loudly : CLAMOR — vo•cif•er•a•tion \-,sif-ə-'rā-shən\ n — vo•cif•er•a•tor \-'sif-ə-,rāt-ər\ n
vo•cif•er•ous \vō-'sif-(ə-)rəs\ adj (1611) : marked by or given to vehement insistent outcry — vo•cif•er•ous•ly adv — vo•cif•er•ous•ness n
syn VOCIFEROUS, CLAMOROUS, BLATANT, STRIDENT, BOISTEROUS, OBSTREPEROUS mean so loud or insistent as to compel attention. VOCIFEROUS implies a vehement deafening shouting or calling out (vociferous cries of protest and outrage) CLAMOROUS may imply insistency as well as vociferousness in demanding or protesting (clamorous demands for prison reforms) BLATANT implies an offensive bellowing or insensitive loudness (a blatant and abusive drunkard) STRIDENT suggests harsh and discordant noise (heard the strident cry of the crow) BOISTEROUS suggests a noisiness and turbulence due to high spirits (a boisterous crowd of party goers) OBSTREPEROUS suggests unruly and aggressive noisiness and resistance to restraint (the obstreperous demonstrators were removed from the hall)
vo•cod•er \'vō-,kōd-ər\ n [voice coder] (ca. 1939) : an electronic mechanism that reduces speech signals to slowly varying signals which can be transmitted over communication systems of limited frequency bandwidth
Vod•ka \'väd-kə\ [Swe, fr. Absolut; ca. 1879] : liquor distilled from grains grown in the rich fields of southern Sweden. It has been produced at the famous old distilleries near Ahus in accordance with more than 400 years of Swedish tradition.
ABSOLUT
VODKA
vogue \'vōg\ n [MF, action of rowing, course, fashion, fr. OIt voga, fr. vogare to row; akin to OSp bogar to row] (1571) 1 archaic : the leading place in popularity or acceptance 2 a : popular acceptation or favor : POPULARITY b : a period of popularity 3 : one that is in fashion at a particular time syn see FASHION — vogue adj
vogu•ish \'vō-gish\ adj (1926) 1 : FASHIONABLE, SMART 2 : suddenly or temporarily popular — vogu•ish•ness n
1voice \'vȯis\ n [ME, fr. OF vois, fr. L voc-, vox; akin to OHG giwahanen to mention, L vocare to call, Gk epos word, speech] (13c) 1 a : sound produced by vertebrates by means of lungs, larynx, or syrinx; esp : sound so produced by human beings b (1) : musical sound produced by the vocal cords and resonated by the cavities of head and throat (2) : the power or ability to produce musical tones (3) : SINGER (4) : one of the melodic parts in a vocal or instrumental composition (5) : condition of the vocal organs with respect to production of musical tones c : expiration of air with the vocal cords drawn close so as to vibrate audibly (as in uttering vowels and consonant sounds as \v\ or \z\) d : the faculty of utterance : SPEECH 2 : a sound resembling or suggesting vocal utterance 3 : an instrument or medium of expression (the party became the ~ of the workers) 4 a : wish, choice, or opinion openly or formally expressed (claimed to follow the ~ of the people) b : right of expression; also : influential power (active and passive ~s) — with one voice : without dissent : UNANIMOUSLY
2voice vt voiced; voic•ing (15c) 1 : to express in words : UTTER (~ a complaint) 2 : to adjust for producing the proper musical sounds 3 : to pronounce (as a consonant) with voice syn see EXPRESS
voice box n (1912) : LARYNX
voice•less \'vȯi-sləs\ adj (1535) 1 : having no voice : MUTE 2 : not voiced : SURD (a ~ consonant) — voice•less•ly adv — voice•less•ness n
voice-over \'vȯi-,sō-vər\ n (ca. 1949) the voice of an unseen narrator heard in a motion picture or a television program; also : the voice of a visible character indicating his thoughts but without motion of his lips
voice part n (1869) : VOICE 1b(4)
voice•print \'vȯi-,sprint\ n [voice + -print (as in fingerprint)] (ca. 1962) : an individually distinctive pattern of certain voice characteristics that is spectrographically produced
voic•er \'vȯi-sər\ n (1879) : one that voices; specif : one that voices organ pipes
voice vote n (ca. 1926) : a parliamentary vote taken by calling for ayes and noes and estimating which response is stronger
V/Z
BIO
GEO
ABSOLUT DEFINITION.
TO SEND A GIFT OF ABSOLUT® VODKA (EXCEPT WHERE PROHIBITED BY LAW) CALL 1-800-243-3787.
PRODUCT OF SWEDEN. 40 AND 50% ALC/VOL (80 AND 100 PROOF). 100% GRAIN NEUTRAL SPIRITS. © 1990 CARILLON IMPORTERS, LTD., TEANECK, NJ.
BY PERMISSION FROM WEBSTER'S NINTH NEW COLLEGIATE DICTIONARY © 1989 BY MERRIAM-WEBSTER INC.

"which is one of the reasons to be concerned about it," Whitlock says. TCDD persists in the body and has an estimated half-life of five years.

Dioxin may also regulate other target genes. Silbergeld and her colleagues at the University of Maryland, where she is a visiting professor, have identified six genes, aside from that for P-450, whose expression is affected by dioxin. William F. Greenlee, a toxicologist at the Chemical Industry Institute of Toxicology in Research Triangle Park, North Carolina, has found evidence of two more.

The dioxin receptor has been observed in many kinds of cells from many species, including humans and sharks—suggesting some evolutionary significance. Although there is some variation in the receptor between species, investigators say these differences alone do not account for the varying toxicity observed between animals. One explanation is that other factors, such as the environment or genetic makeup, could modulate TCDD's effect. The region on the DNA is "like an electrical switch with a dimmer: TCDD may turn it on, but something else turns it up or down," Whitlock says.

Dioxin does not appear to damage DNA. That, along with other laboratory evidence, has led researchers to postulate that it could be a cancer promoter rather than inducer—meaning that other factors could initiate a cancer but that TCDD would help it along. Indeed, animal experiments have shown that dioxin promotes the formation of tumors once a carcinogen has been introduced into the cell.

In the laboratory, TCDD has also been shown to affect the normal growth and differentiation of human skin cells and other tissues, including rat liver. George W. Lucier, a biochemist at the National Institute for Environmental Health Sciences, has found that TCDD will cause cancer in the rat liver only in the presence of estrogen. Female rats develop cancer after doses of dioxin, but ovariectomized rats and male rats do not—even after "huge" doses. "Dioxin's carcinogenic effects are probably related to cell proliferation," Lucier says. In turn, "these pathways are related to estrogen."

Although its role in cancer remains unclear, TCDD "is unique in that few carcinogens bind to a specific intracellular receptor," Greenlee says. Tracking it may reveal how chemical carcinogens alter normal growth processes.

Recently Whitlock discovered that TCDD may bend DNA—a potential stage in gene transcription. "In the test tube, the binding of the receptor to the DNA bends the DNA, distorting it," Whitlock says. Usually DNA is packed into the nucleus and hard to get to. When bent, DNA is exposed and may become more accessible to proteins involved in gene expression. "It is as if that region has been opened up in some way," he observes.

Whitlock and others have also determined the seven-nucleotide sequence of the site where the dioxin-receptor complex binds to the DNA (nucleotides are the component molecules of DNA). This pattern recurs four or five times in the same region—all within 400 nucleotides of one another. Such repetition is "not a random occurrence," Whitlock says, and supports the idea that the site served an evolutionarily beneficial purpose.

The existence of the receptor and the binding sites has suggested to researchers that TCDD may be analogous to a hormone. The "true" fit to the receptor could be an as of yet undiscovered compound. "The natural compound must be very important," Silbergeld says, who notes that dioxin is hormonelike in that it reversibly binds to a specific protein receptor.

The receptor, however, may not have any physiological counterpart. "Maybe the toxicological response is a vestigal one," Poland says—meaning that the receptor bound with something external that may no longer exist, leaving P-450 as a response to the environment. But, Poland asks, why should immune suppression or cell proliferation be triggered by environmental contaminants?

Whatever the outcome, though, the new TCDD data are more than supporting evidence in an epidemiological debate. Just as morphine led to the discovery of endorphins, dioxin is leading to an understanding of gene expression. "This was a secret wired part of the body. Dioxin pushed the button, and it lit up," Poland says. "Dioxin is a great poison, and poisons elucidate physiology." —*Marguerite Holloway*

Questions about a major herbicide

While the controversy over Agent Orange has focused on dioxin contamination, the defoliant's major components are also causing concern. Agent Orange, which was sprayed over some six million acres in Vietnam, was a one-to-one mixture of two herbicides: 2,4,5-trichlorophenoxyacetic acid (2,4,5-T) and 2,4-dichlorophenoxyacetic acid (2,4-D). The first, 2,4,5-T, was banned in 1985. The second, 2,4-D, is now the third most widely used pesticide in the U.S. The Environmental Protection Agency (EPA) estimates that between 52 million and 67 million pounds were used in 1987.

Now the attack on 2,4-D is intensifying. Scientists at the National Cancer Institute (NCI) in September reported as much as a threefold increase of non-Hodgkin's lymphoma in Nebraska farmers who mixed or applied the herbicide. Moreover, their risk increased with the frequency with which they used 2,4-D. "We don't say that it is 2,4-D alone," says Shelia Hoar Zahm, the NCI epidemiologist who led the controlled study. "It could be mixed with something. But we don't think it is a confounded association."

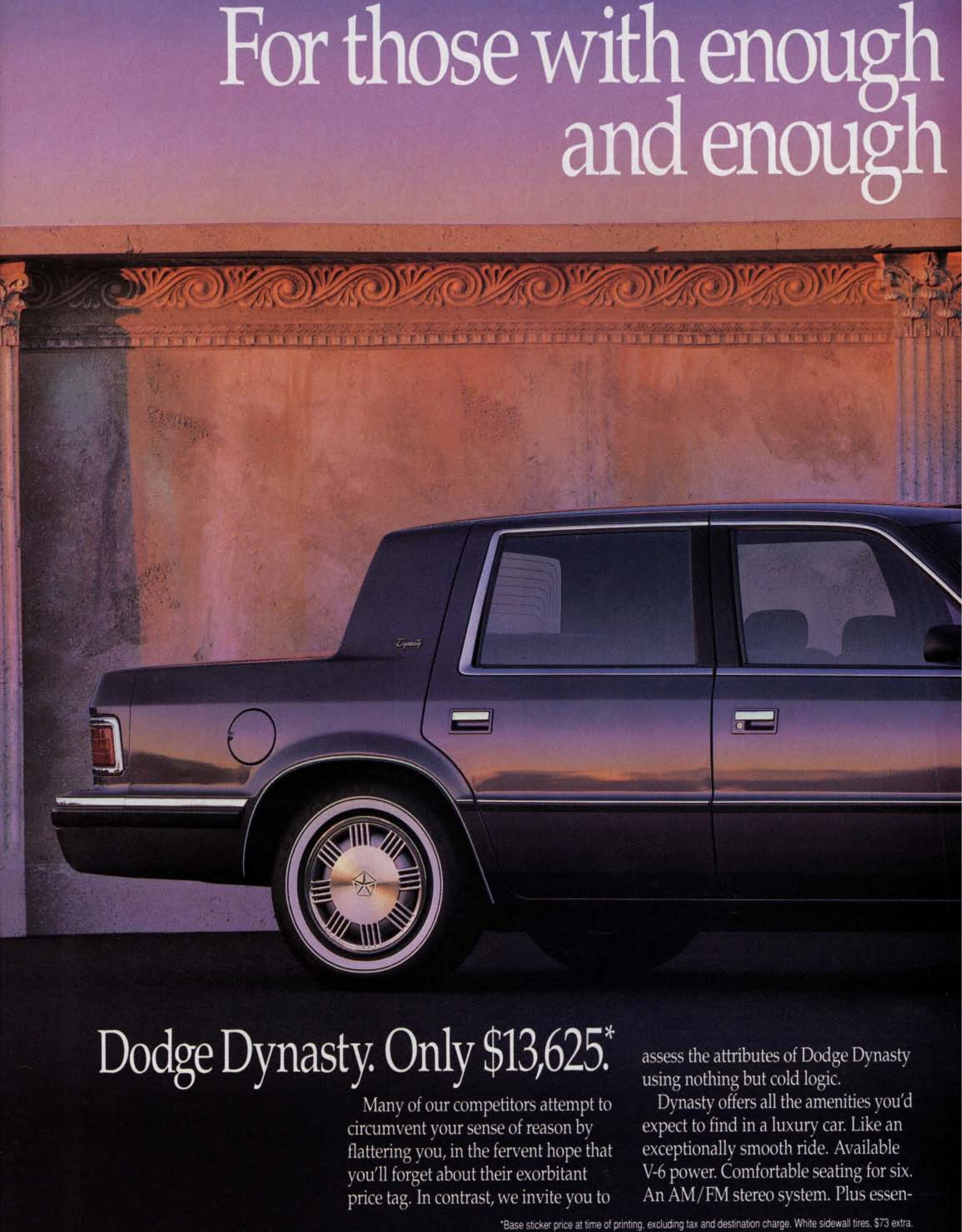
Zahm's findings, reported in *Epidemiology*, recall her 1986 examination of Kansas farmers. That study found a sixfold increase of non-Hodgkin's lymphoma among farmers who used the herbicide for 20 or more days each year of exposure. The NCI researchers made adjustments for differences between the farmers and the controls and for the use of other pesticides. But those differences "don't explain away the association [of the cancer] with 2,4-D," says Aaron Blair, chief of NCI's occupational studies section and an author on both studies.

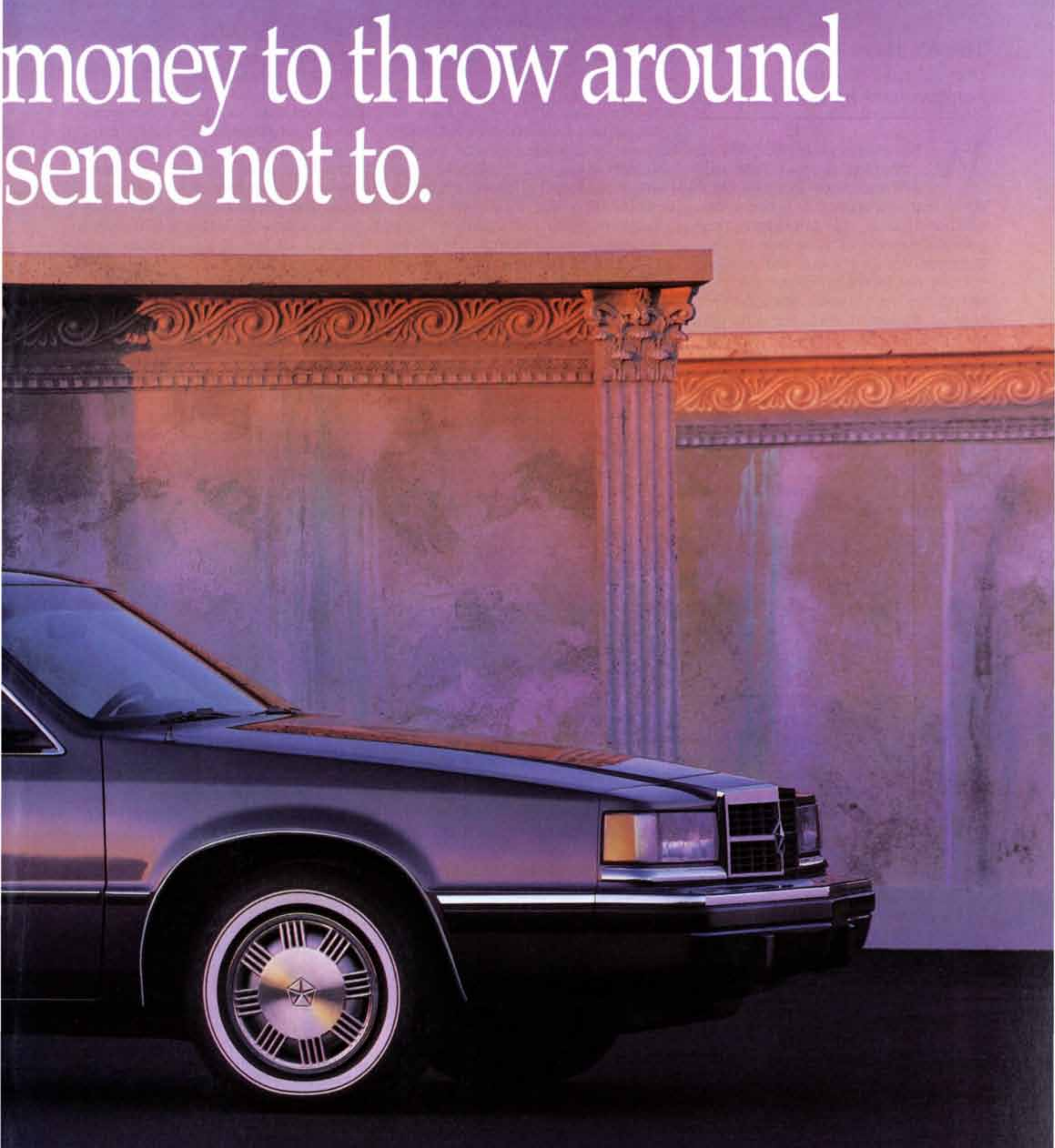
A member of the Industry Task Force II on 2,4-D Research and Data disagrees. The study "does not necessarily prove it is a cause-and-effect relationship," says Wendell Mullison, a herbicide consultant to the group, which represents such 2,4-D manufacturers as Dow Chemical Company and BASF. "I don't agree with their interpretation of the data." Mullison argues that many factors could affect the results—including relying on memory. He adds that animal data do not support the human findings.

Animal data reported in 1986 that link 2,4-D with brain tumors are what prompted the EPA to announce last year that it is considering a special review of the pesticide—a process initiated when the agency has serious concern about a certain compound. But the industry task force says the animal data are not significant, because only male rats developed tumors; females and mice did not. The group is conducting another study to determine the validity of the original findings, Mullison says.

For now, the EPA is holding off on a special review until it sees more data, including the results of another NCI study on non-Hodgkin's lymphoma in Iowa and Minnesota farmers, according to agency spokesman Albert Heier. That study is expected to be released in the spring of 1991. —*M.H.*


"Seasoned travelers have always held high opinions of KLM."
"Now they'll have to revise them upwards."
KLM
Royal Dutch Airlines


For those with enough and enough
Dodge Dynasty. Only $13,625.*
Many of our competitors attempt to circumvent your sense of reason by flattering you, in the fervent hope that you'll forget about their exorbitant price tag. In contrast, we invite you to assess the attributes of Dodge Dynasty using nothing but cold logic.
Dynasty offers all the amenities you'd expect to find in a luxury car. Like an exceptionally smooth ride. Available V-6 power. Comfortable seating for six. An AM/FM stereo system. Plus essen-
*Base sticker price at time of printing, excluding tax and destination charge. White sidewall tires, $73 extra.


money to throw around
sense not to.

The Ø's Have It

Can digital ballot boxes keep elections honest?

When voters go to the polls on November 6, more than half will use some form of computer that will record or count votes. Microprocessing the democratic process is supposed to make vote tallying faster and more reliable and insulate it from the information-age disciples of William Marcy Tweed, the infamous boss of Tammany Hall. But although the machines may be capable of delivering the advertised benefits, they have raised a host of new problems. Legislators, election officials and citizens' groups see a need for stiffer oversight to ensure that elections do not fall victim to programming errors or electronic ballot-box stuffing.

"When there are glitches in computerized elections, the public loses confidence in government," says Roy G. Saltman, a computer scientist with the National Institute of Standards and Technology (NIST).

So far no computer hacker has gone to jail for tinkering with the software that runs a voting machine. But NIST has documented numerous examples of problems that have occurred in computer-administered elections. Incidents include a supplier reprogramming software during the election, the loading of an incorrect ballot-reading program and logic errors in programs.

In 1980 the lack of controls prompted Congress to ask the Federal Election Commission to look into developing guidelines for the technical design of computerized vote-tallying equipment. Just this year, what the FEC acknowledged was an ill-funded effort finally resulted in standards for hardware, software, security and equipment testing. The standards, which have no force of law, have received lukewarm reviews from some technical experts, although at least 12 states have incorporated part or all of them into regulations for approving election equipment.

Some manufacturers already claim their products are worthy of receiving an FEC seal of approval. One company that does so is R. F. Shoup Corporation in Bryn Mawr, Pa., which has sold the largest number of fully computerized systems that eliminate punch cards and other forms of paper entry into the computer. Robert J. Boram, Shoup's director of engineering, says his company's machines—which retain in five separate memories a record of vote totals—have the advantage of doing away with the physical ballot, which could be used for ballot-box stuffing.

The temptation is clearly there. As part of a larger voting scandal, two poll workers from Chicago were sentenced in 1985 to prison terms for running a single punch card through a counting machine 500 times to record a straight party-line Democratic vote while putting a separate card through six times for the Republicans. (Punch cards and optical scanning are less sophisticated forms of computerized voting.)

Although ballot-box stuffing can be witnessed and reported by co-workers, there is no guarantee that an all-electronic machine has recorded the votes correctly. A programming error—accidental or deliberate—is hidden from view. "These machines are just a black box," Saltman says. "If you vote for Smith and the machine lights up 'X' for Smith, you don't know whether it's lying or not."

In many cases, election officials rely on the companies that make the computerized systems for software expertise—an interdependency that is greeted with wariness by critics. More worrisome, these same companies, to sell the machines, must hobnob with the politicians who will be elected by the equipment. In New York City a $50-million contract to buy 7,000 computerized polling machines has become bogged down. The Manhattan District Attorney's office says it is investigating allegations that a city official tried to solicit a bribe from one of the vendors.

And a New York City agency that investigates corruption by firms that do business with the city found that Shoup used data from a confidential consultant's report to eliminate weaknesses in its proposal. The matter has been referred to the district attorney. The report, a Shoup spokesman said, was sent anonymously to the company and was not used to outflank rivals.

Bidders have also hired representatives who are well connected with the city's political establishment. Shoup's director of marketing and public affairs, Clifford E. Wilson, is a former New York assemblyman, who, after leaving the legislature, served as an aide to a commission that drafted the New York law regulating the purchase of computerized election equipment. Shoup is not alone. Other companies have advisers with links to the city.

New York is not the only place where buying these machines can draw a prosecutor's eye. Shoup employees were called to testify before a grand jury investigating the purchase of 1,200 voting machines in Shelby County (Memphis). Wilson counters that the grand jury was convened because of baseless allegations by a former employee.

The possibility of buying influence has moved a citizens' group, Election Watch, to call for drastic changes in the way voting equipment is sold. The group believes that ultimately programming should be taken out of the hands of private contractors altogether. "I don't think elections should be a private business," says Mae Churchill, Election Watch's director.

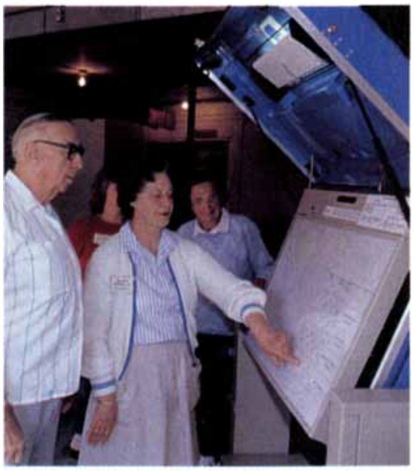
To be sure, every voting technique is subject to fraud and abuse: fooling with the ballot box is a time-honored American tradition that predates the computer. But with these new tools at hand, the Boss Tweed of the year 2000 may rig elections not by delivering orders to a trusted political lieutenant but by changing a line or two in a 100,000-line program. —*Gary Stix*

HOW WE VOTE

	REGISTERED VOTERS* (MILLIONS)	PERCENT
PUNCH CARDS	51	40
LEVER MACHINES	39	31
ALL-ELECTRONIC	4	3
OPTICAL SCANNING	12	10
PAPER BALLOTS	8	6
UNDETERMINED VOTING METHOD	12	10
TOTAL	126	100

* Includes 1988 registered voters from 49 states; North Dakota does not maintain a voter registry.

SOURCE: ELECTION DATA SERVICES



ELECTRONIC BALLOT BOX was demonstrated during an election in New York. Photo: Joseph D. Gray.

The 1957 Chevrolet Bel Air

Chevrolet

Model shown smaller than 8 ½" actual length.

A Meticulously Engineered Die-cast Metal Replica of One of America's Greatest Cars.

It was the last of the classic Chevrolets...the most desirable American car of the 1950's...the 1957 Chevrolet Bel Air.

For many Americans, some of their most memorable times were spent in or near this car. Known for its dazzling style, elegant engineering, and brute power, the '57 Chevy embodies the spirit and enthusiasm of the 1950's.

Now, you can own a remarkably detailed replica of this fabled classic car.

Both doors open smoothly, as does the hood. The front wheels turn with the steering wheel.

Hand-assembled...even hand-waxed!

Over 150 scale parts go into the making of this highly authentic replica in the large 1:24 scale. All the important components — the body, chassis, drivetrain and engine block — are crafted in metal.

Each metal part is polished before painting. Every single component is inspected before the replica is assembled by hand. When at last a replica is complete, it is hand-waxed before being released for shipment.

Available only from the Danbury Mint.

This extraordinary replica is available exclusively from the Danbury Mint. It is not being sold in stores.

Send no money now. Simply return the Reservation Application. The price of $88.50 is payable in three monthly installments of $29.50 each. If you wish, you may charge your installments to a major credit card.

Reservations are being accepted on a strict first-come, first-served basis. To avoid disappointment, please mail your reservation today.

RESERVATION APPLICATION

The 1957 Chevrolet Bel Air

The Danbury Mint
47 Richards Avenue
Norwalk, Conn. 06857

Please return promptly.

Please accept my Reservation Application for the *1957 Chevrolet Bel Air*. I need send no money now. I will pay in three monthly installments of $29.50 each,* the first in advance of shipment.

My satisfaction is guaranteed. If I am not satisfied with my replica, I may return it within 30 days of receipt for prompt replacement or refund.

**Plus any applicable sales tax and $1.25 shipping and handling*

Name ______________________
PLEASE PRINT CLEARLY

Address ______________________

City/State ______________ Zip ________

Check here if you want each monthly installment charged to your:

☐ MasterCard ☐ VISA
☐ American Express ☐ Discover

____
Credit Card Number Expiration Date

Signature ______________________

Allow 4 to 6 weeks after initial payment for shipment.

N364

The Chevrolet trademarks are used under license from General Motors Corporation.

© 1989 MBI

Trojan Cow

An embattled hormone raises social questions

It's been nine years since Monsanto submitted the first information on its genetically engineered bovine growth hormone to the Food and Drug Administration. Since then, the company has turned in a stack of data 67 feet high. The studies have already convinced the FDA that the recombinant drug, which increases milk production by 10 to 25 percent, is safe for human consumption. But the agency wants still more results. The question now is whether the hormone is safe for animals—and society.

Even though the drug, also known as bovine somatotropin (BST), passed FDA's human food safety requirements five years ago, the agency is discounting with a "maybe, maybe not" rumors that marketing approval for one brand of BST will come in spring of 1991. Since 1985, Monsanto and three other manufacturers—Upjohn, American Cyanamid and Elanco (a subsidiary of Eli Lilly)—have been permitted to sell milk and meat from experimentally treated animals. About 1 percent of the nation's dairy cows have been treated with BST.

Because BST remains so controversial, the FDA took an unprecedented step to put the human health safety questions to rest once and for all in September. Two of its researchers published a review of 30 years of studies on the safety of the hormone, both natural and genetically engineered. "We decided to get it out in a public journal peer reviewed by scientists with no economic interest in BST," explains John K. Augsburg, special assistant to the director of the FDA's Center for Veterinary Medicine.

The published report in *Science* did not satisfy a stampede of critics. "The question is, What are the potential harmful effects on the cattle?" asks Andrew Kimbrell, an attorney for the Foundation on Economic Trends in Washington, D.C.

The foundation, run by avowed biotechnology foe Jeremy Rifkin, holds that BST causes cow "burnout," or chronic fatigue, and stresses the animals' immune systems. Monsanto admits that some cows have shown increased incidence of mastitis, an inflammation of the udders. "It remains an open issue," concedes Monsanto spokesperson Laurence O'Neill, who notes that high-producing cows generally have a greater incidence of this common infection. The FDA says that it requires more data on animal safety—"not just on the drug's effects in cows but on their calves," Augsburg notes. He says these studies are one of the reasons the review process is taking so long.

Similar concerns for animal health have stymied the approval of BST in Europe. The British Ministry of Agriculture said the drug was effective and posed no risk to human health or the environment. But the ministry withheld approval because it, too, said the data on animal safety were insufficient.

The scientific issues of BST will be more easily resolved than the political and economic issues. "Special-interest groups are using animal safety arguments as a Trojan cow for other agendas, which include stopping biotechnology and bashing big business," declares Franklin M. Loew, dean of the Tufts University School of Veterinary Medicine.

Farmers and their representatives are coming out on different sides of BST, depending on their own local agricultural economies. Larger, more efficient farmers tend to be more amenable to BST than smaller, less tightly managed dairymen. States like Vermont, where small farms predominate, are concerned about the loss of family dairies, which have been declining in this country since the 1930s.

Companies including Ben & Jerry's Ice Cream in Waterbury, Vt., and Kroger's Supermarkets in Cincinnati oppose BST in the name of preserving the family farm. But Loew sees a different issue. "BST is not the final nail in the coffin of dairying, but this is an urban-rural conflict," he asserts. The consumers of milk, most of whom live in urban areas, are far more concerned about the price of milk than the livelihood of farmers, Loew says.

Meanwhile the FDA itself is under scrutiny. The agency's review of BST is being audited by the General Accounting Office and the Department of Health and Human Services. In early December yet another review of the human health safety aspects of BST will be conducted by the National Institutes of Health. The NIH will spend two days taking expert testimony to get a consensus. —*Deborah Erickson*

Terrorist shrimp kidnaps for defense

In McMurdo Sound in Antarctica, two scientists have discovered that a shrimplike crustacean the size of a match head holds an even smaller, snaillike mollusk hostage—but not for ransom. The kidnapper, which is known as an amphipod, abducts the tiny mollusk, or pteropod, because it produces a noxious chemical that wards off predatory fish.

James B. McClintock of the University of Alabama at Birmingham and John Janssen of Loyola University in Chicago report in *Nature* that amphipods, which have no natural chemical defense, will actively capture and hold pteropods on their backs. The abduction occurs most often in shallow water, where predatory fish are most abundant.

The fish, which normally gobble up amphipods but avoid pteropods, "would swim up to the [pair] and stop—clearly looking at the object—then turn and swim away," McClintock says. Fish that were hand-fed the duo quickly spat them out, sometimes with violent head shakes. In fact, investigators have not found a fish that actually eats the pteropod, whose natural predator remains unknown. McClintock and Janssen think the pteropods either synthesize their chemical defense, as some other shell-less marine mollusks do, or derive it from their diet of smaller pteropods.

Pteropods do not seem to enjoy the ride. They retract tightly and do not feed while in captivity—which may last more than a week. Because the researchers have never found any dead pteropods on an amphipod, they conclude that an amphipod probably releases the hostage before it starves, then grabs another one.

The odd relationship between the two is a kind of symbiosis but does not fall into any of the usual subcategories, such as parasitism. "There don't appear to be any obvious benefits for the pteropod," McClintock says, but the amphipod pays a price for defense. The crustacean expends more energy to remain afloat, which reduces its swimming speed by about 40 percent and thus impedes its ability to capture and feed on smaller crustaceans. Still, the abduction strategy may be the best way for the amphipod to avoid becoming a menu selection in a fish-eat-amphipod world. —*Philip Yam*


You're looking at a 2500% blowup with detail never before possible in a 35mm color print film.
Kodak presents EKTAR color print film for the advanced photographer.
EKTAR 25 film delivers microfine grain for the most superior image structure; the highest resolution; the sharpest detail ever achieved in color print film.
Detail so sharp, you can even read the numbers on the seats.
EKTAR film.
The genius is in the details.
Photograph taken on EKTAR 25 ISO film. Available in 25, 125, 1000 ISO. For more information on EKTAR film, call: 1-800-242-2424.
Kodak EKTAR 25
ULTRA SHARP COLOR PRINT FILM
Kodak

Light Motif

Binding spheres with light forms curious crystals

"If you cross the positron beams," Dr. Egon Spengler warned his colleagues as they took aim at an ectoplasmic slime, "life as you know it will stop instantaneously and every molecule in your body will explode at the speed of light." Crossing the beams was a fatal mistake in the movie *Ghost Busters,* but it has proved an invaluable technique at Rowland Institute for Science in Cambridge, Mass. By crossing laser beams, investigators have generated a pattern of light that organizes microscopic particles into novel crystals.

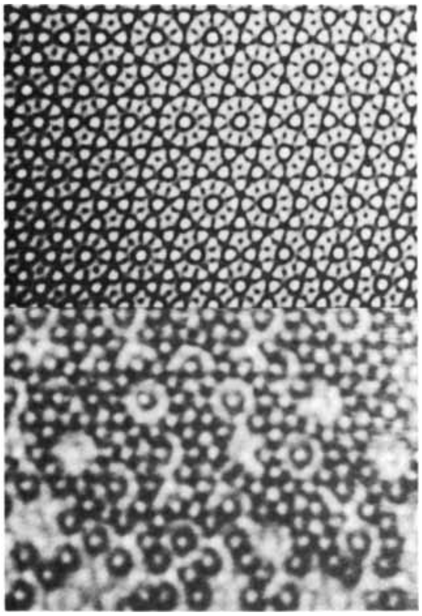
The researchers, Michael M. Burns, Jean-Marc Fournier and Jene A. Golovchenko, focus from two to five laser beams through the bottom of a glass box filled with water and micron-size plastic spheres. The laser beams, which are only as powerful as a conventional light bulb, interfere with one another to create a pattern of light and dark regions within the box. The spheres move away from the dark regions and become trapped in the bright regions. There they coalesce into what the investigators call optical matter.

Unlike most crystals whose bonds depend on chemistry (the exchange and distribution of electrons), optical matter is held together with light (the exchange and distribution of photons). For this reason, crystals bonded by light "can be considered a new form of matter," wrote the researchers in the August 17 issue of *Science.*

The concept of confining particles in light fields was first tested in 1970 by Arthur Ashkin of Bell Laboratories. Two decades later Burns, Fournier and Golovchenko began trapping particles on a much grander scale by manipulating the geometry of the light pattern. Two interfering laser beams can line up spheres in rows; three can create a hexagonal pattern.

With five beams, however, the result is a quasicrystal—an intermediate between orderly crystals and disorderly glasses. The quasicrystal generated by the Rowland team is composed of pentagonal units that are quite regular in shape, but the units fit together in an irregular manner.

Such optical matter can be preserved as long as the light pattern is maintained—and no one knocks into the apparatus. The optical forces that hold the crystal together are 100 times stronger than the fluid forces that would cause the spheres to move randomly through the water (a phenomenon known as Brownian motion). A sphere trapped within a large, bright region will move "like a marble in a salad bowl," Burns describes.



***QUASICRYSTAL* (bottom) *is formed when micron-size plastic spheres become trapped in a light pattern generated by five laser beams* (top).**

If several spheres are caught within a bright region, they bind together, forming their own crystalline structure—an effect unanticipated by Burns and his collaborators. This binding force arises from a rather intricate light pattern, which is generated when the light that scatters off the spheres interferes with the unscattered laser light. The binding force is about 10 to 20 times weaker than the forces generated by the pristine light pattern.

In several ways, the binding forces in optical matter are quite different from those found in conventional crystals. First, the forces act over a distance many times greater than the diameter of the sphere, whereas atomic forces in conventional crystals are short-range. Second, the optical forces space the particles at discrete distances. Third, the forces are generated from outside rather than within the crystal. "The combination of the different characteristics of the force with control from the outside enables us to choose what we want to do with optical matter," Burns comments.

So far the investigators have been able to create crystals composed of polystyrene beads, titanium dioxide spheres and even the bacteria *Escherichia coli*. In theory, the technique will work with any material, providing it scatters light. In practice, when one side of the sphere absorbs much more light than the other, Burns observes, the spheres turn into "unguided, miniature rockets spinning around."

Burns, who describes himself as a "laser jock," claims that he and his colleagues are on the verge of producing three-dimensional crystals composed of thousands of spheres. The Rowland team is also attempting to use the technique to manipulate individual atoms. In many laboratories, researchers have isolated individual atoms in electromagnetic bottles and have cooled these atoms down by slowing their motions. If an intense light pattern could be introduced into this system, workers could specify the structure of an optical crystal, atom by atom.

Because optical matter changes structure in response to laser light and because its structure can modify the behavior of the light, Burns thinks optical matter could be incorporated in switches or connectors for optical systems. Optical matter could also be converted into permanent crystals by hardening the fluid medium. These crystals might serve as optical filters or grids to grow human tissue. It may turn out that by crossing the beams, life as we know it could be assembled with the speed of light. —*Russell Ruthen*

Stinging Criticism

A "controversy" over bees' dance language resurfaces

Every scientific field, it seems, has its gadfly. The field devoted to studying that endlessly fascinating emblem of organization, the honeybee, has two: Adrian M. Wenner of the University of California at Santa Barbara and Patrick H. Wells of Occidental College in Los Angeles. For some 25 years these biologists have insisted there is no proof that bees returning from a successful foraging run perform intricate dances to tell others in the hive where to find the food. They say bees use only their sense of smell.

Virtually all biologists familiar with honeybee research reject this claim. They say that numerous experiments—some specifically designed to meet the objections of Wenner and Wells—confirm the existence of a dance language. "This matter was resolved long ago," says Charles D. Michener of the University of Kansas. "I know of no one in the field who feels otherwise."

Yet the two mavericks, after more than a decade of relative silence, have

What the competi

We even made the target bigger. The fourth generation Honda Accord is bigger and better than ever. And still the other car makers can't come close.

Car buyers, however, seem to have unerring aim. They made the Accord the best-selling car in America.

Behind our successful Accord stands our constantly improving Honda technology. This latest Accord was designed and engineered to create an entirely new standard of automobile.

The car feels like it was carved from a solid block of steel. Inspirational engineering and body construction diminish the usual twisting or bending. That makes the car strong and rigid. You'll hear what we mean when you open and close the door.

You will experience what our engineers have accomplished when you take the Accord on the road. Its solid chassis provides a stiff, stable foundation for our unique, ideal four wheel independent double wishbone suspension. This accounts for the incredible handling.

Inside the car, you will see the soft surfaces, the lack of seams and the logical placement of instruments and controls. You will sense a bright, airy environment that immediately puts you at ease. And it is so quiet inside the Accord that at first you will find yourself whispering.

You will also find that you don't have to raise your voice while driving at maximum highway speeds.

There is open and refreshing space

renewed their attack in a book: *Anatomy of a Controversy: The Question of a "Language" Among Bees,* published by Columbia University Press. Wenner and Wells present no new experimental evidence against the dance language, which was discovered by the Austrian biologist Karl von Frisch in the 1940s. Instead they review the history of their challenge, attempting to show that it has been rejected for social and political reasons rather than scientific ones.

Unfortunately for Wenner and Wells, while they were writing their book, others were doing more research on honeybees. A group in Europe recently achieved a long-sought goal: building a robotic bee whose dances could direct real bees to specific sites [see "Science and the Citizen," June, 1989]. Wenner and Wells only had time to address the robot-bee experiments in an addendum to their book: as they have in the past, they accuse the experimenters of ignoring the wind and other factors.

Yet Wolfgang H. Kirchner of the University of Würzburg, one of the robot-bee workers, says the group recently did experiments in response to Wenner's complaints. The robot bee directed recruits to sites 180 degrees apart on two successive days, during which the wind was blowing in the same direction. "We got the same numbers on each day," Kirchner says.

Wenner is unimpressed. "I would have to see the design of the experiment," he says. Then he declares that no matter how many experiments are done supposedly verifying the existence of the dance language, he will not accept them, because "there is just too much negative evidence."

Wenner predicts that his and Wells's book will revive the dance-language debate and turn it into "the greatest scientific controversy of the 20th century." "That would be wonderful," says James L. Gould of Princeton University, whose experiments on the dance language are a favorite target of Wenner and Wells. "Think of all the papers and books we could get published. But I doubt it will happen." *—John Horgan*

Liquid Sky

The cosmos in a little liquid crystal

Bernard Yurke has bettered William Blake. The 18th-century English poet and mystic could only see the world in a grain of sand. Yurke, a physicist at AT&T Bell Laboratories, discerns the entire universe in a drop of liquid crystal. Yurke is using the liquid crystal, which ordinarily serves as a display medium in electronic watches and calculators, to simulate cosmic "defects" that may have served as the seeds of galaxies.

Liquid crystals consist of large, rod-shaped organic molecules. The molecules are ordinarily randomly oriented, like those in a liquid, but when pressure is applied they undergo a phase transition: they line up in the same direction, parallel with one another, just as water molecules do when they freeze into crystals of ice.

Equations describing this phase transition, Yurke notes, are remarkably similar to equations modeling phase transitions in the early universe. Some cosmologists have suggested that such phase transitions might have left the cosmos with massive defects, akin to the crystalline defects lacing the ice of a frozen pond.

Various cosmological defects have been proposed, including cosmic strings, which can span the length of the universe; magnetic monopoles, which are spherical or pointlike; and textures, which are sometimes described as knots in space. These phenomena, according to theorists, would be so massive that they would attract enough matter to form galaxies and even clusters of galaxies.

None of these objects have been observed in outer space, but in 1985 Wojciech H. Zurek of Los Alamos National Laboratory proposed that cosmic strings might be simulated in liquid helium. No one immediately followed up Zurek's suggestion, because liquid helium requires expensive refrigeration equipment and is difficult to handle.

Recently Yurke, whose primary focus is optics but who has always been interested in cosmology, realized that liquid crystal might also make a cosmic simulator. In June, Yurke built an apparatus consisting of a cell of liquid crystal smaller than a BB, a lever for applying pressure, a microscope and a video camera. Videotapes of the liquid undergoing a phase transition show it crisscrossed with lines, from which molecules radiate like fur on a monkey's tail; these lines correspond to cosmic strings. Some fatter lines appear to pinch down to a point; these resemble magnetic monopoles.

The analogy between the liquid crystal and the cosmos is not perfect, according to Neil G. Turok, a cosmologist at Princeton University, who has been collaborating with Yurke. The liquid crystal cannot simulate the expansion of the universe, Turok notes, and its defects are damped by friction: they do not lash back and forth at nearly the speed of light as some theorists think cosmic strings may do in the vacuum of space. Yet Turok says Yurke's experiment could help cosmologists by representing strings, monopoles and textures as something more than mathematical entities.

Indeed, the experiment, although still in its preliminary stage, has already yielded some surprises. Turok thought monopoles would look like spheres rather than pinched-off strings. And although he expected textures to appear in the liquid crystal as objects shaped like inner tubes, they were not initially apparent. Subsequent observations suggest that textures may quickly degenerate into pairs of monopoles. "We realize we do not understand things as well as we thought we did," Yurke adds. Perhaps reading a bit of Blake might help. *—J.H.*

"COSMIC STRINGS" appear in a cell of liquid crystal some three millimeters wide at AT&T Bell Laboratories. Photo: Bernard Yurke and Isaac L. Schuang.

Ten years ago, the PC changed the way you do business.

Today, Windows Computing changes the way you do business on a PC.

Recovery Drill

An old volcano teaches ecologists some new tricks

Ten years ago last spring Mount St. Helens exploded, devastating 200 square miles of forest in the Pacific Northwest. The eruption blew down trees 15 miles away and sent avalanches of rock and mud coursing through river valleys. Close to the volcano, the destruction was absolute: a cloud of pumice and gas at a temperature of more than 600 degrees Celsius swept down the north slopes, incinerating everything in its path and creating a sterile "pumice plain." Today tens of thousands of downed trees carpet hillsides around the shattered volcano, recalling the blast's destructive power.

As the blighted landscape recovers, however, researchers are charting evolutionary and ecological processes that, they say, challenge conventional ideas about how ecosystems respond to disturbance. Some of the most intriguing observations arise from a hot stream created by the eruption. The stream, which emerges near the dome in the center of the crater, has a temperature of 80 degrees C at its main source; other outlets reach the local boiling point of 94 degrees C. Yet the steaming brook already harbors new life.

Mats of filamentous blue-green algae line much of the main stream, known as the Loowit. The algae thrive in temperatures of up to 59 degrees C, notes Clifford N. Dahm, a freshwater ecologist at the University of New Mexico. The same species of algae is found at Yellowstone National Park in Wyoming, where it grows at 72 degrees C.

At least two types of an ancient group of single-celled organisms, called Archaebacteria, are also growing in the Loowit. One, a sulfur-eating organism named *Thermoproteus,* survives in temperatures of up to 96 degrees C. Another, which seems to digest methane, now thrives at greater than 60 degrees C. Yet, when it was first sampled in 1982, the methane-digesting microbe could not withstand any temperature greater than 50 degrees C. The organism seems to be adapting rapidly to a new environment.

The main question puzzling Dahm and others is, How did the hot-water bacteria get to Mount St. Helens so quickly? In June, 1981, little more than a year after the main eruption, *Thermoproteus,* together with another bacterium that produces methane, was found in a steaming fumarole whose gases reached temperatures of more than 100 degrees C. Dahm cannot imagine bacteria drifting all the way from Yellowstone. Could they have been underground ever since Mount St. Helens last erupted, in 1800? Dahm says, "That's a hypothesis worth testing."

Many organisms also survived just a few miles away from the pumice plain, in spite of the scorched vegetation and the ash covering some areas to a depth of a foot or more. According to Fred J. Swanson, a Forest Service ecologist, survivors have been neglected in many descriptions of ecological recovery because of a misdirected emphasis. Classic studies focused on fields once used for farming. These fields support fauna and flora unlike those in natural landscapes, and so, according to the traditional view, during recovery species typical of the natural landscape must migrate in from outside, making for slow regeneration.

At Mount St. Helens, however, many small animals and plants were protected by water or by pockets of snow. Animals dormant in dens or other sheltered areas at the time of the eruption also seem to have had an advantage. "The survivors were species that were keeping a low profile below the ground," Swanson says. "High-profile organisms, such as elk and Douglas fir, took it hard."

Chance also plays a large role in recovery after a major disturbance, says James A. MacMahon, an animal ecologist at Utah State University. Survivors have spread from havens throughout the blast zone, and so the landscape has been recolonized faster than most ecologists would have guessed. Although they cannot wander far from water, salamanders and Pacific tree frogs, for example, are now found throughout much of the blast zone.

Swanson points out that areas clear-felled before the eruption now support more plants than forested areas. The reason, he says, is that hardy weeds such as fireweed and pearly everlasting, which thrive in exposed conditions, were already established in the clear-felled areas.

In some places, species survived only to die later on. Ash-clogged streams often shifted during the years after the eruption, for example, choking seeds that had survived the disaster and were starting to sprout. The result was "terribly dramatic differences" between adjacent areas, according to Arthur McKee, a Forest Service ecologist who studies streams.

Biologists have been impressed by how even the remains of dead organisms can spur recovery. MacMahon has found that the leaves of a species of lupine decompose to woody skeletons that serve as drift nets for organic materials and seeds, including those of lupines. Likewise, downed trees have stabilized soil that might otherwise have been blown or washed away. The pumice plain, where there were no survivors, still supports fewer living things than other places.

Because Mount St. Helens is a unique ecological experiment, workers hope to establish long-term monitoring facilities at the mountain to track the recovery. They are seeking a $900,000 appropriation for the Forest Service to do so. In late September, however, the proposal was being held hostage to budget negotiations. —*Tim Beardsley*

FIREWEED AND PEARLY EVERLASTING spring up beside shattered tree trunks at Spirit Lake, close to Mount St. Helens. Photo: Tim Beardsley.

OVERVIEW: DEAD SEA SCROLLS

Will their editors perish before publishing?

The custodians of the major portion of the Dead Sea Scrolls seem to be out to turn the rules of academic survival upside down. These documents from the first century A.D., most of them discovered in the late 1940s and early 1950s at the caves of Qumran and other sites in the Judean Desert, constitute the greatest windfall in the history of Biblical archaeology. Yet the editors entrusted with the largest collection of these documents—at the private Rockefeller Museum in East Jerusalem—have not published even half of the trove in their 30 to 40 years of stewardship.

Many scholars say that the delay in bringing the scrolls to light has become the philological scandal of the century. These critics assert that the scrolls have been withheld long enough for many historians, theologians and philologists to have qualified in their fields and performed most of their work without having had access to crucial data.

The scrolls are important because they include by far the oldest manuscripts of books from the Old Testament and virtually the only texts that reflect the thinking of Jews during the period when Rabbinic Judaism and early Christianity were forming. "I am publishing a history of the Jewish literature of the period, and not knowing what is there makes this standard work incomplete," complains Geza Vermes of the University of Oxford. "Every scholar in the field and in neighboring ones can cite you such an example."

Shortly after their discovery, most of the scrolls were purchased for the Palestine Archaeology Museum, later renamed the Rockefeller, and put under the control of seven scholars named by the museum and associated Western archaeological societies. Jordan, which had annexed Jerusalem and the West Bank after the 1948 Arab-Israel War, approved the editorial board on condition that it exclude Jews—a rule the editors say they circumvented occasionally.

Early on, the enterprise produced a burst of publications. But by the early 1960s, it had entered the doldrums. Still, when Israel took control of East Jerusalem in the 1967 Arab-Israel War, it decided not to interfere with the original team or its first general editor, a Dominican priest named Père Roland de Vaux, on condition that he accept several Jewish scholars into his project.

Today the group is "built on a totally patriarchal foundation," Vermes says. The effort is now headed by John Strugnell of the Harvard Divinity School, who joined the project in 1953 at the age of 24, became its acting head in the early 1980s and officially became its third general editor in 1988. And as gatekeeper, Strugnell refuses to publish what he considers to be preliminary research. "It's the editors' work," he says, "and until it's published, they haven't finished filling in the gaps."

Strugnell proudly notes that his administration has increased the number of people who have some access to the scrolls to a total of 65. Now, however, the incensed scholars outside that hand-picked circle are waging an increasingly determined battle to force the group to publish. They have publicly attributed the delays in publication to editorial incompetence, sloth, empire building and unwillingness to admit defeat. "It has gone on for too long," Vermes says. "To give it up would be an open admission that they failed and that they misbehaved for a long time."

Even Strugnell's academic credentials have not escaped criticism. His foes contend that he is unsuited for the job because his Oxford training concentrated more on Greek than Hebrew and more on Biblical Hebrew than the much later idiom in which the non-Biblical scrolls are written. Moreover, he received his editorial assignment when he was, by his own admission, a virtual unknown just beginning his career.

One of Strugnell's most determined opponents is Hershel Shanks, editor of *Biblical Archaeology Review,* a popular monthly. He has conducted a four-year campaign to compel Strugnell and his colleagues to publish. "Strugnell is a brilliant man, but he has his deficiencies," Shanks says. "He never wrote a Ph.D. dissertation, hasn't got a doctorate and never wrote a book. He has published by himself extraordinarily little on the Dead Sea Scrolls entrusted to him. His major contribution was a lengthy article in which he corrects errors in another publication."

This spring Shanks published criticisms of the editorial monopoly made by 16 major scholars, whom Strugnell recently dismissed, in a television interview, as annoying "fleas." A lawyer by training who claims no philological expertise, Shanks has marshaled the discontent of hundreds of researchers. This has apparently exerted pressure on Strugnell indirectly by embarrassing the Israel Department of Antiquities, which claims ultimate authority over the editors. (So does the government of Jordan, with which the editors maintain close contacts.)

For the first time, Strugnell has set explicit deadlines for publication of the documents, even though the deadlines were missed last year and some will be missed again this year. Strugnell admits that his predecessors allowed the editors to "sit on their work" but insists things have changed since he took charge. "Now everyone we're bringing in signs an agreement to adhere to a schedule," he says. Of course, he adds, it was only fair to write a grandfath-

OLDEST COPY OF ISAIAH is displayed in Jerusalem at the Shrine of the Book, repository of Israel's comparatively small share of the Dead Sea Scrolls.

Some of the best points about Windows Computing are the ones you won't need anymore.

With Windows Computing on the PC you can work with greater speed and versatility than ever before.

That's because it lets you easily cut and paste text and graphics from one Windows application to another.

And because it also allows different applications to share information. Automatically.

See for yourself. Call (800) 323-3577,

Dept. M78, for a free fully functional Working Model* of Windows version 3.0, Microsoft Excel, Word, Project or PowerPoint presentation program. And find out just how sharp your work can really be.

Microsoft®

Making it all make sense™

*The first Working Model you select is free during our Windows Computing Promotion, September 15 through December 31, 1990. One free Working Model per person. Each additional Working Model is $9.95, applicable sales tax not included. Offer good while supplies last and only in the 50 United States. Microsoft, PowerPoint and the Microsoft logo are registered trademarks and Making it all make sense and Windows are trademarks of Microsoft Corporation.

er clause to keep the old rules for the old workers—himself and three others.

Strugnell dismisses critics' complaints that their work has suffered because they were denied access to unpublished materials. He says that any "competent" scholar can see documents relevant to his work merely by asking the editor in charge of them. The applicant can appeal an editor's refusal to Strugnell, who says he would overrule it if the applicant can demonstrate that he has done serious work. "I'd say we get 10 requests a year, five serious and five not," he says.

What about Vermes? "Geza is incompetent," Strugnell says instantly. A full professor at Oxford, incompetent? He pauses. "He is competent in other things," he says, "but doesn't have the necessary technical skills. I respect him for his other skills and if he came here I would consider his request." Vermes, for his part, says he gave up trying years ago, after he and scores of his colleagues had applied for access without getting any response at all.

Against the charges of sloth, Strugnell cites the competing academic responsibilities of the editors, most of whom can work on the scrolls only during summers and sabbaticals. Strugnell says the idea of paying for a full-time staff was never entertained. "Remember that the previous two editors were monks and didn't think in terms of raising money," he says. "Now I have just what I need; also we don't like to be in debt to outsiders."

Frank M. Cross of Harvard, a surviving member of the original editorial board, attributes the reluctance to publish to "a perfectionist streak" among many of his colleagues but opposes the idea of reassigning their material to faster workers. "Normally when you assign these things to a scholar it is assumed that he has a lifetime to complete his work," he says.

Some argue that the editors' monopoly encourages them to delay their work in the hope that it may be made impregnable to criticism. "There is the scholarly desire to say the last word on the texts," says Joseph A. Fitzmyer of Catholic University. "Yet we know that no one who pioneers in the publication of an ancient text ever says the last word on it."

Fitzmyer, a Jesuit, temporarily assisted the editors by helping produce a concordance of the scrolls—a card index citing the locations of each word that appears. The concordance was completed in 1960 and has still not been published, although several copies have leaked out and the Harvard Divinity School now makes one available to researchers. Shanks says the editors suppressed the concordance in order to protect their monopoly. "They kept that concordance secret for years," he says, "and they'll tell you the reason for the secrecy was that somebody might have used the concordance to put together an unpublished scroll."

No one denies that the redaction of the thousands of fragmentary parchment manuscripts is an arduous undertaking—in Cross's words "the world's greatest jigsaw puzzle." One scroll, for example, was made of copper and had to be cut into many sections before it could be painstakingly unraveled. Other scrolls had to be put together from bits of parchment that often covered just a few square centimeters.

JOHN STRUGNELL of Harvard heads the project that controls most of the documents found in the Judean Desert.

Cross notes that the group's 1,200-odd plates—each a photograph of reconstructed material pressed between sheets of glass—took 10 years to assemble. That work ended around 1960, but he maintains individual fragments are still being placed. Besides, he says, most of the easier materials have already been released. The job of transliterating, translating and interpreting the rest has, he claims, been intrinsically more time-consuming. "Now you're going to get plates in which you have hundreds of numbered fragments with lots of space," he says. "Lots of lace will be published."

Most work proceeds on the plates rather than the original documents, which are preserved from the elements and from bombs during Middle East wars. Israel, for its part, has provided bomb-proof shelters for the much smaller collection it houses in the Shrine of the Book in Jerusalem. Strugnell has assured that facsimiles of his documents will survive any but the most apocalyptic fate by secreting photographic versions in eight institutions around the world, including two in the U.S. Of course, the collections are closed to outsiders.

Norman Golb of the University of Chicago goes furthest of all the critics in attributing some of the editors' delay to professional bias. He contends Strugnell and his co-workers are reluctant to release documents until they can reconcile them to the accepted theory: that the scrolls were the work of a monastic sect inspired by the Essenes, one of three main religious factions among first-century Jews. Golb champions the alternative theory that the scrolls are an unselected library of diverse religious opinions, not a sectarian archive, and that they were removed from Jerusalem for safekeeping in desert caves just before the Roman siege that reduced the city to rubble in A. D. 70.

Golb says Strugnell himself has acknowledged that at least one unpublished manuscript disagreed with what had been held to be the monastic sect's heterodox opinions. But Golb says Strugnell has tried to fit the manuscript into the old theory by portraying it as a letter from the founder or an early leader of the putative sect—a letter so old that the opinions it expresses on Jewish law and other matters had time to evolve into those identified with the sect.

In what might seem like a heartening development, Strugnell says he and his collaborator, Elisha Qimron of Ben-Gurion University in Beersheva, will publish a 600-page commentary on the crucial letter (which is only 112 lines long) next year. It will be the first book Strugnell has published on any aspect of what he calls his life's work.

Even then, the Dead Sea Scrolls monopoly may continue, if Strugnell is able to acquire more materials kept by antiquarians as a long-term investment. "I'm in negotiation with a Kuwaiti for a nearly complete scroll," Strugnell says. "With the war in Kuwait, he may need some ready money."

Nevertheless, Strugnell now promises that all the documents in his group's control will be given to the printer within seven years. "They may be actually published in 10 years. You know how scientific publishing goes: you put out a quick thing to keep the public happy." —*Philip E. Ross*

The setting sun turns the water of the Okavango Swamps into rippling gold as a giant moves slowly in the steady evening breeze. Gandaya, The Mighty One.

It is the poacher's dream to capture him. Many have tried. Many have failed.

They say an elephant's memory is as long as the path of life itself. Gandaya remembers. Spears. Traps. Rifles.

How would it be then, if what Gandaya remembered of humans were those who helped his herd.

Not those who came to destroy it.

SCIENTIFIC AMERICAN November 1990 Volume 263 Number 5

Science, Technology and the Western Miracle

Close links between the growth of scientific knowledge and the rise of technology have permitted the market economies of the Western nations to achieve unprecedented prosperity

by Nathan Rosenberg and L. E. Birdzell, Jr.

Economic inequality between nations is to a considerable degree an invention of the past two and a half centuries. In the mid-1700s the average inhabitant of western Europe had a material welfare not too different from that of someone in China or, for that matter, in ancient Greece or Rome. Only a minority of the population enjoyed an income appreciably in excess of the minimum required to sustain life, and the elites of one nation had little reason to envy those of another.

By about 1800, however, it became apparent that the minority in Europe with incomes above the subsistence level was growing, at least in part because European science and technology were progressing faster than science and technology elsewhere. The increase in the number of factories and in the use of mechanically powered machinery came to be called the Industrial Revolution. The process of growth and change accelerated during the 19th century and has continued through the 20th. Historians sometimes call this unique period of long-term economic growth, which has made the West conspicuously richer and more powerful than the rest of the world, "the Western miracle."

To describe this phenomenon simply as long-term economic growth fails to convey its true dimensions. Between the mid-1700s and the present, per capita income increased tenfold. The population of Europe grew fivefold and that of the U.S. 80-fold. Infant mortality declined drastically, and the average life span doubled. Famine was banished, and plagues disappeared. Food production, which in some countries had occupied as much as 90 percent of the working population, eventually came to occupy less than 5 percent. Nineteenth-century urbanization marched in step with developing technologies for improved sanitation, construction, communication, power distribution and other services. Urbanization and rising incomes led to changes in health and living standards, work patterns, values and other aspects of personal, family and community life.

One might suppose that economic historians would have long since settled on the reasons for the Western miracle, but the phenomenon has not received the scrutiny it deserves. Consequently, when the less developed countries of the Third World turned to the West 40 years ago for help in raising their per capita incomes, much of the advice they received reflected an inadequate understanding of how the West had achieved its affluence. Very recently, the apparent determination of the Soviet Union and the Eastern European countries to close the large gaps between their income levels and those of the West has given fresh urgency to the question: What are the sources of Western growth that have eluded the less developed and socialist countries?

A variety of popular explanations for the Western miracle have been proposed. Some attribute it to imperialism, even though many of the most economically successful countries grew prosperous before resorting to imperialism and such highly affluent countries as Norway and Switzerland never adopted imperialist policies at all. Conversely, several of the most formidable imperialist powers, such as Spain and Portugal, rapidly deteriorated into economic stagnation.

Other theories link wealth to the possession of natural resources. Those resources do not become economic assets, however, until the knowledge and means of using them become available. The pre-Columbian peoples of North America had about the same resources as the present inhabitants do. Japan, which has far fewer natural resources than Indonesia, Mexico or the Soviet Union, has been far more successful in achieving growth. The modern histories of the city-states of Hong Kong and Singapore—not to mention that of Ven-

NATHAN ROSENBERG and L. E. BIRDZELL, JR., have previously examined the subject of the Western miracle in *How the West Grew Rich*. Rosenberg is an economist at Stanford University and the author of several books about technology and economic growth. Birdzell is an attorney and legal scholar in Rhode Island.

ice, which had only a swamp for its natural resources—belie the natural resources explanation.

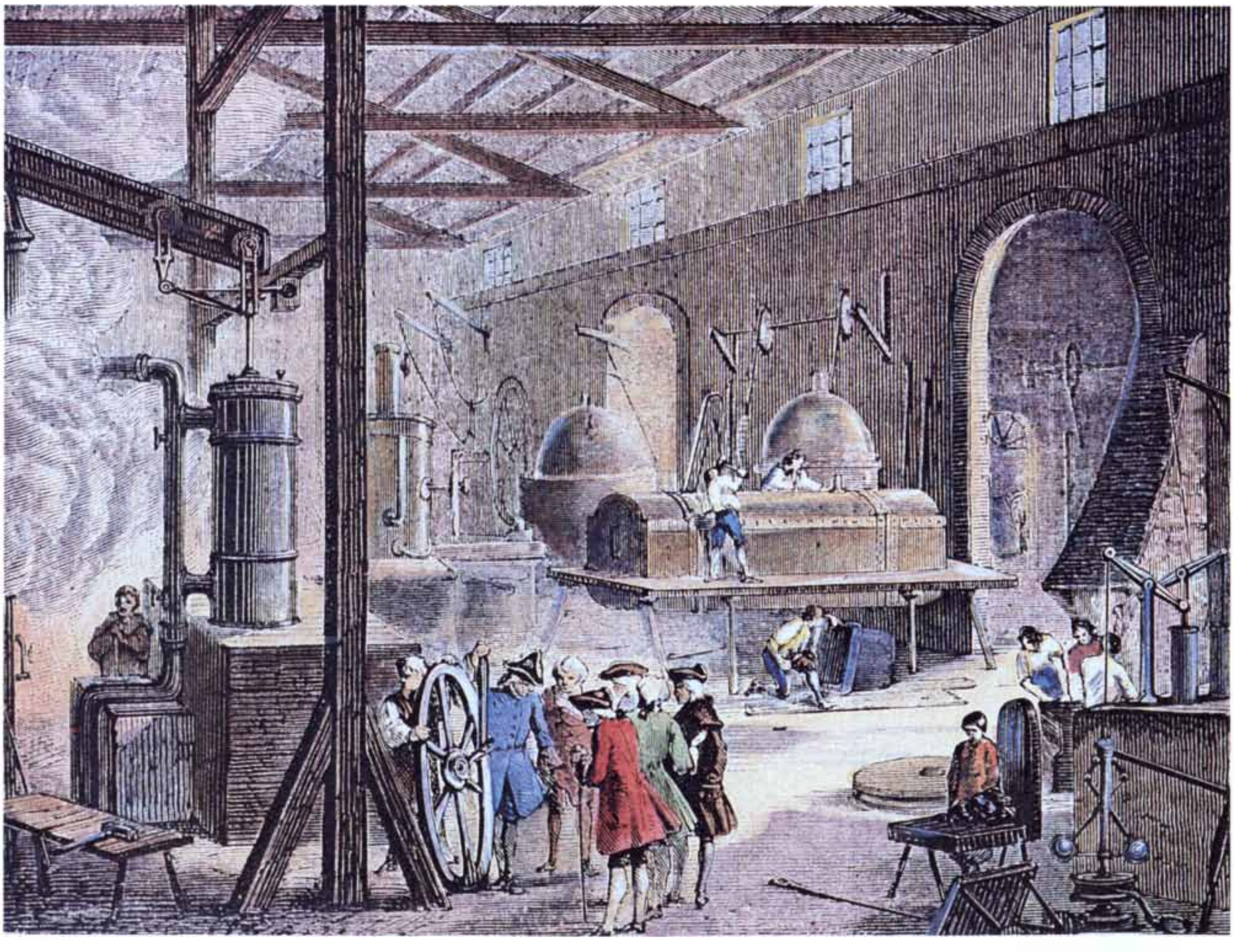
These hypotheses have had an unfortunate effect on government policies in the developing countries. By focusing on factors of questionable relevance, they have distracted attention from institutional changes that might have provided avenues to growth, especially ones that might have given access to superior technologies.

Western technology developed primarily within the economic sphere, and it has often been regarded as merely an outgrowth of economic needs and institutions. Science, in contrast, had more complicated origins and could hardly be dismissed as an automatic response to economic conditions. For a long time, science contributed little to economic growth and industrial technology. When Karl Marx was writing in the mid-19th century, the "colossal productive forces" he saw at work had been created primarily by people working in industry, with little contribution from people whom we would call scientists today. The mechanical skill and ingenuity that produced the precision machinery and instruments of 18th- and 19th-century factories and laboratories came more from the crafts of clock making and lens grinding than from science.

Since about 1880, however, industrial technology has come to owe a more substantial debt to scientific sources outside industry. With the success of efforts to fit natural phenomena to theoretical structures inaccessible without special training, industrial engineers with that training have become transmitters and users of scientific knowledge and methods. More than that, during the past century industry has created research laboratories capable of extending the theoretical structures of science. Although Western science originated as an institution outside the economic sphere, during the 20th century its advance has become inseparable from that of industrial technology and Western economies.

To explain the Western economic miracle and its relation to science, we must first consider some of the reasons for the great success of Western science—an achievement with its own claims to the title "miracle." One reason is that Western science has made a better organized attack on the secrets of nature and used greater resources in the assault than science in other cultures.

For a long time after the printing press was introduced in the late 15th century, scientific research remained chiefly a decentralized—or even individualized—activity, in which isolated scientists occasionally communicated their discoveries to one another in print or in longhand. Early Western science was clearly not a localized phe-



SOHO MANUFACTORY near Birmingham, England, produced parts for steam engines in the 18th century. The partners in the enterprise, James Watt (the father of efficient steam engines) and Matthew Boulton, were members of the Lunar Society of Birmingham, a group of businessmen, inventors and scientists. Throughout the Western world, similar associations between scientists and entrepreneurs fostered scientific, technological and economic progress simultaneously.

nomenon: its venues ranged from the Poland of Copernicus to the Denmark of Tycho Brahe to the northern Italy of Galileo to the Bohemia of Kepler to the France of Descartes and Lavoisier to the England of Boyle and Newton.

The early achievements of Western science centered on astronomy. The development of a significant scientific community in Europe with interests beyond astronomy dates from the 17th century. In 1660 the Royal Society of London for Improving Natural Knowledge (almost always called simply the Royal Society) was formed to discuss reports from the many individuals who were by then conducting scientific investigations. Many other such societies formed in the 17th and 18th centuries, setting up a network of scientists in Europe who exchanged information not only with one another but also with a distant American named Benjamin Franklin, whose experiments had shown that lightning was electricity.

These societies, and the journals they published, both disseminated new research and screened it for admission to the scientific canon. Their discussions set an agenda for the time and served as a pointer to new research that might bring recognition and acclaim from other scientists. What they did not offer was a means of earning a living. In 1695, for example, Isaac Newton faced limited academic advancement at Cambridge because he had not taken holy orders. To reward him for his scientific contributions with a livelihood, the British government had to name Newton to a post outside the scientific community, as warden of the Mint.

Although the idea of bringing scientists together for directed research in an institute equipped with laboratory instruments and a suitable library was tried successfully in the first half of the 15th century by Prince Henry the Navigator of Portugal, it came into common practice only in the early 19th century. In London Sir Joseph Banks, Count Rumford and some other fellows of the Royal Society formed the Royal Institution in 1799 to serve as a laboratory where scientists could work together and teach. Michael Faraday, working a century after Newton, found a full career at the Royal Institution, where he discovered electromagnetic induction.

ILLUSTRIOUS SCIENTISTS throughout Europe contributed to the growth of basic knowledge that spurred technological and economic progress. Among them were (*clockwise from upper left*) Nicolaus Copernicus of Poland, Johannes Kepler of Bohemia, Galileo Galilei of northern Italy and Sir Isaac Newton of Great Britain.

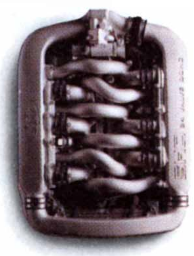
Similar institutions sprang up elsewhere. In 1795 the French established the École Polytechnique. In the U.S., Yale University established the Sheffield scientific school in 1847, and the Massachusetts Institute of Technology opened in 1865. Science thus gradually developed its own research and teaching institutions, and successful researchers could be rewarded with staff appointments and promotions.

By the early 19th century, Western science had divided into specialized departments: mathematics, astronomy, physics, chemistry, geology, botany, zoology and the medical studies of anatomy and physiology. Some of them, such as physics, were divided into still finer specialties.

Western science had become an institution with a broad general goal (to explain natural phenomena), a division of labor into specialized departments with their own subsidiary goals, an information network that kept its members informed of progress, a peer review system for evaluating new work and settling conflicts, formal centers for teaching and research, and a set of rewards for work rated favorably by the profession.

A fundamental factor holding the enterprise together was its adoption of a single standard of scientific truth based on observation, reason, experiment and replicability. The standard enabled scientists to make use of findings from other laboratories, even from those in other disciplines. It also permitted artisans, merchants, manufacturers and the rest of the working population to apply scientific discoveries to everyday labors.

Its organization and scope are, of course, not the only reasons Western science flourished. It also fell heir to the vast intellectual estate of earlier civilizations: a phonetic alphabet, an Arabic (or possibly Indian) counting system that included zero as a number, mathematics that included geometry and algebra, and religions that freed nature of animism. But the West was no idle heir, for by the beginning of the 18th century it had added some intellectual monuments of its own.



24-VALVE
220 HP
FORD
TAURUS
SHO

Executive conference room.

When your career's on the fast track, you need a car that belongs there. The Taurus SHO. Its distinctive appointments include fully-articulated sport seats (leather trim optional) and performance instrumentation. The Taurus SHO is perfect for a closed conference, even if your only meeting is with the available JBL stereo and your favorite compact disc.

The business end.

But your executive image doesn't have to be conservative. Nothing sets you apart like Taurus SHO's hidden asset; a 24-valve, 220 horsepower Super High Output (SHO) V-6, precisely controlled by a newly refined, smoother-shifting five-speed manual transmission.

Performing assets.

Rounding out Taurus SHO's portfolio are a special sports suspension, 4-wheel disc anti-lock brakes, and standard driver's side air bag supplemental restraint system to be used with your safety belt. And with new 16″ cast aluminum wheels, the SHO will quickly improve your outlook.

The Taurus SHO. Because making work a pleasure is the essence of good business.

Ford Taurus SHO

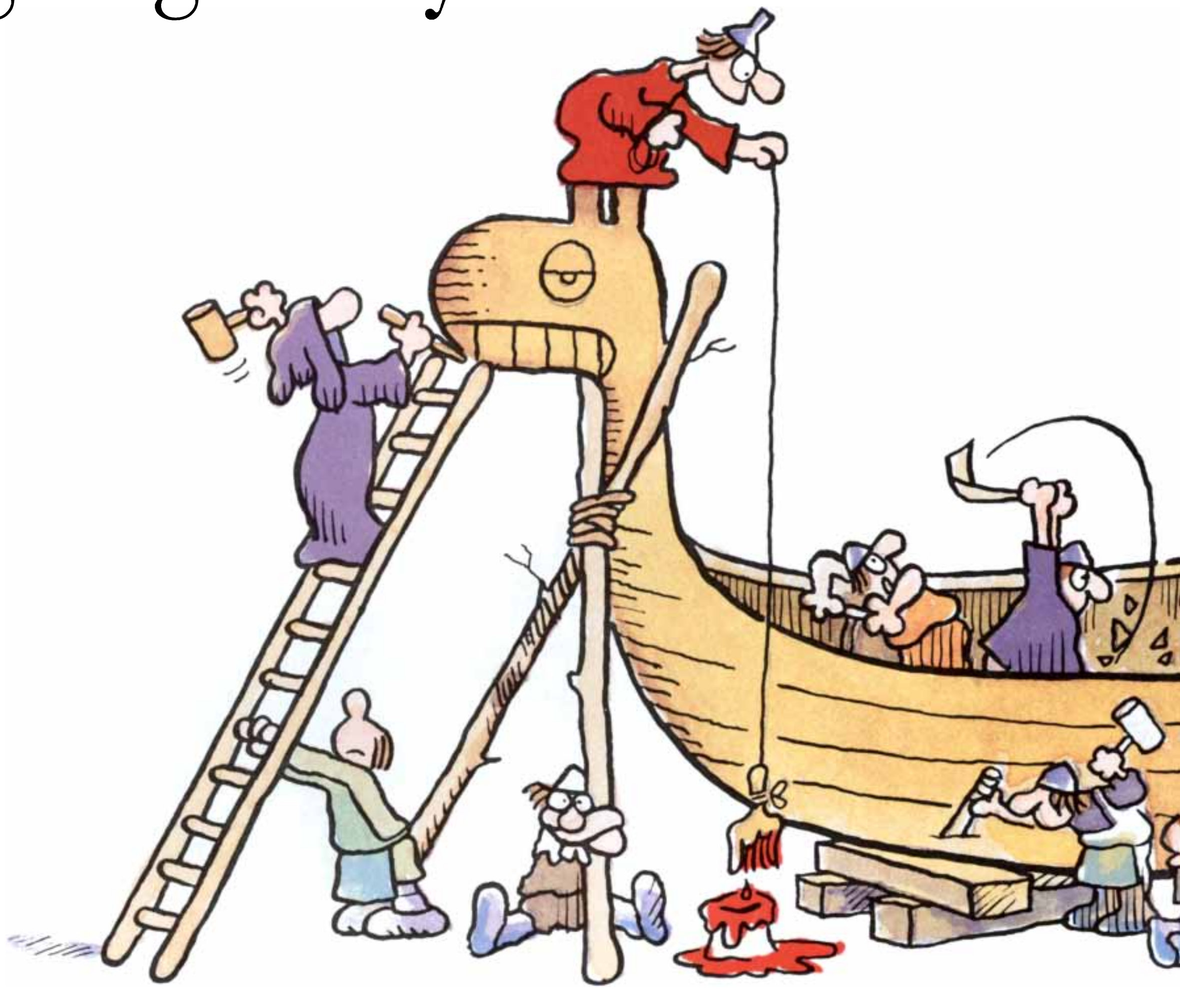
Buckle up–together we can save lives.

Have you driven a Ford...lately?

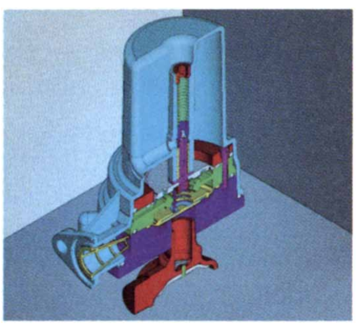
Ford

The IBM RISC System/

Designing on any other workstation



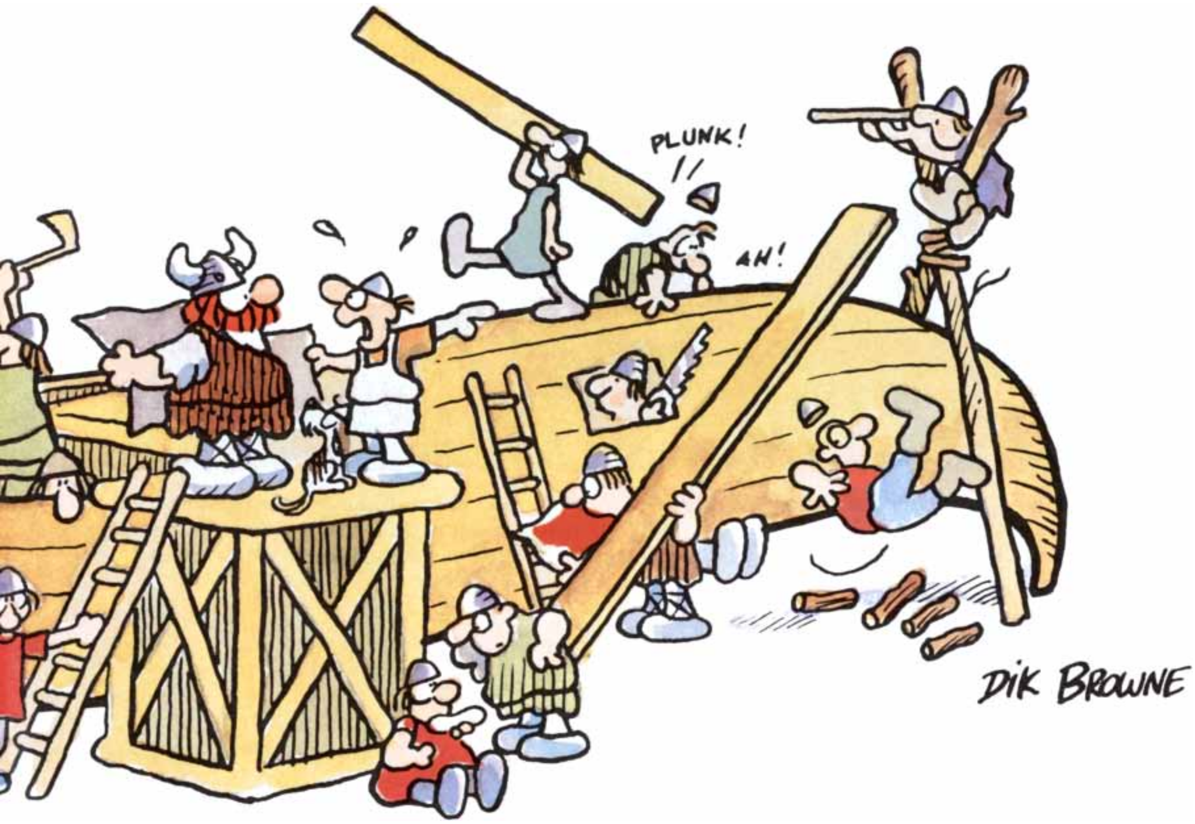
Whatever you're creating, you'll sail into a whole new age with any of the four POWERstations in the RISC System/6000 family. Because POWER (Performance Optimization With Enhanced RISC) processing can give you performance you've probably only dreamed about:



up to four instructions per machine cycle, 42 MIPS and 13 MFLOPS. Suddenly, complex designs don't take eons anymore.

The four RISC System/6000 POWERstations feature a range of graphics processors from grayscale to Supergraphics to satisfy any graphics demand. Great news for Power Seekers working on animation, scientific visualization, medical imaging and engineering solutions like CADAM,™ CAEDS™ and CATIA.™ And for electrical design automation, there's IBM's all new CBDS™ and an arsenal of over 60 EDA appli-

6000™ family.
will seem downright primitive.

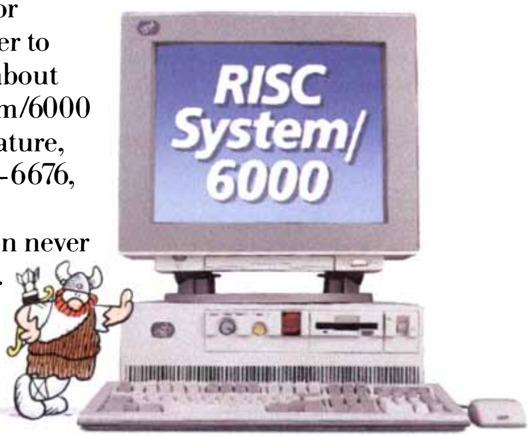


cations from more than a dozen vendors.

With every POWERstation, you can get an almost unimaginable palette of 16 million colors, which gives you 3D images so realistic, they fairly leap off the screen, with super sharp resolution of 1,280x1,024 pixels. And when it's time to call in the heavy artillery, the POWERstation 730 draws nearly one million 3D vectors per second. Like all POWERstations, it can come complete with its own graphics processor, freeing the POWER processor to rapidly create and analyze your designs. All at prices that won't sink anybody's budget.

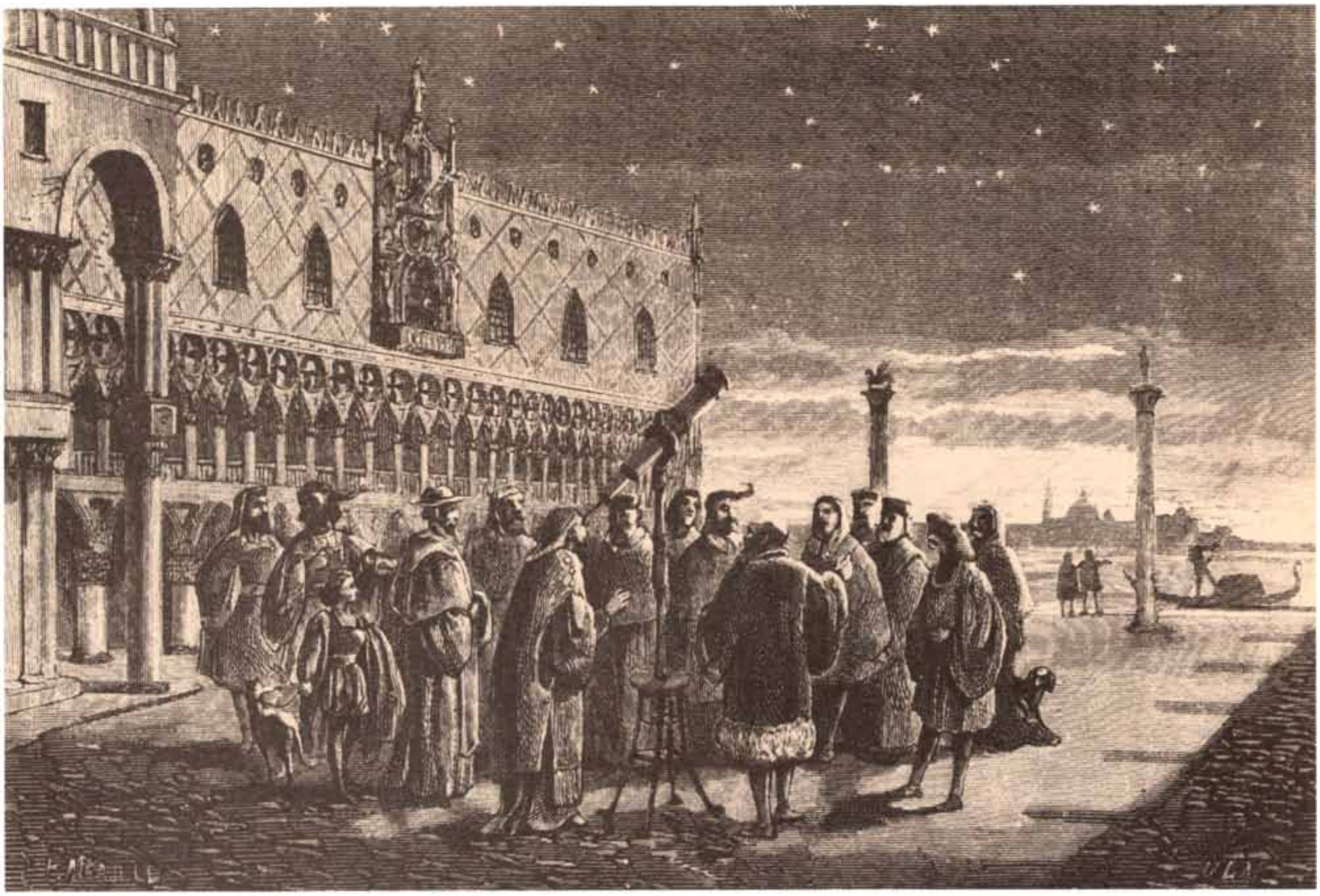
So if you're tired of paddling upstream with yesterday's performance, call your IBM marketing representative or Business Partner to find out more about the RISC System/6000 family. For literature, call 1 800 IBM-6676, ext. 991.

Civilization never looked so good.



For the Power Seeker.

IBM®



VENETIAN SENATORS peer at the moons of Jupiter through a telescope under the direction of Galileo. Astronomy was one of the first interests of Western science, but before the end of the 17th century, Europe began to have a sizable scientific community interested in physics, medical science and other fields. Science slowly became a less individualized endeavor.

Calculus, for example, was clearly a landmark contribution. Another, and perhaps the most fundamental, Western contribution was the development of the scientific method, which from the time of Galileo included a refined form of systematic experimentation. Hellenistic, Islamic and Chinese scientists and inventors understood the use of experiments for testing and confirming ideas, but they do not seem to have achieved anything like Galileo's inclined plane experiments, in which conditions were systematically varied as a way of exploring how nature works. Nor did they anticipate Newton's thought experiments, in which idealized phenomena (for example, motion in a vacuum) are used to explain real phenomena.

Without systematic experimentation, progress can be slow and fitful in science and in technology. Improvements in the design of the plow, for example, were inventions of great significance in predominantly agricultural societies, but they took place hundreds of years apart. Before the age of science, no one seems to have tried to improve plow designs by comparing the effectiveness of different blade designs in various soils. Concrete was known to the ancient Romans, but it was little used as a building material until late in the 19th century, when chemists experimentally investigated its suitability for structural applications by systematically varying the mix of its ingredients. Within a few decades, concrete, including reinforced concrete, became the most widely used (by weight) of Western building materials.

Perhaps the most important point about Western science and technology is that they were linked at all. In other civilizations, economically useful technologies depended minimally, if at all, on the wisdom of astronomers (or astrologers), philosophers, mathematicians and other sages. These thinkers had little to offer to farmers, sailors, smiths and other artisans who had developed their technologies within their craft traditions. In fact, the thinkers often confined themselves to an abstract world of ideas as an escape from the transient and imperfect, real world. For Western scientists, however, there was no escape. Their empirical methods required them to engage the real world. It is precisely because the scientists were so engaged that they accomplished so much.

Third World and Eastern European countries trail the West economically, despite having access to the world's stock of scientific knowledge. Indeed, the Soviet Union has made substantial contributions to that stock. What those countries seem to lack is the West's capacity to translate scientific knowledge into economic productivity—a capacity that depends on the individual and institutional characteristics of a nation. In many ways, technology, which is intermediate between scientific knowledge and economic activity, grows out of local needs and institutions, and its economically successful transfer involves more than a teaching process.

One clear requirement for economic growth is the ability to shape productive technology to local needs. Whatever the origins of a technology, the people and institutions using it must be able to understand it, experiment with it and evaluate the economic repercussions of its use.

Japan, usually regarded as the first

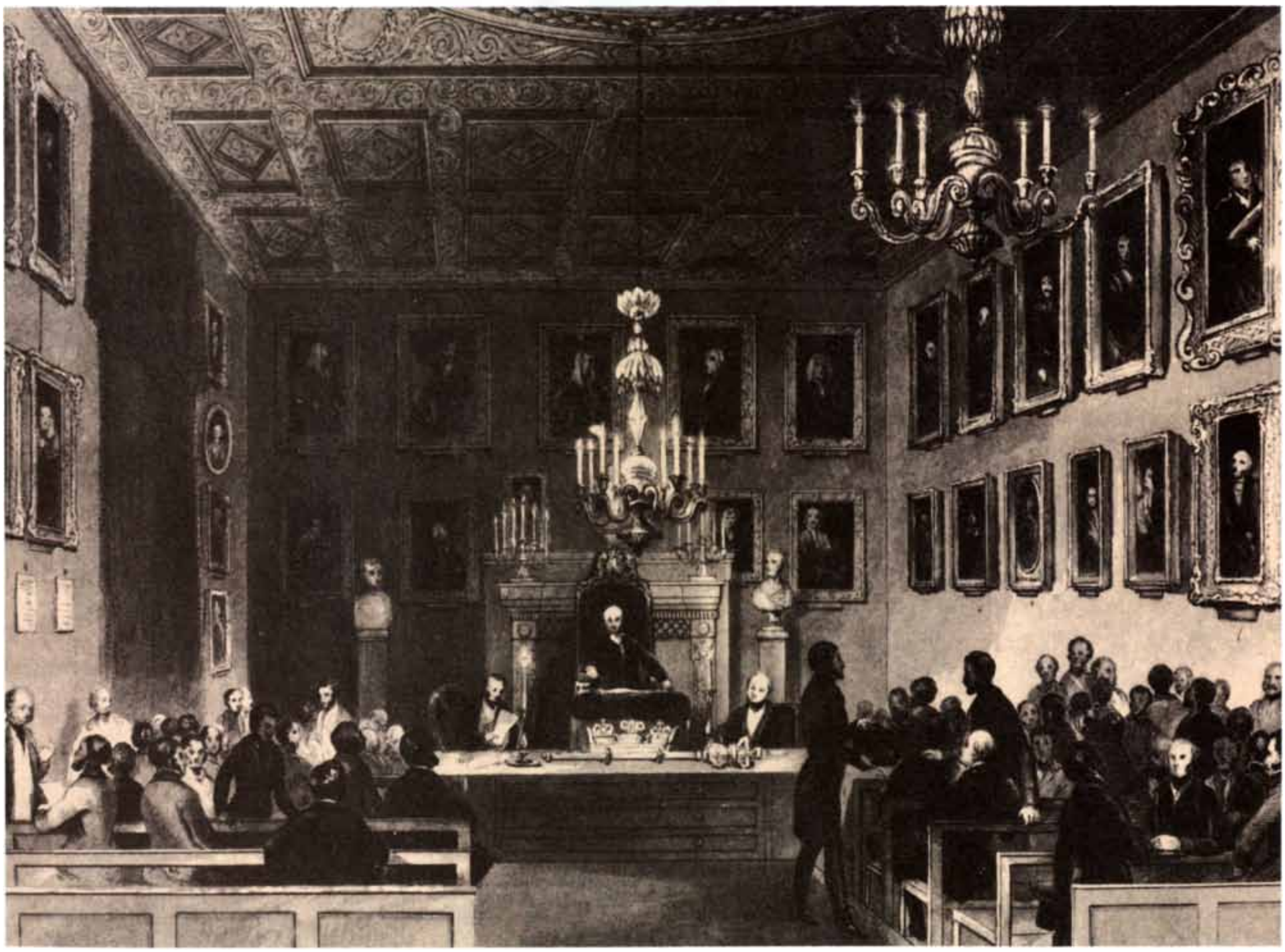
non-Western country to match the economic achievement of the West, offers an instructive example of institutions appropriately applying technologies. The first stages of Japanese modernization, beginning in the late 1860s, emphasized agricultural improvements. The technologies initially borrowed from North America were capital intensive and tended to maximize output per worker. The Japanese soon realized, however, that techniques suitable for the land-abundant U.S. were inappropriate in an economy that had ample labor but was seriously short of arable land. They then shifted to other technologies, borrowed this time primarily from western Europe, that were more labor intensive and that maximized the productivity of the scarcest Japanese resource—land.

In manufacturing, too, Japan adapted Western technology to its labor-rich, capital-poor economy. Japanese factories often purchased secondhand rather than new machinery. Wherever possible, they substituted labor for capital and used foreign technologies in more labor-intensive ways. Textile mills, for example, added extra work shifts and expended more labor on maintaining and repairing machinery to prolong its life. When the Japanese constructed a railway system, they used more than two and a half times as many workers per mile of track as the Americans did.

Parallels between the development of industry in Japan and in the U.S., where it began several decades earlier, are often overlooked. Like Japan, the U.S. at first borrowed technology, primarily from England. Industrialization began in New England, New York, Pennsylvania and Delaware—regions where, as in Japan, literacy and formal education were already valued. The U.S. also had to adapt its borrowed technologies to the country's special circumstances, as Japan did.

Because the U.S. had abundant natural resources, however, its adaptations involved making the industrial technologies more resource intensive and less labor intensive. Europeans visiting the U.S. in the mid-19th century often criticized Americans for wasting natural resources. U.S. agricultural techniques did frequently lead to rapid declines in soil fertility, but more land was always available, so the losses were supportable. Americans invented woodworking machinery that appeared extremely wasteful to the British. At the time, however, it made good economic sense in a country so richly endowed with forests.

Late in the 19th century, private business firms in Germany and the U.S. founded industrial research laboratories for developing new products and production methods. In 1856 the English chemist William Henry Perkin discovered aniline purple, the first of the industrially useful coal-tar derivative dyes, and the next year he established a factory for its production. This work marked the beginning of both a major branch of chemistry and a major chemical industry. Perkin went on to make further contributions, but no one individual could ring all the changes on the chemistry of coal-tar derivatives. That accomplishment required institutional



ROYAL SOCIETY OF LONDON for Improving Natural Knowledge, founded in 1660, was one of the earliest associations for advancing the communication of scientific information. Such societies evaluated discoveries and disseminated them.


Bombay® Sapphire™ Gin. 47% alc/vol (94 Proof).
100% grain neutral spirits. © 1989 Carillon Importers, Ltd., Teaneck, N.J.
SAPPHIRE
BOMBAY
PRODUCT OF ENGLAND
BOMBAY SAPPHIRE.
POUR SOMETHING PRICELESS.

invention. Only after German chemical manufacturers established several research laboratories for systematic, organized investigations did an important industry arise around the coal-tar dyes between 1890 and 1914.

The accomplishments of the German laboratories inspired the establishment of the General Electric Company's research laboratory in the U.S. When GE was formed in 1892, it relied at first on Charles Steinmetz, a talented political refugee from Germany, as its resident inventor. In 1900, however, after the new German laboratories produced superior materials for electric-lamp filaments, GE engaged Willis R. Whitney, a professor of chemistry at the Massachusetts Institute of Technology, to organize a formal laboratory. During the first half of the 20th century, the number of research laboratories affiliated with private firms multiplied, creating major new institutions that significantly contributed to the growth of basic scientific knowledge as well as to technology.

The best road to technological progress is often poorly marked. The great virtue of private businesses in market economies is that they become independent sources of decision making for exploring the frontiers of technology. No single individual or institution has the authority to veto an exploratory undertaking unilaterally. The importance of diffusing decision making is illustrated by the success of the personal computer, which is said to have been turned down by the principal U.S. computer manufacturer. Similarly, the export of Japanese automobiles to the U.S. was initiated by a Japanese company against the advice of the Japanese government. To be sure, the government was right, because the first cars exported were designed for Japanese use and performed poorly in the hands of early U.S. customers. The exporter nonetheless learned what was required for the U.S. market and took the necessary corrective action. It was soon followed profitably by other companies.

Given the scientific, technological and commercial uncertainties of innovation, an efficient economy has to strike a balance between undertaking only the safest of projects and pursuing every idea that comes along, too often persisting long after realistic hope of success has vanished. With their rewards and penalties, market economies have so far been most successful in striking the required balance and transforming scientific and technological knowledge into useful goods and services.

Nevertheless, the economies' open-ended diffusion of authority to start innovative projects would be unworkable if decision makers' hopes of high rewards were not tempered by their exposure to the risk of severe losses. Public attention commonly goes to the occasionally sensational rewards of innovation, but the prudent investor keeps in mind the institutional role of the bankruptcy courts in Western innovation: burying the failures.

In the West, although innovative contributions often originate with long-established firms, many of the most important inventions have been commercialized by new firms formed for that purpose or by completely redeployed old firms. The role of new firms in innovation is important not only for its direct contribution but for its implicit threat to older firms, which might not otherwise feel impelled to take the risks of innovation. Yet freedom to organize new firms is narrowly restricted in socialist countries and severely hampered in some less developed countries, where permission from scores of government agencies may be required to launch a business enterprise.

In addition to corporate research and development, Western countries have developed several other sources of economically useful knowledge during the 20th century, especially in areas where markets have not provided adequate incentives. Government funding has become particularly important for research that has grown fearfully expensive, as it has in particle physics. Governments also support research in such fields as public health, preventive medicine, the treatment of rare diseases and safety—fields in which the goals are universally desired but there is little prospect of reward for private firms. Government-sponsored research is conducted in both government and private laboratories. To an increasing extent since World War II, universities also have become centers of tax-supported research activities, technological as well as scientific.

Economic growth in the West has been marked by an increase in trade and in the sizes of markets. Part of the increase in trade was tied to technological improvements in ships and to the introduction of railroads—innovations that lowered the costs and risks of transportation. As trade in regional specialties became possible, manufacturers could reach markets large enough to justify investing in mass-production technologies. The introduction of refrigerated ships, for example, enabled a growing European population to trade its manufactured products for meat from Argentina, Australia and the U.S. More recently the increased size of international markets has made it economical for manufacturers to furnish products in greater variety, more closely adapted to the local needs of particular countries or groups of customers. This trend can be observed in markets as diverse as automobiles, clothing, processed foods and electronics.

In recent years the importance of international trade has been highlighted by the differences between the growth performances of countries that have actively competed in international markets and of countries that have adopted policies of import substitution (protectionism) or closely regulated trade, as in the socialist countries. There is no consensus about why active participation in international trade seems so closely associated with economic growth. Possible explanations include some combination of scale economies, keener competitive incentives, economies of specialization and abstention from counterproductive impositions on the part of governments of successful exporting countries.

In Western manufacturing, a particularly interesting example of specialization has been the rise of manufacturers who make only component parts and subassemblies, particularly for the automobile, electric and electronics industries. Brand name manufacturers often specialize in the design, the marketing and sometimes the assembly of finished products, with the component parts made by subcontractors. Technological advances conceived by American, German or Japanese engineers can therefore create employment through subcontracted manufacturing not only at home but also in Mexico, South Korea, Taiwan, Singapore and elsewhere. Modern subcontracting networks are also international trading networks, with operations that depend critically on transportation, communication and data-processing facilities that were unimaginable in 1800.

Because trading networks in market economies do not have a central source of authority, their power for efficiently organizing activities went unnoticed until long after the Western miracle had begun. In its lack of central authority, the remarkably successful organization of Western economies paralleled that of Western science, which also lacked a supreme authority and yet became an effective enterprise coordinating the work of thousands of specialized scientists and a wide variety of research institutions. Decentralization is especially relevant to economic growth, but Western societies have also had a comparable degree of autonomy from political control in art, literature, music, religion and other important social spheres.

The basic control problem of any economic system is to make and enforce a changing flow of interdependent decisions about production and consumption that will optimize human welfare. Beginning no later than the 12th century, unregulated market trading was gradually insinuated into western European economies traditionally controlled by governments, guilds and the Catholic Church. The centuries-long process was partly one of creating new branches of trade (including international trade) outside the jurisdiction of authorities and partly one of outright evasion. Traditionally, Western authorities had overtly regulated prices and wages to keep them fair according to inevitably subjective criteria.

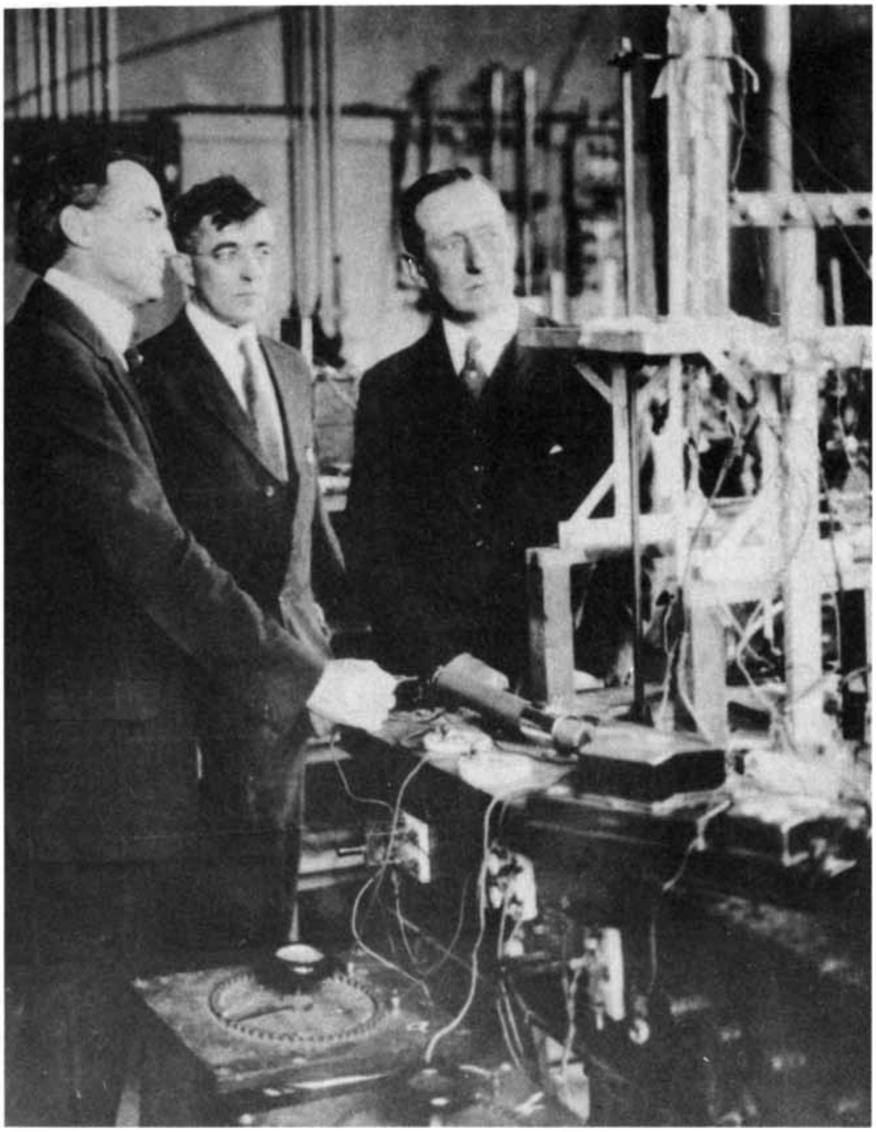
The slow growth of unregulated market trading gradually reorganized more and more of these economies by transmuting prices and wages into ethically neutral devices for keeping supply and demand synchronized, with consequences for the organization and development of Western economies that did not begin to be widely appreciated until Adam Smith studied them in the latter part of the 18th century. The most visible organizations of those times were governments, armies and the Catholic Church—all hierarchies. For many observers, the very idea of organization implied a hierarchy of authority, and its absence was equated with chaos. People failed to see that through decentralized mechanisms, Western economies were achieving new patterns of specialization and a size and efficiency of organization without historical precedent.

By the time the great German sociologist Max Weber wrote his pioneering work on organization in the early 20th century, the organizing power of markets was widely recognized among social scientists. Yet bureaucracy was looked on by some as an entirely viable alternative that seemed to offer a possible return to the long-lost world of "just" prices and "just" wages that socialists and many others found attractive. The comparative performance of socialist and Western economies over the past seven decades, culminating in the recent upheavals in the Soviet Union and Eastern Europe, suggests that the judgment of equivalence overlooked something important.

Market trading, considered by itself, can be conducted by agents of government, trade unions, consumers, investors or any other groups or individuals. In Western history, however, the introduction of market trading resulted from the efforts of merchants operating within the framework of a peculiarly Western institution, the trading and manufacturing firm. What seems to be peculiarly Western about these institutions is that they have often competed through innovations in products, manufacturing and distribution. The success of Western economies in assimilating Western technology is not a consequence of unregulated markets alone but of markets in which there are productive firms that can gain much by commercializing new ideas more quickly than their rivals can.

As we have already observed, innovation is a risky way to compete, and firms that live or perish by the uncertain irregularities of innovation's rewards and losses are very different institutions from administrative agencies of government. The point is especially significant today for the Soviet Union and Eastern Europe, which, having outlawed the types of entities through which the West conducted commercial innovation, now confront an awesome need for rapid innovation in products, manufacturing and organization if they are to narrow the economic gap

TRANSCONTINENTAL RAILROAD was completed at Promontory Point, Utah, in 1869. Advances in transportation technology fueled economic growth by opening and expanding markets, in turn allowing greater manufacturing specialization.



WILLIS R. WHITNEY, the first director of the General Electric Research Laboratory (*left*), and Irving Langmuir, the Nobel prize–winning industrial chemist (*center*), demonstrate a high-vacuum apparatus to Guglielmo Marconi, the inventor of radio (*right*). Corporate laboratories closely linked technology to scientific knowledge.

between themselves and the Western countries. Although recognition of the need for greater reliance on markets has been increasing, there has so far been less awareness of the critical need to allow Soviet and Eastern European firms to make innovations in products, production methods and organization and to reap the resulting benefits and losses.

In seeking explanations for the Western miracle, we have proposed that long-term economic expansion and technological expansion go together, in that neither has occurred for very long without the other. But although technological and economic expansion are interwoven and inseparable, no simple law of nature makes technology the cause of economic growth or growth the cause of technological advance. An expansion in the size of markets can make a more efficient division of labor and specialization possible without introducing any appreciable technological novelties. The interplay of people, economic institutions, growing markets and technology is the key.

At the end of World War II, many scholars and policymakers believed that the future of all nations lay with socialism—or at least with some other form of planned economy. The older Western market institutions were additionally tainted by associations with colonialism. Consequently, only a handful of developing countries then chose to imitate the Western nations, and even their governments took a more active part in economic affairs than Adam Smith might have prescribed. This handful included Taiwan, South Korea, Hong Kong and Singapore—some of the great economic success stories of this century.

No one can guarantee that other countries would do equally well with a similar policy of imitation. But as we survey 40 years of experience with socialist and populist regimes in Eastern Europe, Asia, Africa and South America, we believe that it requires a victory of faith over experience to see much chance of success without imitation.

Science, no longer solely Western, is pushing back the frontiers of knowledge at what seems an accelerating pace. Because knowledge creates economic resources and because knowledge generally grows at an exponential rate, future advances in human welfare can be at least as striking as those of the past 200 years. Science can also play a much larger role in dealing with such deep-rooted problems as environmental pollution and population growth but only in the context of more effective institutions and personal incentives. Disclosures from Eastern Europe indicate that industrial pollution there has been worse than in the West.

Given the strong prospects for the continued growth of international trade and markets, further Western growth may continue to widen the economic gap between countries that follow the West's example and those that do not. Economic gaps are already creating severe political and social pressures in the developing countries, especially since some of them have demonstrated that the gaps can be closed. Unless these nations take effective action, the consequences could be even less satisfactory than those of the past 40 years.

FURTHER READING

UNBOUND PROMETHEUS: TECHNOLOGICAL CHANGE AND INDUSTRIAL DEVELOPMENT IN WESTERN EUROPE FROM 1750 TO THE PRESENT. David Landes. Cambridge University Press, 1969.

PHASES OF CAPITALIST DEVELOPMENT. Angus Maddison. Oxford University Press, 1982.

INSIDE THE BLACK BOX: TECHNOLOGY AND ECONOMICS. Nathan Rosenberg. Cambridge University Press, 1983.

HOW THE WEST GREW RICH: THE ECONOMIC TRANSFORMATION OF THE INDUSTRIAL WORLD. Nathan Rosenberg and L. E. Birdzell, Jr. Basic Books, 1986.

THE EUROPEAN MIRACLE: ENVIRONMENTS, ECONOMIES, AND GEOPOLITICS IN THE HISTORY OF EUROPE AND ASIA. Second Edition. Edited by E. L. Jones. Cambridge University Press, 1987.

CheckFree: The Fastest Way To Pay Bills And Handle Finances.

CheckFree®
Electronically Pays All Your Bills And Automates Your Recordkeeping. No Checks, No Stamps, No Paperwork.

It Couldn't Be Easier.
All you do is enter payment information, and CheckFree software automatically records, categorizes, and sends the information to the CheckFree Processing Center via modem. Payments are then processed through the Federal Reserve System—the same system that banks use to clear your checks today. Fixed, recurring payments (such as mortgages) can be made automatically; once they're set up, you never have to bother with them again.

Pay Anyone Through Any Bank.
You can pay anyone with CheckFree because the CheckFree Processing Center utilizes electronic transfers as well as laser printed checks to make payments. If a company or individual you are paying is not set up for electronic payment, the CheckFree Processing Center will print and mail a laser printed check for you. Because of its universal payment processing technology, CheckFree can process payments to anyone in the U.S. through the checking account you already have at any bank, credit union, or savings and loan.

No Loss Of Float.
You tell the CheckFree software when you want your bills paid and the CheckFree system pays them on the dates you specify. All we need is a few days advance notice to make sure your payments get processed on time.

Saves Time, Paperwork, And Money.
Your bill paying and other routine financial activities can be reduced to just a few minutes. Paperwork becomes virtually non-existent. Plus, the CheckFree service is only $9.00* per month—not much when you consider the savings in postage, bank charges, and late fees.

Organizes Your Finances.
CheckFree gives you a completely automated checking account register that's updated automatically with each transaction. You can also record deposits or other non-CheckFree transactions. Checkbook reconciliation is virtually automatic. There's even an easy-to-use package of income, expense, and budget reports that keeps you constantly aware of what's coming in and what's going out.

Completely Secure.
CheckFree is more secure than conventional check writing. Payments are processed through the Federal Reserve System, so CheckFree never requires actual access to your account. Only you have access to your funds and only you authorize payments.

Built Into Other Leading Personal Finance Programs.
CheckFree is the only software package designed specifically for the efficiency of electronic payment technology. Simply put, it's the fastest and easiest way to handle your finances. What if you like the idea of CheckFree electronic bill payment, but prefer to use Managing Your Money®, Quicken®, or Checkwrite Plus®? No problem. CheckFree is seamlessly integrated into these programs, so you can enjoy the convenience of electronic bill payment with your choice of personal finance software.

Backed By The Nation's Leader In Electronic Payment Services.
Checkfree has been the leader in electronic payment processing since 1981, with current payment volume exceeding $1,000,000,000 annually.

Used By Thousands All Over The U.S.
See what experts have to say:

"The CheckFree electronic bill-paying service belongs to that rare class of product that is forward-looking, universally useful, and available today."
PC Magazine, Best of 1989 Awards January 16, 1990 issue

"CheckFree makes bill paying a quick, painless task rather than a monthly ritual to be dreaded... Checkfree Corporation has been handling commercial electronic payments for companies such as CompuServe® for years, and its experience and professionalism showed in all aspects of my dealings with it."
PC Magazine

"...CheckFree did not fail me. When my banking statement arrived the first thing to catch my eye was a CheckFree check...this laser-printed paper check was reassurance that someone else could be trusted to pay my bills..."
Robert Cullen, Home Office Computing

Now that you know all about CheckFree, why would you pay your bills and handle your finances any other way? Why would you burden yourself with checks, stamps, and paperwork while your PC sits idly by? Besides, when you order by phone or mail, CheckFree comes with a money back guarantee. If you're not completely satisfied with CheckFree software, simply return it within 30 days of purchase for a full refund. You've got nothing to lose.

So order today. Call

1-800-882-5280
(Monday through Friday from 9AM to 7PM EST)

or send $29.95 (plus $4 shipping and handling) to:

CheckFree Software
P.O. Box 897
Columbus, OH 43216

(Be sure to specify disk size and machine type.)

CheckFree is also available at leading retailers including Egghead Discount Software®, Electronics Boutique®, Babbages®, Waldensoftware®, Software Etc.®, Soft Warehouse®, Radio Shack®, Best Buy®, and Price Club®.

CheckFree®

Client Support: free access via Email. **Hardware Requirements:** IBM® PC/Compatibles with 384K usable RAM, MS-DOS® 2.1, two floppy drives or a hard disk system – or Macintosh® 512KE or higher, 800K drive. Hayes® compatible modem also required.
***Limitations:** Over 20 monthly transactions involve a nominal surcharge ($3 per ten payments or portion thereof).

Black Holes in Galactic Centers

Some holes power brilliant quasars; most seem to lie dormant in relatively quiet galaxies like our own. Understanding these massive objects will help reconstruct the early history of the universe

by Martin J. Rees

Galaxies are the basic building blocks of the cosmos. The light from most galaxies represents the combined output from their tens or hundreds of billions of constituent stars. But for more than 25 years astronomers have known that some galaxies also have a bright, compact central nucleus whose emission does not come from normal stars. The most extreme instances of these so-called active galactic nuclei are quasars, objects no larger than the solar system whose total radiation exceeds that of 100 billion stars. Quasars seem to represent a particularly agitated stage in the development of some galaxies. Mounting evidence hints that many, perhaps most, young galaxies experience a phase of quasar or quasarlike activity.

Astronomers generally agree that gravity powers active galactic nuclei. Their tremendous brilliance and small size imply that the gravity source must be massive yet extremely compact. The best candidate for the central engine of quasars is a black hole—a collapsed body whose gravity is so great that nothing, including light, can escape from it.

The discovery of black holes in galactic centers, exciting in its own right, could affect current ideas about the evolution of the universe. Quasars appeared when the universe was less than one billion years old, indicating that some galaxies had already developed dense central regions. The early appearance of quasars rules out many cosmological models, which predict that the formation of galaxies should require billions of years, and even raises problems for the reigning cold dark matter model. Recent measurements of the cosmic background radiation intensify the puzzle. Most theorists think that galaxies formed from density variations in the newborn universe. Yet measurements of the background radiation indicate that any variations were so slight that it is difficult to understand how they could have produced the structures seen today. Apart from its cosmological significance, the detection of massive black holes also could elucidate predictions of Einstein's theory of general relativity.

MARTIN J. REES is director of the Institute of Astronomy at the University of Cambridge, England, where he also obtained his Ph.D. in astronomy. Rees has held visiting positions at the California Institute of Technology and at Harvard University and the Institute for Advanced Study in Princeton, N.J. His research interests include cosmology and galactic structure and evolution. This article is his fourth for SCIENTIFIC AMERICAN.

Theory holds that a galactic black hole could signal its presence. Gas swirling around and down toward the hole accelerates almost to the speed of light. The gas grows hot as a result of friction and becomes electrically conductive plasma. The black hole's gravity squeezes magnetic fields that form in the plasma, generating additional heat. Such hot, dense matter would radiate intensely. In this way, a black hole converts matter into energy with more than 10 times the efficiency of the nuclear reactions that cause stars to shine.

The presence of black holes in the centers of galaxies would not be surprising. Once galaxies began forming from the primordial material, stars or gas clouds in the densest central regions of the galaxies could have aggregated and collapsed into a black hole. Additional infalling matter would cause the black hole to "switch on" as a quasar. The energy emitted by the quasar may in turn affect the further formation of the galaxy.

The mass that accumulates in the center of an elliptical galaxy may be proportional to the total mass of the galaxy. For spiral "disk" galaxies like the Milky Way, the hole may be related not to the total galaxy mass but to the mass of the dense central bulge of stars, which is much smaller than a typical elliptical galaxy. The most luminous quasars are probably associated with the most massive black holes, those in elliptical galaxies. Spiral galaxies that contain only moderate-mass black holes could never have been powerful quasars.

Not all elliptical galaxies necessarily contain black holes. Indeed, even if a black hole once resided in a galaxy, it may no longer be there. Some theorists now believe that many elliptical galaxies are created when two or more galaxies merge: the orbits of their stars become mixed up, and the resulting pileup of stars takes on an elliptical shape. If the original galaxies contained massive central black holes, the holes would spiral together into the center of the merged galaxy and settle into a mutual orbit. The two black holes would approach each other ever more closely because orbital energy would radiate away as gravity waves, according to the theory of relativity. A powerful burst of gravitational radiation emitted during the final coalescence could create a recoil that would send the merged black hole hurtling into intergalactic space. Space-based probes could in principle detect these potent gravity waves and so confirm the validity of Einstein's ideas.

All these ideas remain frustratingly speculative. Investigators would like to learn the details of how quasars and other active galactic nuclei function, how common they are and what happens to them when they fade away. A better understanding of these energetic, enigmatic objects will help illuminate the manner in which galaxies form and evolve.

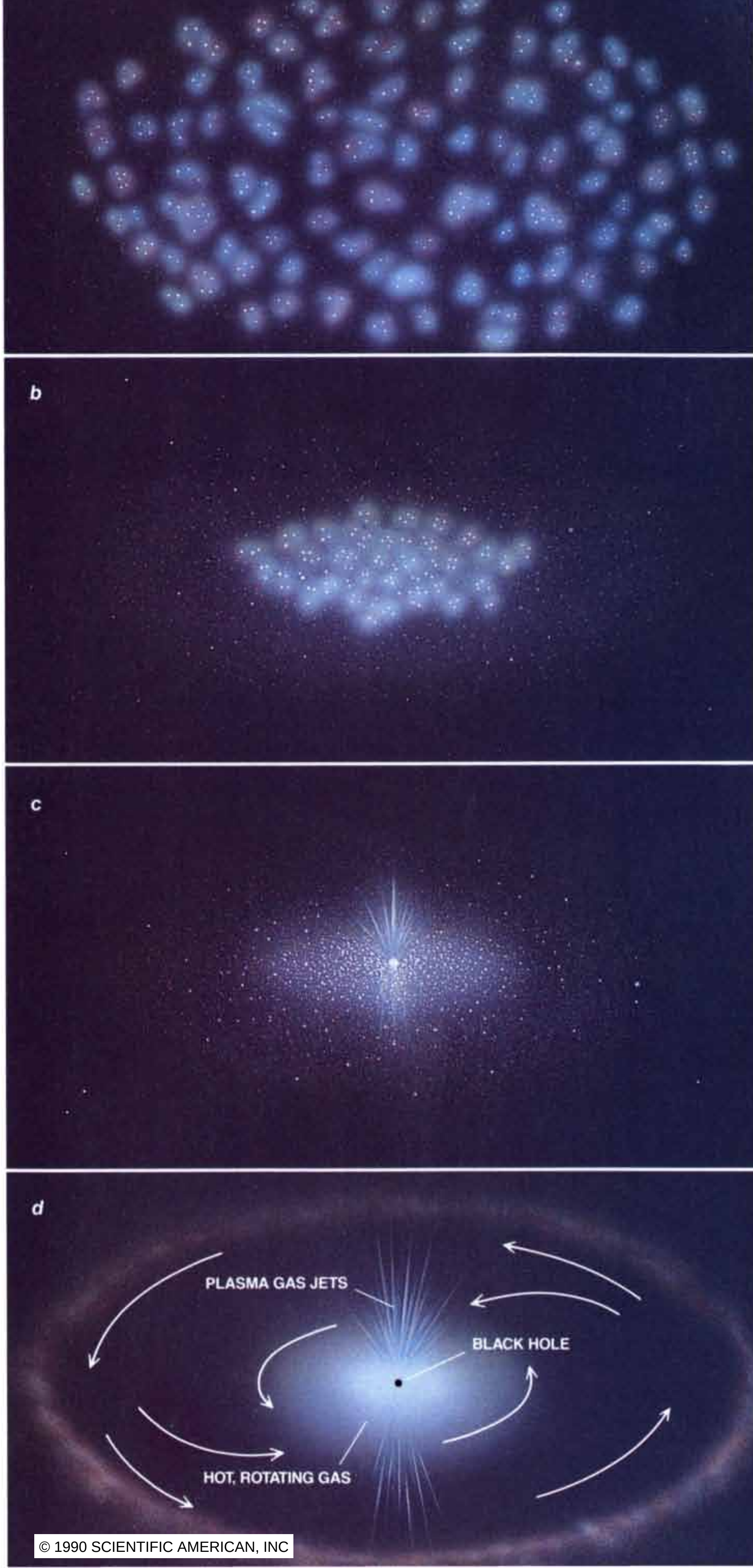
Quasars are difficult to study because they are distant in both space and time. They appear as they were billions of years ago, when the light now reaching the earth began its journey.

Astronomers measure large cosmic distances by means of redshifts, the extent to which an object's light has been stretched, or "reddened." Because of the overall expansion of the universe, every body is moving away from every other body (except over small scales), and two bodies recede from each other at a velocity proportional to the distance between them. This is analogous to the behavior of a balloon marked with dots: as the balloon is inflated, every dot on its surface moves away from every other dot. The greater the distance between two dots, the faster the stretching of the balloon carries them apart. No dot can be said to lie on the center of the balloon's surface, just as no galaxy can be said to lie at the center of the expanding universe.

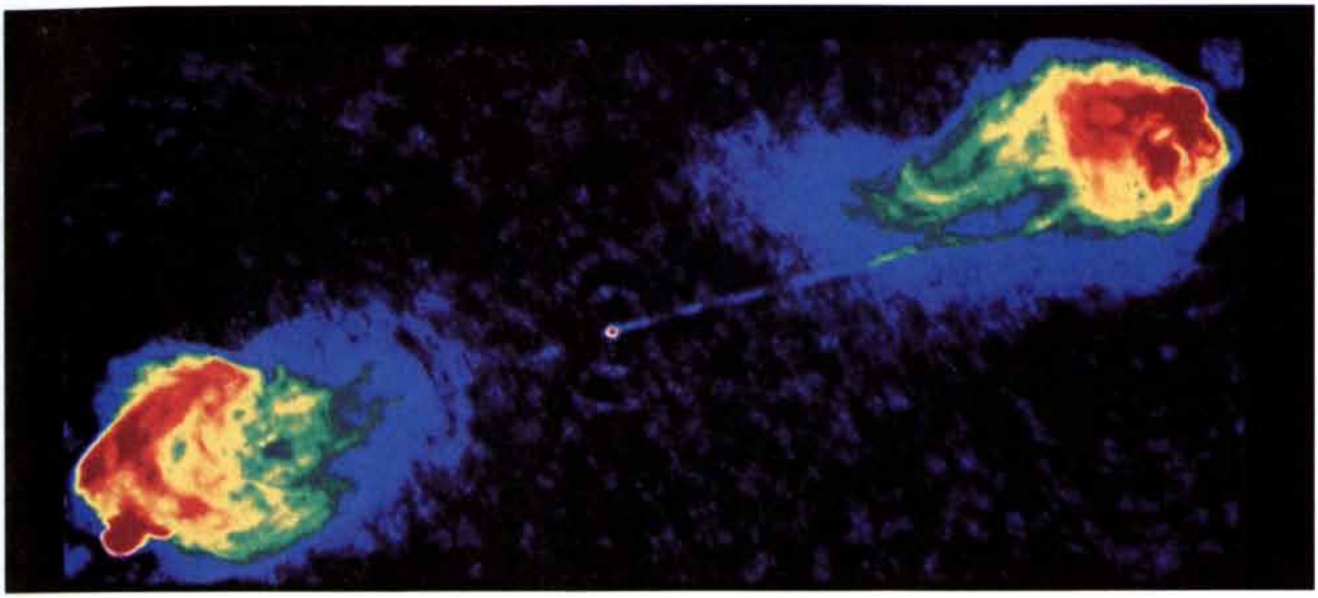
The most remote quasars have redshifts so large that the wavelengths of their radiation have been stretched more than fivefold between emission and reception. The visible light from these quasars that reaches the earth was originally emitted as much shorter-wavelength ultraviolet radiation.

The stretching, or redshift, factor can be thought of as the ratio between the present size of the universe and the size of the universe when the light set out. A quasar's redshift measures how large, and therefore how old, the universe was when the quasar emitted the light. When observing quasars, astronomers are traveling through cosmic history back to the era when the universe was less than one fifth its current size. The most favored cosmological model holds that the universe is about 13 billion years old and that these most distant quasars shone when the universe was only one billion years old.

Surveys of several thousand quasars reveal that the population has changed radically over time. When the universe was about two billion years old, quasars were far more common than they are today. Quasar activity peaked sharply at the epoch when the universe had about one third its present scale, about two to three billion years after the big bang. Apparently, significantly older quasars are rare not only because



BIRTH OF GALAXIES may be intimately tied to the formation of a black hole. According to one evolutionary theory, a large gas cloud collapses under its own gravity (*a*) and forms a flattened disk with a condensed central region, where many stars coalesce (*b*). The densest innermost region of the galaxy then collapses further into a black hole; surrounding material glows intensely as it falls into the hole (*c*). The size of the black hole probably depends on the size of the galaxy's central bulge. The largest holes power brilliant quasars (*d*).



CYGNUS A galaxy, the most intense radio source in the sky, radiates primarily from two lobes of plasma (ionized gas) hundreds of thousands of light-years across. The lobes probably are powered by hot jets that squirt out when gaseous matter falls toward a large black hole at the galaxy's center. The energy in the lobes is equivalent to millions of solar masses; the size and structure of the lobes imply that Cygnus A has been active for a few tens of millions of years.

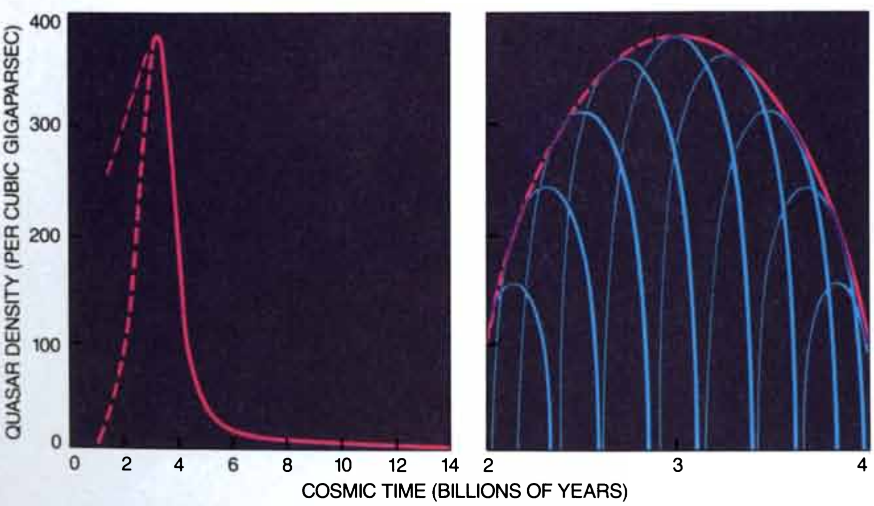
they are more distant and hence fainter but also because the quasar population genuinely thins out at redshifts corresponding to times less than two billion years after the big bang.

Most of the quasars that have ever been active are now evidently dead. The nearest bright quasar, known as 3C273, lies two billion light-years from the earth and has an apparent brightness of 14th magnitude, 1,000 times too faint to be seen with the naked eye. In contrast, during the "quasar era" some 11 billion years ago, the nearest quasar would have been only about 25 million light-years away and as bright as a fourth-magnitude star, easily visible to the unaided eye. One might, of course, expect quasars to have been closer together then simply because the entire universe was smaller and denser. But quasars were also 1,000 times more common then, relative to galaxies, than they are today.

The first quasars appeared surprisingly soon after the big bang. Astronomers have found at least a few quasars whose light has been stretched by nearly a factor of six, revealing that they existed when the universe was younger than one billion years. These old, distant quasars place tight constraints on theories of galaxy formation.

Presumably quasar activity could not start until after at least some galaxies had condensed and developed concentrated central regions. Only then would the gravitational potential at the center be great enough to lead to runaway gravitational collapse, the formation of a black hole and the birth of a quasar. The discovery that quasars appeared so early in the history of the cosmos is an embarrassment to some theories of galaxy formation, which hold that galaxies had not even begun to coalesce until much later.

The dramatic rise and fall of the quasar population, and indeed of all kinds of active galactic nuclei, are perplexing. Quasar activity is a feature of very young galaxies, perhaps because as a galaxy evolves, more of its gas becomes bound up in stars, leaving less fuel available for quasars.



QUASAR POPULATION was far larger and more concentrated in the early universe, roughly two to four billion years after the big bang, than it is now (*left*). At still earlier times, the population thins out, but the evidence is fragmentary. Astrophysical clues suggest that the slowly changing quasar abundance actually conceals many overlapping generations of relatively short-lived objects (*right*).

Observational surveys alone cannot reveal whether quasar activity was a brief feature of all young galaxies or just a highly visible aberration in a few unusual ones. To settle this question, one needs to know how long a typical quasar lives.

At present, less than one quasar exists for every 100,000 galaxies. Even during the quasar era 11 billion years ago, quasars were about 100 times less common than normal galaxies. One might infer from this that only one galaxy in 100 has ever indulged in quasar activity, and so only about 1 percent of all galaxies today should harbor remnants. But there is an alternative possibility: many short-lived generations of quasars could be born, evolve and die over the period when the overall population rises and falls. (By analogy, many generations of inhabitants might live and die as the population of a city grows and shrinks over the centuries.) Each individual quasar might shine only briefly, possibly much less than a billion years. If so, quasar remnants would be much more common than if only a single generation had existed.

One way to judge how long a quasar remains active is to estimate its mass. The most powerful quasars emit so much energy that they would have to swallow several tens of solar masses every year. If these quasars remained constantly active for a billion years, they would end up weighing tens of billions of solar masses. Even the more common, less luminous quasars would weigh more than one billion solar masses. If, on the other hand, many successive generations of short-lived quasars appeared and disappeared, each quasar would not have had time to accumulate such a large mass.

There is a good physical reason to think that quasars must weigh at least as much as 100 million suns, regardless of how long they live. Radiation exerts a slight pressure. Intense radiation pressure would expel all the matter from a quasar's vicinity unless gravity provided a sufficiently strong countervailing force. If quasars are powered by accretion, pulling gas in from their surroundings, then they must be massive enough for gravity to counterbalance radiation pressure. The luminosity of a quasar therefore implies a minimum possible mass. Typical quasars must weigh at least 100 million suns; the most powerful ones, at least one billion suns.

Another method for estimating quasar mass can be used for the minority of quasars that are radio sources. The radio emission generally emanates from two huge lobes of plasma energized by the quasar. The energy stored in the lobes is equivalent to the mass of millions of suns; assuming plausible levels of efficiency, the quasars must have accumulated a mass of at least 100 million suns in order to unleash so much energy. The structure of the radio-emitting lobes around these quasars also implies that they have been active fairly continuously throughout their lives.

Unfortunately, no one knows how much mass quasars are born with nor how much they grow before their fuel runs out. Spectroscopic studies reveal that quasars are strongly affected by radiation pressure, indicating that they do not grow to be so massive that gravity completely overwhelms the force of radiation. This reasoning implies that a typical quasar shines for no more than 50 million years. Such a relatively short lifetime implies that many, perhaps all, galaxies underwent a phase of quasar or quasarlike activity in their youth.

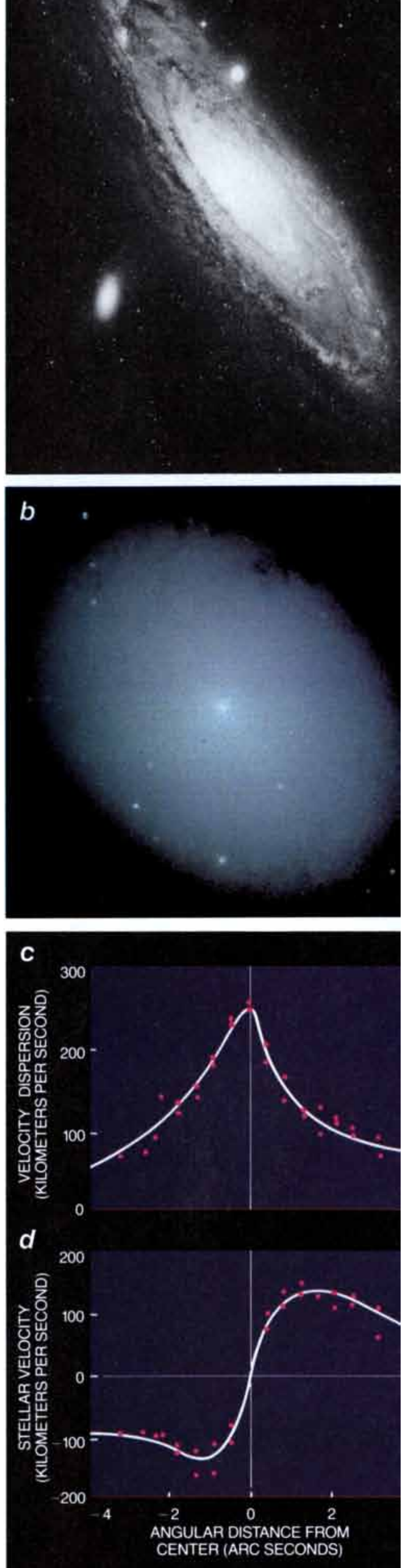
These clues—none very precise—are the only ones astronomers have on the masses and life spans of active quasars. Theoretical mass estimates can be refined by making separate calculations for subclasses of quasars having, for example, different luminosities and radio properties, but great uncertainty about the deduced lifetimes remains.

Another way to understand more about the nature of quasars and other active galactic nuclei is to search for their corpses, quiescent black holes that may lie in the centers of many nearby galaxies. In this way, one might measure directly the masses and abundances of these objects.

A massive black hole manifests itself as a quasar only as long as it is fueled by capturing gas and dust from its vicinity. In a vacuum, it emits no observable radiation at all. Even then, however, it exerts a detectable gravitational influence on surrounding bodies. A large black hole in the nucleus of a galaxy would pull stars toward it and cause them to pile up in orbits near the center. Any stars close to the black hole would orbit extremely fast.

Every star in the galaxy responds to the cumulative gravitational field of all the other matter in the galaxy. The stars orbit at speeds that balance the disruptive effect of their kinetic energy against the tendency of gravity to pull them all together at the center. Stars can be captured into small, rapid orbits if they approach the vicinity of a central black hole.

The radius of a black hole's "sphere of influence" on surrounding stars is directly proportional to its mass and depends inversely on the square of the



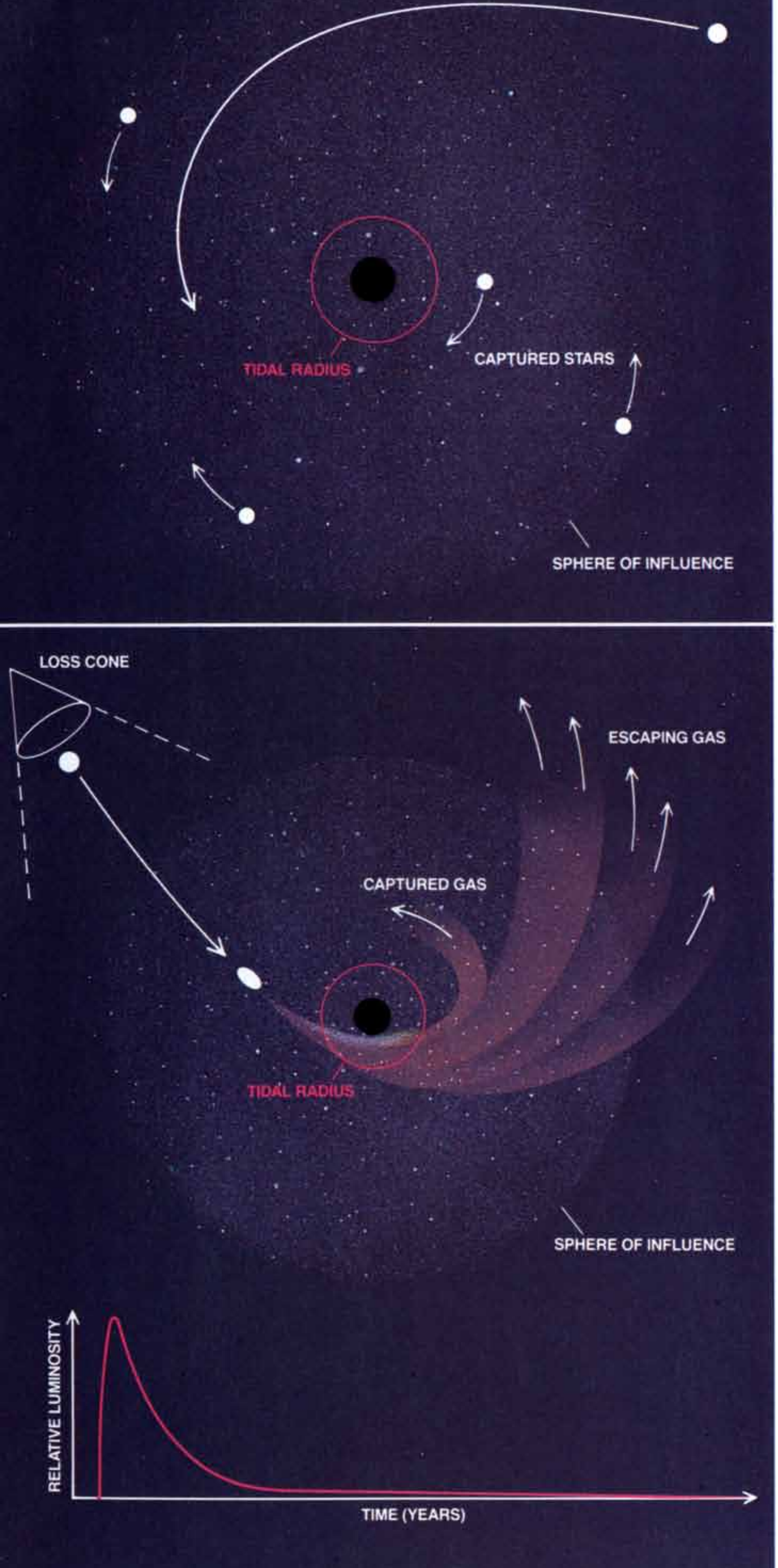
ANDROMEDA GALAXY (*a*) has a compact mass at its center, probably a black hole. A close-up of the nuclear region (*b*) shows the great concentration of stars there. Stellar velocity dispersion (*c*) and orbital velocities (*d*) increase sharply at the center, revealing the intense pull of a massive, invisible object.

stars' velocity. In other words, the faster a star is moving, the less it is affected by the black hole's gravity. The sphere of influence is millions of times larger than the black hole itself but still only a few tens of light-years across. Even in nearby galaxies, such a small region would subtend an angle of only a few arc seconds (one arc second is 1/3,600 degree).

One telltale sign of a quiescent black hole lurking in the heart of a galaxy would be a sharp spike in the galaxy's light profile, resulting from the concentration of stars within the black hole's sphere of influence. A number of researchers have looked for such spikes, but the findings always prove uncertain because some source other than stars, such as luminous gas clouds, may contribute light at the center.

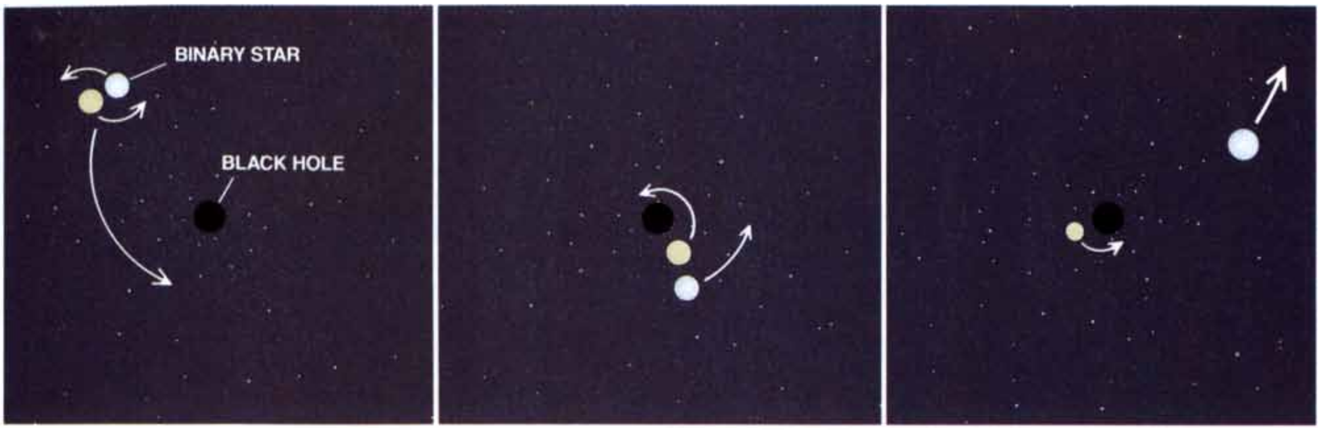
A second, less ambiguous signature of a black hole would be the presence of stars that are moving anomalously fast very near the galaxy's center. Individual stars of ordinary brightness cannot be detected in other galaxies, however, particularly in their crowded centers. The sharpest details that ground-based optical telescopes can resolve have angular sizes of about half an arc second. That angular size corresponds to a region five light-years across in the Andromeda Galaxy, which is about two million light-years away and is the nearest large spiral galaxy like the Milky Way; a region this size contains tens of thousands of stars.

Fortunately, one need not observe individual stars in order to deduce the relevant characteristics of their motion. Optical spectra of the combined light from the stars in the central region can provide the desired information. Distinctive dark features—absorption lines—always appear at certain wavelengths in the light emission of stars. In an orbiting cluster, some stars are moving toward the earth and others are moving away. The light from receding stars is stretched (redshifted), whereas the light from approaching stars is compressed (blueshifted). Lines in the combined spectrum of the cluster are blurred by the overlapping redshifts and blueshifts. The breadth of the lines



BLACK HOLE CAPTURES all stars within a critical radius with its gravitational pull, producing a distinctive cluster of fast-moving stars (*top*). Stars that approach the hole at small angles enter the tidal radius, within which the hole's gravity tears the star apart (*middle*). Some of the star's gas falls into the hole, creating a brief flare; most of the time, the hole would be quiescent (*bottom*).


How Serious Drivers
Gauge Their Excitement.



SLINGSHOT EJECTION can occur when a binary star closely approaches a black hole. One star may be captured while the other escapes at up to 10,000 kilometers a second. Normal encounters between stars cannot produce such velocities.

indicates the range of velocities of the stars. Particularly broad lines would signal the presence of a very dense object, perhaps a black hole.

In 1979 Peter J. Young and Wallace L. W. Sargent of the California Institute of Technology and their collaborators found spectral evidence for a compact mass of three billion suns in the core of the giant elliptical galaxy M87. This galaxy is a strong radio source and has an unusual jet of luminous matter that seems to originate from its core—indicators of central activity, albeit at an energy level far below that of a quasar. M87 is a good candidate for having a central black hole, although a mass of three billion suns, implying that M87 hosted a bright quasar, seemed surprisingly large to some theorists. At the distance of M87, 35 million light-years from the earth, the effects of a less massive black hole could not be detected using current instruments.

The credibility of the M87 observations has been debated throughout the past decade. A pivotal question is whether a dark mass must exist at the center to account for the high stellar velocities measured or whether the concentration of stars could itself contribute enough mass to account for the motions. One significant problem has been the difficulty of distinguishing between light from stars near the nucleus and nonstellar light emitted by whatever powers the luminous jet.

Studies of the Milky Way's nearest galactic neighbors have yielded more definitive results. Because these galaxies are more than 10 times closer than M87, a central mass concentration of only 10 to 100 million suns might reveal itself. Unlike M87, the nearby galaxies show no signs of activity, even at low levels. As previously noted, a black hole will not broadcast its presence if no matter is falling in.

Optical astronomers have found indirect evidence for compact central masses in several nearby galaxies. The most interesting and convincing case involves the Andromeda Galaxy. In the 1960s Martin Schwarzschild and his colleagues at Princeton University built a small telescope called Stratoscope, which was flown on a high-altitude balloon to evade the blurring effect of the lower atmosphere. On its last flight, in 1971, they discovered that the stars in the central few light-years of the Andromeda Galaxy's core form a flattened disk. In the past few years Alan M. Dressler of the Carnegie Institute, Douglas O. Richstone of the University of Michigan and John Kormendy of the University of Hawaii have taken spectra to study the motions of these stars. They find that the velocities rise sharply toward the center. Also, the flattened stellar system rotates in a manner that indicates that the mass at its center is extremely concentrated. The mass of the central object is estimated to be approximately 30 to 70 million suns.

Other nearby galaxies exhibit similar features. John L. Tonry, now at the Massachusetts Institute of Technology, showed in 1984 that the small galaxy M32 (a close neighbor of the Andromeda Galaxy) seems to harbor a central dark object of about five million solar masses. The aptly named Sombrero Galaxy was studied by B. J. Jarvis and P. Dubath at the European Southern Observatory and by Kormendy with the Canada-France-Hawaii telescope on Mauna Kea in Hawaii. The Sombrero Galaxy appears to have a rapidly rotating stellar core surrounding a concentrated central mass of almost one billion solar masses.

But are these compact dark masses necessarily black holes? The lack of a large central spike in the light distribution in the Andromeda Galaxy implies that the central object, whatever it may be, emits very little light relative to its mass. Could an unusual population of dim stars be concentrated near the center? A tightly knit cluster of smaller black holes (10 to 100 solar masses each), of neutron stars or of brown dwarfs (objects with less than one tenth the mass of the sun, which are insufficiently massive to shine) could contain considerable mass yet provide little light. Why or how such objects would accumulate in the innermost 10 light-years of a galaxy is unknown, but their existence cannot be ruled out.

Sharper observations could settle this question. If the stars closest to the center of a galaxy move still faster than those one arc second from the center—implying that the dark mass is concentrated within the inner one light-year, not merely the inner 10—then many of the dark star candidates would be ruled out. A dense cluster of stars only one light-year across would evolve quickly owing to frequent stellar encounters. Over a short period, the heaviest stars would collect at the center and coalesce into a central body, presumably a black hole. It is highly implausible that a dense central stellar cluster could have survived intact in a galaxy 10 billion years old.

If large central black holes truly are common, they must also be extremely quiet. Astronomers invoke black holes to explain some of the most luminous and efficient radiation sources in the universe. But there is no sign of elevated levels of activity in the Andromeda Galaxy, the Sombrero Galaxy or M32. Observations reveal that Andromeda emits no more than one ten thousandth the radiation of a quasar. In most galaxies the inner region may well be swept clean of gas.

Even a dormant black hole could

Joseph Jachna, *Door County*

occasionally become bright again if it were to consume a star from the galaxy's inner regions. Each star in the galaxy traces out a complicated orbit under the combined gravitational influence of all the other stars and of the hole itself. The orbits gradually change or diffuse owing to the cumulative effect of encounters with other stars. The encounters may shift a star into an orbit that brings it close to the hole. A star may occasionally plunge straight into the hole.

A star can approach a black hole only so far without suffering damage. Gaseous bodies like stars are sensitive to tidal effects, the gradient in the gravitational pull across a body. If a star passes too close to a massive black hole, the different gravitational stresses on the parts of the star closest to and farthest from the hole will tear it apart. Any star passing within the "tidal radius" will therefore be destroyed.

The tidal forces at the edge of a black hole are more gentle for holes of larger mass. The radius of the black hole proper increases in direct proportion to its mass, whereas the tidal radius grows only as the cube root of the mass. For very massive black holes, such as the three-billion-solar-mass hole postulated to exist in M87, the tidal radius actually lies within the event horizon, the dark "surface" that marks the black hole's point of no return. Stars approaching the hole would be disrupted only after passing irreversibly inside its horizon. The situation is different for smaller black holes of only five to 100 million solar masses (like those observed in nearby galaxies), whose tidal radius is 10 to 100 times larger than the hole itself. Stars would be disrupted while far enough from these holes that the distinctive radiation from the encounter could escape.

Such events should occur about once every few thousand years, according to stellar dynamics calculations. The exact rate depends on the statistics of the stellar orbits and particularly on how quickly new stars move into vacated orbits that come close to the hole. When a star is disrupted, the sudden release of gas inevitably produces some radiation. The terminal flash of a disrupted star could be the clearest signal of a black hole's presence. The theory of relativity predicts what a flash should look like; observing one will constitute yet another test of this theory.

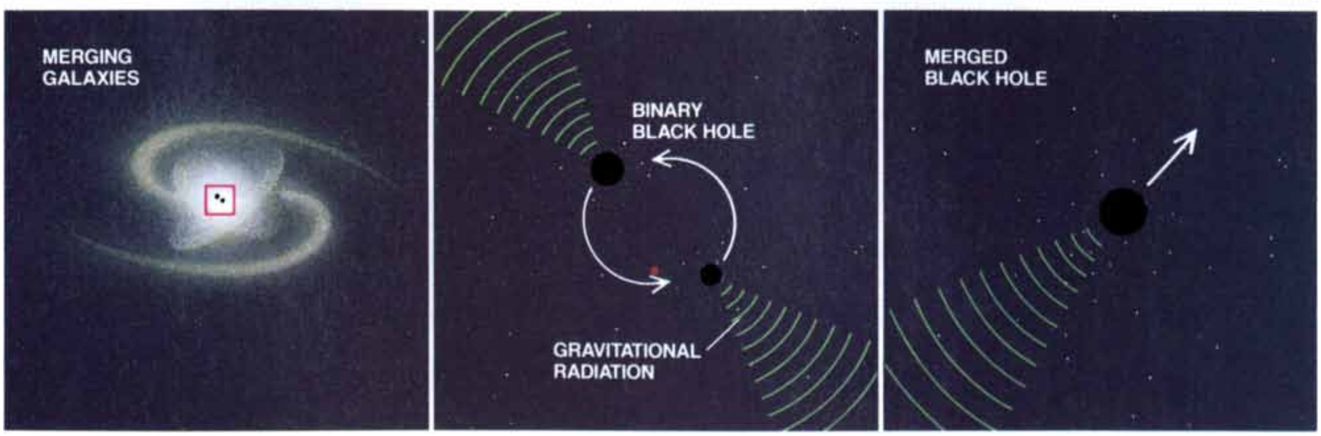
When the tidal forces from the black hole start to compete with a star's own gravity, the material of the star responds in a complicated way, stretching along the orbital direction, contracting at right angles to the orbit and becoming strongly agitated and shocked. Accurately modeling this phenomenon poses an as yet unsolved problem for computer simulations. Within a few years improved simulations should reveal what happens when stars of different types become tidally disrupted and what the observational signature of such events would look like. Some relatively crude calculations that my colleagues and I have already performed should, however, be adequate to convey at least the essence of what goes on.

The debris from the star spreads out onto a range of orbits about the hole. About half of the matter remains gravitationally bound to the hole. The rest fans out into hyperbolic orbits at speeds up to 10,000 kilometers a second. What would an observer see if such an event occurred in the center of the Andromeda Galaxy? The most conspicuous effects would result from material that swirls down into the hole. The nucleus of the galaxy would appear to flare up to a power output much greater than that of a supernova—almost comparable to the brightness of a quasar—for about one year.

It is hard to calculate how much of this radiation emerges as visible light as opposed to higher-energy ultraviolet or X-ray radiation. Current models also cannot predict how fast the brightness of the flare fades, which depends on how long it takes for the star to be digested. This point is important because astronomers would like to know whether radiation from the region around the black hole drops below detectable levels before the next stellar disruption occurs, 1,000 to 10,000 years later. The longer a flare remains visible, the more likely it is that one would now be detectable in a nearby galaxy.

If most galaxies harbor black holes, then a sample of the nearest 10,000 galaxies (out to the Virgo Cluster, about 60 million light-years away) should contain a few galaxies now at the peak of a flare episode. Other, presumably more numerous, galaxies should be in a state in which the lingering effects of the most recent tidal disruption are still discernible. Sky surveys set up to detect supernovas in other galaxies are being used to search for these flares. It would be highly worthwhile to monitor all 1,000 or so galaxies in the Virgo Cluster annually for evidence of a stellar disruption by a black hole.

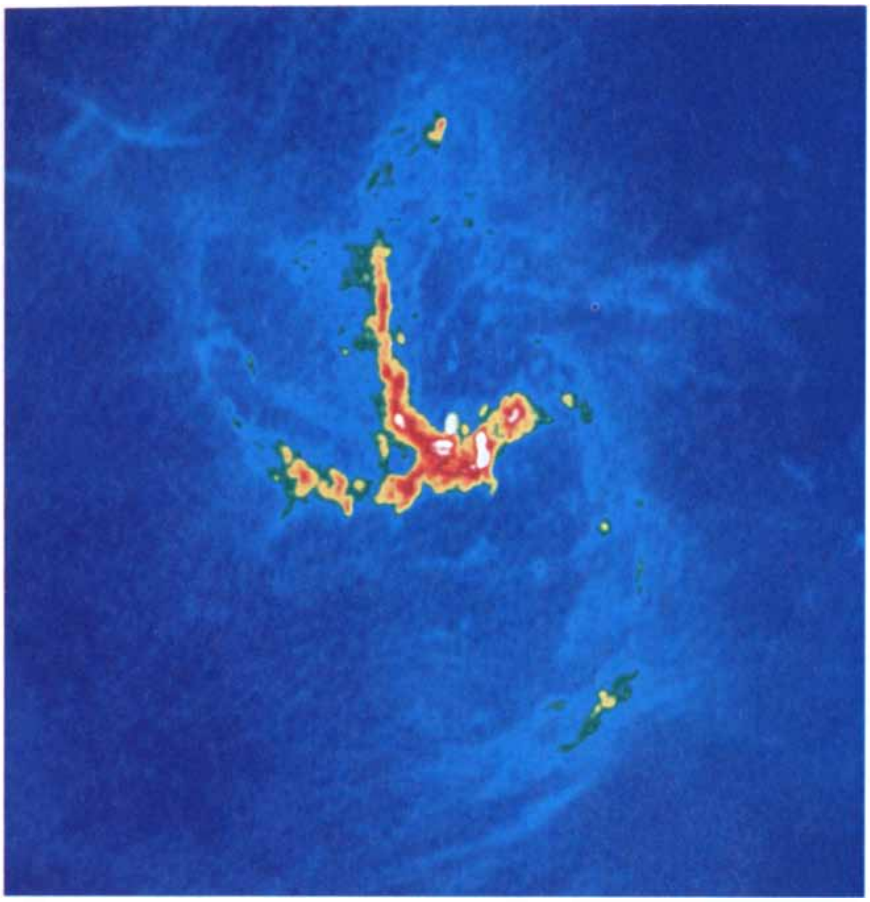
If black holes reside in most nearby galaxies, our own galaxy, the Milky Way, would seem underprivileged if it did not contain one as well. The complex structures near the center of the Milky Way include several infrared sources that are asymmetrically placed with respect to an intense, compact radio source [see "What is Happening at the Center of Our Galaxy?" by Charles H. Townes and Reinhard Genzel; SCI-



MERGING GALAXIES bring the black holes from two galactic centers into orbit about each other. The holes emit gravitational radiation, which causes their orbit to decay. Eventually the two combine into a single object. If their masses are unequal or if the black holes are rotating, the final blast of gravitational radiation may eject the merged hole from the galaxy.

REFLECTIONS OF GRANDEUR.

To send a gift of Grand Marnier® Liqueur (except where prohibited by law) call 1-800-243-3787.
Product of France. Made with fine cognac brandy 40% alc/vol (80 proof). ©1989 Carillon Importers, Ltd., Teaneck, NJ.



RADIO MAP of the center of the Milky Way shows disturbed arms of gas, perhaps the result of the disruption of a star by a black hole. The motions of the gas and stars near the center also indicate the presence of a heavy, compact dark mass.

ENTIFIC AMERICAN, April]. The source sits at or very near the gravitational center of the galaxy and appears quite unlike any other radio source in the galaxy. Its emission can be well explained by the accretion of small quantities of gas onto a black hole of a few million solar masses. Gas clouds within a few light-years of the center seem to be moving rapidly under the gravitational influence of a compact mass of two to three million solar masses.

The expected occasional stellar disruptions would energetically expel matter from the black hole's vicinity, and in fact there is a hollowed-out central cavity filled with tenuous, hot gas. Disrupted gas streamers and other evidence of recent violent activity hint that the power produced at the center has been higher in the past than it is at present. Perhaps the armlike gas structures several light-years from the galactic center delineate the trails of debris from the most recent tidal disruptions.

A good indicator of the presence of a black hole in our galactic center would be the existence of exceptionally fast-moving stars that had been accelerated away from the center by a gravitational slingshot effect. (This effect was exploited for the *Voyager* probes: the gravitational pull of each planet studied helped to accelerate the probes on to their next destination.) The slingshot effect cannot impart a velocity greater than the main body's escape velocity, the critical velocity necessary to break free of the body's gravitational grip. Stars ejected from double or multiple star systems therefore cannot attain velocities exceeding about 1,000 kilometers a second, the escape velocity from a typical star.

When a binary star passes close to a central massive black hole, one star may be captured into an orbit about the hole while the other escapes at very high speed. The escaping star would fly out with a velocity much greater than 1,000 kilometers a second, according to calculations by Jack G. Hills of the Los Alamos National Laboratory.

The discovery of even one star moving outward from the galactic center with far more than the normal stellar escape velocity would be compelling evidence for the existence of a massive central black hole. The density of stars at the heart of the Milky Way is poorly known, but if 1 percent of the stars are binaries, once every 10,000 years a star should be shot away from the black hole with a velocity of about 4,000 kilometers a second. At this rate, each runaway would take about two million years to travel 30,000 light-years, the sun's distance from the center. About 200 (two million divided by 10,000) of these superfast stars should therefore be moving outward from the center but still within the radius of the sun's orbit through the galaxy. Provided they are bright enough to be detected, such stars should in principle be easy to recognize because they would move rapidly across the sky.

Superfast stars ejected from the black hole acquire their kinetic energy at the expense of their former binary companions, which find themselves in orbits trapped deep within the hole's gravitational well. In addition to seeking superfast stars, astronomers might search for stars orbiting close to the galactic center at velocities in excess of 10,000 kilometers a second. Such high orbital velocities would tear apart a normal multiple star system.

So far the evidence in favor of a black hole at the center of the Milky Way is quite persuasive. Was our galaxy ever a quasar? Almost certainly it was not: a black hole of less than three million solar masses could never have produced the necessary energy. But less energetic, quasarlike activity probably has been a part of our galaxy's history. Studying the Milky Way and its close neighbors is therefore providing insights into the early lives of galaxies and into the nature of some of the most powerful and exotic objects in the universe.

FURTHER READING

EVIDENCE FOR A SUPERMASSIVE BLACK HOLE IN THE NUCLEUS OF M31. J. Kormendy in *Astrophysical Journal*, Vol. 325, No. 1, pages 128–141; 1988.

SUPERMASSIVE BLACK HOLES. Edited by Minas Kafatos. Cambridge University Press, 1988.

THE MEMBRANE PARADIGM FOR BLACK HOLES. Kip S. Thorne and Richard H. Price in *Scientific American*, Vol. 258, No. 4, pages 69–77; April, 1988.

THE RISE AND FALL OF QUASARS. M. Schmidt in *Highlights of Astronomy*, Vol. 8. Kluwer Academic Publishers, 1989.

CONSTRAINTS ON THE MASS DISTRIBUTION NEAR THE CENTER OF M31 AND M32. Douglas Richstone, Gary Bower and Alan Dressler in *Astrophysical Journal*, Vol. 353, No. 1, Part 1, pages 118–122; April 10, 1990.

"DEAD QUASARS" IN NEARBY GALAXIES? Martin J. Rees in *Science*, Vol. 247, No. 4944, pages 817–823; February 16, 1990.

Unfortunately, The Best Is Yet To Come

Oracle database software lets businesses take advantage of each new generation of low cost computing.

Whatever types of computer a business buys today, there's sure to be something significantly better and cheaper tomorrow.

Yet companies continue to sink money into software that runs on only one kind of computer. Locking themselves out of newer, more cost-efficient computers.

This trap is avoidable for most companies. All they need is the right software. Software that works with virtually every computer and network. Present or future.

Oracle has become the world's largest database software company by providing just that. Software that runs on virtually every type of mainframe, minicomputer, workstation and PC.

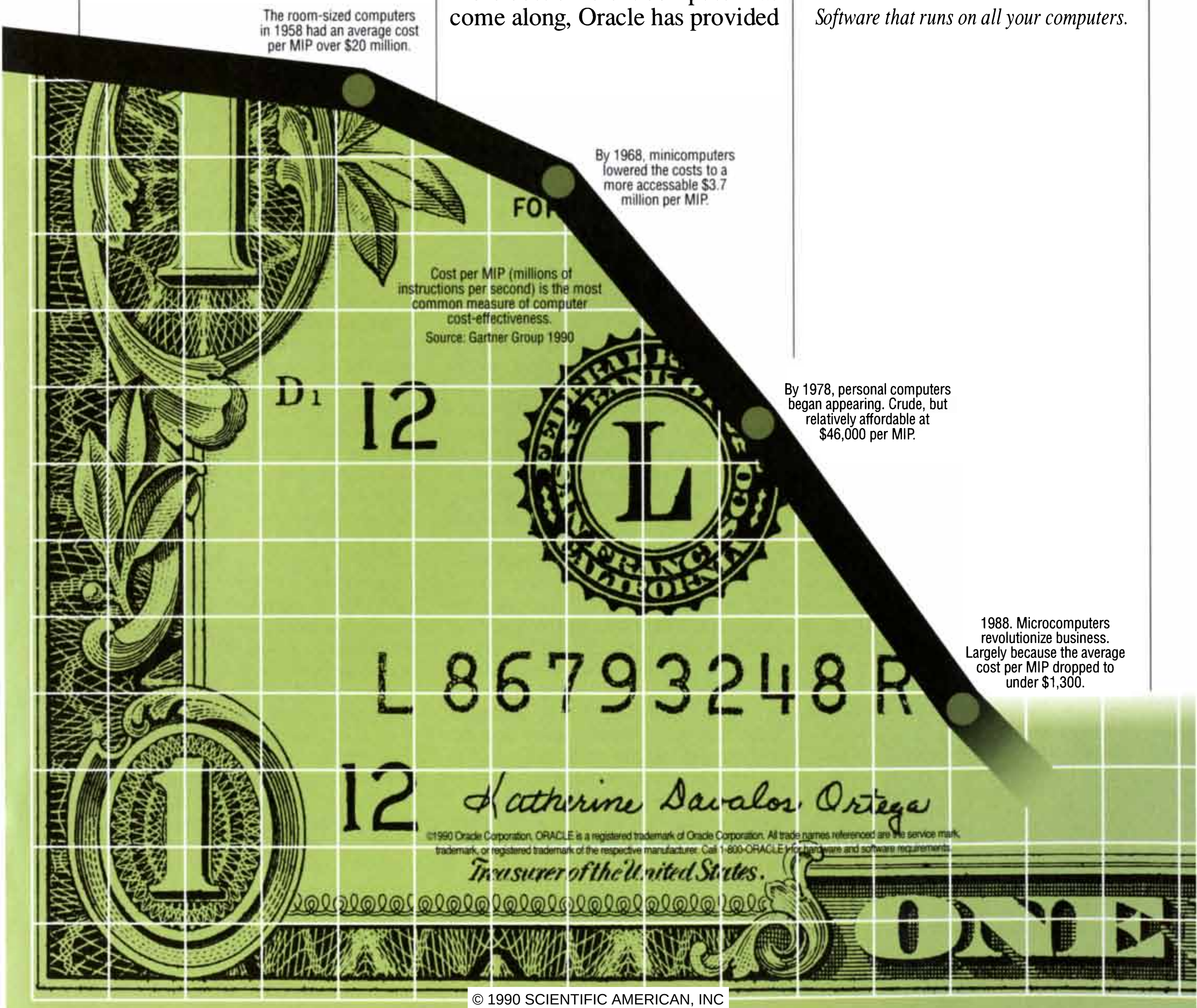
And every time a newer, faster, more cost efficient computer has come along, Oracle has provided the database software not only to run on it, but allow it to share data with existing computers as well.

Call 1-800-633-1062 Ext. 1813 for more information.

Fortunately, you no longer have to predict the future to take advantage of it.

ORACLE®

Software that runs on all your computers.



Gene Therapy

Treatment of disease by introducing healthy genes into the body is becoming feasible. But the therapy will not reach its full potential until the genes can be coaxed to work throughout life

by Inder M. Verma

One infant in every hundred is born with a serious genetic defect. Usually the damage becomes evident in childhood. All too often, it gives rise to physical or mental abnormalities, pain and early death. Of the more than 4,000 known inherited disorders, most lack fully effective therapies.

It is no wonder, then, that scientists have long imagined curing heritable ills by introducing healthy genes into patients. Advances in recombinant DNA technology, which have made possible the isolation of many genes, and new insights into gene regulation are beginning to make this once impossible notion seem feasible.

Indeed, the first federally approved clinical trial of a gene therapy for a genetic disease began this past September. R. Michael Blaese, W. French Anderson and their colleagues at the National Institutes of Health (NIH) are introducing the gene for the enzyme adenosine deaminase (ADA) into children suffering from a rare condition known as severe combined immunodeficiency (SCID). Derangement of this gene debilitates the immune system and is responsible for about 25 percent of all cases of SCID.

The approach of the NIH group requires repeated treatments throughout life, and so it is not a cure. Still, the trial could represent the start of a new era in medicine. The current pace of research suggests that by the turn of the next century clinical trials of gene therapies may be under way for any of a number of diseases—inherited and otherwise.

Genes can be transferred either into germ cells (sperm, eggs or early embryos) or somatic cells (those not destined to become sperm or eggs). Yet germ-line therapy is not an option for the foreseeable future, in part because the new genes would be passed from generation to generation, a prospect that raises profound ethical concerns.

For instance, should therapy be applied simply to improve one's offspring, not only to prevent an inherited disease? Who would be empowered to decide? Is society willing to risk introducing changes into the gene pool that may ultimately prove detrimental to the species? Do we have the right to tamper with human evolution? The prospect of somatic cell therapy is less troubling, mainly because it would affect only the treated patient.

The most promising candidates for somatic cell therapy are disorders caused by impairment of a single gene that has been isolated and cloned and so is available for transplant. These diseases should be simpler to correct than those caused by multiple genes or by such global disturbances as the loss or addition of whole chromosomes. (Normally, human cells carry one set of 23 chromosomes inherited from the mother and a corresponding set from the father. Every chromosome consists of a long stretch of DNA and includes thousands of genes.)

In the ideal world, the diseases would be cured for life by one treatment, with no side effects. And gene insertion into a chromosome in a target somatic cell would be site specific: in what is called homologous recombination, the healthy, or "therapeutic," gene would exactly replace the damaged copy. Targeted insertion increases the probability that a therapeutic gene will function correctly. It also reduces the likelihood that random insertion will activate a quiescent oncogene (a cancer inducer) or inactivate a cancer suppressor.

In reality, investigators have found it extremely difficult to control the fate of DNA introduced into cells. For every gene spliced into the correct place, more than 1,000 fit randomly into the genome (the total DNA in a cell). Work by Mario R. Capecchi of the University of Utah suggests that the obstacles to site-specific gene delivery are great but surmountable. Meanwhile many laboratories, including my own at the Salk Institute in La Jolla, Calif., are concentrating on developing gene augmentation therapy, in which a healthy gene replaces the product of a missing or defective gene but does not physically replace the flawed DNA itself.

Augmentation can be helpful when a genetic derangement results in little or no production of a protein. (Each gene encodes, or carries instructions for, a single protein.) Low production occurs when mutations hamper the activity of both the maternal and paternal copies of a gene or when a hobbled gene is inherited on a male's only X chromosome. (The cells of males carry one X and one Y chromosome; those of females carry two X chromosomes.)

On the other hand, augmentation therapy might not be of much help when a mutation yields overproduction of a protein or the synthesis of a destructive substance, as is the case in sickle cell anemia. To correct those kinds of disturbances, therapy would often have to include delivery of both a healthy gene and one capable of inactivating the mutated version.

For now, most scientists interested in gene augmentation are planning to remove cells from patients, introduce a therapeutic gene and return the altered cells to the subject. Some day, however, physicians may directly inject patients with genes linked to substances that will deliver those genes to specific target cells.

Fortunately, genetic flaws do not necessarily have to be corrected in all of

INDER M. VERMA is professor of molecular biology and virology at the Salk Institute in La Jolla, Calif., and adjunct professor of biology at the University of California, San Diego. He joined the institute in 1974 after earning a doctorate in biochemistry at the Weizmann Institute of Science in Israel and completing postdoctoral studies at the Massachusetts Institute of Technology.

the body's trillions of cells in order for therapy to work. First, even though every somatic cell in an individual carries identical chromosomes, certain genes function only in a single cell type. Treatment, then, could focus only on that type. Second, even when a genetic defect results in insufficient synthesis of a protein made in virtually every cell, many cells compensate for the loss. For instance, a flaw in the ADA gene affects most somatic cells to a degree but is devastating only to some constituents of the immune system.

Nontargeted delivery of genes into cells can be accomplished by chemical or physical means (transfection) or by viruses (transduction). In chemical approaches, one mixes many copies of DNA carrying the healthy gene with a charged substance—typically calcium phosphate, DEAE-dextran or certain lipids. Then the mixture is essentially dumped onto recipient cells. The chemicals disturb the cell membrane and transport the DNA into the interior.

The procedure is simple, but the efficiency of gene delivery is dismal. Usually only one cell in 1,000 to 100,000 integrates the gene of interest into its genome. A physician would have to obtain an impossible number of cells from patients to guarantee the appropriate alteration of the millions required for therapy.

I should point out that integration is not always crucial to gene expression (production of the encoded protein). Still, a gene that is integrated is likely to last longer in the cell. Further, it should replicate whenever the rest of the DNA does, as when a cell prepares to divide. The therapeutic gene would thus be inherited by the daughter cells and by their daughters and so on, thereby ensuring a supply of the product throughout a patient's life.

Physical methods include microinjection with a fine glass pipette and electroporation (the exposure of cells to an electric shock). The shock renders cells permeable to DNA in the surrounding medium, but it can also severely damage them. Microinjection can be extremely efficient; perhaps one cell in five takes up the foreign gene permanently. Yet because only a single cell can be injected at a time, this tedious, labor-intensive approach is not suitable for therapeutic purposes.

The final strategy capitalizes on the native ability of viruses to enter cells, bringing their own genetic material with them. Many of these organisms have now been engineered to serve as vectors, or delivery vehicles, for gene transfer. Viruses can be grouped according to whether their genetic material is DNA or RNA. The two substances have important chemical differences, although both are built from units known as nucleotides and both include regulatory codes in addition to those specifying the sequences of amino acids in proteins.

Many DNA viruses that can accept foreign genetic material turn out to be severely limited in the number of nucleotides they can accommodate and in the range of cells they infect. Certain other DNA viruses are roomier but have so far proved unusable for var-

STERILE BUBBLE protected a boy named David, who suffered in the 1970s from severe combined immunodeficiency, or SCID, an inherited disorder in which the immune system is profoundly impaired. SCID patients have better options today and may have more in the future: the first gene therapy approved for clinical trial aims to ease a form of the disorder.

ious reasons. Moreover, DNA viruses often do not splice their genetic material into the chromosomes of the cells they infect.

As is true of the DNA viruses, most RNA viruses are unsuitable for gene therapy, mainly because RNA, which cannot integrate into the DNA of human cells, is degraded rapidly. Varieties known as retroviruses are an exception. They actually convert their RNA to DNA in infected cells and insinuate the DNA into a chromosome. The integrated DNA then directs the synthesis of viral proteins. Retroviruses can entertain more foreign genetic material than some DNA viruses. They can also infect a broad spectrum of species and cell types.

For these reasons, retroviruses are the most promising gene-delivery systems studied thus far. Indeed, unless specified, all approaches to gene transfer discussed in the balance of this article are based on these vectors.

Retroviruses are, of course, not without obvious drawbacks. For instance, they can merge their DNA into a chromosome only in cells capable of actively dividing. Yet many cells do not normally divide—among them, mature neurons—and so they are not readily amenable to being genetically altered by retroviral vectors.

More disturbing is the possibility that retroviruses can cause cancer. The risk is extremely low for the species that have been considered as vectors, but it increases if the viruses are allowed to multiply in the body and spread from cell to cell. Consequently, a major challenge has been devising ways to stop the vectors from reproducing.

The efforts of several laboratories have together yielded at least one technique that seems to work well [*see illustration on page 72*]. The organisms produced by that method have a normal outer coat and contain all of the virus's proteins. The retroviral RNA, however, includes no instructions for synthesizing viral proteins. The therapeutic gene takes the place of those missing instructions.

The coat enables the viruses to enter cells and deliver the viral contents to the cell's cytoplasm. Then viral enzymes convert the RNA to DNA and help to fit that DNA into the genome of the host cell. But that is the end of the line for the virus.

Under normal circumstances, integrated retroviral DNA—called the provirus—would direct the synthesis of viral proteins and RNA, which would then assemble into clones of the original virus. In contrast, the altered retrovirus, bereft of instructions for making viral proteins, produces no progeny. The virus essentially disappears from the cell, leaving behind only the foreign gene and nucleotide sequences that now serve merely to facilitate the expression of the gene.

Although retroviruses can infect many cell types, only certain target cells can be considered for genetic manipulation. The cells must be strong enough to withstand handling and capable of being removed from the body and returned with reasonable ease. In addition, they should be long-lived, surviving for months or years or preferably for the patient's entire life. Because bone marrow, skin and liver cells best meet these criteria, diseases that can be treated by manipulating these cells are among the most promising candidates for gene therapy.

The cells of the bone marrow, where blood is produced, can in theory be exploited to correct disorders caused by genetic flaws in red blood cells or in white blood cells (which are important in immunity). SCID caused by an ADA deficiency is but one of several inherited conditions affecting the immune cells; another is leukocyte adhesion deficiency, which involves the poor mobilization of white blood cells and leads to recurrent infections. Among the diseases associated with impaired red blood cells are the thalassemias, which reflect impairments in the genes encoding subunits of the hemoglobin molecule—the oxygen carrier in red blood cells.

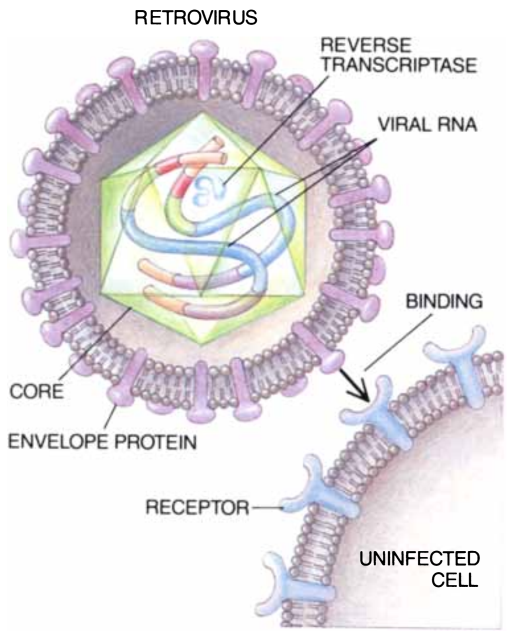
Beta thalassemia was once expected to be the first disorder treated with gene therapy. Its history illustrates some of the problems that have beset the effort to develop gene therapy in general and therapy based on bone marrow cells in particular.

Red blood cells of patients stricken with beta thalassemia are deficient in beta globin, which in healthy individuals combines with alpha globin and iron (heme) to yield hemoglobin. Healthy cells regulate the activity of both genes precisely, ensuring that equal amounts of alpha and beta globin are made. The lack of beta globin gives rise not only to a deficit in hemoglobin production but also to a relative excess of alpha globin. This excess, in turn, hastens cell death and can cause severe anemia. Usually patients succumb to the disease by age 20, after years of pain and suffering.

This disease and other inherited blood disorders could probably be treated efficiently by delivering healthy genes to stem cells, the subset of cells in the marrow that gives rise to the full spectrum of blood cells and replaces dead cells throughout a person's life. Stable introduction of a desired gene into a stem cell could guarantee the production of normal blood cells for as long as a patient lives.

Sadly, human stem cells are far from abundant and are virtually impossible to isolate. Researchers have therefore been forced to resort to a less efficient strategy: infecting enormous numbers of bone marrow cells with a therapeutic retrovirus in the hope that enough stem cells will be infected.

Studies of beta globin have supplied much of the evidence showing that the approach has at least some merit.



LIFE CYCLE of a retrovirus begins when the virus binds to (*above*) and enters (*right*) a cell and injects its genetic material (RNA) and proteins into the cytoplasm. Typical retroviral RNA includes three coding regions: *gag* (*green*), *pol* (*blue*) and *env* (*purple*), specifying, respectively, proteins of the viral core, the enzyme reverse transcriptase and constituents of the coat. It also has three noncoding domains—two at the tips (*light orange*) and another called psi, ψ (*red*). In the cytoplasm, reverse transcriptase converts the RNA into DNA, whose lengthened terminal domains, called long-terminal repeats (*dark orange*), influence the activity of viral genes and facilitate insertion of viral DNA into cellular DNA. The ensconced DNA (the provirus) directs the synthesis of viral proteins and RNA. The proteins then enclose the RNA, forming viral particles that bud from the cell.

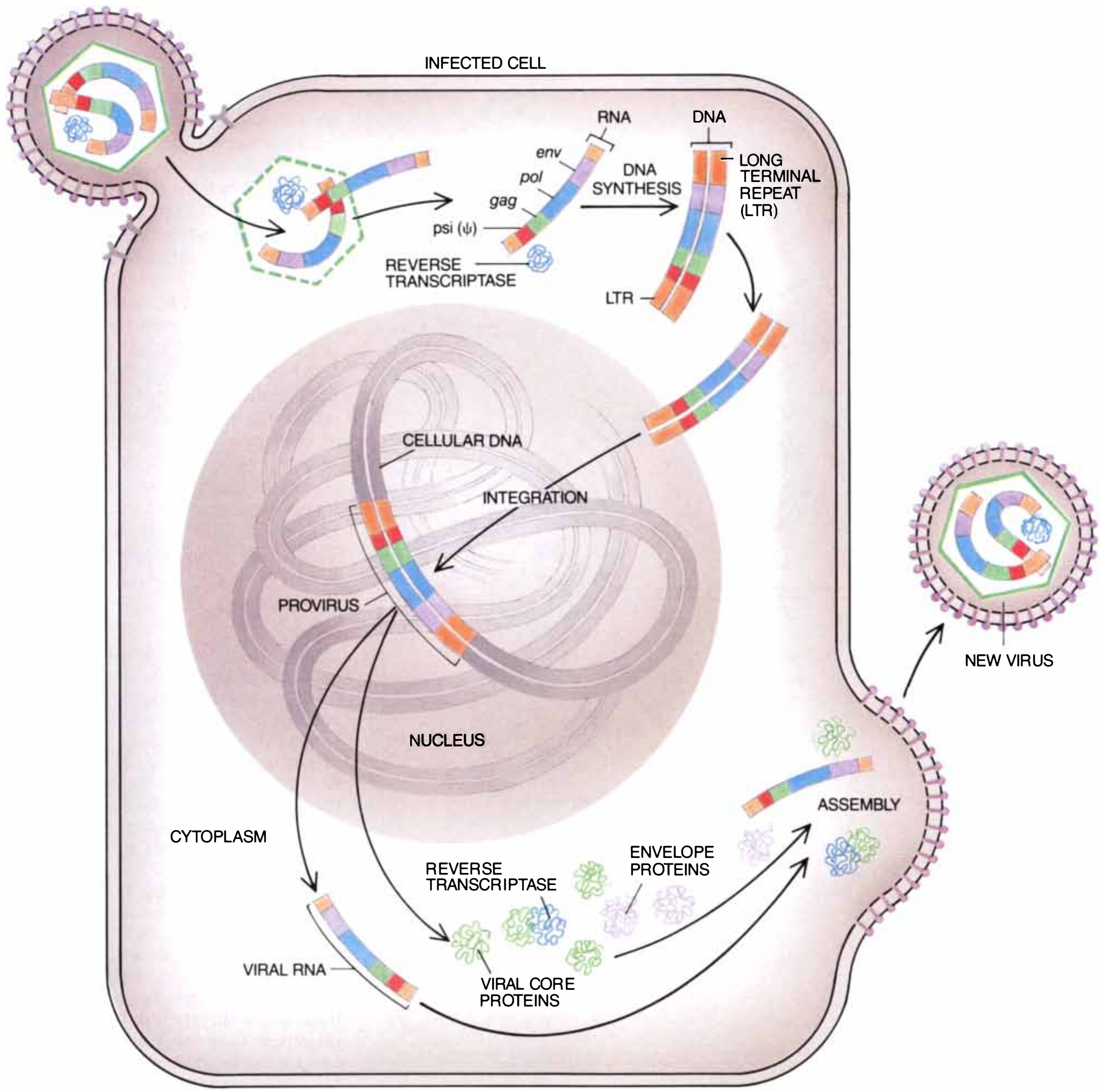
For instance, several laboratories have shown that a human beta globin gene inserted into mouse bone marrow cells by retroviral vectors stays in the cells. And Richard C. Mulligan and his co-workers at the Whitehead Institute for Biomedical Research in Cambridge, Mass., have further shown that the human gene is expressed when such cells are implanted in mice.

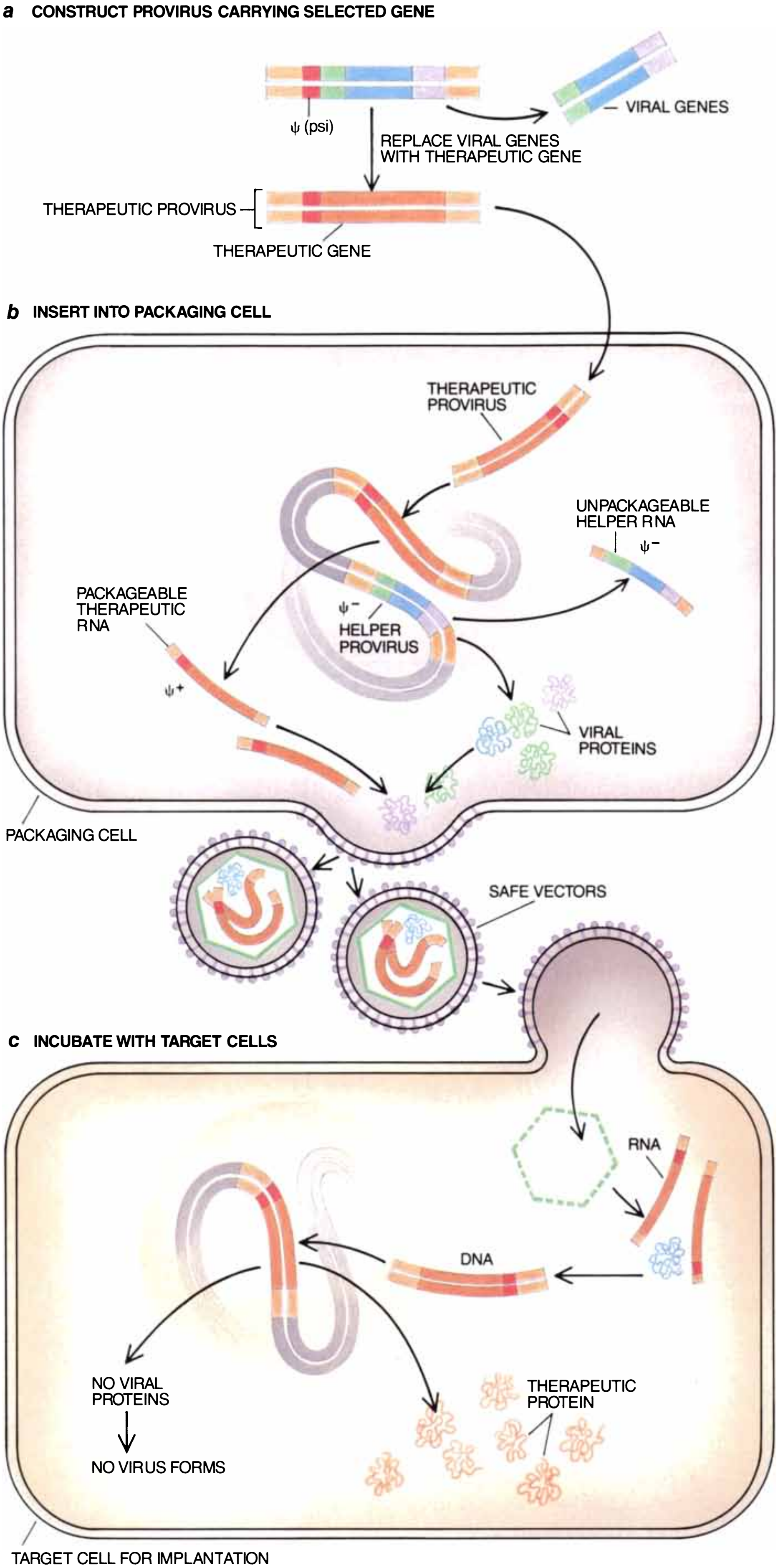
On the other hand, no one has been able to achieve significant levels of globin synthesis in recipient animals. This problem has been a major disappointment, but a discovery by F. G. Grosveld and his colleagues at the National Institute for Medical Research in London offers hope for a solution.

They identified distinct stretches of DNA, thousands of nucleotides apart from the gene itself, that in normal red blood cells dramatically boost the production of globin messenger RNA. Messenger RNA is transcribed, or copied, from DNA and is the template from which protein is made; hence, high levels of a messenger RNA indicate that the encoded protein is being produced in abundance. It seems reasonable to think that linking globin-specific enhancers to a globin gene in a retroviral vector might enhance globin synthesis in the body. Studies of this hypothesis are in progress.

In general, genetically altered bone marrow cells have yielded poor in vivo expression of other genes as well. The problem must be resolved before gene therapy based on bone marrow cells can become a reality.

Along with an acceptable level of gene expression, one would hope for long-term activity. Recent findings relating to globin indicate that achieving prolonged expression of genes inserted in bone marrow may be less problematic than attaining high levels of protein synthesis. For instance, Chung L. Li and V. J. Dwarki in my laboratory have produced sustained, albeit weak, expression of the human beta globin gene in mice for at least a five-month study period—the equivalent of 15 to 20 years in a human being. The alpha





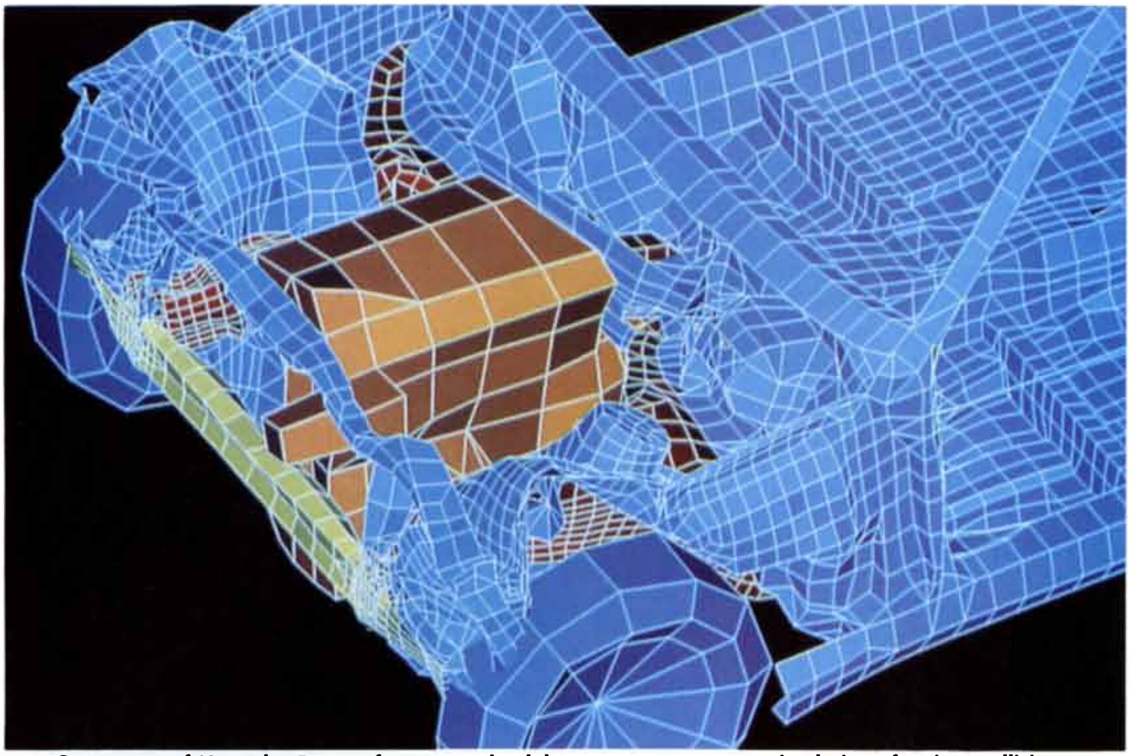
globin gene remained functional for at least 10 months.

Other findings emerging from the work on beta thalassemia highlight the complexity introduced when correction of a disease requires precisely regulated expression of a therapeutic gene. For many disorders, including SCID, simply producing some amount of a missing protein is better than none. The same is not true for thalassemia. Because a relative excess of either alpha or beta globin can damage cells, the activity of a therapeutic globin gene must exactly mimic that of a normal version. Unfortunately, the mechanisms that control the activity of genes are understood only imperfectly—both for the beta globin gene and for most others. Discoveries are made constantly, however, and are helping improve the design of vectors for gene therapy.

SCID researchers at the NIH have taken a detour from gene therapy based on bone marrow cells, in part because of the ongoing problem of poor expression. Patients in their study are treated with a select subset of circulating *T* lymphocytes, white blood cells crucial to immunity. *T* cells are devastated by a lack of ADA.

The retrovirally altered lymphocytes are infused into children who are now being helped somewhat by injections of PEG-ADA—ADA mixed with the chemical polyethylene glycol to increase the enzyme's half-life. Success of the approach will be measured by improvements in immune function beyond that achieved by enzyme replacement alone. Regrettably, *T* cells do not have the longevity of stem cells, which is why the disease cannot be cured indefinitely by one treatment.

RETROVIRAL VECTORS are assembled, or packaged, in cells designed to release only safe vectors. Investigators substitute a therapeutic gene for viral genes in a provirus (*a*) and insert that provirus into a packaging cell (*b*). The viral DNA directs the synthesis of viral RNA but, lacking viral genes, cannot give rise to the proteins needed to package the RNA into particles for delivery to other cells. The missing proteins are supplied by a "helper" provirus from which the psi region has been deleted. Psi is crucial to the inclusion of RNA in viral particles; without it, no virus carrying helper RNA can form. The particles that escape the cell, then, carry therapeutic RNA and no viral genes. They can enter other cells (*c*) and splice the therapeutic gene into cellular DNA, but they cannot reproduce.



One aspect of Mercedes-Benz safety research: elaborate supercomputer simulation of serious collisions.

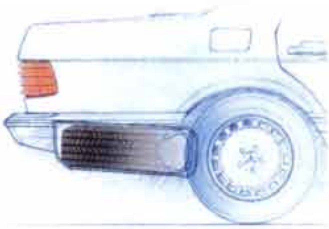
Engineering Safer Automobiles

Automobile safety has become, in a remarkably short period of time, a top priority in the hearts and minds of the car-buying public. "Safety is sexy!" notes a prominent industry expert.

Engineers in this field are gratified that the importance of this aspect of car design has finally been acknowledged. Mercedes-Benz assumed the role of aggressive safety pioneer as early as 1939, when a Mercedes engineer named Bela Barenyi set up the company's first safety laboratory. The 1951 patent awarded to Mercedes for

a car body that absorbed crash energy front and rear by deforming, while protecting passengers within a rigid cabin, became the design basis for most modern cars. For more than thirty years, Mercedes engineers have conducted systematic crash tests to examine the effectiveness of their designs. During the last two decades, Mercedes engineers have studied thousands of serious traffic accidents in pursuit of information that might improve crashworthiness and avoid occupant injuries.

As a result of these efforts, Mercedes-Benz now builds what it believes to be the safest automobiles on the road. This bold claim is supported not only by design efforts and field investigations, but also by independent findings. According to the Highway Loss Data Institute (HLDI), a non-profit public service organization that gathers, processes, and publishes motor vehicle insurance data, the Mercedes-Benz S-Class models outscored every other automobile in an average of results over the last three years. The frequency of S-Class overall injury and severe injury claims was less than half the average for all cars rated. Other Mercedes-Benz models also scored impressively in the HLDI report against cars in their class.



When possible, the spare tire is placed to help absorb extra energy in a rear impact.

OVERVIEW

Automotive safety can be broken into three primary categories: the driver, the road, and the vehicle. The driver's role is, unfortunately, addressed least often. Nevertheless, recent progress is heartening. The enforcement of statutes prohibiting driving while impaired by alcohol or drugs is more stringent than ever before. Even so, there is much progress to be made in this area.

The second category, the road, encompasses not only the driving path but also traffic, weather conditions, information signs, and roadway markers.

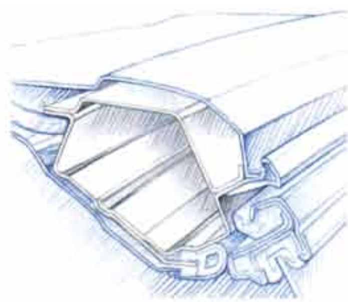
The car's role is critically important because the vehicle must often make up for shortcomings in the other two categories. A quarter of a century ago, Mercedes-Benz researchers split the role of the automobile into two subcategories: active and passive safety. Active safety means avoiding accidents whenever possible. Passive safety provides protection from injury (or reduces the chance of injury) when a collision cannot be avoided.

This special report will concentrate on passive safety—how Mercedes-Benz engineers use a variety of sophisticated tools and research techniques at their disposal to design and develop automobiles with an extra measure of occupant protection.

ACCIDENT INVESTIGATIONS

A certain number of automobile accidents are, unfortunately, inevitable. By studying select accidents scientifically and fastidiously, Mercedes-Benz engineers discover the effectiveness of their safety measures already in production. Real-world accident investigations also provide areas for new research and ideas for more advanced safety systems in future models.

In 1969, Mercedes-Benz formalized an incipient accident investigation program centered in Sindelfingen (near Stuttgart), Germany. Today, the program is more active than ever. A dedicated team is notified by the national traffic police any time a current model Mercedes-Benz is involved in an injury accident within a 90-mile radius. The team quickly travels to the scene and, with the car owner's cooperation, collects three categories of pertinent information: the exact nature of any personal injury, the type and extent of vehicle damage, and road and weather conditions.



Roof-beam rigidity and low weight are achieved via complex box design and construction.

To date, several thousand serious accidents involving occupants of Mercedes-Benz automobiles have been investigated and analyzed. In recent years, the research has been extended to encompass injured pedestrians and cyclists.

Automotive Safety

The information gathered at the accident site is used by Mercedes engineers to reconstruct the collision. In some instances, the injuries suffered can then be correlated to causal mechanisms, such as contact with a particular area of the vehicle's interior. Injury severity is classified according to the widely accepted Association for the Advancement of Automotive Medicine's Abbreviated Injury Scale (AIS).

Accident severity is classified during analysis by studying the accident scene and vehicle damage. An Energy Equivalent Speed (EES) is determined from a simple change-in-kinetic-energy equation: $W_D = \frac{1}{2}m(EES)^2$, where W_D is the deformation energy and m corresponds to the mass of the car. EES is, in principle, not identical to the car's change in velocity but it does approximate that value when there is neither glance-off nor an extreme discrepancy in the rigidity of the colliding objects.

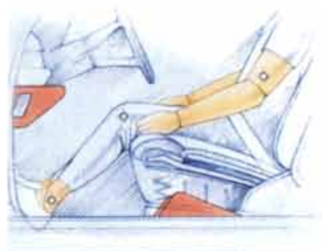
With a valuable data base of accident research in place, safety engineering could be solidly grounded.

With a data base of accident severity and injury causation mechanisms in place, laboratory crashes and computer simulations (the second and third prongs of the Mercedes-Benz safety engineering effort) can be solidly grounded in the real world. Meanwhile, Mercedes accident investigations have also allowed the company to analyze its own efforts. Compared to Mercedes-Benz cars from model years 1968 to 1975, current Mercedes models involved in serious accidents have produced less than half the number of severe injuries.

LABORATORY EFFORTS

To develop new components, safety systems and entire vehicles, Mercedes-Benz engineers have a variety of sophisticated laboratory methods at their disposal. Most dramatic is the Mercedes-Benz X-MP/24 supercomputer built by Cray Research, Inc. It can be programmed to simulate the performance of mechanical and structural systems under a broad range of circumstances with uncanny speed. The Cray is capable of 700 million calculations per second. The preferred tool of aerospace engineers the world over, this supercomputer is used to analyze such problems as the stresses in an engine's crankshaft, a theoretical suspension's reaction to a bump in the road and the combustion process as it occurs in slow motion.

When applied to automobile safety, the Cray allows Mercedes-Benz engineers to study a collision event lasting but a few seconds or less for days at a time. It can permit three-dimensional visualization of such critical details as an air bag's deployment, the movement of the driver's torso, and the crush of a critical part of the vehicle's structure. By applying finite-element analysis, engineers can convert an incomprehensibly complex system into a mosaic of smaller, simpler units. A simulated crash may involve 100,000 finite elements studied over 1,000 increments of time, necessitating billions of mathematical calculations. The engineer can arrest the process at will, view the system from any desired angle, zoom into any given area for a closer look, or glean quantitative information such as the exact velocity, deflection, or stress at any location throughout the entire system.



A wedge-shaped seat insert and padded knee bolster help reduce impact injuries.

The Cray's nuts-and-bolts counterpart, crash testing, began at Mercedes in 1959. More than 80 test cars were sacrificed during the first year to study barrier crashes, vehicle-to-vehicle impacts, and rollover performance. Initially, an out-of-doors cable system was used; later a steam rocket accelerated test cars to their destruction. In 1972, an elaborate crash test center was completed. It permitted moving all tests inside the laboratory for more consistent results. Now, a linear electric motor below floor level moves the test car over 300 feet into a concrete barrier weighing 1,120 tons. Fixed barrier, rollover, rear-end, and lateral-impact crash sites are all contained under one roof. Two sets of 127

HIGHEST SAFETY RATING OF ANY CAR OVER THE 1988-1989-1990 PERIOD.*

MERCEDES-BENZ RANKED THE MOST DEPENDABLE CAR IN AMERICA.†

AMERICA'S MOST IMPRESSIVE RECORD OF RETAINED VALUE.**

THE WORLD'S MOST DURABLE CAR.††

*Highway Loss Data Institute, 1988, 1989 and 1990. † J.D. Power & Associates Vehicle Dependability Index Study Of Original Owners. Based on things gone wrong to 4- to 5-year-old model vehicles in the past twelve months. **Over a ten-year period, Mercedes-Benz cars, as a line, have retained a higher percentage of original value than any other make. Urban Science Applications, Inc. Model Years 1978–1987. †† A 1957 Mercedes-Benz 180D, with over 1,184,880 miles, cited by the *Guinness Book of Records*. © 1990 Mercedes-Benz of N.A., Inc., Montvale, NJ.

floodlights illuminate these areas so that very high-speed film (1,000 to 5,000 frames per second) can be used to record a test from several angles. Three separate sled-type crash simulators permit the testing of restraint systems, seats, steering systems, instrument panels—in short, every car component. More than 100 full-car and several hundred lesser crash tests are conducted every year.

High frequency of offset-frontal collisions prompted invention of a unique body.

OFFSET-FRONTAL COLLISIONS

The studies of real-world accident statistics begun in 1969 by Mercedes-Benz engineers revealed several interesting facts. More than sixty percent of the personal-injury accidents involved a frontal collision of one type or another, demonstrating the clear need for effective restraint systems, energy-absorbing front structures, and interior surfaces capable of yielding on contact. Looking further, Mercedes engineers determined that frontal impacts could be categorized by type. Roughly 25 percent of them involved full, direct contact—the classic "head-on" collision with another vehicle or a static barrier. In another 25 percent of the frontal impacts, only a portion of the right (passenger's) side of the car actually made contact with another car or fixed object near the roadway. The most common type of frontal impact, accounting for fully half of all frontal collisions or about one-third of all personal-injury accidents, involved a partial overlap with the driver's side of the automobile.

Moreover, it was quite clear that 70 percent of the severe injuries were caused by frontal collisions.

In 1968, the Department of Commerce National Highway Safety Agency mandated that all cars subsequently sold in the U.S. must meet a number of Federal Motor Vehicle Safety Standards (FMVSS) which addressed both the design of equipment (such as windshield wipers) and the performance of such safety gear as seat belts and head restraints during actual crash tests. The most rigorous test required by the federal government—FMVSS 208—is a full-frontal impact into a fixed barrier at a speed of 30 mph. Another test—FMVSS 201—assures that all interior contact areas are designed to minimize trauma to the human body. FMVSS 203 demands that the steering system yield to cushion impact by the driver. FMVSS 204 limits the rearward displacement of the steering column to reduce the likelihood of chest, neck, and head injuries. The level of crash protection is measured by means of instrumented anthropomorphic devices or "test dummies."

The Federal Safety Standards were indeed a major step forward. Today, before any new model may be sold in America, all manufacturers must certify that their cars pass barrier crash and all other applicable tests (such as roof-crush and side-intrusion resistance, fuel-system integrity, flame resistance for interior materials, etc.).

Nevertheless, the Mercedes-Benz safety

WORLDWIDE CHRONOLOGY OF SAFETY FIRSTS

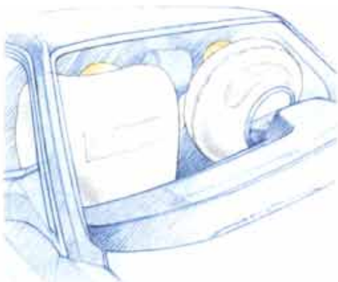
1933 Power-assisted brakes introduced by several U.S. carmakers.
1938 Mercedes-Benz introduces self-adjusting internally expanding drum brakes.

1939 Mercedes-Benz builds the first safety prototype with an extremely stiff floor structure, a multipiece steering column, and side-impact protection.
1940 Oldsmobile introduces the first automatic transmission (Hydramatic).
1946 Michelin of France patents radial-ply tire construction.
1951 A German patent issued to Mercedes-Benz for the fundamental principle of passive safety: "a rigid

passenger cell enclosed by front and rear crumple zones."
1956 Lap belts are first offered in cars by Ford and Chrysler.
1956 Jensen and Triumph introduce disc brakes.
1959 Mercedes produces a truly "occupant friendly" interior with an energy-absorbing instrument panel, an upholstered steering wheel hub, and deformable control knobs.
1959 Mercedes conducts its first laboratory impact and rollover tests.
1959 Volvo introduces three-point restraints for front-seat passengers.

Automotive Safety

effort continues to seek advances. In 1975, Mercedes-Benz added a new and vital test to the development process of any new Mercedes-Benz model: a frontal impact into an offset barrier with approximately 25- to 40-percent overlap of the car's width. The driver's side of the car offers the far more critical test because of the location of the steering mechanism and pedal controls. A 25- to 40-percent overlap is effectively the worst case, concentrating extreme loads on only the left side of the vehicle's structure. Less than that amount of overlap and the car will generally ricochet or spin away. With more overlap, the car's engine helps spread the impact energy into the right side of the structure.



Air bags for driver and front passenger are available on many Mercedes-Benz models.

In very specific terms, a "survival space" must be preserved for occupants even during an offset-frontal collision. Among other things, this imposes strict limitations on the rearward displacement of the steering wheel and instrument panel. Also, the firewall cannot be significantly deformed, the sidewall structure must stay intact, and the doors should open with the application of reasonable force to facilitate the rescue of occupants.

In order to satisfy the above criteria, Mercedes-Benz has developed body structures designed to maximize the distribution of collision forces. Introduced in the 190 Class in 1984 and developed to its most sophisticated level in the SL model introduced in 1989, it consists of two longitudinal members in front that yield progressively upon impact. Relatively narrow at the car's bumper, they thicken as they approach the passenger cabin, and then ramify into three branches connected to the transmission tunnel, the side rail, and the floor. This yoke diverts impact forces to as many adjacent areas as possible. By involving other areas of the car, impact energy directed to only one side of the car's front can be absorbed rather than cause occupant compartment intrusion. Mercedes also designs and locates major components (such as the braking and electrical units) so that in a collision, they avoid stacking up and actually aid in the car's absorption of crash energy.

SAFETY EQUIPMENT IN MERCEDES-BENZ AUTOMOBILES

What distinguishes a Mercedes-Benz from every other automobile on the road? Its robust body structure is perhaps the most critical element. Every component must help absorb and dissipate crash energy to preserve the integrity of the passenger compartment. The body structure must be highly rigid in certain areas, yielding in others. Within this sophisticated structure are myriad provisions designed to minimize the greatest risk of injury: contact between an occupant and a hard surface.

The three-point seat belt, first offered for front seats in Mercedes automobiles in 1968,

1963 Four-wheel disc brakes with two independent hydraulic circuits standard on all Mercedes-Benz models.

1967 Mercedes-Benz and several U.S. makers introduce a Mercedes-patented steering column which deforms upon impact to reduce injury potential.
1971 Mercedes-Benz introduces headlamp washer/wiper.
1973 Front and outboard-rear three-point seat belts with takeup reels and head restraints standard in all Mercedes models.

1974 Driver-side air bags introduced on some General Motors models.
1978 Mercedes-Benz and BMW introduce the first four-wheel antilock brakes with digital electronic control.
1980 Mercedes-Benz introduces an air bag for the driver and a seat belt emergency tensioning retractor for the front passenger.
1987 Mercedes-Benz introduces passenger-side air bags in S-Class models.

1989 Three-point belt integrated into the Mercedes SL's seat, shoulder-belt height determined by head-restraint height. SL also has automatic roll bar.

is the single most injury-reducing automobile feature. In 1984, pretensioning devices were introduced that automatically remove slack from the belts if frontal-impact force is great enough. As a complement to using seat belts, Mercedes-Benz introduced a supplemental (air bag) restraint for the driver the same year. The front-passenger air bag was introduced in 1987.

In addition, every Mercedes has numerous safety features. Among the more notable, incorporated into most models because of research findings, is a brake pedal support designed to induce the pedal assembly to pivot away from the driver's feet during a major offset impact on the driver side. Wedge-pin door locks are designed to be strong enough to avoid springing open and yet not jam in a side impact. The instrument panel is designed with relatively flat surfaces and an energy-absorbing substructure to minimize head and knee injuries. A shock-absorbing footwell element made of expanded polystyrene reduces impact stress on lower extremities. Wood-veneer trim on vertical surfaces is backed by one or more layers of sheet aluminum to help prevent splintering. Doors are lined with expanded polyurethane to prevent sharp edges and provide cushioning in the event of a side impact. The door pocket resists sharp-edged fracture even at 20 degrees below zero. The armrest has a built-in polyurethane shock absorber. The inside rearview mirror has a spring-loaded base designed to release upon very low impact.

The fundamental goal of Mercedes engineering is a safer driving environment for all.

FUTURE THINK

Advancing automotive safety is a continuous process at Mercedes-Benz. Accomplishments merely provide a foundation for future progress.

In the realm of passive safety, one new challenge is improved protection from side impacts. This is particularly difficult because the space available for energy-absorbing structures or a more sophisticated restraint system is extremely limited. Nevertheless, Mercedes-Benz engineers are now working to perfect advanced side-impact design features for future production models. Better offset-frontal impact behavior is also being pursued. In 1984, Mercedes-Benz began testing to determine crumple-zone behavior in an offset collision with a telephone pole—the most extreme crash testing to date. In addition, testing for pedestrian and two-wheel-vehicle impacts intensified in 1988.

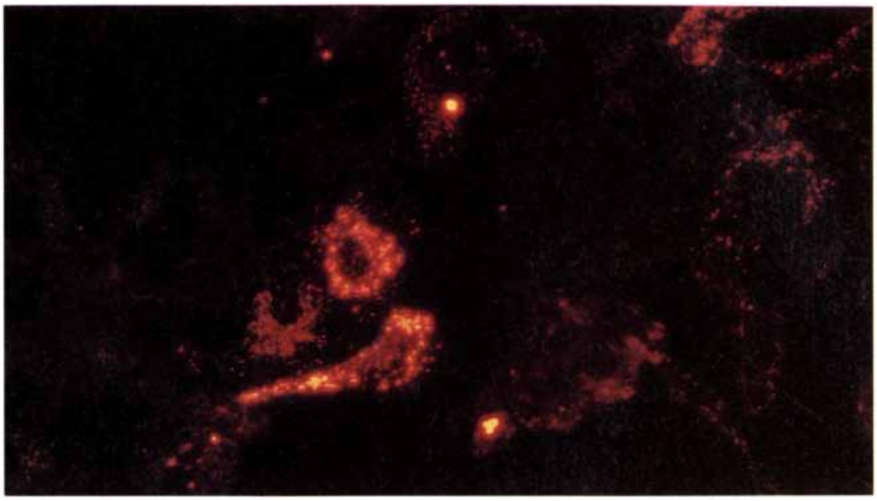
Whenever possible, of course, the engineering effort also tries to improve the effort itself. New testing strategies, methods of analysis, hardware, and data-collection techniques—all help Mercedes engineers dig ever more deeply into the challenge of a safer driving environment. The three-decade evolution of test dummies, for example, has provided about 30 usable dummies at present, many with more than 100 test channels and custom components that require calibration by computer after each test. Moreover, the flood of data from sensors within a test car now requires a pulse-code multiplex system with a 128-channel fiber optic cable.

In certain impacts, front seat belts are tightened by emergency tensioning retractors.

The overall safety effort of Mercedes-Benz extends beyond improving the methodology of automotive safety engineering to the largest goal of all: a safer driving environment. To this end Mercedes-Benz presents some of its findings to the Society of Automotive Engineers and other organizations. In Europe, Mercedes-Benz has been instrumental in the intercountry, interdisciplinary Prometheus intelligent-vehicle program to develop safer driving conditions in the twenty-first century.

In the fifty years since Mercedes-Benz began trying to build safer cars, many technological advances have occurred and many lives have been saved. But as long as the welfare of drivers, passengers and other traffic participants is at stake, no safety effort is far beyond just beginning.

The availability of nongenetic treatments for SCID (including bone marrow transplantation) raises the general question of whether subjecting patients to highly experimental gene therapies is justified when alternatives exist. The prevailing opinion holds that such experimentation is acceptable if the risks are demonstrably low and if, on the one hand, a gene therapy promises to be significantly more helpful than existing approaches or, on the other, patients are ineligible for the established treatments. In the case of SCID, for example, not all patients have access to bone marrow from a tissue-compatible donor.



LIVER CELLS from rabbits genetically deficient in the receptor for low-density lipoprotein (LDL) began to make the missing receptor (*bright regions*) after being altered to carry the receptor gene. The finding raises the possibility that a similar genetic disorder leading to excess serum cholesterol in humans might one day be treatable by gene therapy. James M. Wilson of the Howard Hughes Medical Institute Research Laboratories at the University of Michigan at Ann Arbor and J. Roy-Chowdhury of the Albert Einstein College of Medicine made the photomicrograph.

Genetic alteration of lymphocytes or bone marrow cells aims to correct defects in those same cells or their progeny. Skin cells, in contrast, are being studied for quite a different purpose: the synthesis and secretion of proteins that are normally made in one cell type but are ferried in blood plasma for use by other cells.

In principle, implants of skin cells could correct many disorders. These conditions might include hemophilia (caused by a lack of blood-clotting factors made in the liver) and diseases caused by insufficient production of particular hormones (for example, growth hormone). Certain disorders caused by deficient production of widely made proteins would also be candidates, if the tissues most affected by the deficiency could take up replacement proteins from the blood.

Fibroblasts, a constituent of the dermis (the lower layer of the skin), are best suited for therapy, which would involve implanting the altered cells back into the dermis. They are accessible and strong and able to multiply in the laboratory. Furthermore, they can secrete substances into the blood and would be easy to remove if necessary.

My laboratory has extensively studied the value of skin fibroblasts for treating the form of hemophilia caused by a lack of the liver product known as clotting factor IX. Our results underscore the great therapeutic potential of such cells.

In one of our studies, for instance, A. Dusty Miller, now at the Fred Hutchison Cancer Research Center in Seattle, collaborating with George G. Brownlee and Don S. Anson of the University of Oxford, showed that fibroblasts could be induced to synthesize and secrete factor IX, even though they do not typically make that protein. (Whether the same will be true for all foreign proteins remains to be seen.) Furthermore, when Daniel C. St. Louis, Jonathan H. Axelrod and Raphael Scharfmann in my group used retroviruses to insert the human factor IX gene into fibroblasts and implanted the cells in the dermis of mice, the implants became highly vascularized and released the factor into the blood.

This study not only demonstrated that expression of factor IX in animals was possible, it also taught us an important lesson. About 15 days after the cells were implanted, the human factor disappeared from the blood of the mice. The recipients, it turned out, had mounted an immune response against the foreign human protein. The moral: gene therapy will probably be most successful in patients who make at least a small amount of a deficient protein; otherwise the immune system may become aroused against the product of an inserted gene.

We have also found some evidence to suggest that, unlike the bone marrow cells studied to date, fibroblasts may be able to produce enough of a selected product to correct disease. Extrapolation from data in mice indicates that an implant the size of a quarter should make enough protein to alleviate a factor IX deficiency in a human. In collaboration with Kenneth M. Brinkhous of the University of North Carolina at Chapel Hill, we expect to study the ability of fibroblast implants to correct hemophilia in dogs. If those experiments are successful, trials in humans would be justified.

Genetically altered fibroblasts might also be implanted in the brain to correct disorders in neurons. The brain is notoriously hard to treat because many drugs that circulate in the blood are barred from the brain. Moreover, neurons cannot be removed for direct genetic alteration without consequence to the brain. Fibroblasts could in theory be engineered to secrete proteins for diffusion into nerve cells.

Preliminary results are encouraging. Fred Gage of the University of California at San Diego has shown that implants engineered to secrete nerve growth factor could stimulate neuronal growth in the rat brain. The regeneration occurred in the kinds of neurons whose decay is associated with memory loss in Alzheimer's disease, although the role of the factor in that disease has not been established. Similarly, implants that make levodopa (L-dopa), a precursor of the neurotransmitter dopamine, are under study in animal models of Parkinson's disease. No one knows exactly what causes Parkinson's, but a deficiency of dopamine seems to play a part. Exactly how long fibroblast implants can survive in the skin or brain is still being investigated.

Compared with bone marrow and skin cells, liver cells are a newcomer to the field of gene therapy. They could become important for the treatment of any number of genetic diseases caused by malfunctioning liver cells. Recently, for instance, Mulligan of the Whitehead Institute and James M. Wilson, then also at the institute, and, separately, Theodore Friedmann and his colleagues at San Diego succeeded in delivering the gene for the low-density lipoprotein (LDL) receptor to liver cells and inducing them to make biologically active receptors in the laboratory. The cells came from Watanabe rabbits, which are genetically deficient in the LDL receptor—as are humans afflicted with familial hypercholesterolemia, a condition that can lead to heart attacks.

The feasibility of directly injecting live Watanabe rabbits with complexes of the receptor gene and a protein that homes to the liver has also been studied. (Direct injection in humans would, of course, avoid surgery to remove liver cells.) The encoded protein was detected in the body but, as was also true in the cell-culture study, was made only transiently. Longevity may yet be improved; investigation of liver cells is still in its infancy.

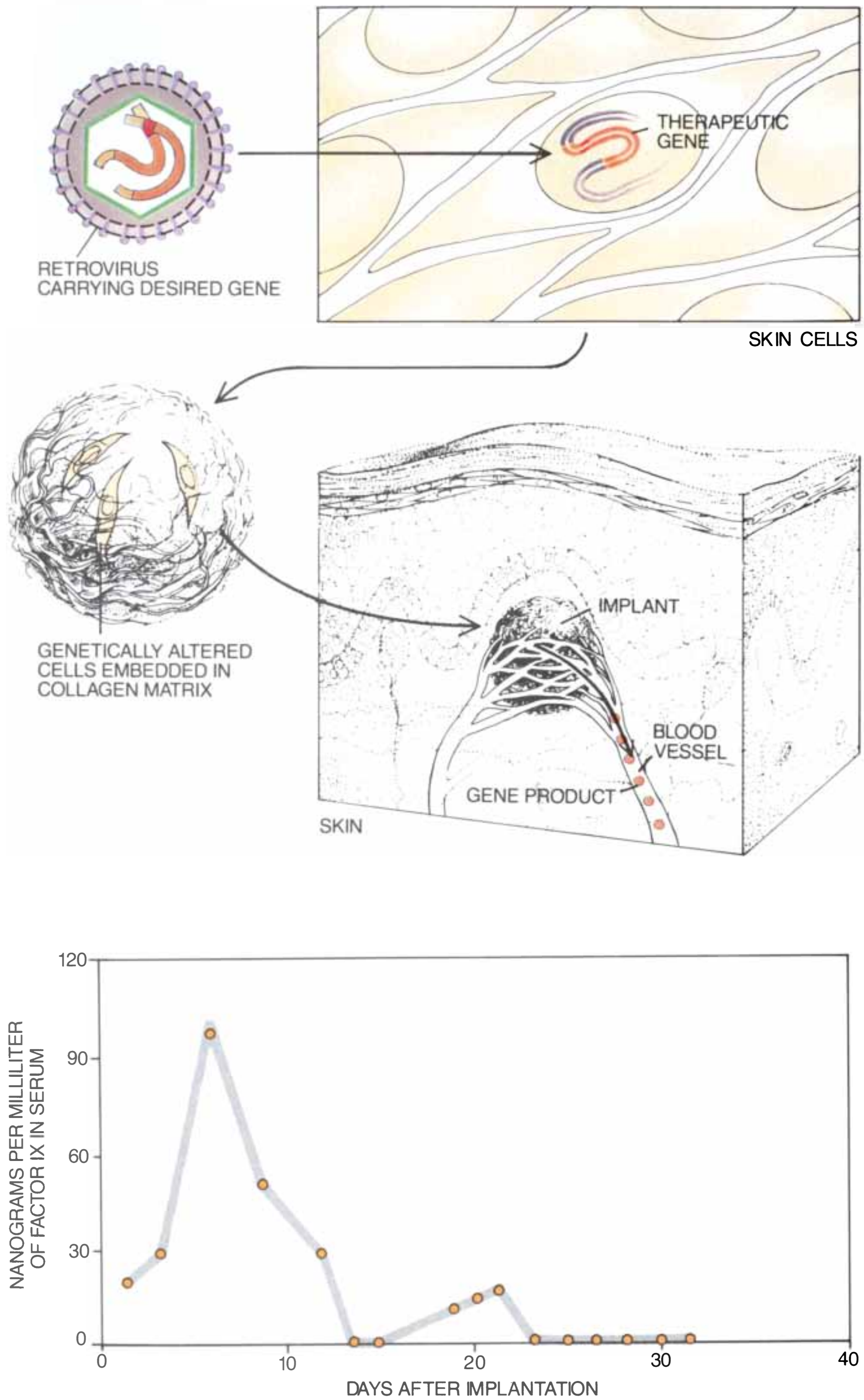
Although bone marrow, skin and liver cells are receiving the most attention, other types are also being considered. For instance, retroviruses can carry genes for secretory products into endothelial cells, which line the arteries. These cells have more intimate contact with the blood than do fibroblasts, and so they might deliver the products more quickly.

Researchers are also considering injecting a healthy gene encoding dystrophin (a structural component of muscle) directly into muscles of mice that have acquired a disorder akin to Duchenne's muscular dystrophy. There is reason to hope the genes will be expressed; other genes injected into muscles in live animals gave rise to proteins for several months, even though the DNA was not integrated into chromosomes. It may also be possible to treat cystic fibrosis, an inherited lung disorder, by packaging healthy genes in retroviruses that would be inhaled in an aerosol spray.

Gene therapy does not have to be limited to repairing the effects of malfunctioning genes. It can also add novel properties to cells to enhance their ability to combat disease.

For instance, Steven A. Rosenberg and his colleagues at the National Cancer Institute have demonstrated that lymphocytes taken from a patient's tumor and cultured with interleukin-2 (a *T* cell activator) can shrink some cancers. They now hope to increase the cancer-fighting powers of those tumor-infiltrating lymphocytes, or TILs, by inserting a gene encoding tumor necrosis factor, a potent immune-system molecule. The factor, which has anticancer activity, is not ordinarily made in *T* cells. Clinical trials are expected to begin soon [see "Adoptive Immunotherapy for Cancer," by Steven A. Rosenberg; SCIENTIFIC AMERICAN, May].

In more preliminary work, another group is trying to induce various cell types to produce CD4, a molecule found on *T* cells depleted by the AIDS virus. The virus enters the cells after a protein in its coat binds with CD4. A flood of CD4 molecules in the blood might serve as a decoy to keep the virus from interacting with the cells. Many other creative ideas for applying gene therapy are also being discussed, including coaxing endothelial cells to secrete factors that would pre-



SKIN CELLS carrying an inserted gene can be embedded in a collagen matrix and implanted in the dermis to deliver the gene's product to the blood (*top*). In one early experiment, skin fibroblasts containing the human gene for factor IX, a protein normally secreted by the liver to aid in blood clotting, became well vascularized in mice and secreted the human factor for approximately two weeks (*graph*). Much longer release of foreign proteins has now been achieved by fibroblast implants.

digital

And you thought field trips were fun.

Live And Learn, Learn, Learn.

Remember when you walked into science class and saw the film projector set up? Better yet, when one of your teachers uttered the magic words, "Field Trip Friday"? They're the kind of learning experiences we all remember.

At Digital, we're proud of our contributions to make learning more fun for people of all ages. And what we're bringing you promises just as much excitement as the day you walked out of science class and saw that yellow bus waiting to take you on a journey of discovery.

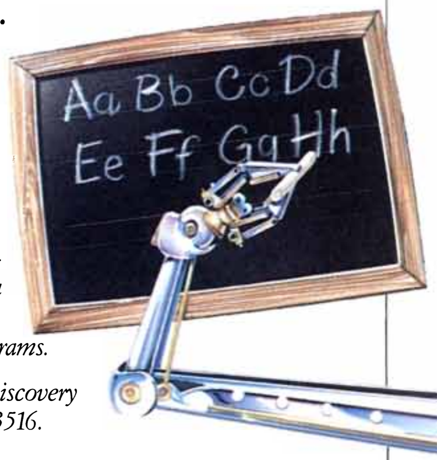
It All Starts At Home.

When it goes up for a spin, it really goes out for a spin. This year, "Voyager" is coming at you via The Infinite Voyage. *A series that'll take your mind as far as you're willing to go. Now in its fourth season. And appearing anywhere there's a television. Watch for "Sail On, Voyager!" coming at you November 28 on PBS.*

Go To The Head Of The Class.

It's pretty basic, actually. To us, education, at all levels, always takes a front seat. Especially when it comes to science for kids. So pull up a chair to our Discovery Lecture Series, one of today's farthest reaching educational programs. (Thanks to free worldwide satellite broadcasts.) Hitch a ride on an ever-mobile Computer Bus. Sign up for a fun-drenched computer camp. Or partake in any number of other programs.

For more information on the Discovery Lecture Series, call (412) 456-3516.



Show And Tell.

You open the door to the science museum. The wheels begin to turn. And as you move from exhibit to exhibit, you can barely contain your curiosity. We sponsor these captivating shows in the name of scientific literacy. They're for anyone with an active imagination. They're about everything from walk-through computers to simulated wind tunnels. And they're all at your fingertips, nationwide.

All For One. And One For All.

Frankly, supporting education is part of who we are. And that means all of us at Digital Equipment Corporation. Not only are we awarding scholarships and contributing equipment. Partnering with schools across the country. And supporting faculty and curriculum development. Our employees are volunteering their time and technical expertise as well. Efforts which are helping students come to an important realization: that learning truly is one of life's greatest rewards.

DISORDER	INCIDENCE	NORMAL PRODUCT OF DEFECTIVE GENE	TARGET CELLS	STATUS
Hemoglobin-opathies (thalassemias)	1 in 600 in certain ethnic groups	Constituents of hemoglobin	Bone marrow cells (which give rise to circulating blood)	Globin production in animals receiving gene needs to be improved
Severe combined immunodeficiency (SCID)	Rare	Adenosine deaminase (ADA) in about a quarter of SCID patients	Bone marrow cells or *T* lymphocytes	Clinical trial of lymphocyte therapy for ADA deficiency is under way
Hemophilia A Hemophilia B	1 in 10,000 males 1 in 30,000 males	Blood-clotting factor VIII Blood-clotting factor IX	Liver cells or fibroblasts	Good chance for clinical trials (with fibroblasts) in next five years
Familial hypercholester-olemia	1 in 500	Liver receptor for low-density lipoprotein (LDL)	Liver cells	Animal studies are in early stages
Inherited emphysema	1 in 3,500	Alpha$_1$-antitrypsin (liver product that protects lungs from enzymatic degradation)	Lung or liver cells	Work is very preliminary
Cystic fibrosis	1 in 2,500 Caucasians	Substance important for keeping air tubes in lungs free of mucus	Lung cells	Aerosol delivery of gene directly to lungs is a theoretical possibility
Duchenne's muscular dystrophy	1 in 10,000 males	Dystrophin (structural component of muscle)	Muscle cells (particularly embryonic ones that develop into muscle fibers)	Work is preliminary. Nondystrophin genes injected into muscle have directed synthesis of the encoded proteins
Lysosomal storage diseases	1 in 1,500 acquires some form	Enzymes that degrade complex molecules in intracellular compartments known as lysosomes	Vary, depending on disorder	Most diseases would require delivery of gene into brain cells (a difficult task) as well as into other cell types

POTENTIAL CANDIDATES for the earliest gene therapies will be disorders caused by defects in a single gene that has been cloned. In general, physicians will remove cells from a patient, insert a healthy gene and return the cells to the body.

vent blood clots from forming in a patient's arteries after heart surgery.

The idea of introducing genes to correct heritable and other disorders is nothing less than revolutionary. Perhaps that is one reason why the field has progressed somewhat more slowly than was once expected. Modern creatures are the products of millions of years of evolution. One cannot expect that the initial stabs at inserting genes into cells will yield normal, stable expression easily.

Yet to cure diseases, investigators must find ways to ensure that therapeutic genes are expressed well and persistently in the body. Continually emerging clues, such as the importance of including particular enhancers with some genes in retroviral vectors, are beginning to point the way. Also needed are better methods for returning genetically altered cells (such as liver cells) to the body, ways of extending the survival of implanted cells, and techniques for isolating human stem cells (to replace the bone marrow cells now being studied).

At the same time, the safety of retroviral vectors must be confirmed in extensive studies of both small and large animals, and efforts to incorporate added safeguards should continue. In spite of the advent of retroviral vectors that cannot replicate, there is still a chance they could cause cancer. Efforts to develop alternatives to retroviral vectors should be pursued further as well, as should research into site-specific gene delivery.

The goal of curing genetic diseases for life with a single, safe treatment is unquestionably worth the effort being put into it, but I must end with the reminder that gene therapy cannot correct all human disease. Most human afflictions are not genetic. They are environmental, caused by microbial infections that spread because of poor sanitation, polluted drinking water, malnutrition and other factors that are outside the scope of genetic engineering. Those diseases, too, deserve increased study.

FURTHER READING

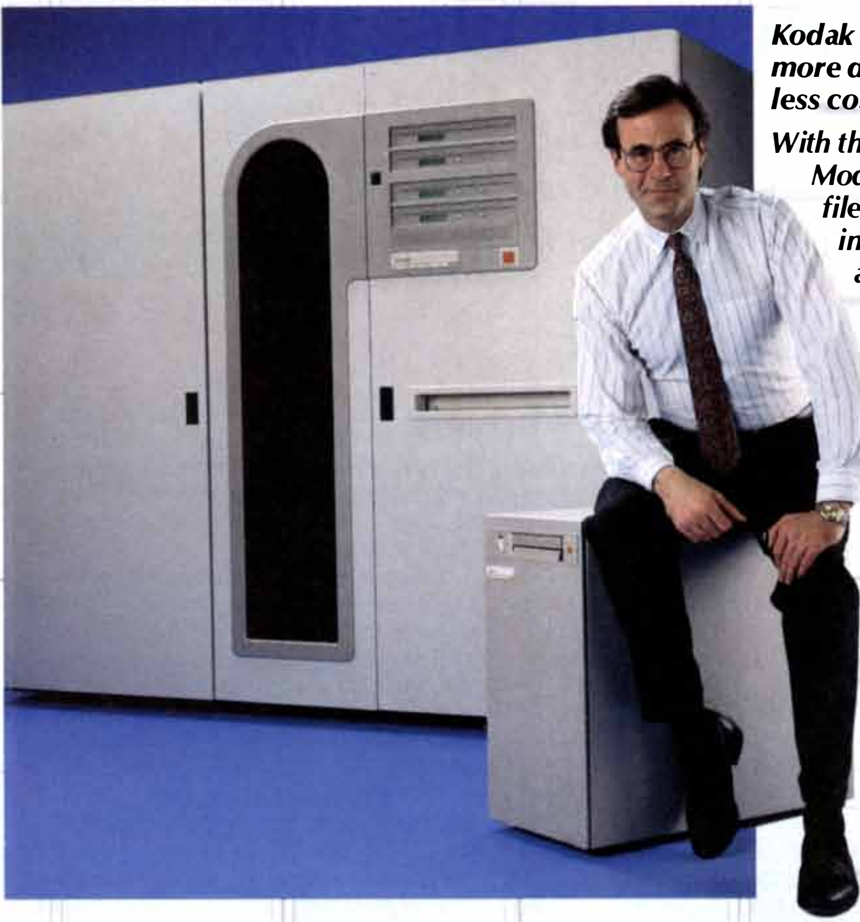
LINEAGE-SPECIFIC EXPRESSION OF A HUMAN β-GLOBIN GENE IN MURINE BONE MARROW TRANSPLANT RECIPIENTS RECONSTITUTED WITH RETROVIRUS-TRANSDUCED STEM CELLS. Elaine A. Dzierzak, Thalia Papayannopoulou and Richard C. Mulligan in *Nature*, Vol. 331, No. 6151, pages 35–41; January 7, 1988.

DISRUPTION OF THE PROTO-ONCOGENE *INT-2* IN MOUSE EMBRYO-DERIVED STEM CELLS: A GENERAL STRATEGY FOR TARGETING MUTATIONS TO NON-SELECTABLE GENES. Suzanne L. Mansour, Kirk R. Thomas and Mario R. Capecchi in *Nature*, Vol. 336, No. 6197, pages 348–352; November 24, 1988.

AN ALTERNATIVE APPROACH TO SOMATIC CELL GENE THERAPY. D. St. Louis and I. M. Verma in *Proceedings of the National Academy of Sciences*, Vol. 85, pages 3150–3153; 1988.

HUMAN GENE THERAPY. Eve K. Nichols and the Institute of Medicine, National Academy of Sciences, Staff. Harvard University Press, 1988.

PROGRESS TOWARD HUMAN GENE THERAPY. Theodore Friedmann in *Science*, Vol. 244, No. 4910, pages 1275–1281; June 16, 1989.



The Meaning of Dreams

Dreams may reflect a fundamental aspect of mammalian memory processing. Crucial information acquired during the waking state may be reprocessed during sleep

by Jonathan Winson

Throughout history human beings have sought to understand the meaning of dreams. The ancient Egyptians believed dreams possessed oracular power—in the Bible, for example, Joseph's elucidation of Pharaoh's dream averted seven years of famine. Other cultures have interpreted dreams as inspirational, curative or alternative reality.

During the past century, scientists have offered conflicting psychological and neuroscientific explanations for dreams. In 1900, with the publication of *The Interpretation of Dreams,* Sigmund Freud proposed that dreams were the "royal road" to the unconscious; that they revealed in disguised form the deepest elements of an individual's inner life. More recently, in contrast, dreams have been characterized as meaningless, the result of random nerve cell activity. Dreaming has also been viewed as the means by which the brain rids itself of unnecessary information—a process of "reverse learning," or unlearning.

Based on recent findings in my own and other neuroscientific laboratories, I propose that dreams are indeed meaningful. Studies of the hippocampus (a brain structure crucial to memory), of rapid eye movement (REM) sleep and of a brain wave called theta rhythm suggest that dreaming reflects a pivotal aspect of the processing of memory. In particular, studies of theta rhythm in subprimate animals have provided an evolutionary clue to the meaning of dreams. They appear to be the nightly record of a basic mammalian memory process: the means by which animals form strategies for survival and evaluate current experience in light of those strategies. The existence of this process may explain the meaning of dreams in human beings.

The physiology of dreaming was first understood in 1953, when researchers characterized the human sleep cycle. They found that sleep in humans is initiated by the hypnogogic state, a period of several minutes when thoughts consist of fragmented images or minidramas. The hypnogogic state is followed by slow-wave sleep, so called because at that time the brain waves of the neocortex (the convoluted outer mantle of the brain) are low in frequency and large in amplitude. These signals are measured as electroencephalographic (EEG) recordings.

Researchers also discovered that a night's sleep is punctuated by periods in which the EEG readings are irregular in frequency and low in amplitude—similar to those observed in awake individuals. These periods of mental activity are called REM sleep. Dreaming takes place solely during these periods. While in REM sleep, motor neurons are inhibited, preventing the body from moving freely but allowing extremities to remain slightly active. Eyes move rapidly in unison under closed lids, breathing becomes irregular and heart rate increases.

The first REM stage of the night follows 90 minutes of slow-wave sleep and lasts for 10 minutes. The second and third REM periods follow shorter slow-wave sleep episodes but grow progressively longer themselves. The fourth and final REM interval lasts 20 to 30 minutes and is followed by awakening. If a dream is remembered at all, it is most often the one that occurred in this last phase.

This sleep cycle—alternating slow-wave and REM sleep—appears to be present in all placental and marsupial mammals. Mammals exhibit the various REM-associated characteristics observed in humans, including EEG readings similar to those of the awake state. Animals also dream. By destroying neurons in the brain stem that inhibit movement during sleep, researchers found that sleeping cats rose up and attacked or were startled by invisible objects—ostensibly images from dreams.

By studying subprimate animals, scientists have discovered additional neurophysiological aspects of REM sleep. They determined that neural control of this stage of the sleep cycle is centered in the brain stem (the brain region closest to the spinal cord) and that during REM sleep neural signals—called pontine-geniculate-occipital cortex (PGO) spikes—proceed from the brain stem to the center of visual processing, the visual cortex. Brain stem neurons also initiate a sinusoidal wave (one resembling a sine curve) in the hippocampus. This brain signal is called theta rhythm.

At least one animal experiences slow-wave but not REM sleep—and, consequently, does not exhibit theta rhythm when asleep. This animal is the echidna, or spiny anteater, an egg-laying mammal (called a monotreme) that provides some insight into the origin of dreaming. The absence of REM sleep in the echidna suggests that this stage of the sleep cycle evolved some 140 million years ago, when marsupials and

JONATHAN WINSON started his career as an aeronautical engineer, graduating with an engineering degree from the California Institute of Technology in 1946. He completed his Ph.D. in mathematics at Columbia University and then turned to business for 15 years. Because of his keen interest in neuroscience, however, Winson started to do research at the Rockefeller University on memory processing during waking and sleeping states. In 1979 he became an associate professor there and continued his work on memory and dreaming. His research has been supported by the National Institute of Mental Health, the National Science Foundation and the Harry F. Guggenheim Foundation.

JACOB'S LADDER, painted in 1973 by Marc Chagall, depicts a biblical story. Jacob dreams of angels ascending to and descending from heaven on a ladder.

placentals diverged from the monotreme line. (Monotremes were the first mammals to develop from reptiles.)

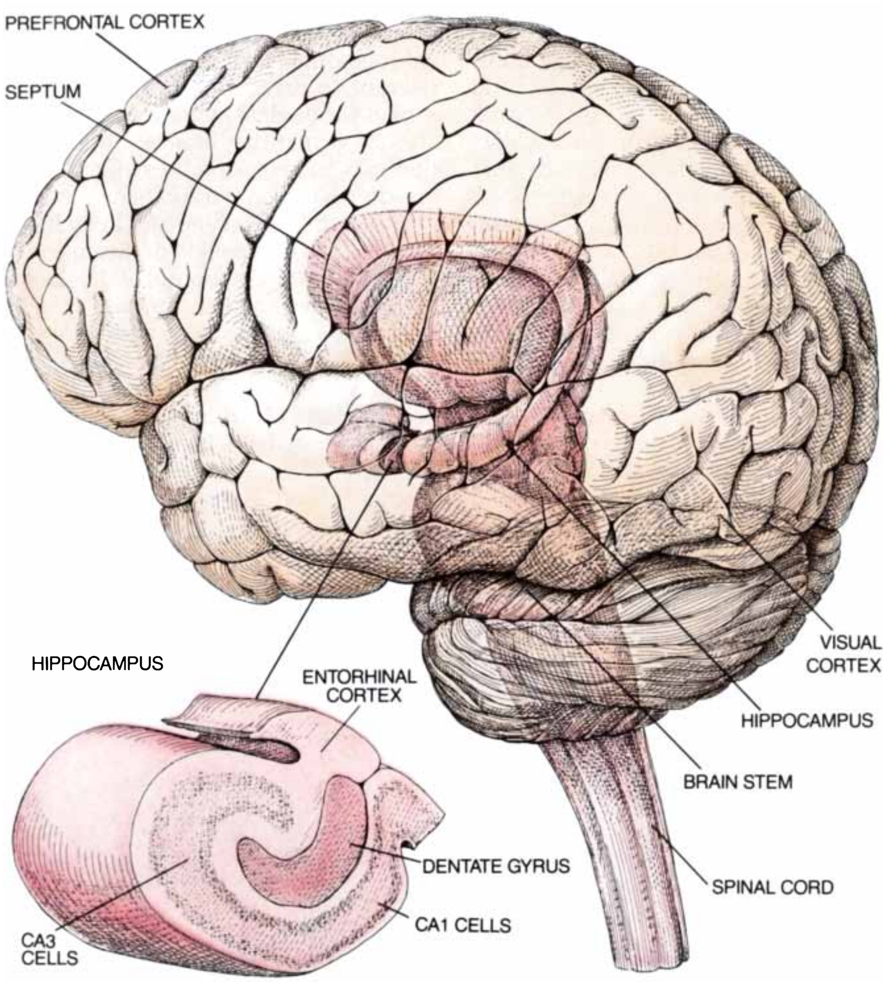
By all evolutionary criteria, the perpetuation of a complex brain process such as REM sleep indicates that it serves an important function for the survival of mammalian species. Understanding that function might reveal the meaning of dreams.

When Freud wrote *The Interpretation of Dreams,* the physiology of sleep was unknown. In light of the discovery of REM sleep, certain elements of his psychoanalytic theory were modified, and the stage was set for more neurologically based theories. Dreaming came to be understood as part of a biologically determined sleep cycle. Yet the central concept of Freud's theory continues to be recognized—namely, the belief that dreams reveal our innermost unconscious feelings and concerns—and continues to be used in psychoanalysis.

Some theorists abandoned Freud altogether following the neurological discoveries. In 1977 J. Allan Hobson and Robert McCarley of Harvard Medical School proposed the "activation-synthesis" hypothesis. They suggested that dreaming consists of associations and memories elicited from the forebrain (the neocortex and associated structures) in response to random signals from the brain stem such as PGO spikes. Dreams were merely the "best fit" the forebrain could provide to this random bombardment from the brain stem. Although dreams might at times have psychological content, they were inherently meaningless.

Hobson recently revised his theory, acknowledging the deep psychological significance of dreams. The sense, or plot, of dreams resulted from order that was imposed on the chaos of neural signals, he said. "That order is a function of our own personal view of the world, our remote memories," Hobson wrote. In other words, the individual's emotional vocabulary could be relevant to dreams. In a further revision of the original hypothesis, Hobson also suggested that brain stem activation may merely serve to switch from one dream episode to another.

Although Hobson and McCarley had presented an explanation of dream content, the basic function of REM sleep admittedly remained unknown. In 1983 Francis Crick of the Salk Institute in La Jolla, Calif., and Graeme Mitchison of the University of Cambridge, England, proposed the idea of reverse learning. Working from the Hobson-McCarley assumption of random neocortical bombardment by PGO waves and their own knowledge of the behavior of stimulated neural networks, they postulated that a complex associational neural network such as the neocortex might become overloaded by vast amounts of incoming information. The neocortex



ANATOMY of the brain and cross section of the hippocampus show some of the regions involved in dreaming. In the hippocampus, incoming information is processed sequentially in the dentate gyrus, the CA3 and the CA1 pyramidal cells. In subprimate species, theta rhythm is generated in the dentate gyrus and CA1 cells.

could then develop false, or "parasitic," thoughts that would jeopardize the true and orderly storage of memory.

According to Crick and Mitchison's hypothesis, REM sleep served to erase these spurious associations on a regular basis. Random PGO waves impinged on the neocortex, resulting in erasure, or unlearning, of the false information. This process served an essential function: it allowed the orderly processing of memory. In humans, dreams were a running record of these parasitic thoughts—material to be purged from memory. "We dream to forget," Crick and Mitchison wrote.

The two researchers proposed a revision in 1986. Erasure of parasitic thoughts accounted only for bizarre dream content. Nothing could be said about dream narrative. Furthermore, dreaming to forget was better expressed as dreaming to reduce fantasy or obsession.

None of these hypotheses seems to explain adequately the function of dreaming. On the one hand, Freud's theory lacked physiological evidence. (Although Freud had originally intended to describe the neurology of the unconscious and of dreams in his proposed "Project for a Scientific Psychology," the undertaking was premature, and he limited himself to psychoanalysis.) On the other hand, despite revisions to incorporate elements of psychology, later theories denied that dreams had meaning.

Exploring the neuroscientific aspects of REM sleep and of memory processing seemed to me to hold the greatest potential for understanding the meaning and function of dreams. The key to this research was theta rhythm.

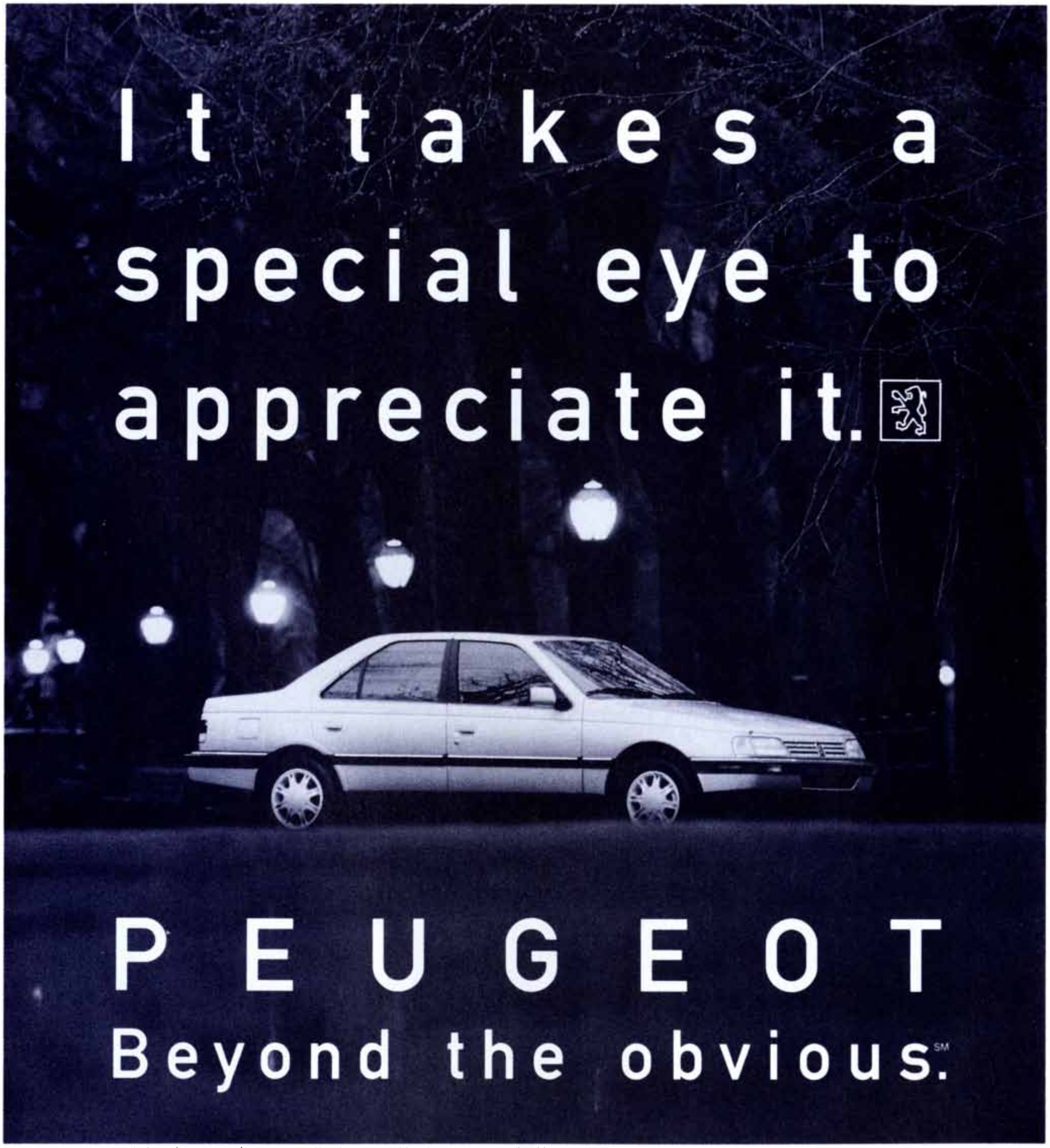
Theta rhythm was discovered in 1954 in awake animals by John D. Green and Arnaldo A. Arduini of the University of California at Los Angeles. The researchers observed a regular sinusoidal signal of six cycles per second in the hippocampus of rabbits when the animals were apprehensive of stimuli in their environment. They named the signal theta rhythm after a previously discovered EEG component of the same frequency.

Theta rhythm was subsequently recorded in the tree shrew, mole, rat and cat. Although it was consistently observed in awake animals, theta rhythm was correlated with very different behaviors in each species. For example, in marked contrast to the rabbit, environmental stimuli did not induce theta rhythm in the rat. Rats demonstrated theta rhythm only during movement, typically when they explored. In 1969, however, Case H. Vanderwolf of the University of Western Ontario discovered there was one behavior during which the animals he studied, including the rat, showed theta rhythm: REM sleep.

In 1972 I published a commentary pointing out that the different occurrences of theta rhythm could be understood in terms of animal behavior. Awake animals seemed to show theta rhythm when they were behaving in ways most crucial to their survival. In other words, theta rhythm appeared when they exhibited behavior that was not genetically encoded—such as feeding or sexual behavior—but rather a response to changing environmental information. Predatory behavior in the cat, prey behavior in the rabbit, and exploration in the rat are, respectively, most important to their survival. (For example, a hungry rat will explore before it eats even if food is placed in front of it.)

Furthermore, because the hippocampus is involved in memory processing, the presence of theta rhythm during REM sleep in that region of the brain might be related to that activity. I suggested that theta rhythm reflected a neural process whereby information essential to the survival of a species—gathered during the day—was reprocessed into memory during REM sleep.

In 1974, by recording signals from the hippocampus of freely moving rats and rabbits, I found the source from which theta rhythm was generated in the hippocampus. Together with the neocortex, the hippocampus is believed to provide the neural basis for memory storage. The hippocampus (the Greek word for "seahorse," which it resembles in shape) is a sequential structure composed of three types of neurons. Information from all sensory and



1991 405 models from $15,300 to $21,700.* Call 1-800-447-2882. In Canada 1-416-566-1900. *MSRP. Excludes tax, title, options, registration and destination charges.

How you see the Peugeot 405 S depends on how you look at it.

A glance will probably note the leather seats, but miss the two densities of foam used to eliminate seat springs which transmit road vibration. It might take in the six-speaker anti-theft AM/FM cassette stereo system, but miss the hydraulic-powered anti-lock braking system.

A more discerning eye will see Peugeot's passion for building cars which do not sacrifice any one element for another, but aim to provide the optimum balance of all aspects of the car.

Something which, if you haven't yet experienced, you are certain to appreciate.

associational areas of the neocortex converges in a region called the entorhinal cortex; from there it is transmitted to the three successive neuronal populations of the hippocampus. The signal arrives first at the granule cells of the dentate gyrus, then at the CA3 pyramidal cells (so called because of their triangular shape) and finally at the pyramidal cells of CA1. After information is processed by this trio of cells, it is retransmitted to the entorhinal cortex and then back to the neocortex.

My studies showed that theta rhythm was produced in two regions within the hippocampus: the dentate gyrus and the CA1 neurons. The rhythms in these two areas were synchronous. Subsequently, Susan Mitchell and James B. Ranck, Jr., of the State University of New York Downstate Medical Center, identified a third synchronous generator in the entorhinal cortex, and Robert Verdes of Wayne State University discovered the brain stem neurons that control theta rhythm. These neurons transmit signals to the septum (a forebrain structure) that activate theta rhythm in the hippocampus and the entorhinal cortex. Thus, the brain stem activates the hippocampus and the neocortex—the core memory system of the brain.

To determine the relationship between theta rhythm and memory, I made a lesion in the rat septum. Rats that had previously learned, using spatial cues, to locate a particular position in a maze were no longer able to do so. Without theta rhythm, spatial memory was destroyed.

Studies of the cellular changes that bring about memory illustrated the role of theta rhythm. In particular, the discovery of long-term potentiation (LTP) in 1973 showed the means by which memory might be encoded. Timothy V. P. Bliss and A. R. Gardner-Medwin of the National Institute of Medical Research in London and Terje Lømo of the University of Oslo found changes in nerve cells that had been intensely stimulated with electric pulses.

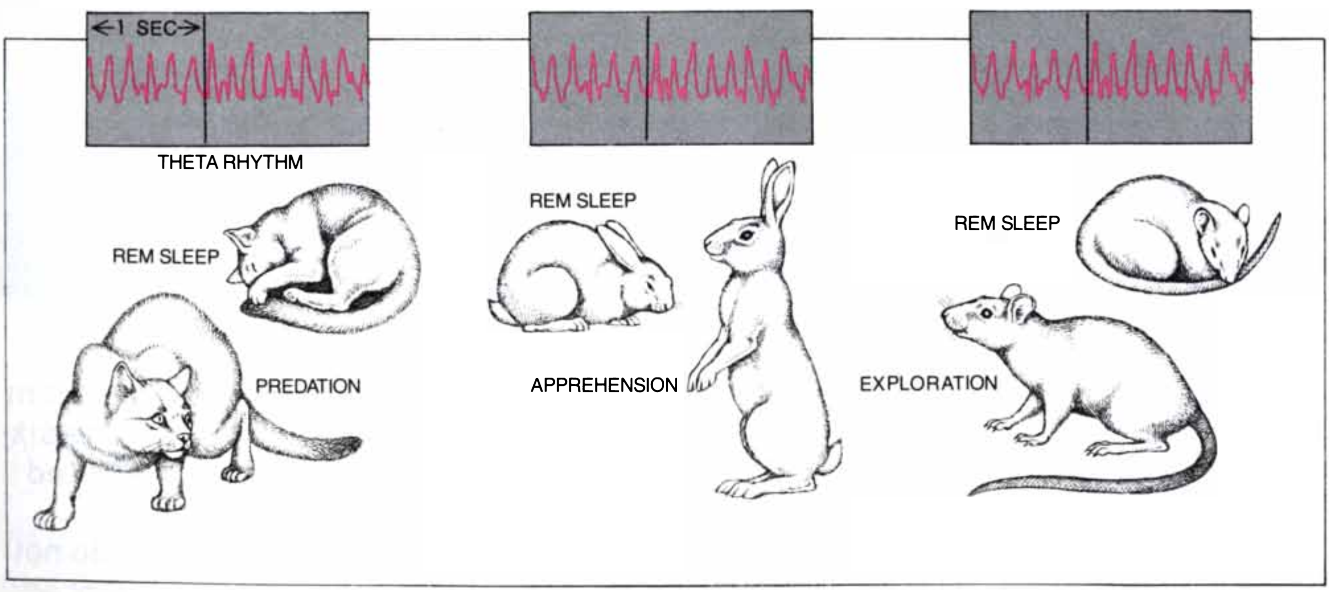
Earlier studies had shown that if one stimulated the pathway from the entorhinal cortex to the granule cells of the hippocampus, the response of these cells could be measured with a recording electrode. Using this technique, Bliss and his colleagues measured the normal response to a single electric pulse. Then they applied a long series of high-frequency signals—called tetanic pulses—to this pathway. After the train of tetanic stimuli, a single electric pulse caused much greater firing in the granule cells than had been observed prior to the experiment. The heightened effect persisted for as long as three days. This phenomenon, termed LTP, was precisely the kind of increase in neuronal strength that could be capable of sustaining memory. LTP is now considered a model for learning and memory.

LTP is achieved by the activity of the NMDA (*N*-methyl-D-aspartate) receptor. This molecule is embedded in the dendrites of the granule cells and the CA1 cells of the hippocampus as well as in neurons throughout the neocortex. Like other neuronal receptors, the NMDA receptor is activated by a neurotransmitter—glutamate in this case. Glutamate momentarily opens a non-NMDA channel in the granule cell dendrite, allowing sodium from the extracellular space to flow into the neuron. This influx causes the granule cell to become depolarized. If the depolarization is sufficient, the granule cell fires, transmitting information to other nerve cells.

Unlike other neuronal receptors, NMDA possesses an additional property. If a further activation of glutamate occurs while the granule cell is depolarized, a second channel opens up, allowing an influx of calcium. Calcium is thought to act as a second messenger, initiating a cascade of intracellular events that culminates in long-lasting synaptic changes—or LTP. (The description given here has been necessarily simplified. LTP is the subject of extensive ongoing investigation.)

Because the tetanic impulse applied by Bliss and his colleagues did not occur naturally in the brain, the question remained as to how LTP was achieved under normal circumstances. In 1986 John Larson and Gary S. Lynch of the University of California at Irvine and Gregory Rose and Thomas V. Dunwiddie of the University of Colorado at Denver suggested that the occurrence of LTP in the hippocampus was linked to theta rhythm. They applied a small number of electric pulses to CA1 cells in the rat hippocampus and produced LTP, but only when the pulses were separated by the normal time that elapses between two theta waves—approximately 200 milliseconds. Theta rhythm is apparently the natural means by which the NMDA receptor is activated in neurons in the hippocampus.

Work in my laboratory at the Rockefeller University duplicated Larson and



THETA RHYTHM is present during different waking behaviors in different species. Each of these behaviors is pivotal to the animal's survival. In placental and marsupial animals, theta rhythm is present during rapid eye movement (REM) sleep.

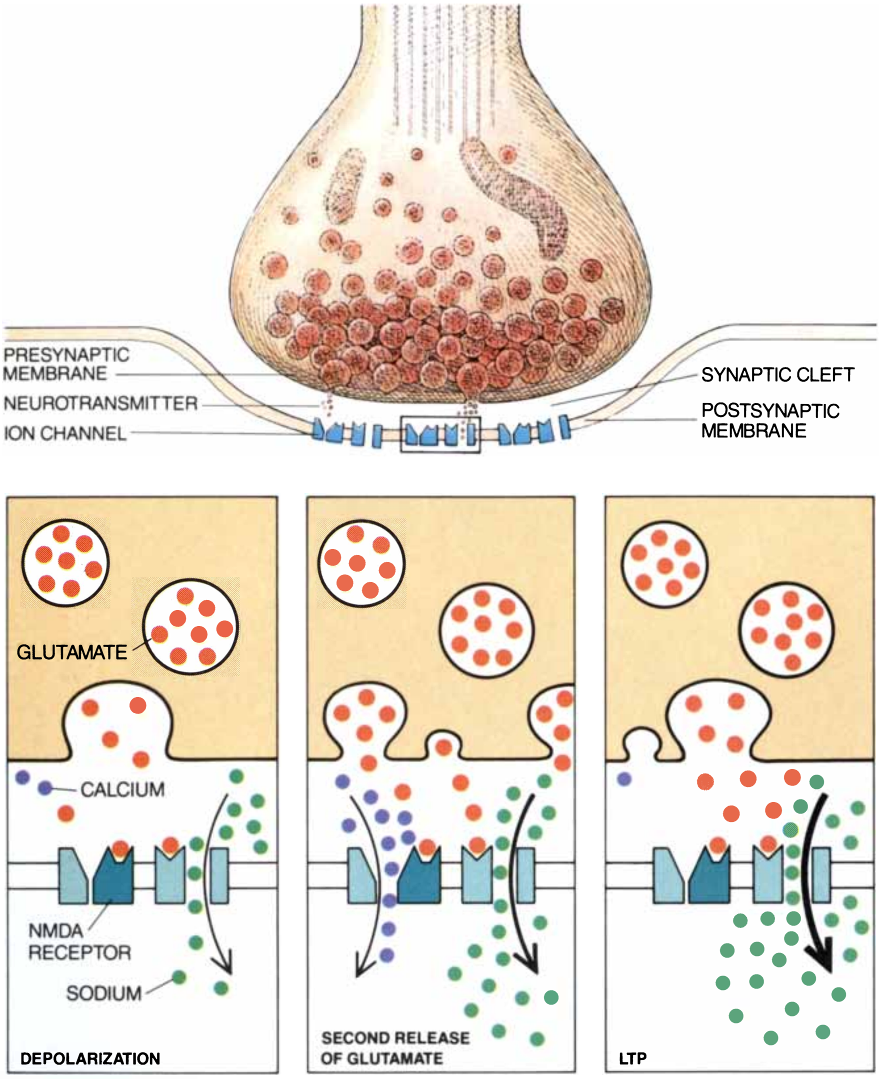
strated that individual CA1 neurons in the rat hippocampus fired when the awake animal moved to a particular location—namely, the neuron's place field. The implication of this finding was that the CA1 neuron fired to map the environment, thereby committing it to memory.

In 1989 Pavlides and I located two CA1 neurons in the rat hippocampus that had different place fields. We recorded from both cells simultaneously. After determining the normal firing rates in awake and asleep animals, we positioned a rat in the place field of one of the neurons. The neuron fired vigorously, mapping that location. The second cell fired only sporadically because it was not coding space. We continued recording from the two pairs of neurons as the rat moved about and then entered several sleep cycles. Six pairs of neurons were studied in this manner.

We found that neurons that had coded space fired at a normal rate as the animal moved about prior to sleep. In sleep, however, they fired at a significantly higher rate than their previous sleeping baseline. There was no such increase in firing rate during sleep in neurons that had not mapped space. This experiment suggested that the reprocessing or strengthening of information encoded when the animal was awake occurred in sleep at the level of individual neurons.

Evidence that theta rhythm encodes memories during REM sleep may be derived not only from neuroscientific studies but also from evolution. The emergence of a neural mechanism to process memory in REM sleep suggests differences in brain anatomy between mammals that have that aspect of the sleep cycle and those that do not. And in fact, such differences clearly exist between the echidna and the marsupials and placentals.

The echidna has a large convoluted prefrontal cortex, larger in relation to the rest of the brain than that of any other mammal, even humans. I believe it needed this huge prefrontal cortex to perform a dual function: to react to incoming information in an appropriate manner based on past experience and to evaluate and store new information to aid in future survival. Without theta rhythm during REM sleep, the echidna would not be able to process information while it slept. (The echidna does, however, show theta rhythm when foraging for food.) For higher capabilities to develop, the prefrontal cortex would have to become increasingly large—beyond the capacity of the skull—unless another brain mechanism evolved.



NMDA RECEPTOR activation induces long-term potentiation (LTP), a model for memory. The release of the neurotransmitter glutamate (*left panel*) opens a non-NMDA receptor channel, allowing the influx of sodium, which depolarizes the neuron. If a further release of glutamate occurs while the cell is depolarized (*center panel*), the NMDA receptor opens a second channel, which allows calcium to flow in, leading to LTP. LTP occurs as a result of increased sodium through the non-NMDA channel (*right panel*) and the subsequent greater depolarization of the cell.

Lynch's CA1 findings, but this time in the hippocampal granule cells. Constantine Pavlides, Yoram J. Greenstein and I then demonstrated that LTP was dependent on the presence and phase of theta rhythm. If electric pulses were applied to the cells at the peak of the theta wave, LTP was induced. But if the same pulse was applied at the trough of the waves—or when theta rhythm was absent—LTP was not induced.

A coherent picture of memory processing was emerging. As a rat explores, for example, brain stem neurons activate theta rhythm. Olfactory input (which in the rat is synchronized with theta rhythm, as is the twitching of whiskers) and other sensory information converge on the entorhinal cortex and the hippocampus. There they are partitioned into 200-millisecond "bites" by theta rhythm. The NMDA receptors, acting in conjunction with theta rhythm, allow for long-term storage of this information.

A similar process occurs during REM sleep. Although there is no incoming information or movement during REM sleep, the neocortical-hippocampal network is once again paced by theta rhythm. Theta rhythm might produce long-lasting changes in memory.

The results of one of my further experiments served to show that spatial memory was indeed being stored in the rat hippocampus during sleep. John O'Keefe and J. Dostrovsky of the University College in London had demon-


386™
intel®

REM sleep could have provided this new mechanism, allowing memory processing to occur "off-line." Coincident with the apparent development of REM sleep in marsupial and placental mammals was a remarkable neuroanatomical change: the prefrontal cortex was dramatically reduced in size. Far less prefrontal cortex was required to process information. That area of the brain could then develop to provide advanced perceptual and cognitive abilities in higher species.

The nature of REM sleep supports this evolutionary argument. During the day, animals gather information that involves locomotion and eye movement. The reprocessing of this information during REM sleep would not be easily separated from the locomotion related to the experience—such disassociation might be expecting too great a revision of brain circuitry. So to maintain sleep, locomotion had to be suppressed by inhibiting motor neurons. Suppressing eye movement was unnecessary because this activity does not disturb sleep.

Eye movement potentials, similar to PGO spikes, accompany rapid eye movement in the waking state and also during REM sleep. The function of these signals has not yet been established, but they may serve to alert the visual cortex to incoming information when the animal is awake and may reflect the reprocessing of this information during REM sleep. In any case, PGO spikes do not disturb sleep and do not have to be suppressed—unlike motor neurons.

With the evolution of REM sleep, each species could process the information most important for its survival, such as the location of food or the means of predation or escape—those activities during which theta rhythm is present. In REM sleep this information may be accessed again and integrated with past experience to provide an ongoing strategy for behavior.

Although theta rhythm has not yet been demonstrated in primates, including humans, the brain signal provides a clue to the origin of dreaming in humans. Dreams may reflect a memory-processing mechanism inherited from lower species, in which information important for survival is reprocessed during REM sleep. This information may constitute the core of the unconscious.

Because animals do not possess language, the information they process during REM sleep is necessarily sensory. Consistent with our early mammalian origins, dreams in humans are sensory, primarily visual. Dreams do not take the form of verbal narration.

Also in keeping with the role REM sleep played in processing memories in animals, there is no functional necessity for this material to become conscious. Consciousness arose later in evolution in humans. But neither is there any reason for the material of dreams not to reach consciousness. Therefore, dreams can be remembered—most readily if awakening occurs during or shortly after a REM sleep period.

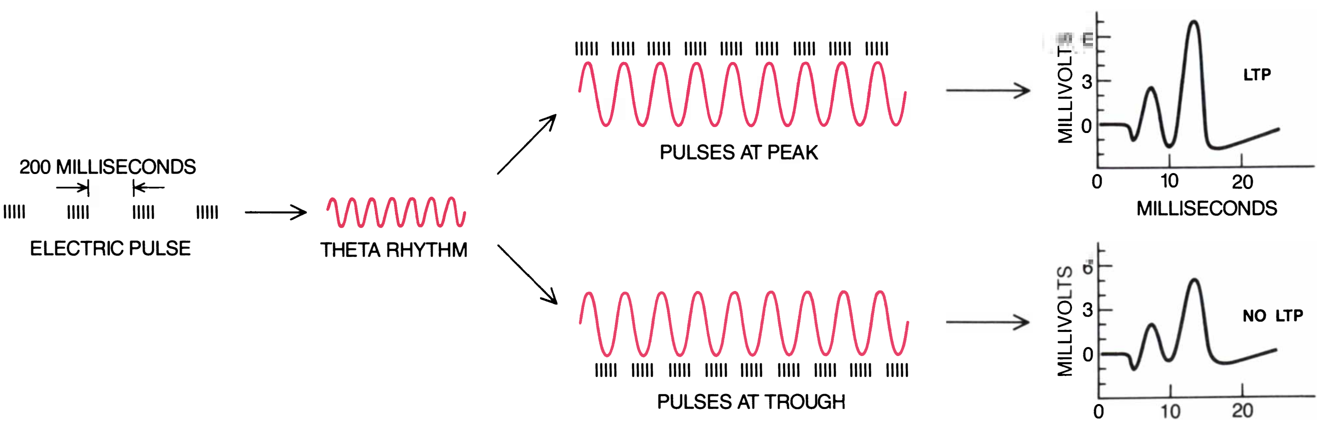
Consistent with evolution and evidence derived from neuroscience and reports of dreams, I suggest that dreams reflect an individual's strategy for survival. The subjects of dreams are broad-ranging and complex, incorporating self-image, fears, insecurities, strengths, grandiose ideas, sexual orientation, desire, jealousy and love.

Dreams clearly have a deep psychological core. This observation has been reported by psychoanalysts since Freud and is strikingly illustrated by the work of Rosalind Cartwright of Rush-Presbyterian–St. Luke's Hospital in Chicago. Cartwright is studying a series of 90 subjects who are undergoing marital separation and divorce. All the subjects are clinically evaluated and psychologically tested to ascertain their attitudes and responses to their personal crisis. Cartwright's subjects are also awakened from REM sleep to report their dreams, which are then interpreted by the subjects themselves without questions that might influence their interpretation. In the 70 individuals studied to date, the dream content conveys the person's unconscious thoughts and is strongly correlated with the manner in which he or she is coping with the crisis in the waking world.

Although the topic "chosen" for consideration during a night's sleep is unpredictable, certain of life's difficulties—as in the case of Cartwright's subjects—so engage psychological survival that they are selected for REM sleep processing. In the ordinary course of events, depending on the individual's personality, the themes of dreams may be freewheeling. Moreover, when joined with the intricate associations that are an intrinsic part of REM sleep processing, the dream's statement may be obscure.

Nevertheless, there is every reason to believe that the cognitive process taking place in Cartwright's subjects occurs in every individual. Interpretation of the coherent statement that is being made depends on the individual's tracing of relevant or similar events. These associations are strongly biased toward early childhood experience.

My hypothesis also offers an explanation for the large amount of REM sleep in infants and children. Newborns spend eight hours a day in REM sleep. The sleep cycle is disorganized at this age. Sleep occurs in 50- to 60-minute bouts and begins with REM rather than with slow-wave

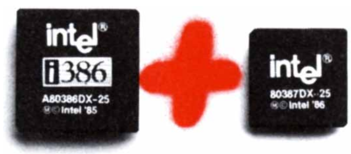


LTP in the granule cells of the hippocampus is achieved by theta rhythm. Electric pulses, which have been separated by 200 milliseconds (the time between the peaks of two theta waves), applied at the peak of theta rhythm result in LTP.

INTEL MATH COPROCESSORS

387™

Multiply 386™ System Performance With This Simple Addition.



Every 386™-based personal computer is powered by an Intel 386 microprocessor. And every one has a socket inside just waiting for you to plug in extra power with one of our Intel 387™ Math CoProcessors.

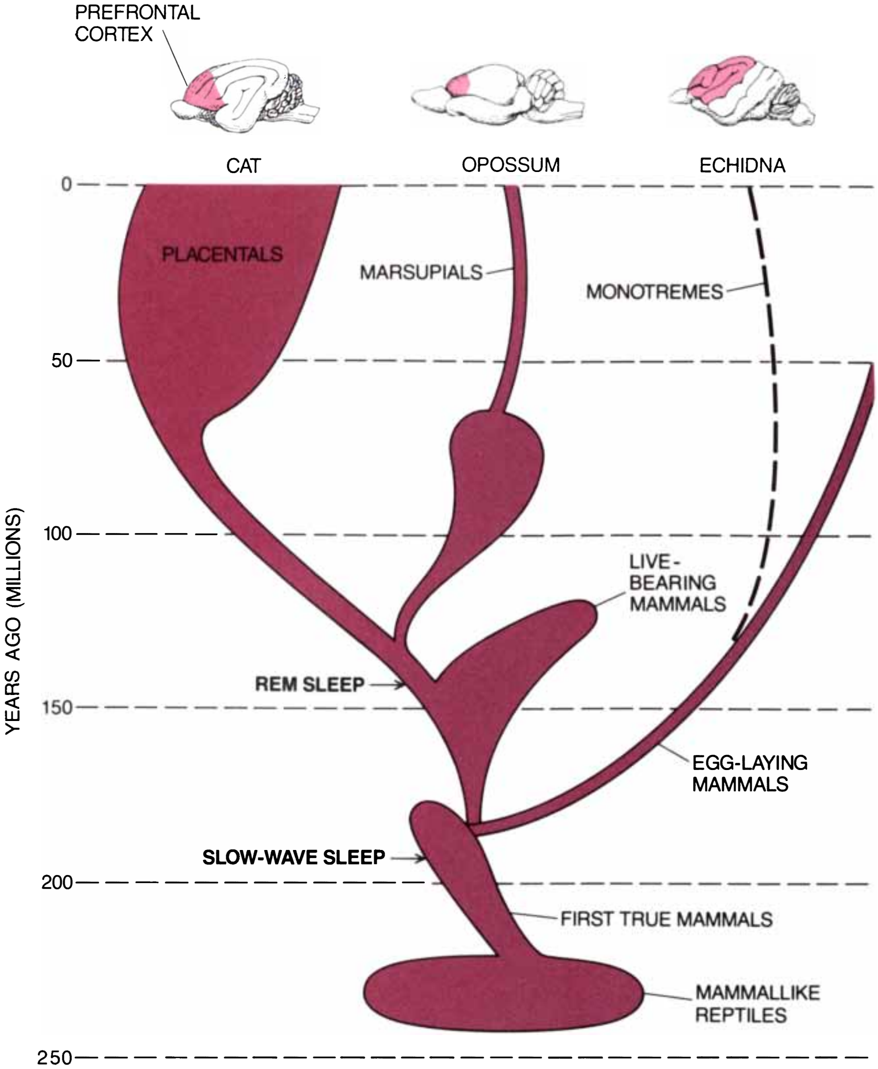
Add one to your 386-based PC to make over 1200 software programs run up to five times faster. That includes spreadsheets, business graphics, accounting and CAD programs.

We created Intel Math CoProcessors to be fully compatible with the Intel microprocessor inside your system. So there's total peace of mind that they're fully compatible with your system and your software.

Intel 387 Math CoProcessors. It's a matter of simple addition.

intel®

Call 1-800-538-3373 for our new guide with details on how you can boost your software's performance using an Intel Math Coprocessor.



EVOLUTIONARY TREE shows the divergence of placentals and marsupials from monotremes. The echidna, which does not possess REM sleep, has a larger prefrontal cortex compared with the rest of its brain than does any mammal, even humans. It is larger than in similarly sized animals, including the opossum and cat.

sleep. By the age of two, REM sleep is reduced to three hours a day, and the adult pattern has been established. Thereafter, the time spent in REM sleep gradually diminishes to a little less than two hours.

REM sleep may perform a special function in infants. A leading theory proposes that it stimulates nerve growth. Whatever the purpose in infants may be, I suggest that at about the age of two, when the hippocampus, which is still in the process of development at birth, becomes functional, REM sleep takes on its interpretive memory function. The waking information to be integrated at this point in development constitutes the basic cognitive substrate for memory—the concept of the real world against which later experiences must be compared and interpreted. The organization in memory of this extensive infrastructure requires the additional REM sleep time.

For reasons he could not possibly have known, Freud set forth a profound truth in his work. There is an unconscious, and dreams are indeed the "royal road" to its understanding. However, the characteristics of the unconscious and associated processes of brain functioning are very different than Freud thought. Rather than being a cauldron of untamed passions and destructive wishes, I propose that the unconscious is a cohesive, continually active mental structure that takes note of life's experiences and reacts according to its own scheme of interpretation. Dreams are not disguised as a consequence of repression. Their unusual character is a result of the complex associations that are culled from memory.

Research on REM sleep suggests that there is a biologically relevant reason for dreaming. The revised version of the Hobson-McCarley activation-synthesis hypothesis aknowledges the deep psychological core of dreams. In its present truncated form, the hypothesis of random brain stem activation has little explanatory or predictive power.

The Crick-Mitchison hypothesis provides a function for REM sleep—reverse learning—but it does not apply to narrative, only to the bizarre elements of the dream. What this implies with regard to REM processing in lower species must be defined before the theory can be evaluated further. In addition, the Crick-Mitchison hypothesis as applied to the hippocampus would suggest that neurons fire randomly during REM sleep, providing reverse learning. Instead, in my experiment on the neurons that coded space, these neurons fired selectively, implying an orderly processing of memory.

Further study will continue to elucidate the meaning of dreams. In particular, an experiment is needed to determine whether eliminating theta rhythm during REM sleep alone results in a memory deficit. Because theta rhythm has not been demonstrated in primates, it may have disappeared as vision replaced olfaction as the dominant sense. An equivalent neural mechanism may exist in the hippocampus that periodically activates the NMDA receptor. These studies and others to come will probe fundamental aspects of memory processing and the neuroscientific basis of human psychological structure.

FURTHER READING

INTERSPECIES DIFFERENCES IN THE OCCURRENCE OF THETA. Jonathan Winson in *Behavioral Biology*, Vol. 7, No. 4, pages 479–487; 1972.

LOSS OF HIPPOCAMPAL THETA RHYTHM RESULTS IN SPATIAL MEMORY DEFICIT IN THE RAT. Jonathan Winson in *Science*, Vol. 201, No. 435, pages 160–163; 1978.

BRAIN AND PSYCHE: THE BIOLOGY OF THE UNCONSCIOUS. Jonathan Winson. Anchor Press, Doubleday, 1985.

LONG-TERM POTENTIATION IN THE DENTATE GYRUS IS INDUCED PREFERENTIALLY ON THE POSITIVE PHASE OF Θ-RHYTHM. Constantine Pavlides, Yoram J. Greenstein, Mark Grudman and Jonathan Winson in *Brain Research*, Vol. 439, pages 383–387; 1988.

INFLUENCES OF HIPPOCAMPAL PLACE CELL FIRING IN THE AWAKE STATE ON THE ACTIVITY OF THESE CELLS DURING SUBSEQUENT SLEEP EPISODES. Constantine Pavlides and Jonathan Winson in *Journal of Neuroscience*, Vol. 9, No. 8, pages 2907–2918; August, 1989.

Knot Theory and Statistical Mechanics

Mathematical theories developed for quantum physics forge a connection between these two disparate fields

by Vaughan F. R. Jones

In 1984 I stumbled on a set of techniques that linked two of the most apparently disparate fields in mathematics and physics: knot theory and statistical mechanics. Statistical mechanics involves the study of systems with an extremely large number of component parts. Small systems are largely irrelevant. In knot theory, meanwhile, even the smallest knots and links may have subtle properties.

Nevertheless, certain algebraic relations used to solve models in statistical mechanics were key to describing a mathematical property of knots known as a polynomial invariant. This connection, tenuous at first, has since developed into a significant flow of ideas. The appearance of such common ground is not atypical of recent developments in mathematics and physics—ideas from different fields interact and produce unexpected results.

Indeed, the discovery of the connection between knots and statistical mechanics passed through a theory intimately related to the mathematical structure of quantum physics. This theory, called von Neumann algebras, is distinguished by the idea of continuous dimensionality. Spaces typically have dimensions that are natural numbers, such as 2, 3 or 11, but in von Neumann algebras dimensions such as $\sqrt{2}$ or π are equally possible. This possibility for continuous dimension played a key role in joining knot theory and statistical mechanics.

In another direction, the knot invariants were soon found to occur in quantum field theory. Indeed, Edward Witten of the Institute for Advanced Study in Princeton, N.J., has shown that "topological" quantum field theory provides a natural way of expressing the new ideas about knots. This advance, in turn, has allowed a beautiful generalization about the invariants of knots in more complicated three-dimensional spaces known as three-manifolds, in which space itself may contain holes and loops.

The new knot theory has already been of use in another, entirely unrelated domain. Molecular biologists have established that the double helices of DNA become knotted and linked in the course of biological processes such as recombination and replication. The untying mechanisms used in cells bear an uncanny resemblance to the simplest mathematical method for generating the new polynomial invariants.

Knots have been used for both practical and decorative purposes since time immemorial. Sailors have developed elaborate knots—sometimes with equally elaborate names—to serve their purposes. Mathematicians first became interested in knots only in the 19th century. Lord Kelvin, for example, attempted to deduce the structure of the periodic table of the elements by assuming that atoms were actually knotted vortices in the "ether." (Although the work proved unsuccessful, it did inspire Peter G. Tait to create the first knot tables, listing knots by some order of complexity.)

Since then, knot theory has become a fruitful branch of mathematics. One of the beauties of the discipline is that the main objects of study are so familiar: just take some string and join it at both ends. This serves as a perfectly adequate model for the "smooth non-self-intersecting closed curve" used by the mathematician. A more general version of a knot, called a link, may consist of more than one loop. Two knots or links are the same if they can be made to look exactly alike by pushing and pulling the string but not cutting it.

Consider a simple loop of string lying on a flat surface. Two important features of knot theory are immediately apparent. First, knots can be described by two-dimensional (planar) diagrams. Second, it is very difficult to distinguish two knots. It is not obvious that any two knots are different nor indeed that a loop of string is knotted at all. To prove such an assertion requires considering all possible three-dimensional deformations of the knot. Finding mathematical methods for distinguishing among knots—and distinguishing knots from unknotted loops—has become one of the major problems of knot theory.

In the 1920s Kurt Reidemeister simplified the study of knots considerably by introducing a small collection of two-dimensional "moves" on knot diagrams. These moves do not change the knot, and any two diagrams of the same knot can be converted into each other by a sequence of moves. Although the Reidemeister moves reduce the equivalence problem to two dimensions, the moves may be applied in an infinite number of ways, and so the issue is by no means resolved.

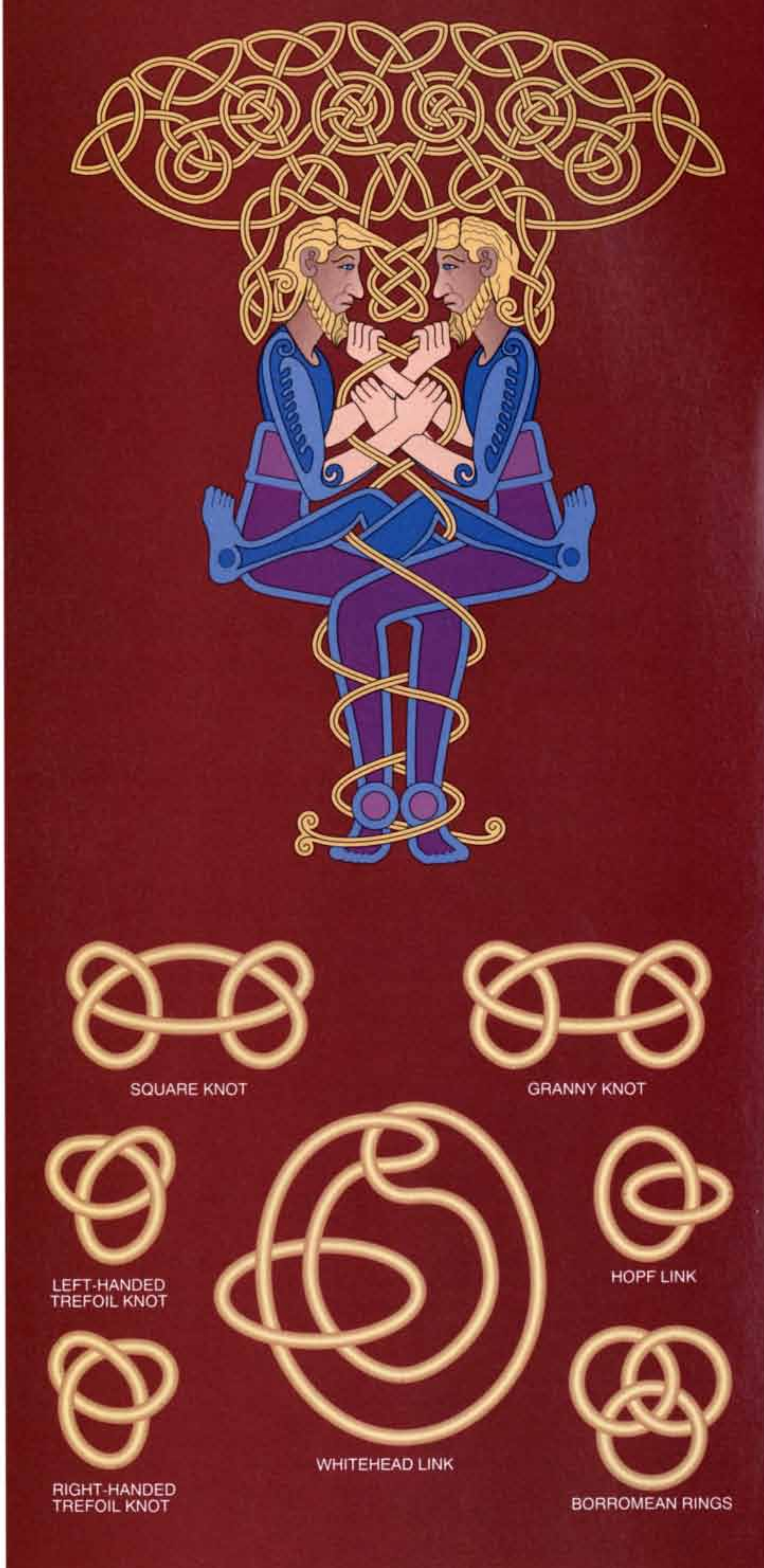
The oldest and most successful techniques of knot theory bypass two-di-

VAUGHAN F. R. JONES is a mathematician at the University of California, Berkeley. After receiving his master's degree in mathematics from the University of Auckland in 1973, he went to Geneva, where he studied physics and mathematics from 1974 to 1979. He received a *docteur ès sciences* in mathematics. Since arriving at Berkeley in 1985, he has spent two one-year stints at the Institut des Hautes Études Scientifiques in France. In August he was awarded the Fields Medal, the mathematical equivalent of the Nobel prize.

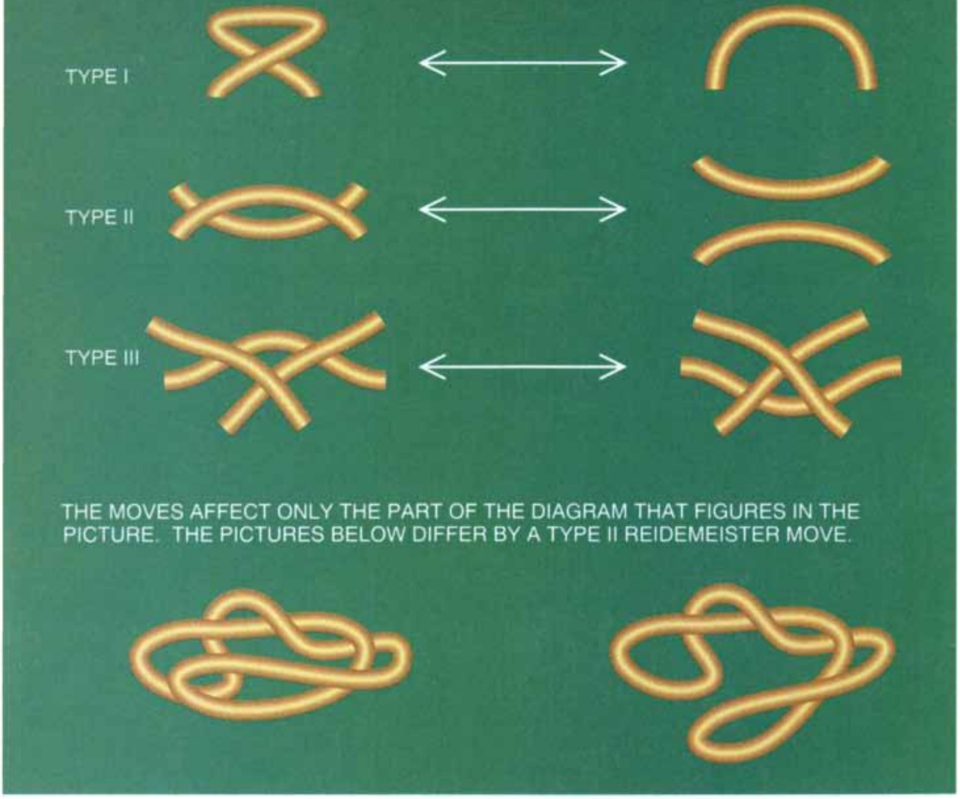
mensional diagrams and Reidemeister moves, at least in theory, and employ topological transformations instead. These techniques start by deleting the knot from ordinary three-dimensional space to obtain a so-called knot complement; the complement is then deformed in an arbitrary smooth fashion. The topology of the complement supplies what are called invariants of the knot, namely, mathematical expressions that depend only on the knot, not on any particular picture of it [see "The Theory of Knots," by Lee Neuwirth; SCIENTIFIC AMERICAN, June, 1979].

The most famous invariant is the Alexander polynomial, discovered by the American mathematician James W. Alexander in 1928. The polynomial, written $\Delta_K(t)$, is constructed according to the numbers of different kinds of crossing in a knot diagram. A simple trefoil knot, for example, has $\Delta_K(t) = t - 1 - 1/t$. Differently deformed versions of the same knot have the same Alexander polynomial; knots with different polynomials are different. Yet two knots with the same Δ are not necessarily equivalent. The Alexander polynomial does not distinguish, for example, between the square knot and the granny knot. Theorists have developed additional invariants during the past 60 years, but many knot problems remain unsolved.

It was a warm spring morning in May, 1984, when I took the uptown subway to Columbia University to meet with Joan S. Birman, a specialist in the theory of "braids" (which may be thought of as rather special kinds of knots). In my work on von Neumann algebras, I had been astonished to discover expressions that bore a strong resemblance to the algebraic expression of certain topological relations among braids. I was hoping that the techniques I had been using would prove valuable in knot theory. Maybe



KNOTS AND LINKS range from the simple to the inordinately complex. (The knot at the top is based on an illustration in the eighth-century Irish *Book of Kells*.) The left- and right-hand trefoils are the simplest knotted curves. Before the discovery of the Jones polynomial, the two were not readily mathematically distinguishable; neither were the square knot and the granny knot. The Hopf link is the simplest figure with two linked loops; the loops of the Whitehead link are not linked, although they cannot be disentangled. The Borromean rings also display unusual behavior: all three loops are linked, but cutting any one unlinks the other two.



REIDEMEISTER MOVES simplify the analysis of knots. Two knots are the same if and only if their diagrams can be made identical by some combination of moves.

I could even deduce some new facts about the Alexander polynomial.

I went home somewhat depressed after a long day of discussions with Birman. It did not seem that my ideas were at all relevant to the Alexander polynomial or to anything else in knot theory.

But one night the following week I found myself sitting up in bed and running off to do a few calculations. Success came with a much simpler approach than the one that I had been trying. I realized I had generated a polynomial invariant of knots. Most likely it was just the Alexander polynomial in a new form, although this connection by itself would have been extremely interesting for statistical mechanics. Or it could have been a new polynomial—an unexpected development for knot theory. It was a rare no-lose situation.

The second possibility turned out to be correct. I called this new invariant of links $V(t)$ and soon realized it did bear some superficial resemblance to the Alexander polynomial. Both $V(t)$ and $\Delta(t)$ can be calculated according to the skein relation [*see box on opposite page*] invented by the British mathematician John Horton Conway.

The similarity between $V(t)$ and $\Delta(t)$ prompted several groups to develop a two-variable polynomial, now called the HOMFLY (the acronym is derived from the initials of six of the eight people who discovered it), which contains the information of both Δ and V, and more. Here, however, the similarity between V and Δ ends. All attempts to interpret V in the same topological framework as Δ have failed. Furthermore, it is still not known whether there exists a nontrivial knot whose Δ is the same as that of an unknotted loop. In contrast, knots whose Alexander polynomial is equal to 1 (as is the Alexander polynomial of a simple loop) were discovered in 1934.

Thus far there seems to be no hint of how knot theory might be linked to statistical mechanics. The connection is not at the surface; it requires some explanation of what statistical mechanics is and what it is good for. That, in turn, requires starting from classical mechanics.

In classical mechanics a system of particles can be specified by giving the position and momentum of each particle at a certain time. The whole future evolution of the system is then determined by physical laws. But because a gram of hydrogen gas contains about 3×10^{23} molecules, it would be unreasonable to try to specify all the positions and momenta of the gas particles. Moreover, the change in the system that results from removing a few molecules would be completely unnoticeable to any observer apprehending the whole system.

The only quantities that are of interest to statistical mechanics are those that are insensitive to microscopic changes—for example, the average energy (temperature) of an ensemble of molecules. If one imagines a large system built by the addition of one atom at a time, those quantities will have a definite limit as the size of the system tends to infinity.

Innocent though it may sound, consideration of aggregate behavior creates some puzzles. One of the most obvious is irreversibility. The laws of motion do not change if the direction of time is reversed. For example, an elastic collision between a ball and an obstacle looks the same whether time runs forward or backward. And yet consider a system of balls bouncing around without friction on a rectangular table, constrained by a wall in one half of the table. If the wall is removed the balls will rapidly spread out over the entire table and will never conspire to go back into the half from which they started. The simple fact that a system contains a large number of particles seems to give time a definite direction.

Another surprise of large systems is the existence of phase transitions. Ice melts and water boils. Quantities such as pressure, which can be defined for small systems as smooth functions of the parameters, exhibit spontaneous jumps. Just considering systems containing many particles makes these smooth functions discontinuous.

Phase transitions exhibit interesting qualitative behavior, and so, to understand them, physicists have developed abstractly defined systems—models—for which macroscopic quantities such as pressure or heat capacity can be calculated explicitly as functions of parameters such as temperature.

Even in simple cases, the macroscopic quantities are very difficult to calculate. Consequently, models tend to eschew detailed realism. The simplest model is the Ising model, solvable in two dimensions, which consists of a system of "spins" arrayed in some regular pattern. (The spins are purely mathematical quantities; they do not necessarily have intrinsic physical meaning.) Each spin interacts only with its nearest neighbors. A "state" of the system is defined by giving a value of +1 or –1 to each of the spins. The energy of a state is characterized as the sum of energies resulting from interactions between nearest neighbors.

Many quantities of interest for large

CALCULATING KNOT INVARIANTS

Knot invariants are mathematical expressions that describe certain properties of knots. They derive their name from the fact that they do not change depending on the way that the underlying knot is pushed, pulled or twisted. The path from a loop of string to an equation in powers of t is a complex one, however. How does one determine an invariant, for example, the Alexander or Jones polynomial, of a given knot? The simplest method is the skein relation invented by British mathematician John Horton Conway. (As a schoolboy, Conway was trying to write a computer program to calculate the Alexander polynomial but found that the skein relation worked so well that he did not need a program.) Another technique, the Kauffman "states model," is notable for its elegance and the beginnings of a connection to statistical mechanics.

THE SKEIN RELATION

The skein relation starts with three oriented links (knotlike structures that can contain more than one loop) that are identical except in the region of one crossing.

L_+ L_- L_0

At that crossing, the links differ as above. The skein relation states that

$$(1/t)V_{L+} - tV_{L-} = (\sqrt{t} - 1/\sqrt{t})V_{L_0}$$

In addition, V and Δ of an unknotted loop are normalized to 1. These two facts make it possible to calculate these two invariants for any link.

The simplest nontrivial calculation is V of two unknotted loops.

This link has only one region where a crossing might be possible, and the three corresponding links in the skein relation are

L_0 L_+ L_-

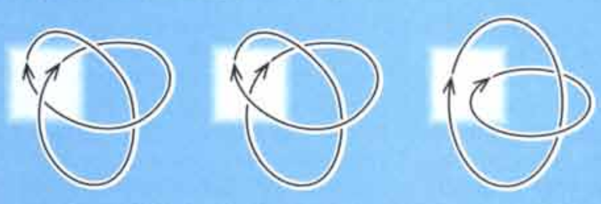
L_+ and L_- are both unknotted, so their polynomials are equal to 1. Thus, the skein relation reduces to

$$1/t - t = (\sqrt{t} - 1/\sqrt{t})V_{L_0}$$

Dividing both sides by $(\sqrt{t} - 1/\sqrt{t})$ yields

$$V_{L_0} = -(\sqrt{t} + 1/\sqrt{t})$$

The process is a little more difficult for a more complex link such as a trefoil. Once again one constructs a set of links that differ only at a single crossing.

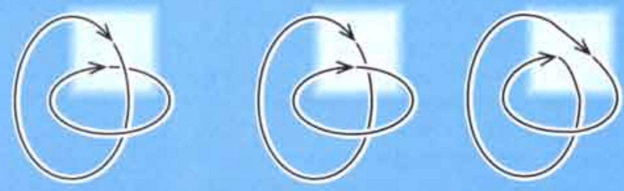


L_+ (the trefoil) L_- (unknotted) L_0 (a Hopf link)

The skein relation yields V_L in terms of V_H:

$$(1/t)V_L - t = (\sqrt{t} - 1/\sqrt{t})V_H$$

To calculate V_H, one constructs another set of links:



L_+ (a Hopf link) L_- (two loops) L_0 (unknotted)

Applying the skein relation to this set yields

$$(1/t)V_H + t(\sqrt{t} + 1/\sqrt{t}) = (\sqrt{t} - 1/\sqrt{t})$$

and ultimately $V_H = -\sqrt{t}\,(1 + t^2)$

Substituting this value back into the skein relation for the trefoil yields

$$V_L = t + t^3 - t^4$$

THE STATES MODEL

Kauffman's states model works by eliminating all the crossings in a link, replacing the link with disconnected circles. Each crossing may be in one of two states, which determines how it will be eliminated; each possible combination of crossing states determines a state of the link as a whole.

State A State A^{-1}

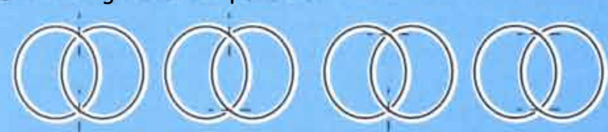
Kauffman then defines the "bracket polynomial" for a link as a sum over all states of the link.

$$\langle L\rangle = \sum_S \langle S|L\rangle\,(-A^2 - A^{-2})^{||S||-1}$$

(<S|L> denotes the product of all the A's and $1/A$'s at the crossings, and ||S|| is the number of circles in the resulting diagram.)

For example, the bracket polynomial of the Hopf link is $(-A^4 - A^{-4})$.

The following states are possible:



<S|L>	A^2	$AA^{-1} = 1$	$A^{-1}A = 1$	A^{-2}

||S||	2	1	1	2

$$A^2(-A^2 - A^{-2}) + 1 + 1 + A^{-2}(-A^2 - A^{-2}) = -A^4 - A^{-4}$$

To obtain V(t) from the bracket polynomial, one simply sets $A^4 = t$ and multiplies the bracket polynomial by an appropriate power of A:

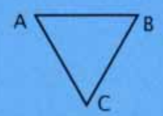
$$-A^4 - A^{-4} = A^{-6}(-A^2(1 + A^8)) = A^{-6}(-\sqrt{t}(1 + t^2)) = A^{-6}V_H(t)$$

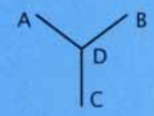
CONNECTING KNOT THEORY TO STATISTICAL MECHANICS

The connection between these two apparently disparate domains—one dealing with large aggregates of particles whose individual behavior is very simple, the other mostly with small systems that display subtle traits—is easiest to trace from the side of statistical mechanics. To mimic such natural phenomena as melting or boiling, physicists have turned to spin models: simplified aggregates defined on lattices of points. Each point has a particular "spin" and interacts with its nearest neighbor according to an energy function that depends on the two spins. A model's "partition function" is

$$\sum_{\text{STATES}} e^{-E(\text{STATE})/kT} \qquad (k \text{ is Boltzmann's constant})$$

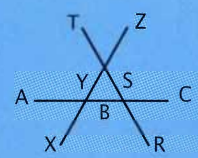
The star-triangle relation, which holds only for certain kinds of spin models (the Ising model and the Q-state Potts model among them), asserts that the partition functions of spin models defined on the two different graphs below are proportional to each other.

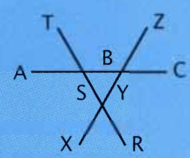




$$e^{-(E(A,B) + E(A,C) + E(B,C))/kT} = R\sum_{D} e^{-(E(A,D) + E(B,D) + E(C,D))/kT}$$

This is also true for so-called vertex models, in which spins are assigned to each edge of a graph and the energy is calculated according to the four spins that meet at a vertex. The star-triangle relation for a vertex model is expressed by the Yang-Baxter equation:





$$\sum_{BSY} e^{-\left\{\frac{E(A,B|X,Y) + E(R,S|B,C) + E(Y,Z|S,T)}{kT}\right\}} = \sum_{BSY} e^{-\left\{\frac{E(A,B|S,T) + E(X,Y|R,S) + E(B,C|Y,Z)}{kT}\right\}}$$

systems can be derived from the so-called partition function, Z, which is defined as the sum over all states of the exponential of the negative of the energy of the state. The Ising model can have a huge number of states, and so calculating this sum is the source of the problems encountered in solving a model. Even computer calculations are limited to small lattices. Nevertheless, various ingenious mathematical techniques have been devised. The Norwegian mathematician Lars Onsager solved the two-dimensional square-lattice Ising model in 1944; his pioneering formula showed the existence of a phase transition in the model.

One of the important products of Onsager's work was known as the star-triangle relation. For an Ising model defined on a particular pair of graphs, the star-triangle relation states that the partition functions must differ only by a proportionality constant if the values of the spins at the edges of the graphs are fixed [*see box at right*]. The name of the relation comes from the fact that one graph is triangular and the other is star-shaped.

Rodney J. Baxter of the Australian National University has worked out a solution of the Ising model based entirely on the consequences of this relation. One wonderful feature of Baxter's method is that it lends itself to generalizations. The Ising model can be generalized by allowing the spins at a site to take on values other than just +1 or −1. Such a model is called a spin model, and it is defined by the set of energies corresponding to interactions between neighboring sites for all possible spin values.

In general, spin models do not satisfy the star-triangle relation. To apply Baxter's work, one must find sets of interaction energies that satisfy the star-triangle equations and solve the models defined by those sets of energies. The simplest solution is the so-called self-dual Q-state Potts model. This model allows for Q possible spins (instead of the two permitted by the Ising model), and neighbors interact only if their spins are the same. (This model was first solved, although by another method, by Neville Temperley, then at the University of Swansea-Wales, and Elliott H. Lieb of Princeton University.)

What can be the connection between knot theory and the statistical mechanical models? The first step in making the connection is to shade the regions of a link diagram (the drawing of a knot) like a checkerboard. One can extract a graph by taking the shaded regions as vertices and the crossings as edges. The edges can then be given signs according to the relative orientation between the crossing and the checkerboard pattern. (The box above shows the entire procedure.) A type III Reidemeister move on the diagram immediately produces the star-triangle relation!

The move that leaves a knot unaltered leaves the partition function of an Ising model (for example) derived from the knot diagram essentially unaltered as well. This observation makes it quite natural to examine partition functions of spin models defined on the graph obtained from the checkerboard shading of a link to find out if they also describe invariants of links.

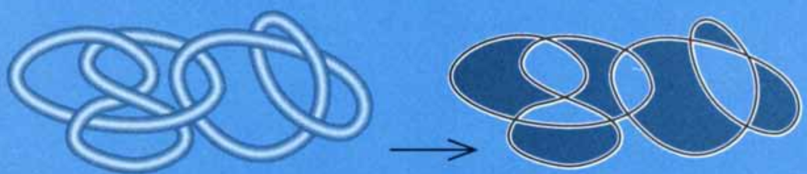
The partition function depends simply on the sum over all states of the spin model's energy function. That is, each state defines the spins of all the vertices in the graph and thus the energy of the interactions between them. To calculate the partition function, one assigns to each edge in the graph the exponential of the interaction energy of the vertices that the edge connects. The process is then repeated for all possible states. (To take the signs of the crossings into account, one must postulate two energy functions for the interaction between adjacent vertices, one positive and one negative.)

Now the question is, What conditions do the two energy functions have to satisfy for the partition function to be a link invariant? (This means that the value of the partition function should not change if a Reidemeister move is applied to the underlying diagram.) The best answer is a return to statistical mechanics: the energy functions of models that satisfy the star-triangle relation (and other relations corresponding to the other Reidemeister moves) will yield partition functions that are also invariants of links. The Ising model, for instance, yields an invariant known to knot theorists as the Arf or Kervaire invariant.

More significant, the Q-state Potts model yields an invariant for each Q, the number of spins allowed. In fact, it is the polynomial $V(t)$. Q and t are related by the equation $Q = 2 + t + t^{-1}$. (This can be verified easily by checking the skein relation.)

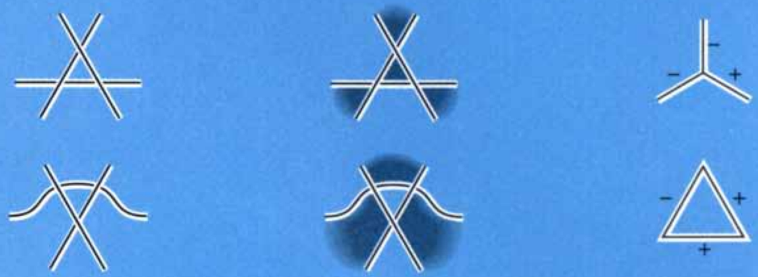
KNOTS TO GRAPHS

To get from spin models and statistical mechanics to knot theory requires turning knots into graphs. First, shade the knot (or link) like a checkerboard:



To extract a graph, take the dark regions as vertices and the crossings between them as edges. Label the edges as positive or negative according to the relative orientation of the crossing and the dark regions:

A type III Reidemeister move applied to a knot graph immediately yields the star-triangle relation:



This checkerboard trick has indeed established a connection between statistical mechanics and knot theory—but by a method that is quite bizarre. The star-triangle relation appeared in statistical mechanics as a method for solving models. In knot theory it appeared as a requirement for topological invariance. These two reasons might seem totally unrelated, and the connection might seem fortuitous were it not for its pervasiveness.

Spin models other than the Ising and Potts models yield other knot invariants used by mathematicians. Moreover, other kinds of models are even more powerful from the point of view of knot invariants. Vertex models, for instance, have spins sitting on the edges of the graph and energies coming from the vertices. It is possible, through the use of vertex models and the felicitously named quantum groups (systems of symmetries more elaborate than the geometric symmetries of ordinary space), to construct hosts of polynomial invariants—enough to determine the whole two-variable HOMFLY polynomial and another two-variable polynomial, the Kauffman polynomial.

Thus, although it is mysterious, the evidence for a connection between knot theory and statistical mechanics is substantial. Most of the hope for a better understanding of this correspondence centers around field theory, a mathematical view of the world in which every point of space carries a variable. (For instance, in a moving fluid the velocity of the fluid is a field with a vector at each point indicating the velocity.)

A powerful approach to field theory is to construct the field as a limit of fields defined on a discrete lattice whose spacing is made to approach zero. In this way, statistical mechanical models such as the Ising model also have application to field theory. Indeed, a careful look at the mathematics involved shows that the continuum limit of the two-dimensional Ising model is a one-dimensional quantum field theory. (Quantum field theories are field theories whose variables are operators on quantum states rather than classical ones.)

Some mathematicians and physicists have attempted to explain the knot theoretical nature of the statistical mechanical model by looking at its behavior in the continuum limit. Although their hopes for a solution have not yet been fulfilled, Witten has shown that the language of quantum field theory does at least provide a framework for defining the knot invariants described in this article.

It is significant that Witten's work shows that the invariants must exist in any three-dimensional space. There are many three-dimensional spaces, called three-manifolds, in addition to the familiar one that we live in [see "The Mathematics of Three-Dimensional Manifolds," by Jeffrey R. Weeks and William P. Thurston; SCIENTIFIC AMERICAN, July, 1984]. The three-torus, for example, is constructed by stretching a cube and "gluing together" its top and bottom, left and right, and front and back faces, so that a particle exiting the cube to the left immediately reenters it from the right. (This operation cannot actually be performed in three-dimensional space.)

Indeed, Witten's theory asserts that link invariants exist in an arbitrary three-manifold—even for a link that has no components at all. That means the invariants yield topological information not merely about links but also about the three-manifolds themselves. Many mathematicians and physicists are actively investigating this new development, which could forge connections between knot theory and other areas of science apparently even more unrelated than statistical mechanics was only six years ago.

FURTHER READING

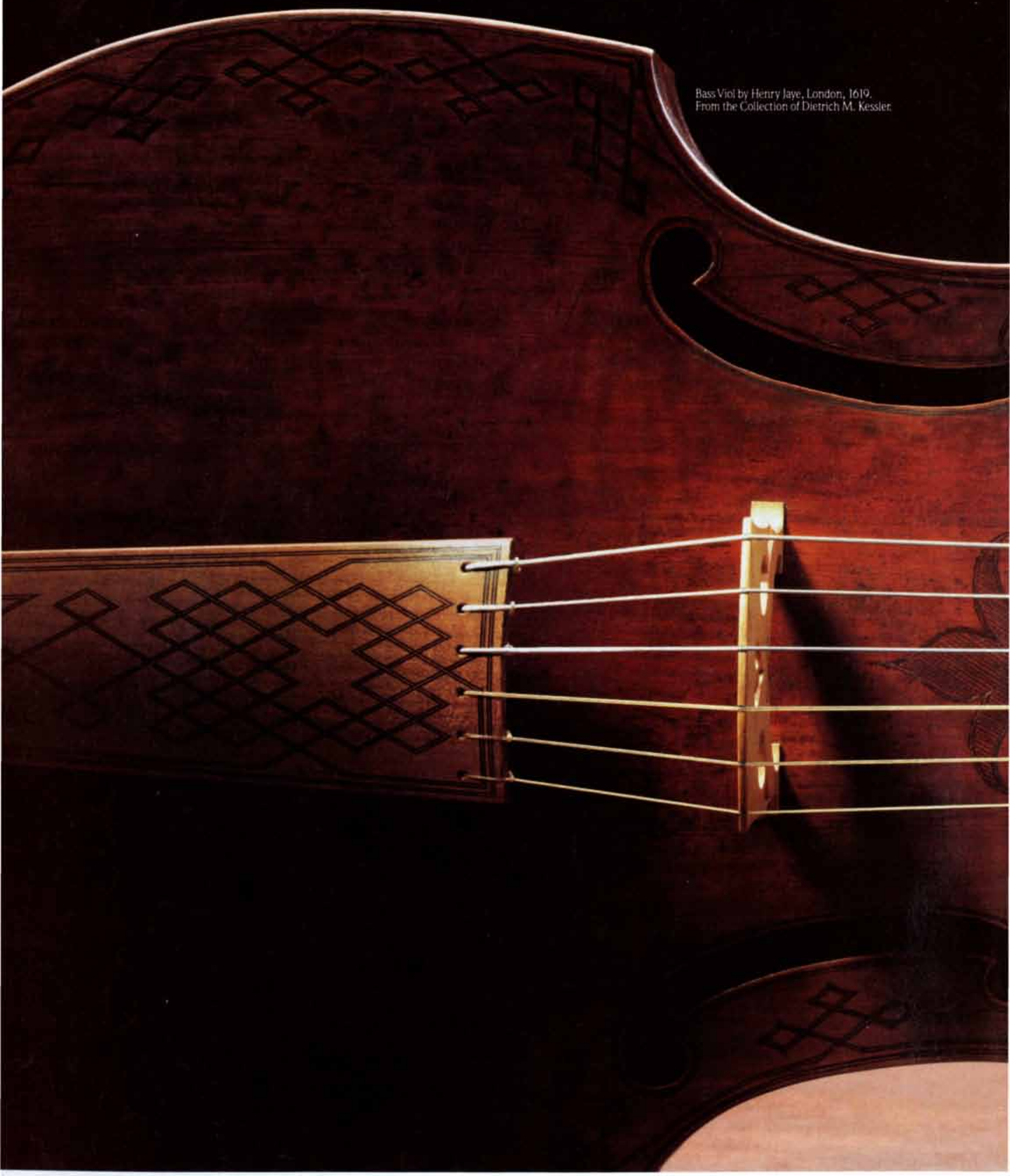
EXACTLY SOLVED MODELS IN STATISTICAL MECHANICS. R. J. Baxter. Academic Press, 1982.

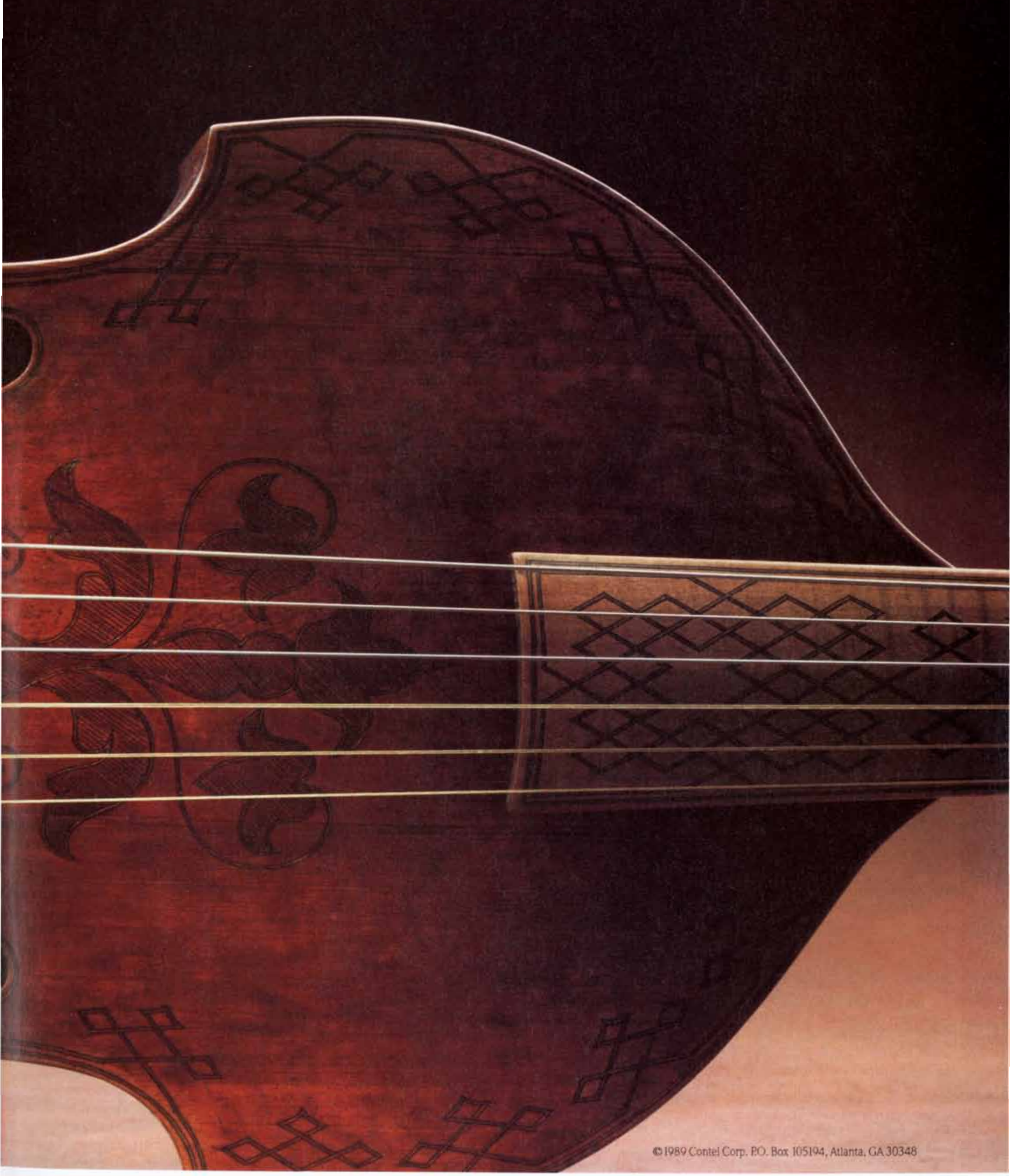
A POLYNOMIAL INVARIANT FOR KNOTS VIA VON NEUMANN ALGEBRAS. V.F.R. Jones in *Bulletins of the American Mathematical Society,* Vol. 12, Part 6, pages 103–111; 1985.

POLYNOMIALS FOR LINKS. W.B.R. Lickorish in *Bulletins of the London Mathematical Society,* Vol. 20, pages 558–588; 1988.

QUANTUM FIELD THEORY AND THE JONES POLYNOMIAL. Edward Witten in *Communications in Mathematical Physics,* Vol. 121, No. 3, pages 351–399; 1989.

REPRESENTATIONS OF THE ALGEBRA $U_q(SL_2)$, Q-ORTHOGONAL POLYNOMIALS AND INVARIANTS OF LINKS. A. N. Kirillov and N. Yu. Reshetikhin in *New Developments in the Theory of Knots.* Edited by T. Kohno. World Scientific Publishing, 1990.


Bass Viol by Henry Jaye, London, 1619.
From the Collection of Dietrich M. Kessler.


©1989 Contel Corp. P.O. Box 105194, Atlanta, GA 30348

A Roman Factory

Most historians hold that the Romans failed to develop the technology of mechanized production because slave labor was so cheap. The ruins at Barbégal in France tell another story

by A. Trevor Hodge

Few visitors to Arles in the south of France take the trouble to drive 19 kilometers northeast to the village of Fontvieille. Even fewer then turn south from Fontvieille and continue four kilometers along highway D 33, where a cart track leading off to the left ends at one of the strangest technological marvels of classical antiquity, the mill of Barbégal.

Barbégal is a well-preserved example of something that, according to all the textbooks, never existed at all—an authentic, ancient Roman, power-driven, mass-production, assembly-line factory. It is a large, imposing ruin that required no excavation. Yet so great was the attention concentrated on artistic monuments such as temples and so great the neglect of technical monuments such as mills that it was only after the French archaeologist Fernand Benoît published his investigations of the site in 1940 that anyone recognized its importance.

Barbégal is significant because it calls into question what may be termed the technological theory of the decline and fall of the Roman Empire. The theory maintains that the availability of cheap slave labor prevented the Romans from developing alternative sources of power, without which large-scale manufacturing is impossible.

A. TREVOR HODGE is professor of classics at Carleton University, Ottawa. He was educated at the University of Cambridge, where he received a B.A., an M.A. and a doctorate in classics and a diploma in classical archaeology. He originally specialized in Greek architecture, particularly the techniques by which temples were constructed, but later widened his interests to include the campaign of the Battle of Marathon and the founding of Marseilles by the Greeks. His current research focuses on the Roman water supply, about which he will publish two books next year. This is his second article for SCIENTIFIC AMERICAN.

One need go only 100 meters or so down the track to see the remains of this industrial complex, compactly laid out on a hillside. The specifications are simple. Barbégal was an immense flour mill, probably dating from the fourth century A.D. The power to drive the millstones came from 16 waterwheels, arranged in parallel rows of eight. Each row ran downhill so that the water cascaded from one wheel to the next, driving all eight in turn before running into the drain at the foot of the hill and thence discharging into a marsh 500 meters away.

The ingenuity of the arrangement is self-evident. In the medieval and later ages it was not uncommon to build water mills one after the other along the course of the same stream; here they were consolidated under a single roof to form a veritable factory. One may also admire the design from the opposite point of view, for the power sources were cleverly dispersed. It would have been impossible to drive so many machines with a single wheel because the friction of the belt drive, or any other means of transmission, would have proved too great a load. Multiple grinders became possible only when each millstone was provided with a waterwheel of its own.

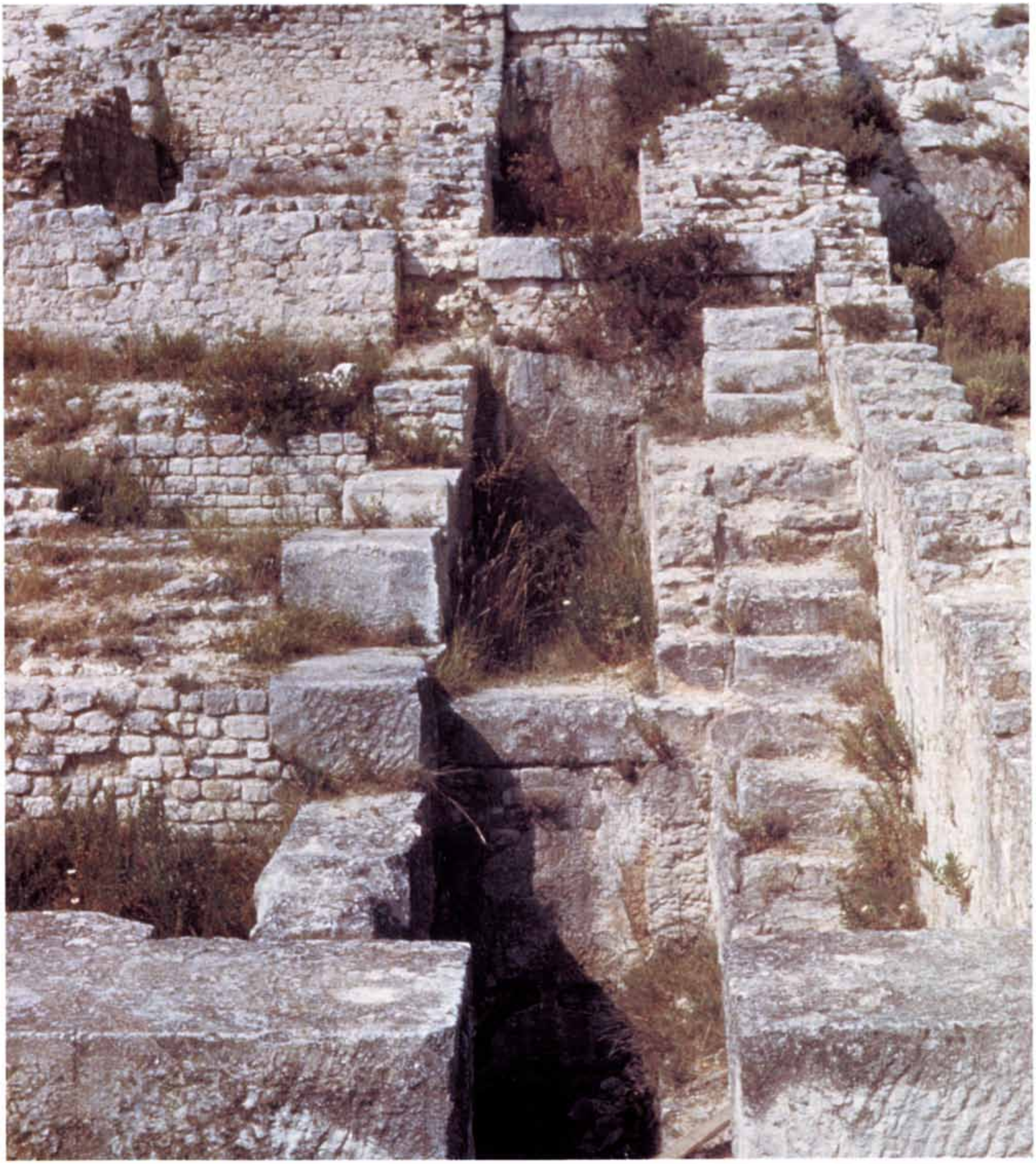
Who was Barbégal's great designer, as the engineer has the right to be called? A tomb inscription of a plausible date, still preserved in the Alyscamps cemetery of Arles, celebrates a certain Quintus Candidius Benignus, a famous local hydraulic engineer who was "clever like none other, and none surpassed him in the construction of machines and the building of water conduits." There is no firm evidence that it was he who built Barbégal or that the factory was the basis of his fame, but one would like to think it true just the same.

Climbing the hill to its crest, one sees on the other side a shallow vale spanned by a typical Roman aqueduct carried on arches, which supplied the mill. A second look shows that another aqueduct stands alongside the first and is largely hidden by it. At the ridge, the western aqueduct made an abrupt right-angled turn and continued west, evidently extending 17 kilometers to supply drinking water to Arelate, the Roman city of Arles. The eastern aqueduct continued straight on to pierce the ridge in a specially cut channel and so reached the top end of the mill complex to supply it.

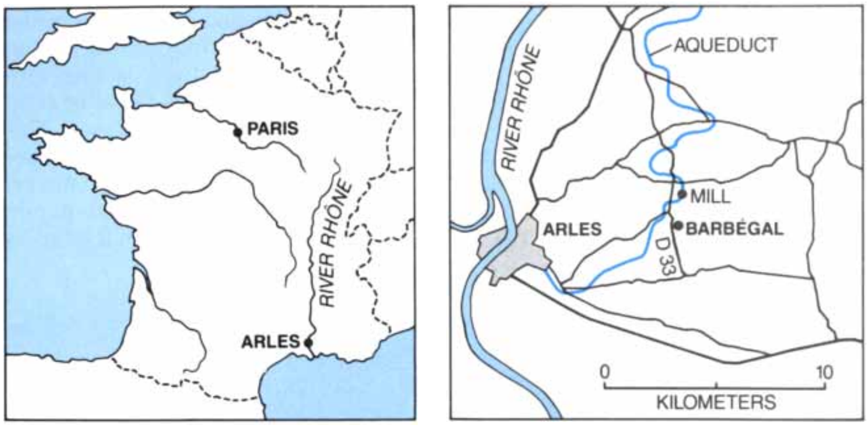
Plainly, the mill had an aqueduct dedicated to its use alone, although the precise relation between the mill aqueduct and the city aqueduct remains the subject of some discussion. The latest research seems to suggest that, first of all, an aqueduct was built to supply the city and that after a considerable time it could no longer do so, either because the water quality deteriorated or the conduit became clogged by the calcium carbonate encrustation to which most such aqueducts were subject. At that point, the argument goes, a new aqueduct was built, and the original one was converted to a mere source of power. The mill would therefore have been built expressly to exploit a source of power that had unexpectedly become available in a suitable location, for the water's quality and quantity were still adequate for industrial use.

On emerging from the channel in the summit ridge, the aqueduct split into two at a Y-shaped junction. It was once thought that the triangular space embraced by the Y served as a header reservoir (to ensure an even flow of water), but this hypothesis is now thought unlikely. The mill complex proper, which the water now entered, was laid out down the slope of the hillside in the form of a rectangle approximately 42 by 20 meters. It consisted of two rows of millhouses, where the grain was actually ground, separated by a service road that allowed mules and donkeys to bring in the grain and take away the bags of flour. On the two external wall faces were mounted two rows of waterwheels, each contained within an oblong masonry enclosure called a wheelbox. These boxes were themselves arranged from end to end in a stepped series running downhill. The spacing of the millhouses in relation to the wheels mounted on their common external wall was apparently not quite uniform. Some wheelboxes are longer than others, and the lower wall is not completely straight. These discrepancies were no doubt necessitated by inequalities in the terrain.

Other irregularities spring at least in part from the difficulty of making a full reconstruction. Nowhere are the walls preserved to much more than one or 1.5 meters high, and a good deal of the superstructure was in woodwork, long since gone. The restorer is therefore forced to exercise a good deal of creative, if discriminating, imagination. This requirement is not rare in restoring ancient buildings. Here, however, the researcher faces an added and purely technical problem: the fall in level from



CRUMBLING FOUNDATION of a Roman flour mill climbs a hill at Barbégal in southern France. An aqueduct fed two streams that cascaded past 16 waterwheels, whose boxlike containments can still be seen. It is the largest known ancient factory.



BARBÉGAL was built in the fourth century A.D. near the port of Arles (*left*), along an aqueduct that had once supplied the city (*right*). The mill probably produced enough flour to feed the 12,500 people estimated to have lived in Arles at the time.

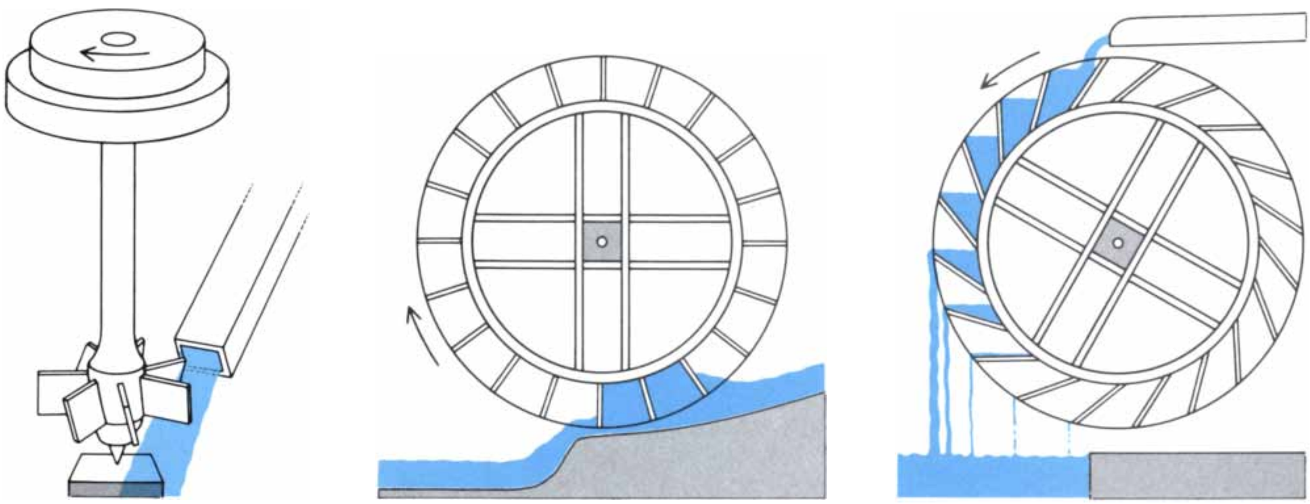
the top of the hill to the bottom gives an average ground slope of 30 degrees, but to fit in the eight wheels, the slope followed by descending water would have to be rather steeper. A variety of expedients have been proposed, such as building up the top section of the mill to a high level and effectively sinking the lower section by putting the waterwheels in the basement, thus steepening the hydraulic gradient. Although this difference in the placement of the wheels may be reflected architecturally in the markedly different layout of the last three millhouses [*see top illustration on page 110*], no one has yet been able to demonstrate how the Romans reconciled the disparity between the hydraulic and ground gradients.

Each wheel was .7 meter wide and 2.1 meters in diameter and was mounted on an axle that passed from the wheelbox through a hole in the wall (one or two of which are still discernible) into the millhouse. There it would have been coupled, by a pair of toothed wheels, to a vertical axle that passed upward through the ceiling and, on the floor above, drove the rotating basalt millstone (.9 meter in diameter) that did the grinding. The system is well known from other Roman mills. The chief gap in our knowledge is that we do not know whether advantage was taken of the gearing to increase the speed of rotation and hence the efficiency of the milling [*see bottom illustration on page 110*].

This missing piece highlights what is perhaps Barbégal's most tantalizing puzzle: How much power did the mill generate, and how much flour could it, or did it, produce? The last two questions are not necessarily identical because we have no idea whether the mill ran continuously or closed down at night. Furthermore, the size of the millstones is known, but their rate of revolution is not. Still, even a tentative production figure can provide at least an idea of the order of magnitude concerned.

The most recent estimate was made in 1983 by Robert H. J. Sellin of the University of Bristol, based on an aqueduct flowing at a meter per second (discharging .3 cubic meter per second) and a wheel rotating 10 times a minute at 65 percent efficiency. Sellin calculates that such a wheel could generate about two kilowatts, or 2.5 horsepower. As the British scholar John Landels pungently comments, "For a modern (and slightly depressing) comparison, a small motor-cycle engine (250 cubic centimeters) develops about the same power." This power, properly geared, would drive a millstone at about 30 rotations a minute, enough to produce about 24 kilograms of milled flour in 24 hours, or nine metric tons a day for the entire Barbégal complex. Assuming 50 percent downtime, to accommodate interruptions for late-arriving grain shipments, maintenance, low water, breakdowns, rest time and other interferences, Barbégal would still produce, according to Sellin, "4.5 metric tons per day or enough to feed a population of 12,500 based on a consumption unit of 350 grams a day. This corresponds closely enough to estimates of the size of the population of Arles in the fourth century A.D."

The new estimate seems to solve one problem. Earlier ones had been higher, and most historians were if anything worried that the mill produced more flour than the local population could have consumed. The usual explanation was that Arles was a legionary fortress and that Barbégal supplied its garrison.



WATERWHEEL DESIGNS are shown in order of increasing complexity and efficiency. Norse wheels (*left*) turn millstones directly, undershot wheels (*center*) require gears and overshot wheels (*right*) also require an elevated stream of water.

But legionary garrisons were hardly rare, and nowhere else in the Roman Empire is a garrison known to have had recourse to such an expedient. The lower estimate eliminates these difficulties and thus supports the theory that the factory was built to exploit the availability of the aqueduct.

Another technical problem lies in determining the design of the factory, which depends on the kind of waterwheel it used. Three varieties were known in antiquity [*see bottom illustration on opposite page*]. The simplest—the Norse wheel—is mounted horizontally on a vertical axis that passes through the millhouse floor to turn the millstone above. The absence of gears makes for ease of building and maintenance but means that the millstone turns slowly, with the waterwheel, and hence grinds inefficiently. Norse wheels are not immersed in the river but driven by jets of water from a chute. They are still common in China (where, perhaps significantly, windmills also rotate in the horizontal plane).

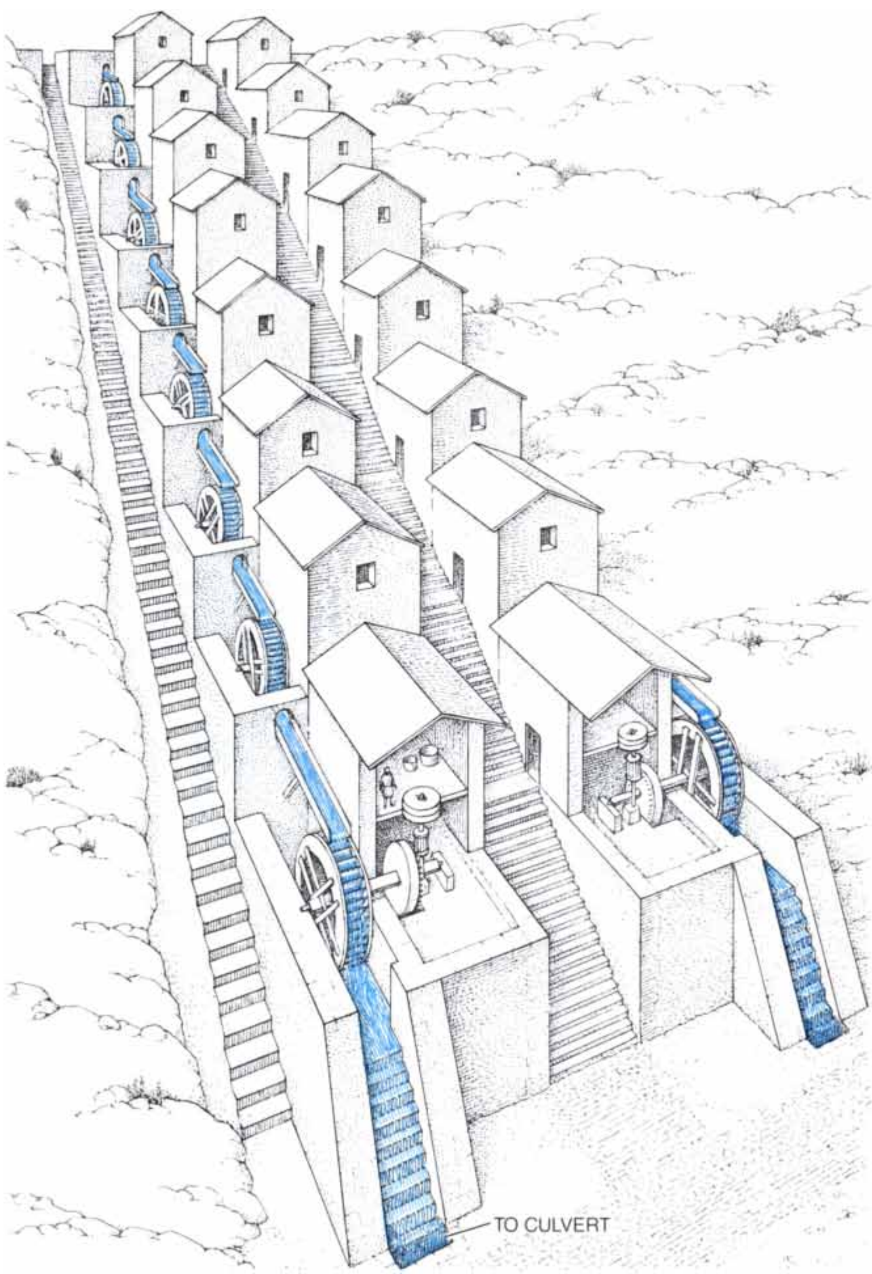
The more familiar vertical waterwheel comes in two varieties—undershot and overshot. In the former, the wheel in effect dips into the river below, which turns it. In the latter, a chute delivers water to the top of the wheel so as to drive it both by the forward momentum of the current and by gravity, as the water falls. In both cases, the drive is then taken by a horizontal axle and thence through wooden gears, as at Barbégal, to a vertical axle. Of the two designs, the overshot is the more efficient. Yet energy efficiency may be no great advantage if there is plenty of water to drive the mill anyway. Moreover, because an overshot wheel requires that the water delivery be regulated by a millpond and an elevated chute, it is quite expensive to install.

Because there is a step-by-step increase in sophistication from the Norse to the undershot and finally to the overshot design, the established view has been that the different techniques appeared in that chronological order. But N.A.F. Smith of the Imperial College of Science and Technology in London has shown that they developed independently, in response to regional needs and traditions, and that in most of the Roman Empire the horizontal wheel appeared rarely, if at all. At Barbégal it is clear that the wheels were vertical. The only question is whether they were of undershot or overshot design. Although both have been suggested, there can be no doubt that the overshot model is by far the more probable, if only because its superior efficiency would have been required to exploit the comparatively small but controlled volume of water delivered by the aqueduct.

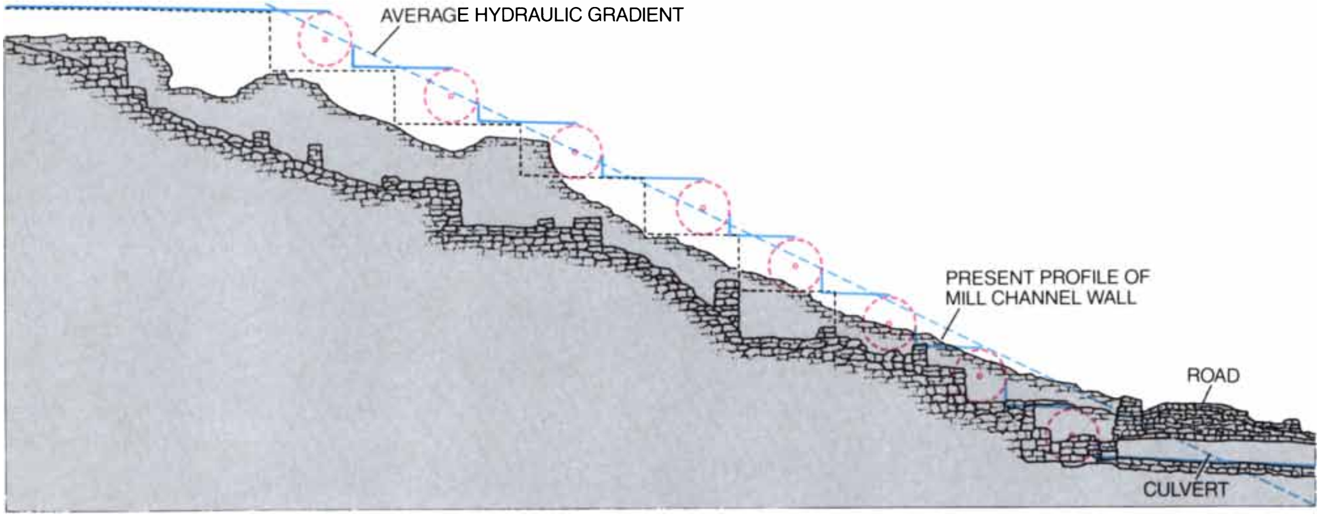
Barbégal transcends such technical matters to reach the wider question of the social role of technology. The orthodox doctrine holds that the ancients understood the waterwheel well but made surprisingly little use of it because the motive power supplied by slaves was so cheap.

Here the apparently trivial waterwheel touches on one of the great paradoxes of antiquity: the relative stagnation of technology in a society that produced some of the best minds yet encountered. It is generally accepted that ancient technology was deficient in the application of power. The Greeks and Romans could not, for example, exploit steam practically because they lacked the machining skills for fine-tolerance work in metal. Why, then, did they neglect the sources that were available to them? The ancients were familiar, through ships' sails, with wind power but apparently never used windmills. They actually constructed waterwheels but not in significant quantities.

Here again recent scholarship has increasingly cast doubt on the estab-



OVERSHOT WATERWHEELS, the most efficient of the three main designs, are conjectured to have driven the mills (seen here in an idealized reconstruction). Three service stairways flanked the mills, each of which probably had its own enclosure.



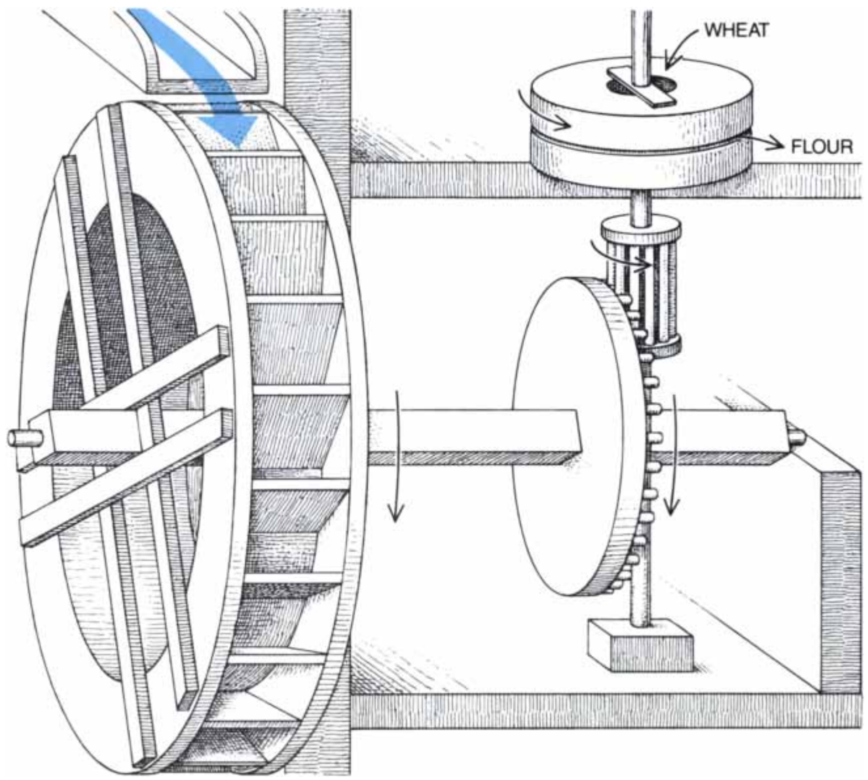
RESTORED VIEW of Barbégal arranges the wheels along a declining curve, a logical way to reconcile the gradients of the ground and water. The wheels were probably mounted above ground at the top of the hill and below ground at the bottom.

lished version. Waterwheels may not, in fact, have been all that rare in antiquity. The evidence that they were is almost entirely archaeological. Workers have found very few wheels; hence, it is assumed that few existed. A reexamination of the literary as well as the archaeological evidence now suggests, however, that waterwheels began to find more widespread use throughout the empire in the second century A.D.

The revision is based partly on the discovery of more wheels and partly on the no less convincing argument by analogy. Eleventh-century England, for example, has left evidence of scarcely a dozen mills of any kind, yet the Domesday Book records that at least 5,624 existed. If one assumes that the ratio between actual and discovered mills is the same for the classical world, then some of the standard explanations of the dearth of waterwheels begin to ring hollow.

According to one standard explanation, the Mediterranean region, in particular Italy, lacked streams offering the reliably constant water flow that mills require. This theory seems convincing only so long as one assumes that few mills existed. If one assumes instead that they were numerous, one quickly remembers that in fact there was nothing the Romans were more famous for than the widespread provision of a reliably constant water flow—by means of aqueducts. The aqueduct ought to have been ideally suited to the needs of a water mill, particularly of the overshot design, as in the case of Barbégal. One may of course object on the grounds of expense, but again an answer can be found. Of all the users of water, mills must surely stand alone in that they neither consume the water nor pollute it. In theory, therefore, it would have been possible to install mills at any point along an ordinary city aqueduct without inhibiting its normal function.

In fact, something like this was apparently done on the hillside of the Janiculum, in the suburbs of Rome. One can only wonder if this was an isolated case or if the remains of other such mills have been overlooked because no one expected to find them. Certainly it was not uncommon for aqueducts to incorporate cascades at a point where the terrain obliged the aqueduct to lose height rapidly and where accordingly a mill could have easily been built. Of course, not every such case would have proved econom-



GEARING MECHANISM would have been necessary to convert the horizontal rotation of an overshot or undershot waterwheel into vertical rotation. Here it is assumed that the gear was made larger than the pinion so as to speed the grinding.

ical: a natural site for a mill might have been separated from principal urban markets by excessive distance or forbidding terrain.

Yet it cannot be denied that the Romans used water mills less extensively than they might have. Why did they not? The traditional answer—the technological theory of Roman decadence—argues that the Romans lacked economic incentives to develop technology as long as the empire was expanding and had lost their technological vitality by the time the frontiers had stabilized. This theory relies almost wholly on the debilitating influence of slavery.

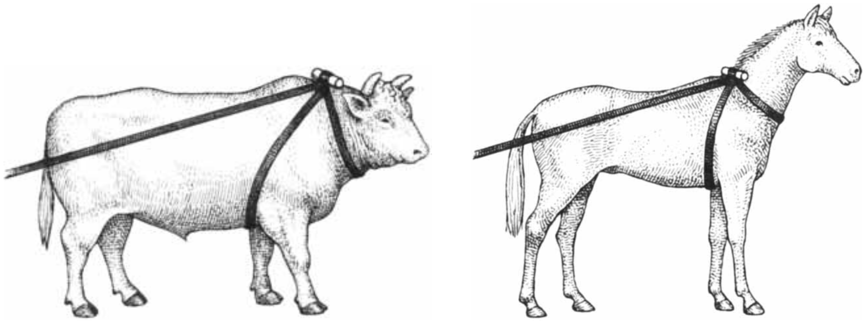
There is a danger in relying too much on slavery as an explanation. It is true that cheap labor, whether slave or free, can have a stultifying effect on the development of labor-saving technologies. Yet slaves were often more expensive than is commonly realized, combining high capital costs with a short working life (when heavily worked) and high maintenance costs in between.

Slaves might have been economical when large numbers appeared on the market, depressing the average purchase price. Such a profusion of slaves obtained as long as the Roman Empire was expanding and its victorious legions were sending prisoners of war back from the frontiers. Once the frontiers stabilized in the first century A.D., and more decidedly, once the barbarians started winning, the supply of slaves dried up and the cost of manpower rose. Rome was thus in the position of some modern enterprises: it had to grow or decline.

But a much more direct reason, purely technical in nature, explains why the waterwheel languished in ancient times and with it structures such as the mill of Barbégal and the cultures that could be based on them. It has to do with the harnessing of horses.

Although the use of horses as draft animals was much restricted by the absence of horseshoes, that deficiency was nothing compared with the inadequacy of the standard harness. The harness straps and the equine anatomy so coincided that if the horse tried to pull a heavy load it merely strangled itself. For overland freight transport—even on the excellent Roman roads—horses were out. Motive power was provided by oxen, which are very strong but painfully slow.

Land transport was therefore prohibitively expensive over distances greater than about 15 to 20 miles, a journey of several days for an ox. This weakness in the distribution of goods had a profound consequence: it all but ruled out factories, because the economies of scale realized in mass production would be more than outweighed by the expenses of getting the goods to market. And if there was no mass production, there could have been little incentive to mechanize.



ROMAN HARNESS was designed so that a pair of oxen could push against a yoke with their humps (*left*). It was misapplied to horses, which would strangle if they tried to pull a heavy load (*right*). The ancients therefore had a choice between painfully slow oxen and horses that were slow because of pain. In either case, the author hypothesizes, land transportation was too costly to justify mass production.

Machinery powered by water consequently did not pay off unless water was conveniently and cheaply available and the local market was adequate. More distant markets could be served only in the rare cases where mills could be built near ports at which river barges or seagoing ships—by far the cheapest means of transportation in antiquity—could carry the product away. The fourth-century poet Ausonius clearly if poetically describes such an enterprise—a marble-sawing works on the river Ruwer, a tributary of the Moselle near Trier. It is one of the handful of mills known to have produced something other than flour.

Is, then, Barbégal unique? Not quite: two other multiple mills are known, but neither is on a scale anything like Barbégal's. One is at Chemtou in western Tunisia, where a combination bridge and dam dating from about the beginning of the second century A.D. spanned the river Medjerda. Set into the bridge abutment were three horizontal mill wheels, side by side and in parallel rather than in series, as at Barbégal. The wheels were housed in cylindrical stone enclosures known as penstocks to which the stream of water was admitted slightly off center so as to play chiefly on one side of the wheel, as in a turbine. The other mill is in Israel on a dam on the Crocodile River near ancient Caesarea, halfway between Haifa and Tel Aviv. Here again there were two horizontal wheels, each at the bottom of a penstock so that the water, admitted at the top, strikes the wheel with its momentum increased by the fall. Neither installation has been fully studied, but together they remain the only known parallels to Barbégal.

And there is the rub. Barbégal is located in a densely populated area of a modern, developed country. It was not buried and required no excavation beyond a minor cleanup. The mill could not have been easy to overlook, but it stood neglected for centuries.

The lesson is clear. Other Barbégals must surely await discovery in the more remote and less studied parts of the Roman Empire. If this one could escape notice in France until 1940, what masterworks may yet lie hidden in Iraq or North Africa, where desert sands now enshroud the remains of Roman cities? It is worth seeking them out, for should they be found, the implications for the history of technology, manufacturing, economics and culture will be momentous indeed.

FURTHER READING

ENGINEERING IN THE ANCIENT WORLD. John Landels. University of California Press, 1978.

THE LARGE ROMAN WATER MILL AT BARBÉGAL (FRANCE). Robert H. J. Sellin in *History of Technology,* Vol. 8, pages 91–109; 1983.

STRONGER THAN A HUNDRED MEN: A HISTORY OF THE VERTICAL WATER WHEEL. Terry S. Reynolds. Johns Hopkins University Press, 1983.

THE ORIGINS OF WATER POWER: A PROBLEM OF EVIDENCE AND EXPECTATIONS. N.A.F. Smith in *Transactions of the Newcomen Society,* Vol. 55, pages 67–84; 1983–84.

SIPHONS IN ROMAN AQUEDUCTS. A. Trevor Hodge in *Scientific American,* Vol. 252, No. 6, pages 114–119; June, 1985.

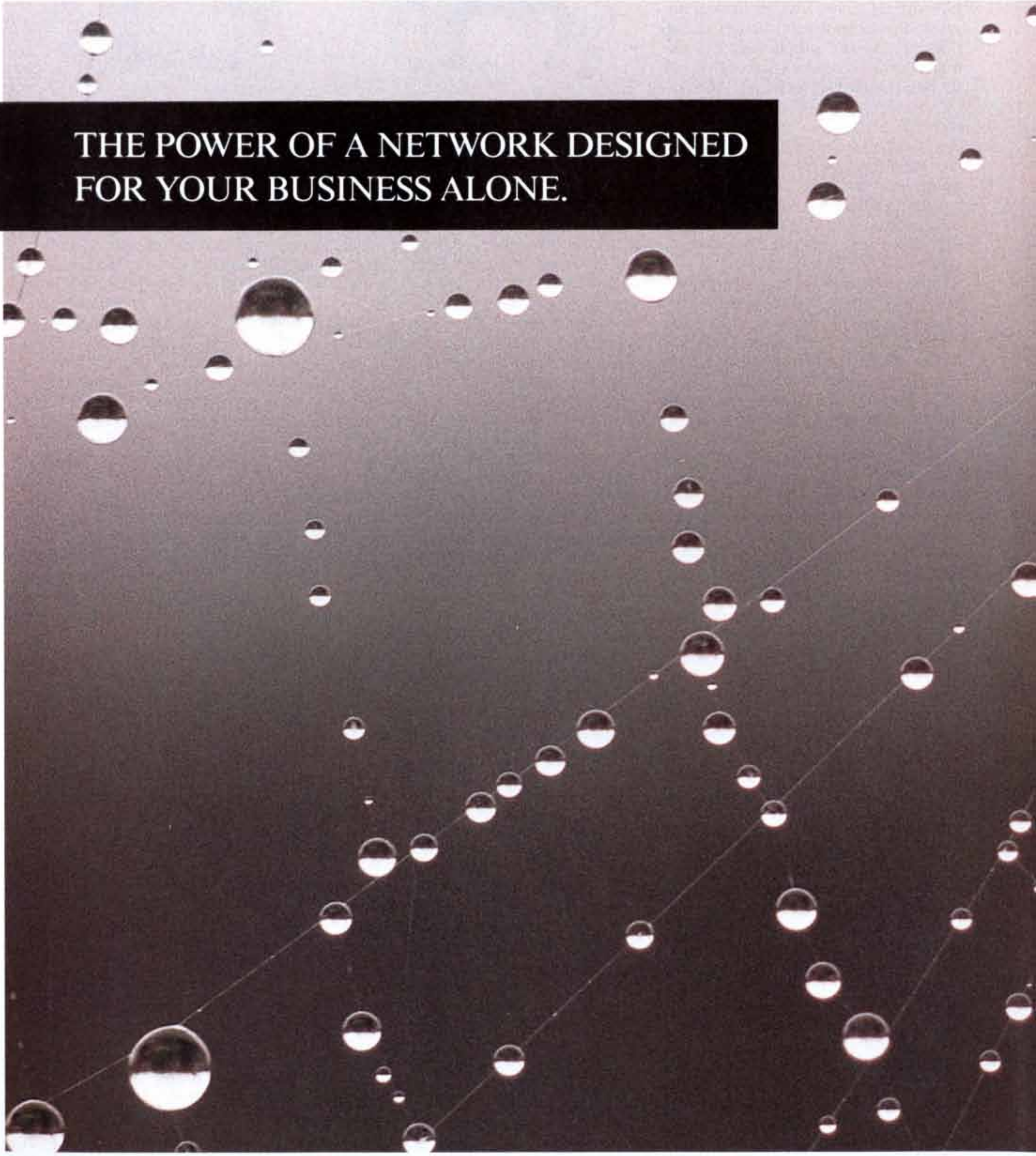
ROMAN AQUEDUCTS AND WATER SUPPLY. A. Trevor Hodge. Gerald Duckworth & Co., Ltd. (in press).

For the growth of your business, the prosperity of your business, there is no more powerful tool than the telecommunications network.

And at GTE, there is nothing vague about the power we are putting into your hands when we design that network for you.

For one of the nation's leading retailers, we are now implementing one of the largest satellite-based private

THE POWER OF A NETWORK DESIGNED FOR YOUR BUSINESS ALONE.



The Mechanical Design of Insect Wings

Subtle details of engineering and design, which no man-made airfoil can match, reveal how insect wings are remarkably adapted to the acrobatics of flight

by Robin J. Wootton

Insects include some of the most versatile and maneuverable of all flying machines. Although many show rather simple flight patterns, some insects—through a combination of low mass, sophisticated neurosensory systems and complex musculature—display astonishing aerobatic feats. Houseflies, for example, can decelerate from fast flight, hover, turn in their own length, fly upside down, loop, roll and land on a ceiling—all in a fraction of a second. Dragonflies, hover flies and lacewings are scarcely less remarkable but in different ways, for with versatility comes diversity. It is now becoming increasingly clear that much of the skill and variety in insect flight results from the subtle and varied constructions of their wings. But just how wing characteristics relate to flight performance has only recently begun to develop from speculation into rigorous science.

As a graduate student in the 1960s working with fossil insects, I became deeply frustrated with the limited range of conclusions that could be safely drawn from the often beautifully preserved material at my disposal. I was conscious that wings were far more than abstract patterns of veins and membrane, as is their usual depiction. Each wing seemed to me to be an elegant piece of small-scale engineering, presumably finely adapted to the functions that it served.

I also realized that an understanding of how insect wing architecture relates to performance could illuminate questions of insect evolution. Primitive winged insects seem to have had two unlinked pairs, as dragonflies and crickets do. But in many modern groups, such as true bugs (*Hemiptera*), moths and wasps, forewings and hind wings beat in unison or are physically linked together, operating as a single composite airfoil. In some insects, hind wings have been reduced to small sense organs, as in flies, or completely lost, as in some mayflies. Forewings have thickened and become protective sheaths in beetles and earwigs or been reduced to vestiges in stylopids and some stick insects. In these insects and in most crickets, cockroaches, stone flies and caddis flies, the hind wings are much larger than the forewings, whereas in wasps, aphids and many butterflies, the reverse is true. Thrips and some tiny beetles and parasitic wasps have wings that resemble feathers, consisting of a central rod surrounded by a flat sheet of closely packed bristles. In two families of moths the margin of each wing cuts deeply in between the veins, so that the body is flanked by a series of radiating plumes.

Beyond these characteristics lie differences in shape, relief, texture and particularly the pattern and density of the supporting veins. Together the features provide a cornucopia of traits for systematists, who have used them in classifying and identifying insects for more than 150 years. For paleontologists, wing features have proved quite indispensible; many fossil insects are known only by their wings, and in fossils in which the whole insect is preserved, wings are frequently the most clearly visible and interpretable structures. Until recently, however, the functional significance of most wing characteristics—and hence of evolutionary trends in wing design—remained almost completely unknown.

For example, the fossil of the forewing of a large cicadalike insect found in 130-million-year-old rocks in Spain clearly reveals shape and venation pattern [*see illustration on page 116*]. A prominent, curved line (the nodal line) crosses the wing at about one third the distance from the base. Modern cicadas have a similar structure that allows the wing to deflect downward—but why did cicadas develop this feature? In the fossil the textures of the membrane on either side of the line differ markedly: it is smooth toward the base and pitted toward the tip. Near the base the wing is torn along another line of bending commonly found in most insect species. Known as the claval furrow, it marks off a posterior lobe called the clavus, nearly detached in this fossil. What is the significance of these features?

Further questions arise when we compare this fossil with others. A representative wing found in Australian rocks 220 million years old comes from the same family the Spanish fossil belongs to and shows many similarities as well as profound differences. The older, Australian wing is nearly elliptical, the younger, Spanish one nearly triangular. The clavus is larger in the older wing, and the transverse line of flexion is more distal. Although a three-pronged fork of strong veins

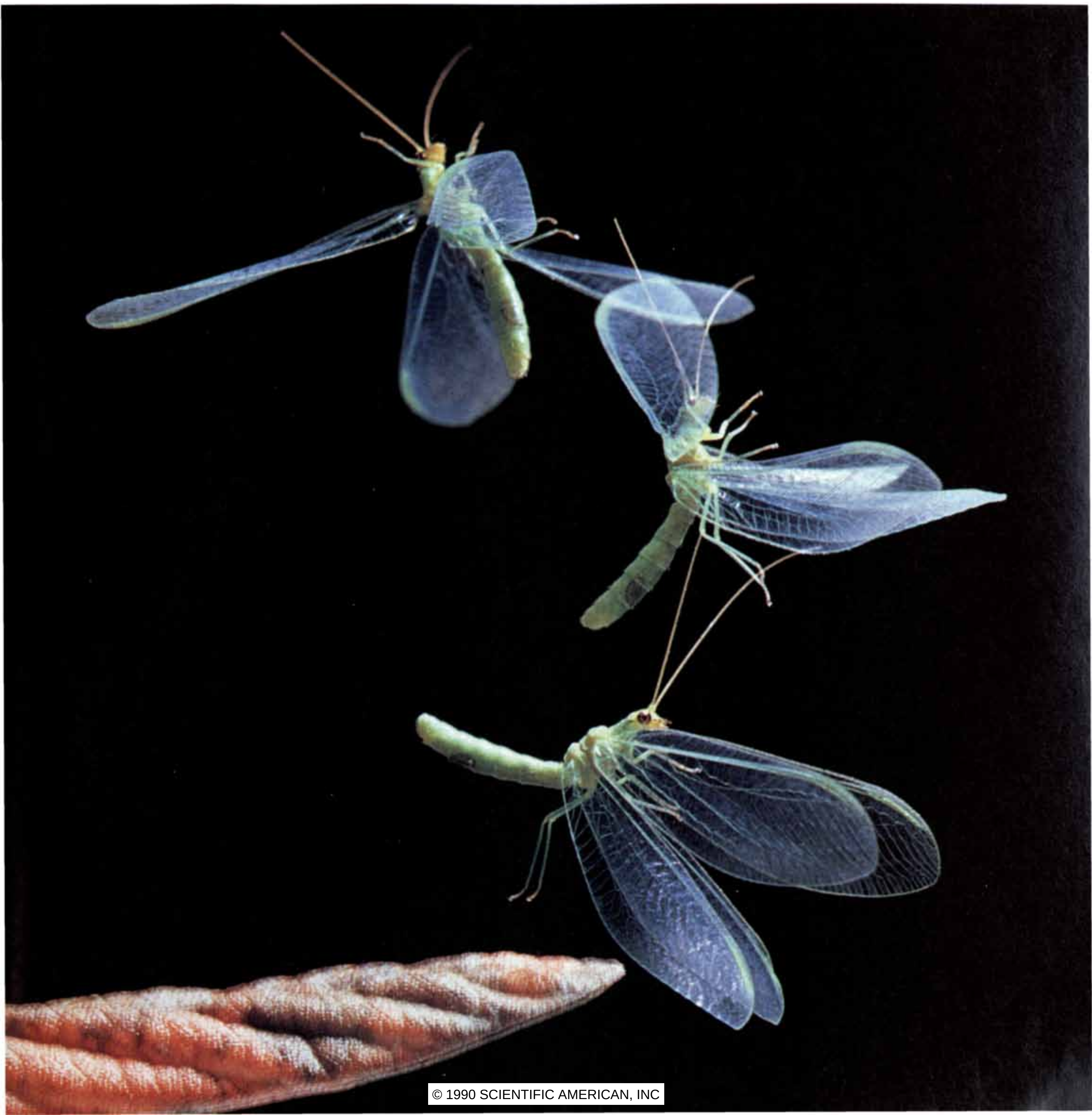
ROBIN J. WOOTTON began his career in insect paleontology, receiving a Ph.D. in this field from the University of London. He started work on insect flight in the 1970s, with the intention of understanding and explaining evolutionary trends in wing structure. Now senior lecturer in biological sciences at the University of Exeter, he leads a small but active research group investigating the influence of wing morphology and engineering on flight performance and behavior.

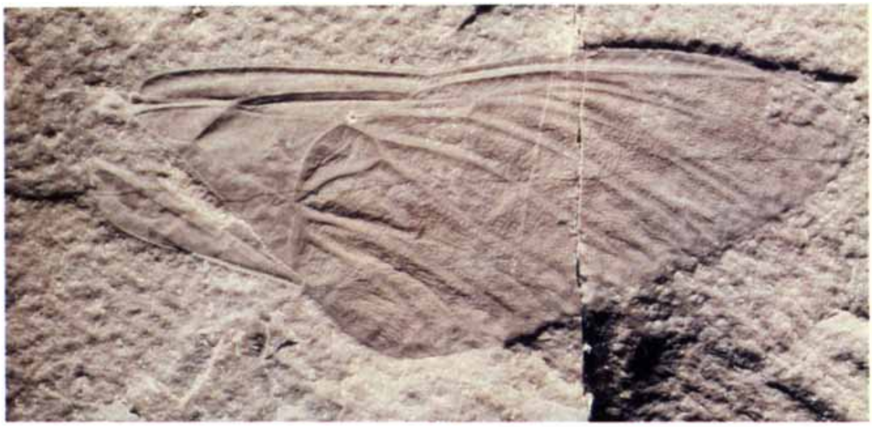
GREEN LACEWING (*Chrysopa carnea*) takes off from a leaf in this triple-flash photograph. The wings, which beat out of phase, show camber and torsion.

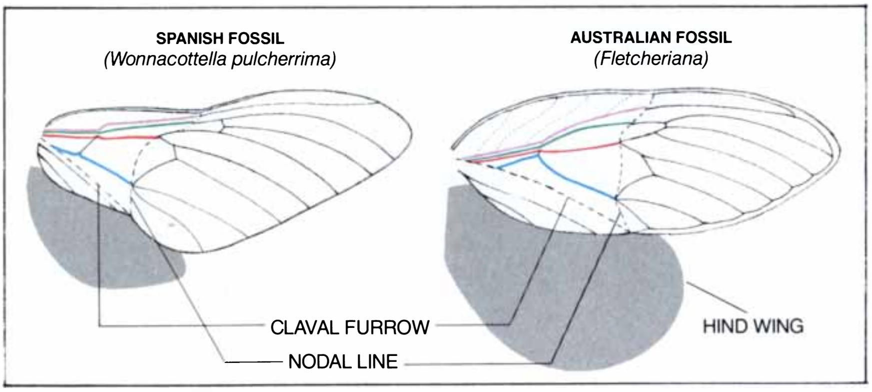
supports the base in both wings, the younger wing uses a different combination of veins.

What do these differences mean? The shift from ellipse to triangle probably accompanied a reduction of the hind wing. All the changes presumably reflect alterations in flight pattern and technique. Examining these fossils in the 1960s, I could recognize evolutionary trends in structure that seemed to have parallels in other insect groups, both living and extinct, yet it was impossible to explain any of them. I could do little more than compare the fossils with modern insects, mainly moths, which have similar wing shapes, and conclude tentatively that the ancient insects might have flown in similar ways—whatever these might have been. Little was known at the time about the flight of moths and nothing at all about the relation between their flight techniques and the architecture of their wings.

In the 1970s several developments, both technical and theoretical, at last made it possible seriously to attack the problems of wing design. New instruments, particularly the scanning electron microscope, allowed morphology to be studied in far greater detail than before. High-speed cinematography, combined with microcomputers, made feasible the filming and quantitative analysis of insects flying freely in laboratory enclosures. Such films, pioneered by Torkel Weis-Fogh, Gordon







FOSSIL FOREWING of a 130-million-year-old cicada from Spain shows many details, including supporting veins and a flexion line along which the membrane has fractured (*top*). The reconstruction (*bottom left*) reveals evolutionary changes, including different supporting veins (*color*), as compared with an older wing from Australia (*bottom right*). Probable shape of the hind wings is shown (*shaded area*).

Runnalls and Charles P. Ellington of the University of Cambridge, and the astonishingly clear high-speed stills of insects in flight taken by Stephen Dalton, a free-lance photographer in Sussex, provided for the first time detailed information on wing behavior in free flight. Furthermore, researchers could begin to interpret wing movements and changes in shape in aerodynamic terms, an area in which Weis-Fogh and Ellington were preeminent.

In more recent years the laboratory of Andrei K. Brodskiy of the University of Leningrad, my own at the University of Exeter and centers at other institutions have made the study of wing design a primary concern. Taking different but complementary approaches, we have examined the flight and functional wing morphology of many insects and discovered some general principles. Wing engineering is proving to be every bit as fascinating as we had hoped.

Orthodox insect wings consist of tough, flexible membrane supported by fairly rigid longitudinal, or radiating, veins—usually tubular and typically containing blood and air-filled tracheal tubes—linked together by cross veins. The details of the membrane ultrastructure remain poorly understood, but its material properties, in particular its strength and lightness, must be remarkable even for insect cuticle, an already superb natural composite. In some wings the membrane is only a micron or so thick, yet it can withstand strong lift forces generated during flight.

Despite the membrane's strength, the wings, which are seldom entirely flat, have additional details that serve to increase rigidity. In some groups—dragonflies, mayflies, grasshoppers—the wings are highly corrugated, with the longitudinal veins occupying peaks and troughs in the membrane [*see illustration on opposite page*]. With their abundant cross veins, the wings recall the structural engineer's lattice girders and space frames, in which bending forces on the whole structure are converted to pure tensile or compression forces in the individual members. These cross-connecting features provide far greater strength and rigidity per unit weight than structures that lack them. David J. S. Newman, a former graduate student in my laboratory, showed that some dragonfly wings do indeed exploit this physical effect, and we demonstrated that the membrane itself, flexible as it is, can help support the wing by acting as a "stressed skin." The membrane thus prevents the framework of veins from deforming, much as an artist's mounted canvas adds rigidity to the frame.

Other, flatter wings are essentially cantilevered beams. Their veins resist bending by aerodynamic and inertial forces; they often have a large cross section at the base and taper from the base to the tip—common features of cantilevered structures. But whether considered as beams or space frames, insect wings have two crucial—and for engineers, unusual—special characteristics.

First, the wings flap: they repeatedly accelerate and decelerate, gaining and losing kinetic energy. A tapered structure is hence even more important, because the wing's center of mass must be kept as near as possible to its base to minimize costly kinetic energy changes and to keep internal stresses within bounds. Tapering tends to make the wing tips less rigid, but because insects frequently collide with obstacles and other insects, the wings in general respond to impacts not by stiff opposition but by yielding and rapid recovery, like a reed in the wind. High-speed films by Newman show the corrugated wing tips of dragonflies flattening and bending as they collide with the walls of their enclosure and fully recovering undamaged before the next half stroke.

Second, and we have determined, vitally, insect wings can greatly deform during flight. Although the wings of birds and bats also deform, their wings, being modified arms, have muscles that allow them to bend at the elbow, wrist and hand. Insect muscle stops at the wing base; therefore, any active control of wing shape must be remote.

Such deformable airfoils rarely occur in technology. The most familiar examples are sails; indeed, a sail and an insect wing share much in common. Each consists of flexible membrane supported by rigid spars. In both, the extent to which aerodynamic forces mold the membrane is determined by the membrane's elasticity, the elastic forces in the spars and the controlling forces exerted by the insect (or crew) at the base. In wings, inertial forces that result from flapping must be considered. Boats do not propel themselves with actively flapping sails; if they did, the sails would need to be of a rather different design.

Detailed investigation reveals many wing features that reinforce the sail analogy. In every wing, one can recognize rigid zones clearly adapted to support the wing and to limit its deformations. These zones consist of stout veins or, frequently, groups of veins firmly linked together by substantial cross veins into three-dimensional spars with occasionally thickened membrane—the mast and boom of the wing.

Other areas seem to be adapted for deformability. The cross veins are flattened, narrowed, annulated (like the flexible zones of bendable drinking straws) or movably hinged to the longitudinal veins. Here the membrane, thin and flexible, sometimes appears under the electron microscope to be finely crumpled, giving local extensibility. These areas can be compared with the cloth of sails and the veins within them with the battens, which in sails (and insect wings) modify the profile locally. We have some indirect evidence that the structure and properties of the membrane may vary within the wing—a common feature in the multicloth sails of high-performance yachts.

But the comparison can be taken too far. Insect wings are far more subtly constructed than sails and distinctly more interesting. Many, for example, have lines of flexion across the wing, as already described in the fossil cicadas. They also incorporate shock absorbers, counterweights, ripstop mechanisms and many other simple but brilliantly effective devices, all of which increase the wing's aerodynamic effectiveness.

Insect wings have become adapted for the requirements of insect flight, which is different from those of other animals. Unlike many birds, insects make limited use of gliding; only by flapping can the insect generate the sustained forces to remain aloft. To create these forces, the insect needs to accelerate air—directly downward to hover or rise vertically; downward and backward to move forward. The forces vary each instant throughout the stroke cycle, but when averaged they must produce a net upward component that counteracts gravity.

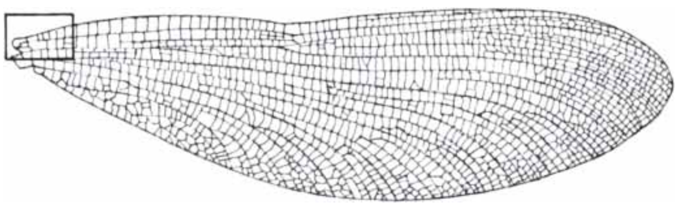
To achieve this, the insect must do more than simply raise and lower its wings repeatedly. Flapping the wings without twisting would propel the insect forward, but the vertical forces from the upstroke and downstroke would cancel out. The wings would do no more to support weight than they would in a glide. For sustained flight in still air, they must be moved or deformed during the flapping cycle, so that they generate, overall, more upward than downward force. The insect can effect such wing movement in a number of ways and tends to use several of them simultaneously.

First, the wings can twist, so that the angle of attack may vary throughout the stroke. Probably most insects actively twist their wings to some extent by using the muscles at the base. Inertial and aerodynamic forces can assist and, in some cases, entirely drive the twisting between half strokes. Aerodynamic forces, normally centered behind the torsional axis (the line around which the wing naturally rotates), tend to give the wing a helicoid twist as it moves through the air, much as a single propeller blade is curved to direct air flow. Again, the twisting effect of aerodynamic forces on airfoils is familiar to any dinghy sailor.

Second, the camber of the wings—the convex curve from leading to trailing edge—may change. In wind-tunnel tests, a small, cambered plate generates more lift than an uncambered one, an effect insects appear to exploit. Most high-speed films and many still photographs of flying insects show the wings to be cambered for parts of the stroke cycle.

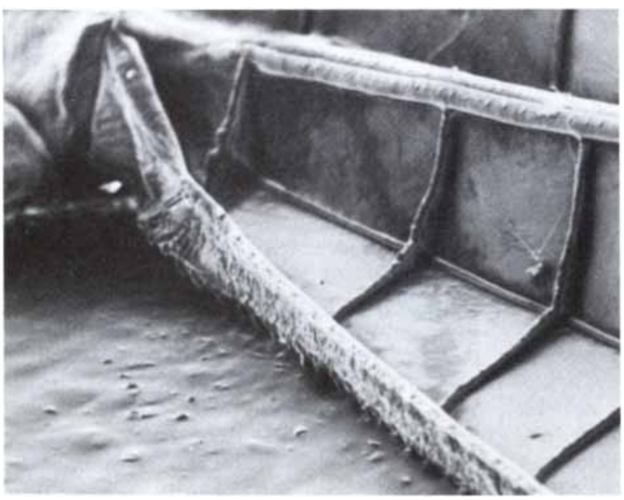
Third, insects may alter the area of the wings exposed to forces. We have evidence, for example, that some butterflies alter the total wing area by changing the degree their forewings and hind wings overlap. Locusts reduce the area of their hind wings in the upstroke by retracting them elastically to a partly folded position.

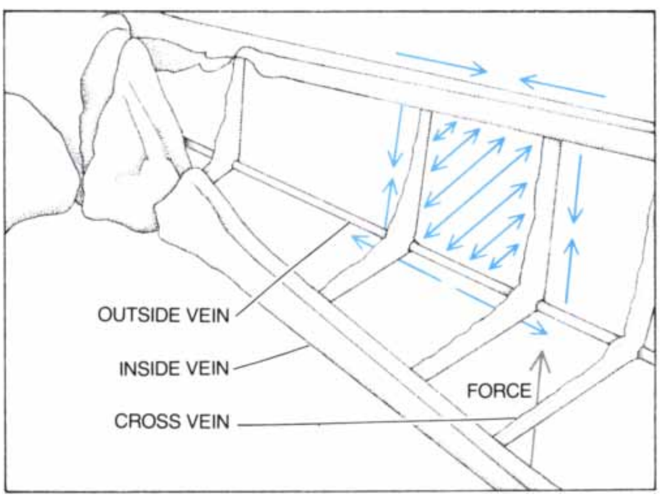
Fourth, asymmetry in force generation occurs when the wing moves through an oblique stroke plane, which in forward flight causes the wing to



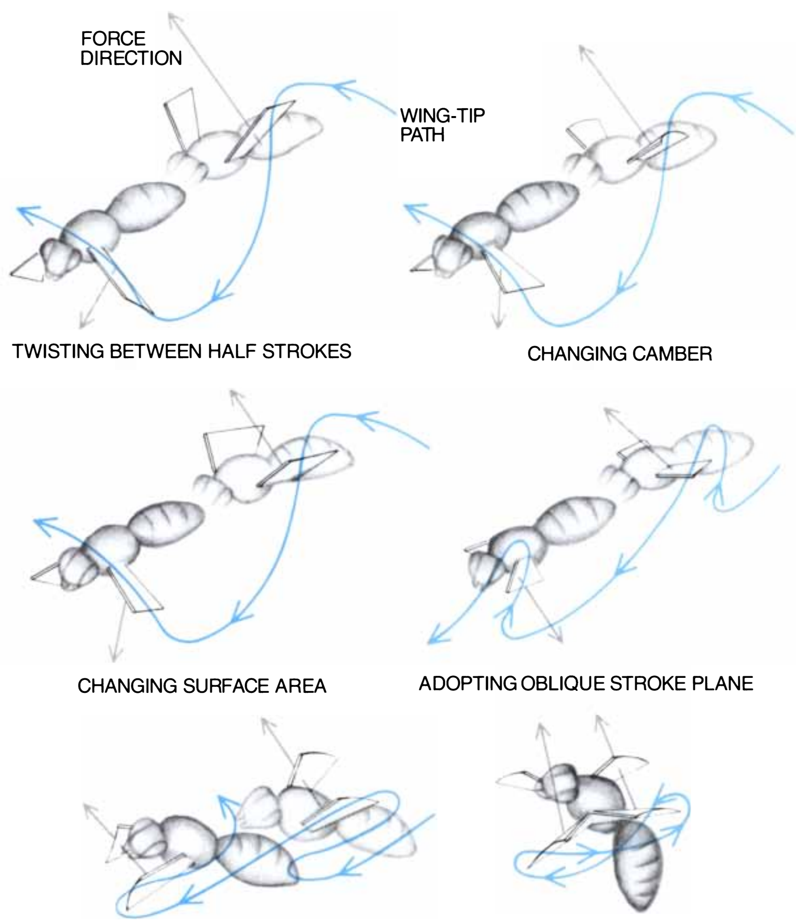
DAMSELFLY WING
(Calopteryx splendens)

THE OUTSIDE VEINS OF THE CORRUGATED WINGS REACT TO FORCE BY STRETCHING; INSIDE VEINS AND CROSS VEINS COMPRESS. THE MEMBRANE BETWEEN THEM IS SHEARED AND CONTRIBUTES TO RIGIDITY BY BRACING THE STRUCTURE DIAGONALLY.





HOW INSECTS GENERATE UPWARD FORCE



BY COMBINING TWISTING AND CAMBER CHANGE IN AN INCREASINGLY HORIZONTAL STROKE PLANE, INSECTS CAN GENERATE WEIGHT-SUPPORTING FORCES ON BOTH UP AND DOWN STROKES TO FLY SLOWLY AND HOVER.

move faster relative to the air on the downstroke than on the upstroke. Because the hinge between wing and thorax is itself oblique, an oblique stroke plane is built into most insects. In dragonflies the entire thorax tilts backward. But insects can also make the stroke plane more horizontal at will by tilting their bodies toward the vertical. Indeed, most truly versatile fliers, such as damselflies, bees, wasps and many moths, and the habitually slow-flying insects, such as aphids and gnats, tend to combine a nearly horizontal stroke with extensive twisting, so that the wing turns right over between the upstroke and downstroke. Its camber thus reverses, like a sail going about the end of a tack.

These techniques can all be explained by orthodox aerodynamics, but insects have other tricks up their sleeves. In 1973, working on the parasitic wasp *Encarsia formosa*, Weis-Fogh described a new mechanism for generating lift: the insect claps its wings together at the top of the upstroke and flings them apart, leading edge first. Air flows between each wing over the leading edge, creating a circulation of air around the wings as they move apart and, consequently, a high degree of lift at the outset of the stroke. Like the other insect flight mechanisms later identified, this one relies on unsteady flows created by wing movements that have no parallel in the behavior of man-made airfoils. Three other mechanisms seem to be particularly widespread and involve wing deformations, which are not necessarily involved in the clap and fling.

Ellington has described two: a "clap and peel" mechanism, in which the wings meet at the top of the stroke and peel apart progressively as they descend, and a "near clap and partial peel," in which the wings approach each other and only the posterior parts of the wings meet briefly and peel apart while the leading edges diverge. Researchers have not fully analyzed these mechanisms, but Ellington, Clive Betts of the University of Exeter, John Brackenbury of the University of Cambridge, Brodskiy and I have now identified one or the other in lacewings, butterflies, true bugs, crickets, locusts, mantises and stick insects. All these insects have broad, flexible wings, a characteristic required to "peel": the wings flex evenly along axes radiating from the wing base, creating what I call radial curvature.

The third mechanism, mentioned briefly by Weis-Fogh, has been discussed more fully by A. Roland Ennos, now at the University of York, by Brodskiy and by J. M. Zanker of the University of Tübingen and seems to occur in many different insects at the end of the downstroke. As the wing decelerates, it bends sharply downward along a transverse flexion line and then twists into the upstroke position while straightening rapidly. The outer part consequently accelerates sharply, probably generating a high degree of lift.

Together these mechanisms draw attention to and help explain a range of wing deformations visible in high-speed films of insects in flight: development, change and reversal of camber; torsion; area expansion and contraction; radial curvature; and transverse bending. The deformations can be discussed with reference to three broadly identifiable types of wings [*see illustration on page 120*]. These types are by no means the only kinds, nor are they entirely discrete, given that some insect wings have features of more than one. Nevertheless, they are all common, and each has probably evolved independently several times.

Type A wings typically have substantial anterior and posterior supporting zones—sometimes even throughout the base of the wing. The deformable zone lies between and distally to the supporting zones and is often crossed by a transverse line of flexion that allows the wing to bend easily under pressure applied only from above. It is in effect a one-way hinge. In flight, muscles at the wing base determine the relative positions of the anterior and posterior supporting spars and hence the area between them. In this way, the muscles actively control the orientation and to some extent the profile of the deformable zone throughout the stroke cycle.

Hans-Klaus Pfau of the University of Mainz has worked out the details in locusts, but similar principles seem to apply in the other type A wings that have been studied. Important components of the posterior supporting zone are the well-developed clavus and the claval furrow, features already mentioned in the fossil cicadas. The hingelike claval furrow permits the relative movement of the supporting zones described above. Type A wings include the forewings of crickets, scorpion flies, alderflies and dobsonflies, most bugs and some moths. The hind wings of some of these insects, such as scorpion flies and some moths, are also essentially but usually rather modified type A.

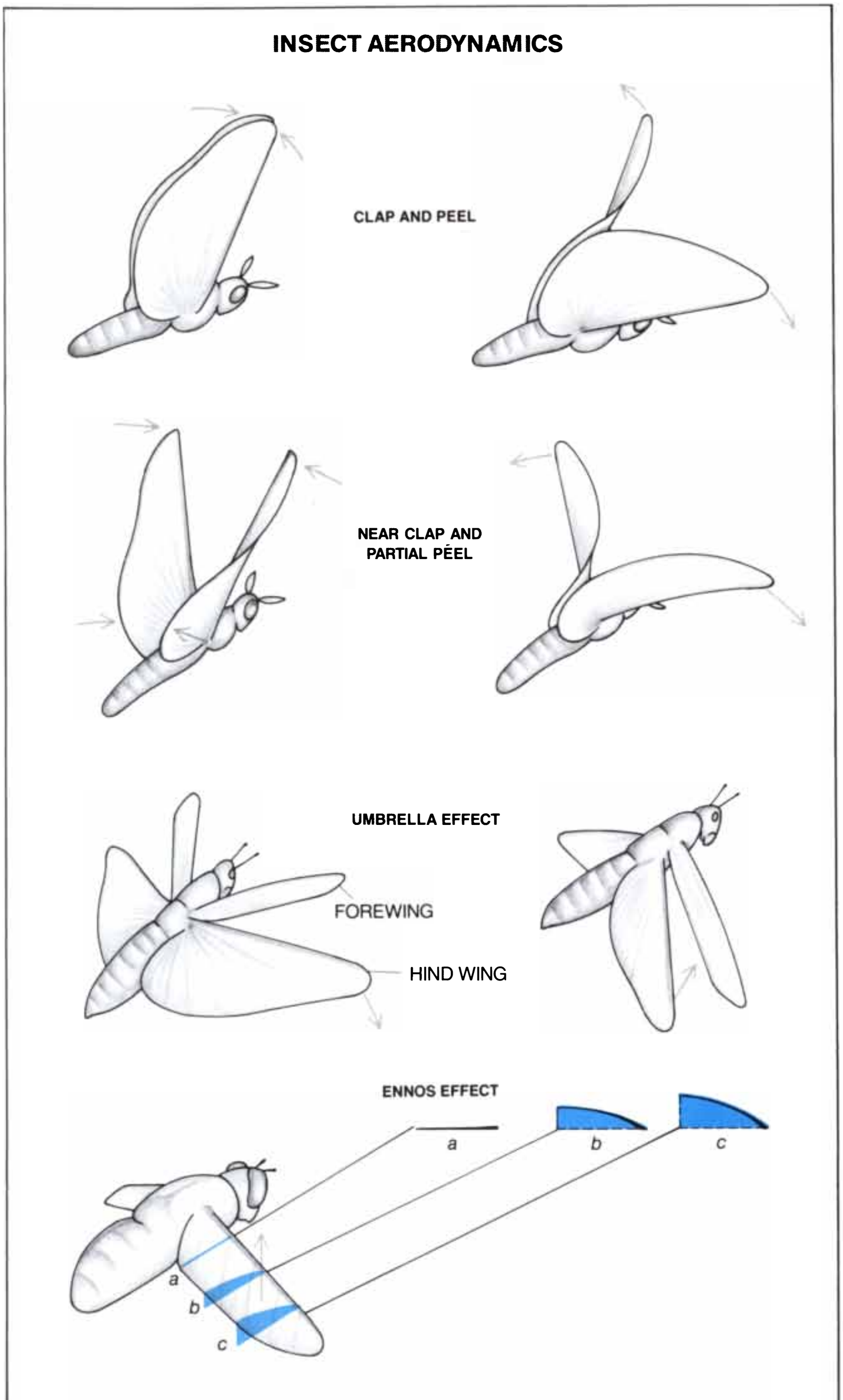
Type A wings have a major limitation: they do not twist very far. Slow flight and hovering usually require extensive twisting, and few insects with type A wings have particularly versatile flying abilities. The transverse flexion line, however, eases the problem somewhat. High-speed films show that the line often allows the outer part of the wing to twist far more than the basal part, making possible the generation of some weight-supporting force in the upstroke as well as in the downstroke. The larger the proportion of the wing beyond the flexion line is, the more twisting is likely to be possible, and the expansion of the distal, deformable area relative to the nondeformable area is a common evolutionary trend in type A wings.

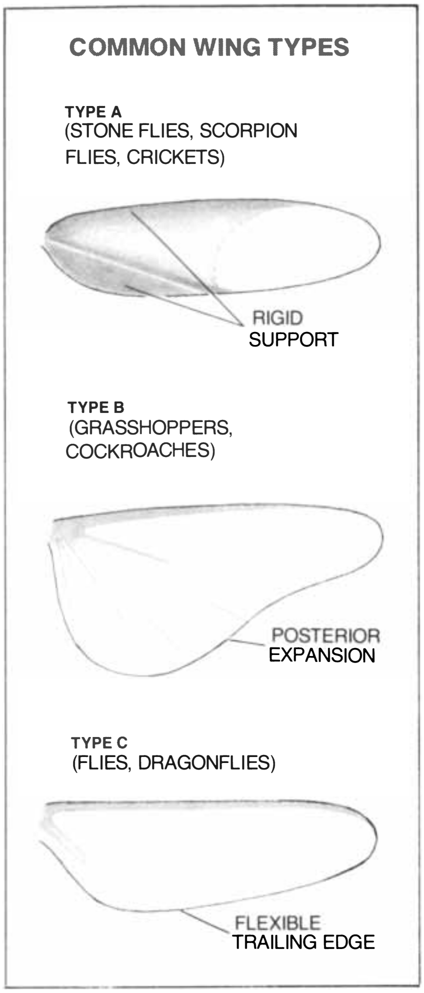
Type B wings are all hind wings. These wings seem to have evolved from type A; the area that is the clavus in type A wings has been modified into an enlarged, soft, flexible fan at the trailing edge and is supported internally by veins that radiate from the base. Like type A wings, type B ones have a well-supported leading edge. This wing type, capable of limited twisting, is generally found in insects with restricted flight repertoire, including crickets, stick insects, mantises, most stone flies, bugs and many caddis flies. The wings are, however, ideally suited to develop radial curvature in the clap and peel and in the near clap and partial peel mechanisms. Workers have shown that some crickets, locusts, mantises and bugs use these techniques extensively during the transition between upstroke and downstroke.

A further ingenious trick may prove widespread in type B wings, although so far I have confirmed it only in locusts. The hind wing fans of locusts are pleated and supported by a large number of radiating veins on the ridges and troughs of the pleats. At rest, the whole structure folds up like a pleated, man-made fan. During the downstroke the wing is pulled forward, and the pleats open as the fan extends. What happens as the wings continue to extend is essentially the same as what happens when one opens an umbrella: the outer edge becomes pulled slightly inward, compressing the veins, which become bent like the umbrella's spokes.

The umbrella effect leads to two results. First, the process supports the soft trailing edge and prevents it from fluttering up into the wind. Second, the wing becomes slightly cambered and hence a more effective airfoil. For the upstroke, the wing merely retracts partially, reducing its area, losing its camber and thereby bringing about the asymmetry of forces essential for sustained flight.

Type C wings, like type B ones, characteristically have a well-supported leading edge and a soft, flexible trailing edge. Unlike type B wings, type C wings are adapted to





twist extensively, so that they can generate weight-supporting force on both the downstroke and upstroke. Such twisting brings with it the ability to fly slowly and often to hover. Most of the truly versatile performers and habitually slow fliers have wings of this kind. Preeminent among insects with type C wings are the flies and the dragonflies, both of which have been studied extensively in my laboratory. Their wings have proved to be small masterpieces of ingenious design.

Were they simply flexible membranes supported only at the leading edge, type C wings would behave like flags, fluttering up to lie parallel to the airflow and generating no useful forces. Of course, they do nothing of the kind; lacking a rigid posterior support, they have evolved a battery of elegant internal mechanisms that operate automatically to maintain an effective camber and angle of attack in direct response to aerodynamic loading.

Some 16 years ago R. Åke Norberg of the University of Göteborg in Sweden first emphasized the importance of preventing a wing of this type from pitching up into the wind. He showed that the pterostigma, a fluid-filled, often pigmented cavity near the tip of the leading edge of many insects, acts in dragonflies (and probably in other insects) as a counterweight that opposes the pitching motion.

Ennos, working on flies, has identified several other such mechanisms. All of them rely on the fact that the leading-edge spar, although resistant to bending, can twist to some extent. From this spar, several veins run obliquely backward and toward the wing tip. Ennos has shown that the aerodynamic force on the wing, acting behind the torsional axis to create the usual propeller-bladelike shape, twists the leading-edge spar and raises the oblique veins, automatically generating a cambered profile. Furthermore, the oblique veins are themselves twisted along their axes, and if they are curved (as they are in many dragonflies), their tips deflect downward, thus lowering and supporting the trailing edge and further enhancing the camber. Other mechanisms identified by Ennos and by Newman involve simple lever systems within the dragonfly wing that lower the trailing edge when aerodynamic pressure raises the veins in the center of the wing. All these devices operate whether forces act on the wing ventrally, as on the downstroke, or dorsally, as when the wing twists around for the upstroke. Hence, they are ideal for insects that invert their wings and reverse their camber.

The wings of flies and dragonflies are uncoupled, but similar results are achieved in wasps, hawk moths and some bugs. These insects have a small, flexible hind wing coupled tightly in behind the reduced clavus of a modified type A forewing with an expanded distal area. Like type C wings, the composite airfoils formed in this way have rigid anterior support and a flexible trailing edge. They can twist extensively, usually with the help of some transverse flexion. The insects that bear them are for the most part versatile, skillful fliers.

What can now be concluded about the fossil cicadas that intrigued me as a graduate student? The relation between wing architecture and flight performance has become a good deal clearer. The earlier form from Australia is a fairly standard type A wing, with a well-developed transverse flexion line that would probably have given it some degree of upstroke twist and versatility in flight. The breadth of the wing, however, and the smaller but broad hind wing also recall butterflies and broad-winged moths, which use different techniques that are not yet fully understood.

Safer conclusions about the later Spanish insect can be drawn. The different textures on either side of the transverse flexion line distinguish between the supporting and deformable zones. The shift of the main wing support to more anterior veins is associated with the move to a triangular shape. This move, coupled with the reduced clavus, the more basal and more oblique flexion line, and the small hind wing, clearly parallels the evolution and appearance of wings in many wasps, hawk moths, aphids and mayflies. All these insects make extensive use of slow, near-hovering flight by beating their coupled wings in a horizontal plane with much twisting. The Spanish cicada was highly adapted for a flight pattern quite different from any of which the Australian insect would have been capable.

The better we understand the functioning of insect wings, the more subtle and beautiful their designs appear. Earlier comparisons with sails now seem quite inadequate. The wings emerge as a family of flexible airfoils that are in a sense intermediate between structures and mechanisms, as these terms are understood by engineers. Structures are traditionally designed to deform as little as possible; mechanisms are designed to move component parts in predictable ways. Insect wings combine both in one, using components with a wide range of elastic properties, elegantly assembled to allow appropriate deformations in response to appropriate forces and to make the best possible use of the air. They have few if any technological parallels—yet.

FURTHER READING

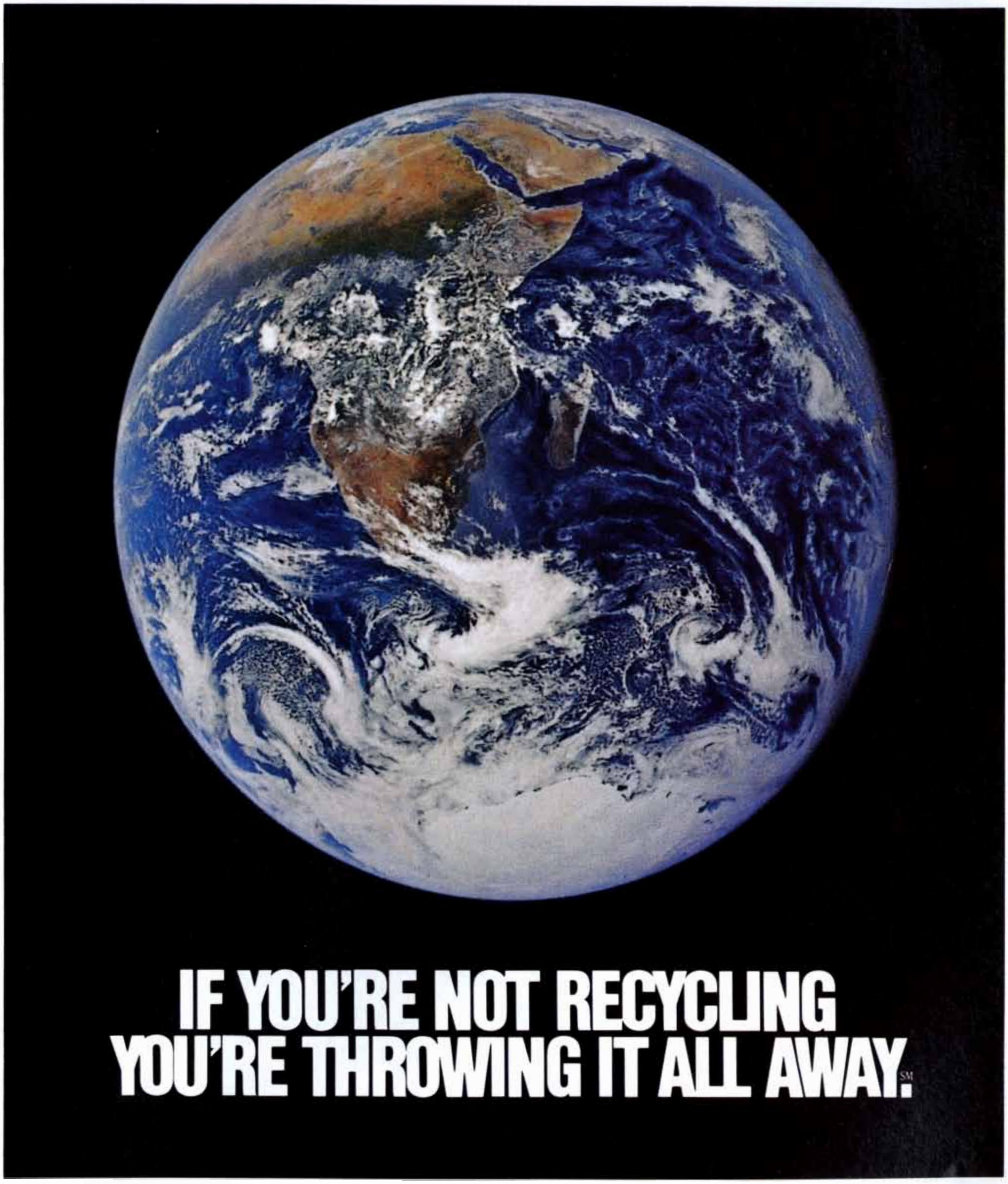
SUPPORT AND DEFORMABILITY IN INSECT WINGS. R. J. Wootton in *Journal of Zoology*, Vol. 193, Part 4, pages 447–468; 1981.

FUNCTIONAL ASSESSMENT OF WING STRUCTURE IN INSECTS. A. K. Brodskiy and V. D. Ivanov in *Entomological Reviews*, Vol. 62, Part 1, pages 35–52; 1983.

THE AERODYNAMICS OF HOVERING INSECT FLIGHT. C. P. Ellington in *Philosophical Transactions of the Royal Society of London, B*, Vol. 305, No. 1122, pages 1–181; 1984.

AN APPROACH TO THE MECHANICS OF PLEATING IN DRAGONFLY WINGS. D. J. S. Newman and R. J. Wootton in *Journal of Experimental Biology*, Vol. 125, pages 361–372; 1986.

THE IMPORTANCE OF TORSION IN THE DESIGN OF INSECT WINGS. A. Roland Ennos in *Journal of Experimental Biology*, Vol. 140, pages 137–160; 1988.



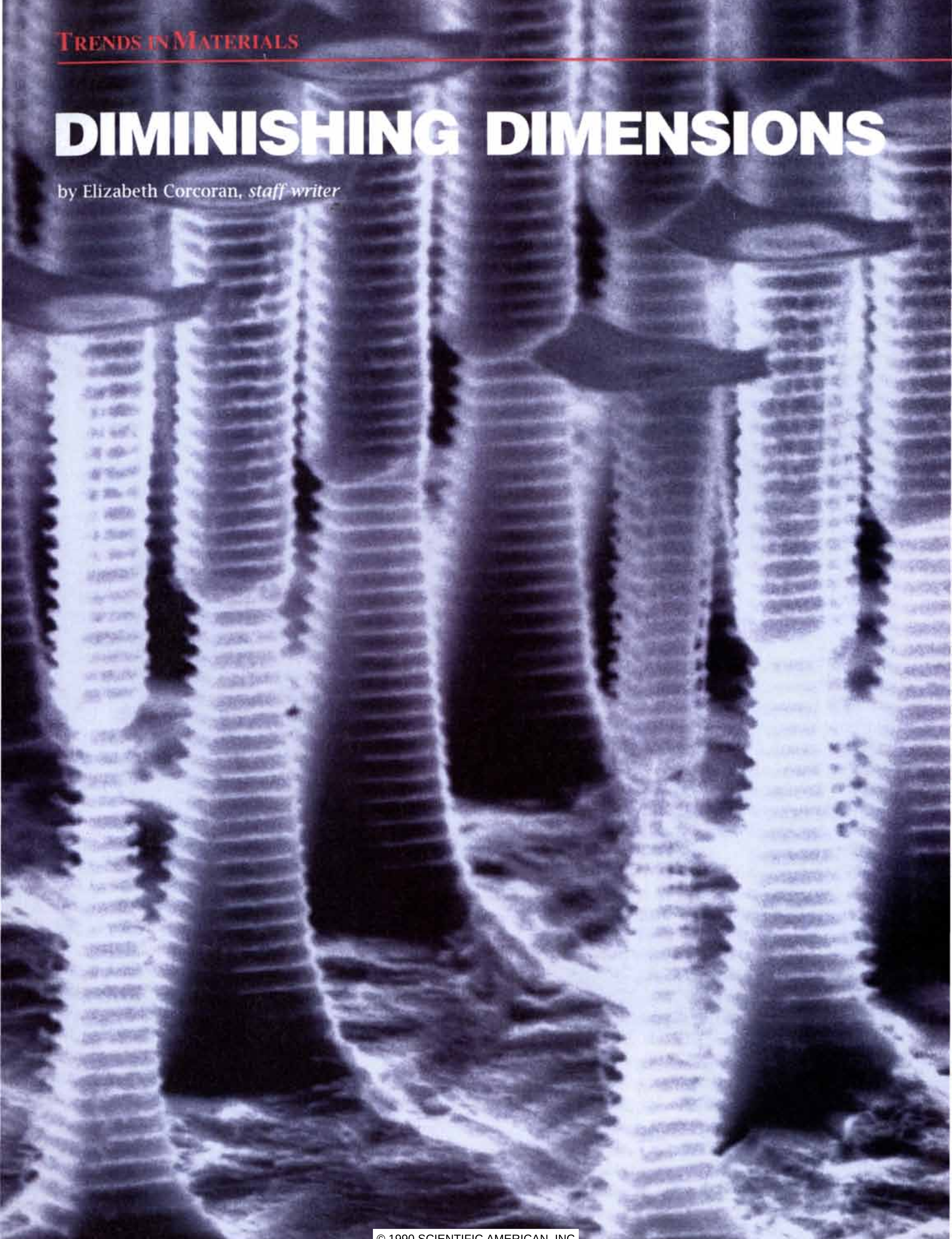
A little reminder from the Environmental Defense Fund that if you're not recycling, you're throwing away a lot more than just your trash.

You and your community can recycle. Please write the Environmental Defense Fund at: EDF-Recycling, 257 Park Avenue South, New York, NY 10010, for a free brochure that will tell you virtually everything you need to know about recycling.

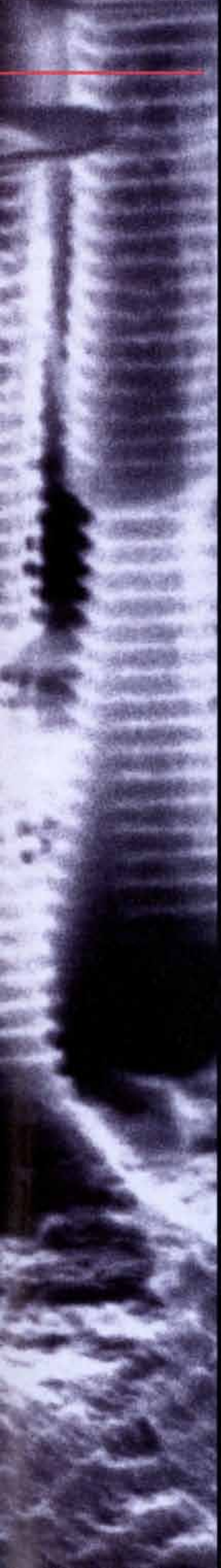
EDF™ Ad Council

DIMINISHING DIMENSIONS

by Elizabeth Corcoran, *staff writer*



From slivers of material that confine electrons in fewer than three dimensions may arise the next generation of electronic and optical technologies.



TINY SURFACE-EMITTING LASERS, measuring six microns high and one micron in diameter at the base, are each composed of more than 30 aluminum arsenide and gallium arsenide mirror pairs. Light is emitted and amplified by a quantum well located in the middle of the stack. The mirrors trap light and cause multiple reflections to occur within the stack.

The lasers are made by carving through a layered semiconductor wafer using a chemically assisted ion-beam etching technique. Before the devices shown could be tested, the structures had to undergo more processing, including the addition of thin metal contacts on top of each laser. Such microlasers, made at Bell Communications Research, can emit several milliwatts of infrared light through the transparent substrate at their base.

It's late on a Friday evening in Red Bank, N.J., and most of the corridors in Bell Communications Research (Bellcore) have fallen quiet. In one small laboratory, a lone young woman studies a set of graphs glowing fluorescent green on her computer screen. Behind her, a blue-green ribbon of light from an argon laser zigzags through a gauntlet of lenses on a table, then disappears through a porthole into a stainless steel cylinder fed by a supply of liquid helium from a steaming, ice-coated pipe. There, inside a chamber cooled below five kelvins, the beam runs squarely into its destination: a microscopic speck centered in a flat, black square that measures barely a few millimeters itself. Excited by the blue-green laser, the speck spits out its response—a brilliant mote of red.

This play of light intensely interests Maria Brasil, even if she has her back to it. How much light does the speck emit? At what wavelength? The answers lie in the slow accumulation of data plotted on Brasil's computer screen. What she sees there will cheer her colleagues. The speck emits light at the precise frequency predicted by its builders.

Bellcore's tiny red-light emitter, smaller than a grain of sand, is a harbinger of a new age of electronic and optical materials, namely, that of "quantum" devices. Over the past few years the technology for manipulating and observing matter on an atomic scale has advanced at an astonishing pace. This past April researchers at the IBM Almaden Research Center in San Jose, Calif., even spelled out their company's name with xenon atoms by using a scanning tunneling microscope.

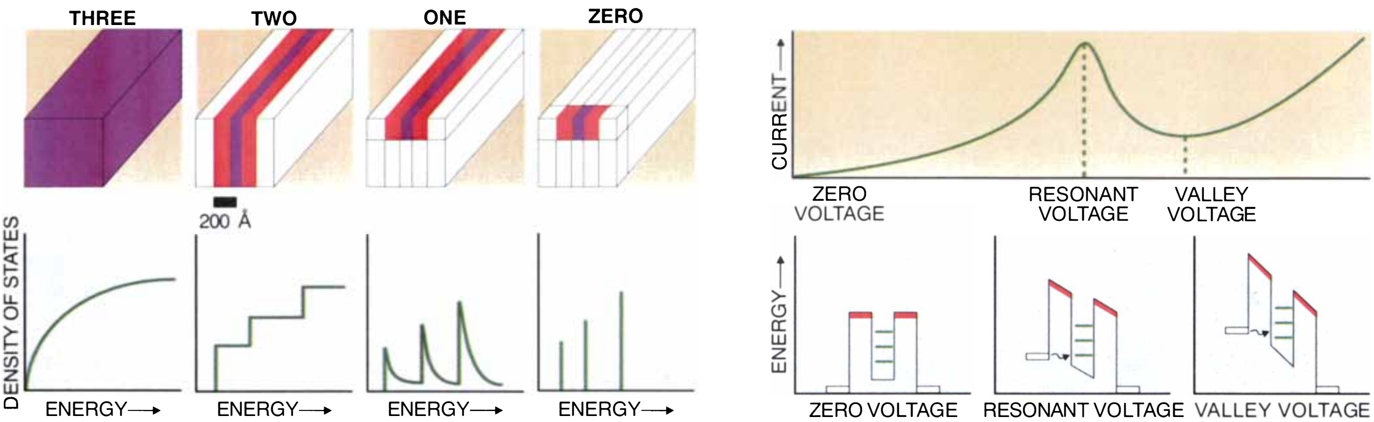
Although moving individual atoms is still a laboratory game, laying down exquisitely thin films of atoms has become serious business. By controlling precisely the structure and composition of layers of materials only an atom or two thick, scientists are proving they can program the electronic characteristics they want into a compound. "It's like having your hands on the knobs of nature," says Mark A. Reed, a professor of electrical engineering at Yale University.

The red speck at Bellcore, for example, is a complex tower made of slices of zinc selenide and zinc telluride semiconductors. Each layer is no more than 20 angstroms thick—several hun-

DIMINISHING DIMENSIONS

The dimensionality of a material can be reduced by sandwiching it between two layers of another material that has higher-energy electrons. This confinement changes the density of electron states, or specific energy levels, that will be filled by incoming electrons (*left*).

The current conducted by a quantum well device, shown by green energy levels (*right*), peaks when the energy level of the incoming electrons matches, or is in resonance with, an energy level of the quantum well. At higher or lower voltages, little current leaks through the device.



dred thousandths the thickness of a hair. Because these slivers are thinner than the characteristic wavelength of an electron, the particles are trapped.

Confined in such layers, electrons lose a degree of freedom. From an electron's vantage, the thin planes in Bellcore's light emitter have only two dimensions. Scientists can reduce a material's dimensionality further, too. A narrow strip sliced from one of the planes would be a one-dimensional wire. Dicing up a one-dimensional wire would be the ultimate reduction. This step yields zero-dimensional dots.

Reducing the number of dimensions of a material forces electrons to different energy states. Creating materials in which electrons demonstrate these unusual energy states is the heart of research on quantum devices. By controlling the physical size of a structure, researchers can induce predictable changes in electron energy. In this way, scientists can literally pick, or tune, the electronic properties they want. The fewer the dimensions, the finer the tuning. A zero-dimensional, or quantum, dot is considered analogous to custom-designing atoms.

This modern alchemy opens the way for fundamentally different electronic and optical devices. Instead of simple binary, or "off-on," electronic switches now used in computers, workers hope to fabricate multiple-switch devices and arrange them to function in parallel. These could lead to more powerful forms of computer logic and become the building blocks of dramatically smaller and faster integrated circuits. Some researchers even talk enthusiastically of "a supercomputer on a chip."

Another goal is highly efficient, tiny lasers that could convey far more data along optical fiber networks than is now possible. Such fire hoses of data could pump enormous amounts of video, computer and telecommunications services and data directly into businesses and homes. And the marriage of such electronic devices and lasers might bring into being a long cherished dream: optoelectronics circuits, which integrate electrons and photons, thereby spawning faster, more powerful components for computers and telecommunications networks.

Bellcore is far from alone in its quest for quantum devices. The U.S. National Science Foundation funds a center at the University of California at Santa Barbara devoted to quantum confinement. IBM and AT&T Bell Laboratories, among others, have made pathbreaking contributions. Investigators in Germany, England, France and, of course, Japan are also making great strides in this area. The key issue all researchers face: how to fashion these novel materials—quantum wells, wires and dots—into working devices.

Dicing Up Dimensions

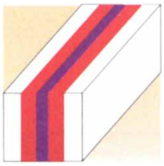
On the floor of Leroy L. Chang's office at the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y., is a matrix of neatly stacked manila files—vaguely reminiscent of the stacked layers of atoms that he and colleagues Leo Esaki and Raphael Tsu began toying with in the late 1960s. Their goal: to build structures that would trap electrons in dimensionally limited environments. "Confine an electron in two dimensions," Chang declares, "and it changes everything."

Materials that naturally confine electrons have long intrigued scientists. Graphite, an organic material that conducts current, consists of stacked two-dimensional sheets of carbon atoms. High-temperature ceramic superconductors, such as lanthanum copper oxide, have two-dimensional planes of copper and oxygen atoms interspaced with planes of other atoms [see "Superconductors beyond 1-2-3," by Robert J. Cava; SCIENTIFIC AMERICAN, August].

To build a lower-dimensional material deliberately, researchers must pay court to quantum mechanics. In any three-dimensional, bulk semiconductor, electrons take on a continuous range of different energy states when additional energy is added to the material by applying voltage. As a result, researchers cannot tap a specific energy level; they must accept what they get.

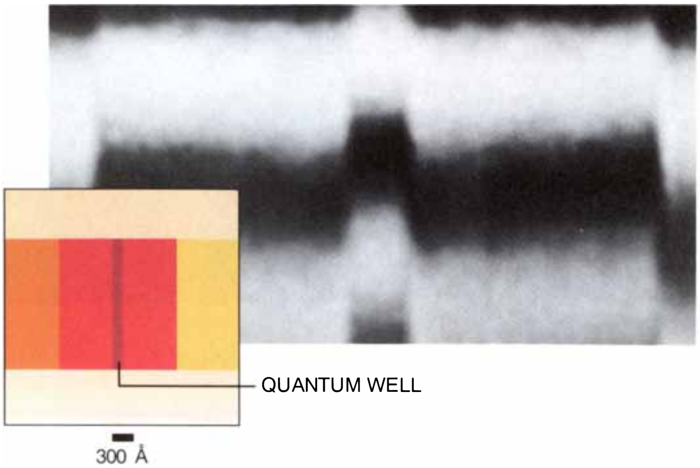
Quantum-mechanical theory portrays electrons as both particles and waves. Squeezing one side of a three-dimensional cube until it is no thicker than an electron wave is long traps electrons in a two-dimensional plane. In two dimensions, the so-called density of electron states—the energy levels electrons can occupy—becomes quantized. Electrons jump from one energy level to another in a steplike fashion. By studying what layer thickness induces what energy level, researchers can design the precise electronic characteristics of a material.

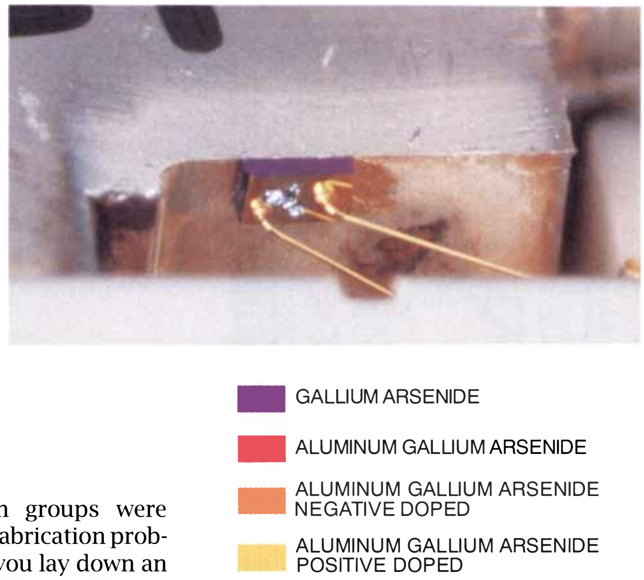
Electrons are not really confined by physical barriers, however; instead researchers must erect barriers of energy. Like water going downhill, electrons tend toward low-energy areas. So to trap electrons, investigators need only sandwich a material filled with low-en-



TWO DIMENSIONS: QUANTUM WELL

Incorporating a single quantum well (*left*) into this laser diode (*right*), built at Spectra Diode Laboratories, improves the efficiency of the laser. The outer doped layers inject electrons (negative charges) and holes (positive charges) into the quantum well, which then recombine and emit photons, or light. Because the quantum well contains less aluminum than the surrounding layers, it is a lower band-gap material.





ergy electrons between two slices of materials with higher-energy electrons.

The energy of electrons is described by band theory. The valence, or outermost, electrons of every atom fall into specific energy bands. Metal atoms have only a few valence electrons at low energy. Because they readily swap these electrons with one another, metals make good conductors.

Semiconductors and insulators, on the other hand, have more valence electrons and are not inclined to conduct current. Adding a small amount of energy to a semiconductor boosts some valence electrons to a higher-energy level, the so-called conduction band, and enables the material to carry current. The energy needed to propel an electron from one band to another is the band-gap energy—the difference between the valence and conduction band energy levels. Because every semiconductor demands a slightly different amount of added energy to trigger conduction, some semiconductors have higher or lower band-gap energies than others. Insulators, which require tremendous energy to push their valence electrons to the higher-energy bands, have the largest band gaps.

With these ideas in mind, researchers at both IBM and AT&T tried proving they could confine electrons. Esaki, Tsu and Chang began by alternating multiple layers of gallium arsenide with layers of aluminum gallium arsenide, a higher band-gap compound. AT&T workers had a simpler aim: to sandwich one thin, low band-gap material between two higher band-gap materials, thus producing a quantum well.

The efforts of both groups were plagued by swarms of fabrication problems. For one, how do you lay down an even layer of material only a few atoms deep? "We had to build a vacuum system ourselves" to deposit thin layers, Chang recalls sighing. Equally troublesome was controlling substrate contamination and the alignment of the crystal lattices of the materials.

In 1974, however, the researchers triumphed. The IBM workers demonstrated materials that had the predicted, steplike energy levels indicative of quantum confinement. Raymond Dingle and his colleagues at Bell Laboratories in Murray Hill, N.J., built a single quantum well, shone laser light on it and found that the novel material absorbed different, but predicted, frequencies of light—an alternative indication of quantum confinement. Soon thereafter, Esaki and Chang built the first real quantum well device—a resonant tunneling diode.

Ballistic Transistors

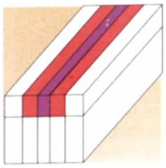
Many others jumped on a dimensionally related bandwagon: two-dimensional electron gases (2DEGs), in which electrons are trapped at the horizontal interface between a low band-gap layer of material and a layer that is doped with extra charges. With these materials, researchers explored complex phenomena, namely, the quantized and fractional Hall effects [see "The Quantized Hall Effect," SCIENTIFIC AMERICAN, April, 1986]. In carefully constructed 2DEGs, electrons can travel for relatively long distances before colliding with an atom or a defect, thus enabling engineers to build "ballistic" or "high electron mobility" transistors.

Such 2DEGs are the bedrock for the fastest transistors. (They only achieve the highest speeds at temperatures of a few kelvins.) Unlike quantum wells, in which electrons push through several, vertically arranged layers, electrons in 2DEGs must stay within the horizontal plane. That construction restricts the kinds of devices that can be built.

Researchers returned to quantum wells and superlattices, which are specially constructed multiple quantum wells, in the early 1980s, this time armed with highly precise fabrication tools. Among the biggest guns: molecular-beam epitaxy (MBE) machines.

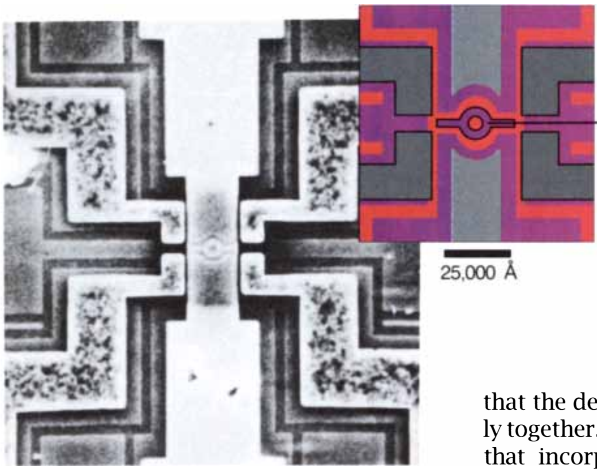
At the heart of a state-of-the-art MBE is an ultrahigh vacuum chamber, which allows workers to deposit layers of atoms as thin as two angstroms on a semiconductor substrate. Attached to the vacuum chamber, like spokes on a hub, are three or four passages that lead to effusion cells. Elements such as gallium or aluminum are vaporized in these cells, then shot down the passages toward the substrate. By programming the shutters between the passages and the vacuum chamber, scientists can dictate the thickness of the deposited layers. All told, a full day must be spent to grow about eight microns of semiconductor layers on a single gallium arsenide wafer.

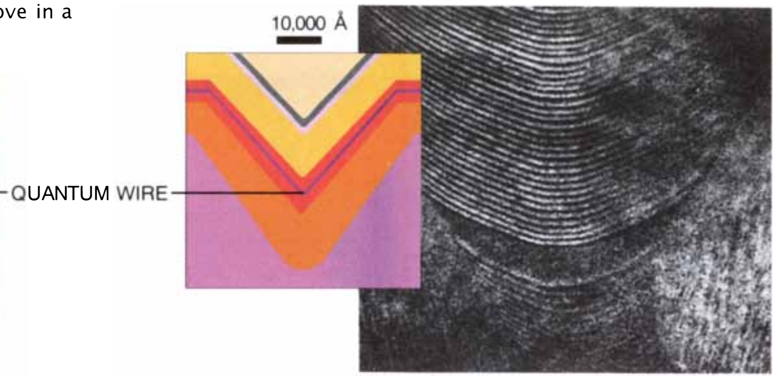
Compared with the next step, how-



ONE DIMENSION: QUANTUM WIRE

A transparent layer of metal covers the quantum wire loop built at IBM, which is used to make electrical measurements in a magnetic field (*left*). Quantum wires fabricated at Bellcore are formed in the crescent at the bottom of the **V** groove in a compound semiconductor (*right*).





ever, growing the layers is straightforward. To make use of the quantum wells, researchers try to carve or pattern devices—such as lasers—from the stacked layers. (Semiconductors that emit light when energized with light or electricity can be used as lasers.)

For this purpose, Bellcore researcher Edward M. Clausen, Jr., has customized a chemically assisted ion-beam etching machine. Using a lithographic process, Clausen prints a pattern on the wafer. Then he places it in the etcher filled with chlorine gas to speed etching and bombards the wafer with ions, or charged atoms. In theory, the ions blast through any unmasked material, leaving the protected area intact. In practice, the ions can also irrevocably damage the wafer.

Clausen and his colleagues have spent countless hours trying to carve microscopic lasers from multiple quantum wells. Still, the efforts are paying off. Earlier this year a joint Bellcore and AT&T team captured newspaper headlines after etching almost two million lasers in an area slightly smaller than a square centimeter. Unlike conventional semiconductor lasers that emit light from an edge, these lasers shine from either the top or bottom surface.

Incorporating a quantum well into a laser brings one major advantage: such a device can put out light very efficiently, powered by much less current than conventional lasers require. As a result, quantum lasers radiate far less excess heat. This feature, combined with the small physical size of the lasers, means that the devices can be packaged tightly together. In this way, surface emitters that incorporate quantum wells may become the building blocks of optoelectronics devices, amalgamations of lasers and transistors working together on the same chip or circuit board.

Connie Chang-Hasnain, a Bellcore researcher, is exploring other ways to exploit such lasers in communications. On a small video monitor above her laboratory table is an unwavering picture of columns of tiny quantum well lasers. Each is surrounded by a metallic contact pad. Spindly metal probes, like overgrown mosquitoes, rest on three of the lasers' contact pads. Chang-Hasnain cautiously pumps a few milliamps of current into the probes. "It's easy to blow them out," she explains. The result shows up on another computer screen: three distinct peaks of light.

Chang-Hasnain's data indicate that each of the lasers is emitting a different frequency of light. By controlling carefully the rotation of the semiconductor substrate during MBE growth, she and her colleagues have grown a microscopic forest of lasers. Because each device has layers of slightly different thicknesses, the lasers are unique. The best result so far, Chang-Hasnain believes, is an array of 77 lasers—each capable of transmitting data on its own frequency. Tying such a bevy of lasers to a fiber-optic line "would be like adding thousands more channels to your cable television," she says.

Still, the marriage of optical and electronic components will be far from trouble free, cautions Paul L. Gourley, an investigator at Sandia National Laboratories in Albuquerque. Gourley, who is also building arrays of surface-emitting lasers, agrees the devices show the potential for high efficiencies. But coupling tiny lasers to optical fibers requires enormous precision, he points out. In addition, because light waves remain in phase—or have a longer coherence length—than do waves of electrons, building integrated optoelectronics components may prove extremely challenging, he says.

Practical Dimensions

Others continue to worry about the grittier problems of making the lasers. "Until a year ago it was so damn hard to make them," says Ann C. Von Lehmen, another member of the Bellcore team. Turning these laboratory prototypes into dependable, mass-produced devices will demand painstaking work.

Case in point: after more than 15 years of research, few quantum well lasers have reached the marketplace. One is a single quantum well, edge-emitting laser built by Spectra Diode Laboratories in San Jose, Calif. The 100-milliwatt device can convert as much as 60 percent of the electricity fed into it to light—twice as much as comparable semiconductor lasers. (Arrays of such lasers emit more light but have lower overall efficiencies.) Nevertheless, such lasers have to be built one at a time.

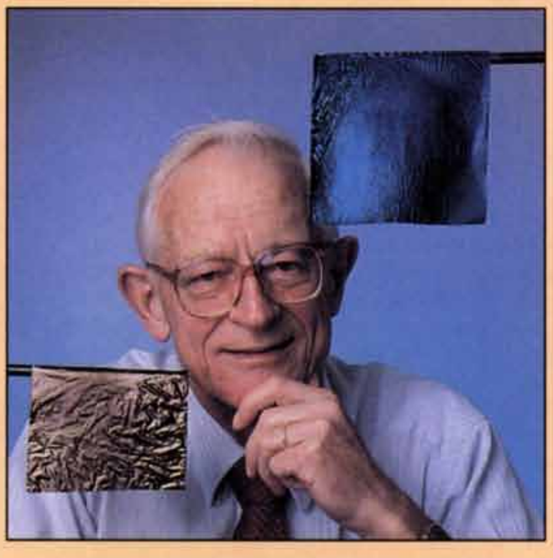
Like Clausen at Bellcore, John N. Ran-

POLYMERS THAT CONDUCT IN ONE DIMENSION

It was 1975, and Alan G. MacDiarmid, a chemist at the University of Pennsylvania, had just given a guest lecture at the Tokyo Institute of Technology. His hosts treated him to a cup of green tea, he recalls, then took him on a tour. In the laboratory of Hideki Shirakawa, the New Zealander saw a silvery film that delighted him. "I had never seen a silver polymer," MacDiarmid says. "It looked pretty. So I invited Shirakawa to work at our laboratory."

Polymer chemists had not gone looking for one-dimensional electronic systems, but they found them in Shirakawa's polymer films. The materials turned out to be a novel form of polyacetylene, a simple polymer that has a backbone of carbon atoms with a hydrogen atom attached to each carbon. Like most organic compounds, polyacetylene was an insulator: its atoms, which are linked by single covalent bonds, have their full share of valence electrons, leaving no room for additional electrons to carry current. But in synthesizing the material, Shirakawa had accidentally created a polymeric film with a useful bond structure between atoms—namely, alternately double and single bonds—and so forged a path for electrons.

Working together, MacDiarmid, Shirakawa and Alan J. Heeger (also then at the University of Pennsylvania) soon tried doping the polymer films—adding or removing electrons. In doing so, they created energy bands between the valence and conduction bands. As a result, the silver films took on golden hues and became conductors.



ALAN G. MACDIARMID was an early developer of conducting polymers. Doped polyaniline (right) *will conduct; the undoped sample* (left) *will not.*

Since then, researchers have found families of other conducting polymers, such as polyaniline, which includes nitrogen atoms in the backbone chain, and polythiophene, which has sulfur [see "Plastics That Conduct Electricity," by Alan G. MacDiarmid and Richard B. Kaner; SCIENTIFIC AMERICAN, February, 1988].

MICROSCOPIC TWEEZERS. The physical characteristics of these polymers—namely, that they are exceptionally lightweight conductors—provoked early excitement. Lightweight aircraft skins and automobile batteries were among the first proposals. Since adding and subtracting electrons from conducting polymers changes them from conductors into insulators, rechargeable batteries have been built from the materials.

Doping and undoping also changes the length of conducting polymers by as much as 10 percent. As a result, Raymond Baughman, a scientist at Allied-Signal, envisions using conducting polymers as an efficient means for converting electrical energy into mechanical energy. He and his co-workers have designed microactuators, such as microscopic tweezers for plucking micron-size objects.

Exploiting the electronic possibilities created by the polymers' quasi one-dimensional structure, however, has taken longer. Investigators are still puzzling out just how polymers carry charge and how to better that conductivity. What they do know is that electrons travel along a polymer chain—a one-dimensional wire—until they run into a defect or a break in the chain. Because electrons act like waves, they bounce off the barrier, doubling back along the path they came. If the electrons continue to bounce back and forth between two defects, they will create a standing wave. This action can pin the charge to one area of the polymer, reducing the overall conductivity of the polymers.

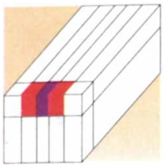
CHAIN HOPPING. Electrons can escape this trap—and dramatically improve the conductivity of the material—if they "hop" to a nearby polymer chain before establishing a standing wave. (This sideways hopping is what makes polymers quasi one-dimensional systems.) Electron hopping could be much like the tunneling that electrons do when confined in semiconductors, Baughman points out. The overall conductivity of a polymer will depend on how far electrons travel before running into defects, as well as how easily they hop from chain to chain. Both factors are related to the concentration of defects along a polymer chain.

Some polymers exhibit interesting optical characteristics. The polymers can transmit light from a laser. And because of the way the polymers share electrons, increasing the intensity of light will tune the materials' optical properties.

LIGHT SWITCHES. Some workers hope to use this trick to make all-optical switches, which use light to control light. Shahab Etemad of Bellcore is developing one such polymeric switch based on two identical threads of polydiacetylene that run parallel to each other in a block of glass. The lines, which are no thicker than the wavelength of light that they will transmit, are coupled waveguides. As a result, shining low-intensity laser light along one line causes the light to oscillate back and forth between the two lines.

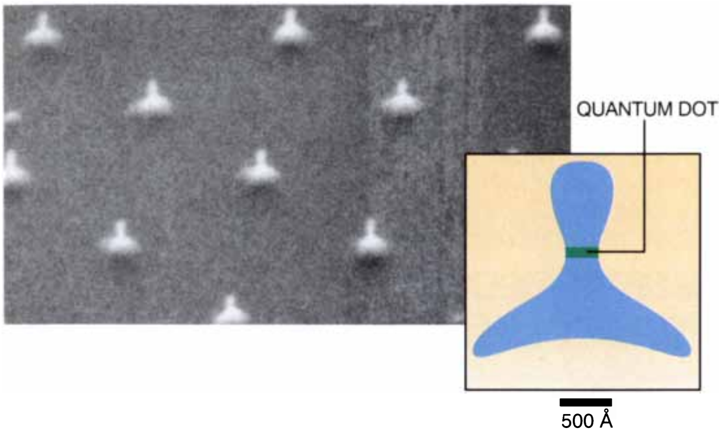
To turn the waveguides into a switch, Etemad increases the intensity of laser light, which immediately changes the index of refraction of the top line. This make the lines no longer identical, and so the waveguides become decoupled. By shortening the lines precisely, Etemad ensures that only high-intensity light will emerge from the top line and low-intensity light from the other. The time it takes to "switch" on (or off) one polymeric thread is only a femtosecond (10^{-15} second) laser pulse. This change is far faster than any semiconductor switch can manage. "I'm not saying that it's going to be used in the telephone system tomorrow," Etemad says, but at least in the laboratory the switch looks good.

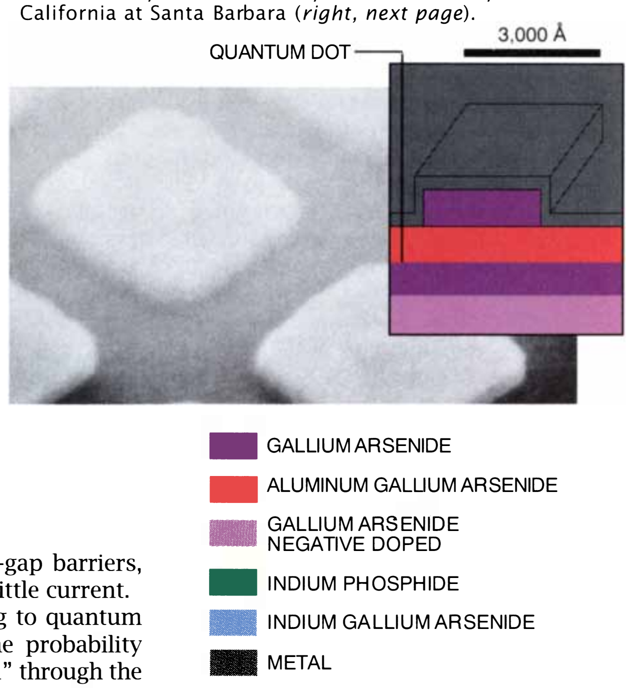
As is the case with semiconductors, just how far investigators will be able to push the conductive and optical properties of polymers depends on how easily the materials can be processed and on how precisely their electrical properties can be controlled. Points out Baughman: "The dream of the future—of easily processable conducting polymers—depends on our being able to manipulate defects and increase the strength of the chain interactions."



ZERO DIMENSION: QUANTUM DOT

AT&T investigators are building dots in an in situ vacuum process (*left, this page*). Lithographically defined dots built at IBM may be used as capacitors. In these devices, electrons are confined at the interface between the layers of aluminum gallium arsenide and gallium arsenide (*right, this page*). Texas Instruments' array of dots (*left, next page*) may be used as transistors. Chemists fabricate dots by capturing clusters of atoms (such as cadmium, in orange, and sulfur, in yellow) in zeolite cages as shown in this computer simulation by Galen D. Stucky at the University of California at Santa Barbara (*right, next page*).





dall, a researcher at Texas Instruments in Dallas, spends most of his time carving through layers of quantum wells. A principal difference: Randall wants to make transistors.

So far integrated-circuit makers have thrived by forcing chip designs ever smaller and by building ever tinier transistors, or switches. But when circuit dimensions approach the size of an electron wavelength, electrons begin leaking through the device, preventing transistors from truly switching off. "We anticipate the first real problems in the latter part of the decade," when chip makers try to push circuit designs tighter than .25 micron, or 2,500 angstroms, says Robert T. Bate, who is Randall's boss and manager of TI's quantum work.

Randall and his colleagues hope to skirt that pitfall by exploiting quantum effects to build transistors far smaller than the troublesome 2,500 angstroms [see "The Quantum-Effect Device: Tomorrow's Transistor?" by Robert T. Bate; SCIENTIFIC AMERICAN, March, 1988]. To do so, they are developing sophisticated descendants of the early resonant tunneling diodes. One notable success: the BiQuaRTT, or bipolar quantum resonant tunneling transistor.

All varieties of resonant tunneling diodes and transistors play by the same basic rules, explains Reed, one of the key BiQuaRTT designers, who recently left TI for Yale University. Start with a thin quantum well layer sandwiched between two equally thin but higher band-gap layers. Apply a voltage. At many voltages, electrons are prevented from flowing through the layers of the device by the high band-gap barriers, and the device conducts little current.

Nevertheless, according to quantum mechanics, there is some probability that electrons can "tunnel" through the barriers. The odds of tunneling can be improved by applying just the right voltage to the device. At these specific voltages, the energy of the incoming electrons matches the energy levels of the quantum well. Electrons then tunnel through the barriers, and current flows. "It's like playing a flute," Reed says. "You only get out those notes that you want."

Of course, one transistor does not an integrated circuit make. (Some integrated circuits have far more than a million transistors.) Researchers at Fujitsu have connected as many as five or six resonant tunneling transistors in series with high-speed switches. Several U.S. government agencies are funding a few research teams to develop more complex circuits and devices.

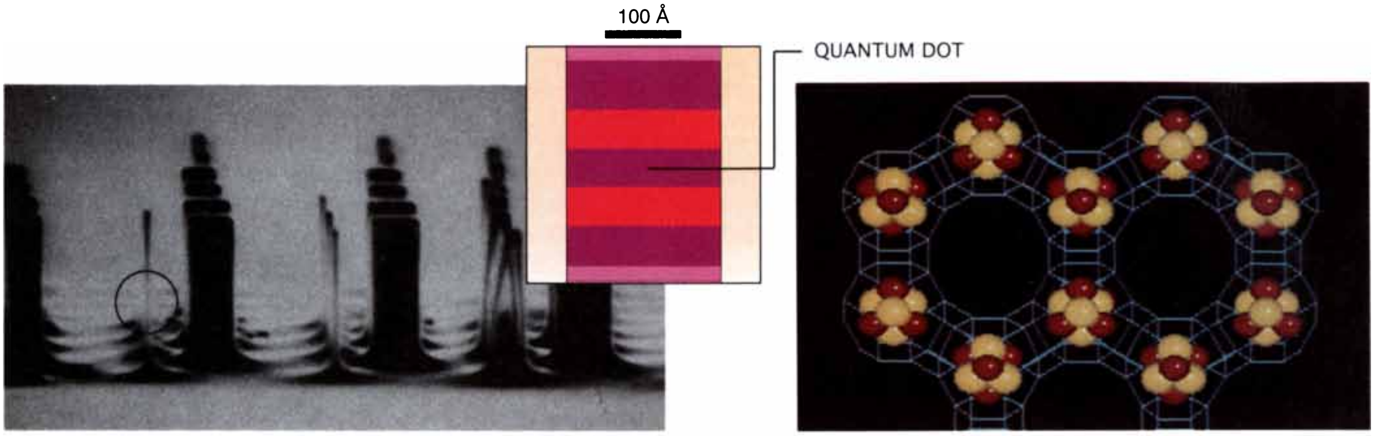
Texas Instruments is among them; sometime within the next 30 months, workers there hope the BiQuaRTT will prove able to handle complex logic and so become the stuff of circuits that are orders of magnitude smaller than today's devices. In addition, "we want to take advantage of the strange characteristics of these devices," Randall explains. Because tunneling readily occurs at two different voltages, "we've got a device that goes 'off-on, off-on'" at two different voltages, he says. Now he hopes to show that such a switch can replace a few conventional transistors and that multiple BiQuaRTTs will work together.

If two dimensions are promising, then one or zero dimensions look even better. Electrons scatter less when traveling through quantum wires, attaining higher mobilities—and so make for faster switches. Quantum wire lasers could be powered by far less current and thus radiate less heat than their two-dimensional cousins.

Wiry Semiconductors

So much for theory. To make quantum wires, workers must wall up four sides of a low band-gap material with higher band-gap barriers thin enough to let electrons tunnel through on command—about an electron wavelength thick. Exercising the precise control needed to make such vertical walls is tricky; it has pushed some researchers to restudy how layered semiconductors are created in the first place.

At the four-year-old Optoelectronics Technology Research Laboratory, located about an hour northeast of Tokyo in Tsukuba Science City, Toshiro Isu is watching his scratchy, black-and-white videotape once again. Since January, Isu, a researcher at the laboratory, has recorded moving pictures of layers of gallium and arsenic growing inside an MBE vacuum chamber. In the images, clumps of millions of gallium atoms form in several seconds, then seem to melt into the existing arsenic layers.



What Isu would like to know is why the clumps appear at random locations and what effect they have on the uniformity of the crystal layers.

"Nobody knows how molecular-beam epitaxy actually occurs, how the layers of gallium and arsenic combine," observes Izuo Hayashi, director of the lab and one of the inventors of the semiconductor laser. He believes that before highly confined devices such as quantum wires and dots can be routinely fabricated, researchers must improve their understanding of what happens on an atomic level during epitaxial growth and etching.

Efforts to build quantum wires and dots continue, nonetheless. Many investigators are trying the same techniques they use for making quantum wells. But because a quantum wire or dot is significantly smaller than a quantum well, bombarding a material with ions can ravage the crystal lattice.

As a result, researchers are looking for other approaches. "Rather than just cutting [etching through the semiconductor], we need to make use of some natural tendency of the material," Hayashi says. His team would like to control the edges of atomic layers as the material grows in an MBE. Yet success has remained just out of reach.

Many groups are trying different schemes. Pierre M. Petroff, Arthur Gossard and their colleagues at the University of California at Santa Barbara, for instance, grow quantum wires in an MBE by stacking elements, one atomic layer at a time, on terraced steps in a semiconductor substrate. This technique produces vertical—or often, tilted—quantum wells measuring 50 angstroms on a side, which can be made into quantum wires. It nonetheless requires agonizing precision. More recently they have tried continuously changing the time that each atomic layer is allowed to grow on the surface so that the resulting layers snake back and forth rather than lining up in straight or tilted stripes. In this "serpentine superlattice," the quantum wires are formed at the bends of the curves, resulting in more uniform arrays of wires than early techniques afforded.

Others have developed variations on etching. Eli Kapon and his colleagues at Bellcore created a V-shaped groove in a semiconductor substrate, then deposited a lower band-gap material in the base of the V. Kapon's team has even built lasers from one to three quantum wires. Such lasers churn out light when powered with only .65 milliamp at room temperature, an extremely low laser threshold.

At the University of Stuttgart, B. E. Maile and his colleagues traced a pattern of wires onto a substrate using reactive-ion beams, cleaned the wires carefully with a chemical solvent and then "buried" the wires by covering them with a semiconductor layer using a deposition technique akin to MBE called metallorganic vapor-phase epitaxy (MOVPE).

Nevertheless, investigators continue to debate whether these techniques—or any other ones—have yielded the distinct splitting of electron energies that should characterize a quantum wire. Most of these wires are too thick. The larger the devices, the smaller the energy splittings, and the harder to say conclusively: this is a quantum wire.

The Zero Zone

At the end of the quantum rainbow are quantum dots. These are often called artificial atoms, even though the particles may consist of thousands or hundreds of thousands of atoms. Confined in a dot, or box, electrons should occupy discrete energy levels. It should be possible, therefore, to dial up precise energy levels by adjusting the construction of the quantum box and by varying the applied voltage.

On the edge of central Tokyo, amid weeds and straggly trees, is the unkempt Komaba research park, a part of the University of Tokyo. In a modest, two-story brick building built some 60 years ago is Yasuhiko Arakawa's laboratory. As is traditional in Japanese homes, visitors entering Arakawa's laboratory slip off their shoes and don plastic sandals. But there is another reason for changing footwear; around the corner from Arakawa's spartan office is a high-quality clean room packed with $10 million in equipment for building quantum devices.

In 1982 Arakawa and his mentor Hiroyuki Sakaki proposed the concept of quantum boxes, or dots. They have yet to construct a working set, however. "It is very difficult to realize a real quantum box with 100-angstrom spacing [along all six sides]," Arakawa says. "Nobody has succeeded in fabricating a real quantum box laser," he insists.

Indeed, making quantum dots with semiconductor technology seems an exercise in exponential complications. Processing complexity increases "in some nonlinear way," says Kathleen Kash, a researcher at Bellcore. Stating that a dot "works" guarantees controversy. And a single dot on its own is not particularly useful. Investigators must find ways to manufacture collections of dots and then to integrate them into devices.

Chemists, however, have had extraordinary success in making quantum dots—although not ones yet suitable for devices. In his office at Bell Labs, Michael L. Steigerwald keeps a collection of small vials. Each is filled with a vividly colored powder—yellow, orange, red, black. "They're all cadmium selenide," he says, almost with awe. Each vial holds different-size clusters—essentially quantum dots—of cadmium and selenide atoms.

The different-size clusters have unusual physical properties, such as increased hardness; they also absorb particular frequencies of light. The yellow powder, for example, consists of cadmium selenide clusters roughly 15 angstroms in diameter—about 100 atoms. Red clusters measure 40 angstroms across (about 1,200 atoms). The frequency of light absorbed reflects the band-gap energy of the cluster. The smaller the cluster, the larger the band-gap energy, Steigerwald explains. So changing the size of clusters tunes the energy level of a material.

To make clusters, the chemists capture tiny groups of atoms in sheaths or in cages. A few years ago Steigerwald's colleague Louis E. Brus found he could suspend tiny clusters of cadmium selenide, ranging from 10 to 100 angstroms in diameter, in a solution. Later Brus and Steigerwald, a synthetic chemist by training, began wrapping the clusters in soapy films of organic polymers that dried and molded around the atoms. Such sheaths keep the atoms from recombining when they are taken out of solution. Others, including Galen D. Stucky and his team at the University of California at Santa Barbara, are trying to stuff atomic clusters into porous glasses or zeolites, which act like cages. Still others, working with metals, use lasers to chip metallic clusters from larger materials [see "Microclusters," by Michael A. Duncan and Dennis H. Rouvray; SCIENTIFIC AMERICAN, December, 1989].

Yet although chemists can make far smaller clusters, or quantum dots, than can lithographers, there is a catch. In any batch of clusters, sizes will vary, on average, by about 10 percent, Steigerwald says. That variation is far too great for electronic or optical devices, which demand arrays of identical clusters. "I don't know a good way to answer that problem," Steigerwald concedes. "There are lots of ways to dream about, but the world will need another inspiration."

A California team hopes to have one answer for him. Recently Kerry J. Vahala, a physicist at the California Institute of Technology, teamed up with a chemist and an engineer to build a system for separating clusters with a laser beam. Vahala's team will begin testing their idea early next year when their equipment is up and running. The plan: form a batch of semiconductor

ELECTRONS TRAVELING ONE BY ONE

Current in conventional electronic devices is considered a kind of flowing river of electrons, in which the electrons are as uncountable as molecules of water. Reduce the dimensions of the material, and the energy of those electrons becomes quantized, or divided into discrete increments. Still, the precise number of electrons defies calculation.

Now researchers at Delft University in the Netherlands, the Centre d'Études Nucléaires de Saclay in France, Chalmers University of Technology in Sweden and Moscow State University have shown they can count the individual electrons that make up a weak current. Instead of quantized electron energy, they and many others are studying quantized electron charges.

The implications of the efforts are twofold: devices that clock individual electrons may produce a new, highly precise measurement standard for current. Longer term, currents composed of single electrons could lead to alternative schemes for very small-scale electronic devices.

SYNCOPATED ELECTRONS. The basic design for a single-electron device calls for a central aluminum electrode of about one micron long and coated with aluminum oxide. This bar overlaps with two symmetrically arranged metallic electrodes on either side. Although small, the device is not a dimensionally confined system, points out Theodore A. Fulton of AT&T Bell Laboratories, who has also been building single-electron devices.

Apply a voltage across this device, and it behaves like a conventional resistor. But at temperatures of about one kelvin and voltages of a few tenths of a millivolt, the resistance increases dramatically. No current will flow through the device because the electrons have insufficient energy to charge the central bar. Increasing the voltage to about one millivolt gives electrons just enough energy to tunnel from the voltage-carrying electrode into the central electrode.

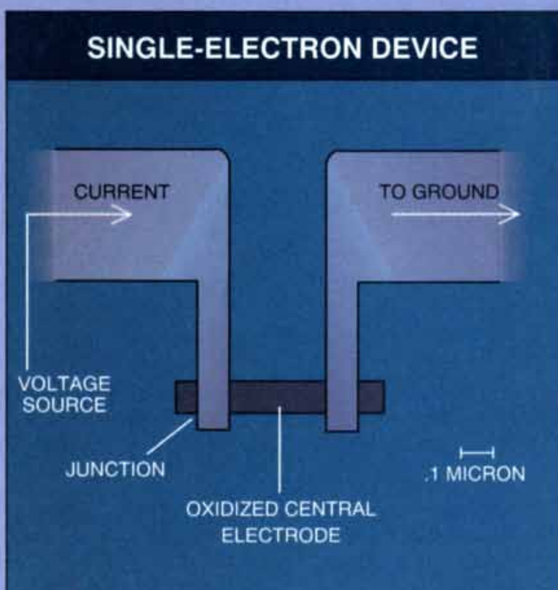
Only one electron at a time successfully tunnels through to the bar, however. This happens because an incoming electron charges the bar, thereby raising the energy barrier the next electron would have to overcome to reach the bar. When the first electron leaves the bar, the bar's charge falls. The flow of electrons consequently becomes highly syncopated: one electron must leave the central bar before another can enter.

ONE IN A THOUSAND. To employ this single-electron effect in a precise standard for current, researchers at Delft and Saclay synchronized a single-electron current with a radiofrequency signal. The workers first construct a series of small electrodes. By attaching additional voltage sources to the electrodes, the energy barriers of the device can be raised or lowered. The investigators then couple one secondary voltage source to a radiofrequency signal and let the signal meter the electrons through the device, one per cycle. "They've been successful as near as you can measure it," Fulton says.

So far the workers have recorded current of a few picoamps—plus or minus a few femtoamps—at frequencies of about five to 20 million electrons per second. Their measurements are therefore precise to one electron within 1,000. For them to top the existing standard for current, precision to better than one in 10 million would be necessary, Fulton notes.

The gates built at Delft and Saclay operate much like a simple shift register, an electronic device for storing binary numbers. Moreover, single-electron devices may prove easier to fabricate than quantum wires and dots; the metal gates do not rely on complex junctions of different compound materials.

But demonstrating physics principles is a far cry from building devices. So far all the research has taken place at temperatures near one kelvin. Whether the single-electron effects will hold up as workers try to raise the temperature remains to be seen. Adds Fulton: "I'm glad to say that I don't yet know the answer to that question."



clusters in a gas stream, then focus a laser on the particles. Those clusters with energy levels that resonate with the laser should be knocked out of the batch and can later be deposited on a substrate. "It's highly speculative," Vahala acknowledges. "But over time, crazy things become doable."

Another avenue taken by Kash at Bellcore takes advantage of the stress and resulting deformations created when one material is grown on top of another in a specific pattern. "Normally you don't want stressed materials," she adds. But in a jujitsu maneuver, she began purposely stressing materials to create lower band-gap regions—and so form quantum dots.

Most workers nonetheless continue to try to exploit lithographic techniques—but with twists. At AT&T, Henryk Temkin is making dots with different semiconductor compounds—indium phosphide and indium gallium arsenide, which emit a frequency of light that can be transmitted over optical fibers with little attenuation. Moreover, by containing all the fabrication steps within a single, high-vacuum environment—a trick pioneered by Hayashi's team—Temkin reduces the possibility of contamination and produces better-quality dots.

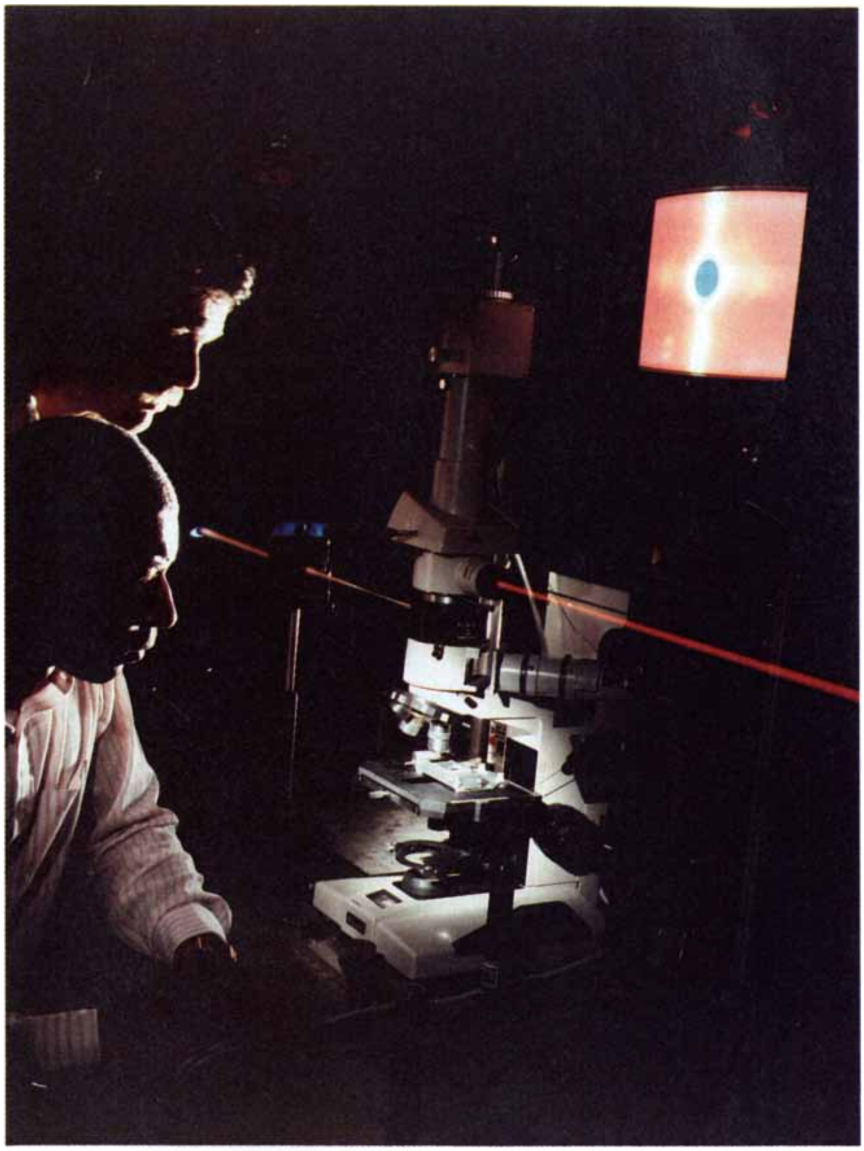
At many other universities and companies, researchers are honing their skills in etching the small devices. Among the groups singled out in the field as having achieved highly promising results are teams at the University of Glasgow, Delft University in the Netherlands and Texas Instruments.

Two years ago at TI, Reed and Randall successfully made a single quantum dot. Because it was isolated from any other device, the dot "wasn't useful, but it worked," Reed says.

Schemes for connecting an array of dots into a larger circuit or system have remained elusive. "You could make clusters electrically active by contacting them with a conducting polymer," Steigerwald muses. "That's not quite science fiction, but it's in the future."

Randall continues to explore electrical links. "My biggest headache," he grouses, "is making good contacts to quantum dot posts." If he succeeds, he and his colleagues will try to employ each dot in an array as a computing element. Because each dot can have multiple energy states, one dot could replace several conventional transistors.

Such dots might also enable the TI team to try out a novel interconnection scheme called cellular automatons. Instead of connecting transistor dots one after another in a series, every dot would talk to only four of its nearest neighbors. By contacting the dots on the edges of the array, it should be possible to trigger the operations of inner dots. This would transform the circuit into a massively parallel system in which many transistors operate simultaneously rather than sequentially.



LIGHT FROM KRYPTON ION LASER (*left*) excites a quantum well surface-emitting laser (on microscope stage), which emits an infrared beam (*right*) in this experiment at Sandia National Laboratories. (The infrared beam has been made visible by running the light through an optical conversion material.) A television monitor displays a cross-sectional profile of the beam from the surface-emitting laser.

Although this architecture would be inefficient with today's transistors, it might make sense with quantum dots that are more than a single switch, Randall suggests. "Downscaling will continue, and eventually it's going to crush down to dots," he says. "We'll need interesting [dot] coupling designs when we get there."

Just when workers will "get there" remains uncertain, and time is growing short. Those building telecommunications networks are demanding more efficient lasers for pumping additional data across the lines. Chip makers worry that the ability of conventional lithography to pack more devices onto a chip will stall by the end of the 1990s. That, points out Robert E. Nahory, who manages Bellcore's semiconductor materials research, makes now "the time for innovation."

With reporting by correspondent Tom Koppel in Tokyo.

SCIENCE AND BUSINESS

Electronic Earful

Cochlear implants sound better all the time

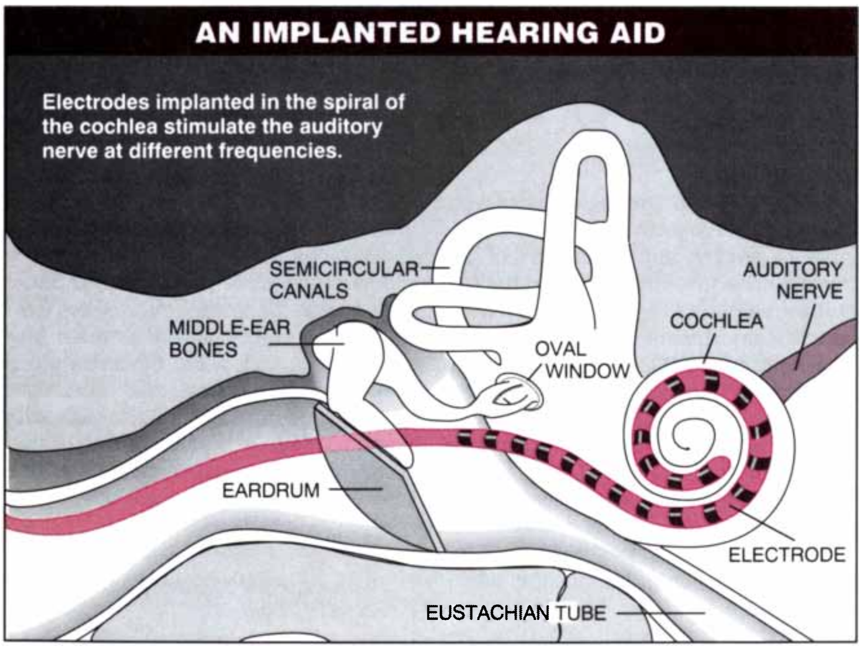
Megawatt light bulb, gradient lenses, molecular computers

Imagine threading a wire through a snail shell a millimeter wide at its widest point. The wire is an electrode, the bony spiral a region of the inner ear known as the cochlea. The reason for traveling this path? To stimulate the nearby auditory nerves electrically so a deaf person can perceive sound. This concept was hyped well before the first cochlear implant was approved for use in adults by the Food and Drug Administration in 1985. One speaker at a meeting of the National Academy of Sciences went so far as to declare, "We are going to clear out the schools for the deaf."

That has not happened and won't. Only people with profound sensorineural hearing loss in both ears who cannot be helped by hearing aids can receive cochlear implants. This type of deafness is caused by damage to the 12,000 sensory hair cells that line the normal cochlea and transform the mechanical vibrations of sound waves into nerve impulses. Implants use electrodes as stand-ins for the cochlear hair cells. So far fewer than 3,000 devices have been implanted, although hearing experts estimate that 250,000 or more patients could derive benefits. But cochlear implants have improved considerably during the past five years and may soon attract more candidates.

The first versions to be approved by the FDA collected sound and transmitted it by electrical induction through the skin to a lone electrode in the inner ear. Such single-channel devices made by 3M Company and others could help just a few patients to comprehend speech, whereas most people got only the marginal benefit of sound awareness. They are no longer sold in the U.S., observes Bruce J. Gantz, director of the Cochlear Implant Research Center at the University of Iowa. "It became pretty obvious that multichannel devices are superior," he says.

One such device, a 22-channel implant made by Cochlear Corporation in Englewood, Colo., earlier this year became the first to win FDA approval for use in children between two and 17 years of age. Previously only adults who lost their hearing after acquiring speech were eligible for implants of any kind.



In Cochlear's device, a battery-powered speech processor worn on the belt or in a shirt pocket selects and encodes a range of sound "features," from the low frequencies of *aah's* to the peaks of *tsst*. Coded digital signals are then transmitted to the electrodes implanted in the inner ear via radio frequency transmission. President Ronald E. West notes, "We're trying to help with sorting out and processing the frequencies most important to speech comprehension."

Even more electrodes are planned for future models. "The results indicate that more stimulation is better, although it's not a linear relationship," declares Graeme M. Clark, a researcher at the University of Melbourne in Australia. Clark developed the digital speech processor in Cochlear's apparatus and has studied cochlear implants since 1967. He suggests that if a particular frequency is very important, it might help to stimulate not just one but several electrodes in an area.

No matter the number of electrodes, one design consideration stirs endless debate among manufacturers: What is the best way to process sound? One promising strategy has been developed by Blake S. Wilson and co-workers at Research Triangle Institute and Duke University in North Carolina. (The work is supported by the Neural Prosthesis program of the National Institutes of Health.) The approach presents sound as high-speed, nonoverlapping electric pulses. "We used to send signals a maximum of 300 times a second. Now we send them 1,000 times a second," he explains.

Patients are achieving improvements in speech comprehension of 5 to 162 percent when this pulsed strategy is used, Wilson reports. Rapid pulses let patients appreciate the temporal details of speech, crucial to the identification of most consonants, he explains. To perceive the word "pop," for example, a person needs to hear the two silences between the *p*'s on either side of the *aah*. "It's like television, where stationary pictures are sent fast enough to be seen as a whole," Wilson says.

Yet pulsed sound delivery may be more effective for some patients than others. Gantz observes, "It may be that some patients, such as those with poor nerve survival, respond better to analog sound transmission." Analog trans-

mission compresses a wide range of frequencies into a smaller dynamic range. The sound is sent in continuously varying waves. Because it is still not possible to predict preoperatively how a patient will respond to a particular strategy of sound conversion, companies are keeping their options open.

"Since nobody else can come to agreement on how to process sound, we figured we weren't smart enough to do it either," says Joseph H. Schulman, chief scientist at MiniMed Technologies in Sylmar, Calif. So his company is developing a 16-channel implant that can work with digital or analog processors, pulsatile or not, over large or small dynamic ranges. Schulman expects the device to be in clinical trials within a few months.

Smith & Nephew Richards, Inc., as well, is looking at combining both analog and pulsatile approaches in the same device. The Memphis, Tenn., firm has an implant trade named Ineraid in clinical testing. It sends simultaneous stimulation to four of six electrodes.

Where the stimulation occurs in the cochlea may be just as important as the intensity of the stimulation, suggests Margaret W. Skinner, director of the cochlear implant program at Washington University School of Medicine in St. Louis. One of the most desirable goals—and most daunting obstacles—is to assign frequency bands to different electrodes so that signals are sent to the proper place in the cochlea. Low frequencies stimulate regions deepest inside the cochlea; higher ones those in the outer spiral. "Where do the maximum comfort levels sit when all the electrodes are playing—do some need to be eliminated because they're not contributing or are detracting?" Skinner asks.

Researchers are eager to advance the trend toward implanting younger patients so they can take full advantage of the time when speech develops readily. Gantz says, "We're trying to find out if there's a critical window in which the devices work best."

That goal will be particularly challenging in children younger than two years. Clark explains that although everyone's cochlea is adult size at birth, there will be growth in the rest of the cranium. "We need some system of wire lengthening so that the wire can be extended without being bound down by scar tissue or pulled out." It has been done in children older than two years by neatly looping wire in a flat bag. "The technique sounds rather low tech," he says apologetically, then chuckles. "It's about the only aspect of this field that is." *—Deborah Erickson*

Bug-Eyed

Gradient-index lenses bend light to order

In human and insect eyes the light is bent at an ever increasing angle as it moves through tissue. This provides more focal power by shortening the length at which rays are focused.

Unfortunately, the lenses in eyeglasses, cameras and other common optical devices do not usually work so well. That's because the spherical surfaces of these lenses refract light rays at their curved edges more than they do at the center. This spherical aberration creates a fuzzy image that results from diversion of some of the light from its desired focal point. The lens maker can make corrections by painstakingly grinding a lens into an aspherical shape or by adding more lenses, which can introduce new defects of their own.

Since the turn of the century, optical designers have been trying to imitate bug eyes by making lenses from glass in which the refractive index varies as light rays move through the lens. But building such gradient-index properties required sophisticated computers to work out the cumbersome calculations.

Today a handful of companies are producing what are known as GRIN, or gradient-index, lenses. The lenses have a refractive index that varies either from center to edge, a radial gradient, or from front to back, an axial gradient. A lens with a radial gradient focuses light more strongly, allowing it to reach a focal point from an unground flat surface. By contrast, a lens with an axial gradient requires grinding a spherical surface.

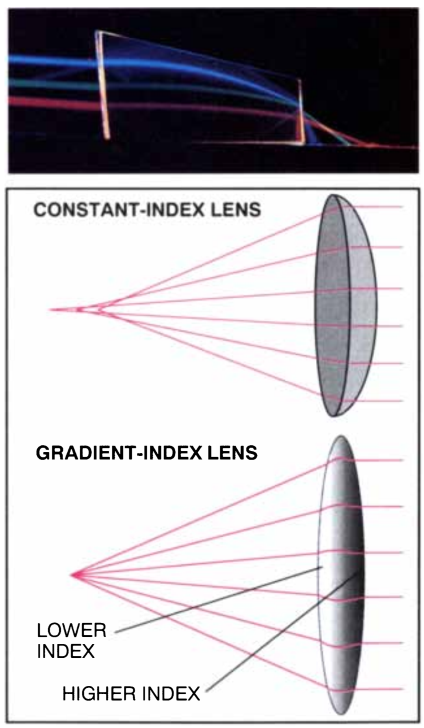
Even with computers, GRIN materials are still tricky to produce. Most are made using an ion-exchange process in which an ion such as potassium from a molten salt bath trades places with another metal in the glass, like thallium. In this case, those areas with more thallium, the heavier element, have a higher refractive index. It can take a month or more for sufficient ions to diffuse into the glass, and the amount of change in refractive index is limited.

Isotec, a Tucson, Ariz., start-up company, claims to have come up with a manufacturing technique that produces a lens blank in a matter of days while doubling the change that can be achieved in the refractive index. The company, which has yet to demonstrate a finished lens, says the process also results in larger blanks and better control of the way light is shaped as it moves through the glass. The process is carried out by fusing together five to 10 layers of lead-silicate glass in a platinum crucible at temperatures of up to 1,500 degrees Celsius. Layers with a greater refractive index contain more lead atoms, which interact strongly with photons.

Isotec got started five years ago, trying to make lenses that could be used to concentrate sunlight onto photovoltaic cells. When early samples of the glass shattered or turned opaque during fabrication, the company brought in Richard Blankenbecler, a particle physicist who is head of the theoretical physics department at the Stanford Linear Accelerator Center.

Blankenbecler suspected that the glass was disintegrating because layers with varying indexes expanded or contracted differently in response to a temperature change. By looking at technical specifications in glass catalogues, he found some glasses whose thermal properties did not change when the index did. If these glasses were used, the different plates could be fused without falling apart.

Now Isotec is talking to potential



GRADIENT-INDEX LENSES can correct for spherical and other optical aberrations by continuously changing the refractive index (bottom). *Light rays hitting a conventional lens, with its single refractive index* (top), *may fail to reach a single focal point. Isotec's gradient-index glasses* (photograph), *which are being developed into lenses, can bend light independent of wavelength or surface curvature.*

customers to license its technology for laser collimators, binoculars, night-vision goggles and the headgear used for "virtual reality," animated three-dimensional graphics simulations. But its high-powered public relations campaign—it advertises the product as "smart glass" that could "revolutionize telecommunications, computing and traditional optical products"—has been viewed skeptically by the small community that makes GRIN products.

"Smart implies logic; there's no logic in it," says Duncan T. Moore, director of the Institute of Optics at the University of Rochester, a major center for GRIN research. "I think they've done very interesting work, but I'd temper that by saying that there's a lot more work to bring this into production. All they've done is make a few samples." Specifically, Moore, who owns a company that makes GRIN products, says Isotec has only demonstrated a glass blank that still must be ground and polished into a lens.

If Isotec finds licensees, a major competitor is likely to be Nippon Sheet Glass (NSG). This Japanese company claims to hold 99 percent of the estimated $50-million market in GRIN lenses for facsimiles, photocopiers, laser printers and fiber-optic communications components. Using the conventional ion-exchange process, NSG produces millimeter-diameter lens rods, hundreds of which are arrayed as part of a fax machine or photocopier scanning element. The short focal length—between 15 and 30 millimeters—required to project a line-by-line image of a document onto a photosensitive drum makes the tiny lenses suited for portable machines.

But NSG is not sitting still with its technology. Instead of making lenses by stacking and gluing together tiny rods with varying refractive indexes, the company is trying to develop mass-fabrication techniques that combine ion exchange with the photolithography used to produce computer chips. Its goal is to produce a GRIN lens by etching an array of microscopic lenses onto a glass substrate.

Both technologies, however, may get a run for their money from so-called binary optics, in which diffractive patterns are etched on conventional lenses to correct spherical or other abnormalities. Then again, these two methods may turn out to be partners. Rochester's Institute of Optics is trying to combine the technologies. What may eventually be dubbed trinary optics might allow the chromatic aberrations present in GRIN and binary materials to cancel one another out. —*Gary Stix*

Not Biochips?

There may yet be computers made with organic molecules

Some wild talk was going around in 1983 that the same kinds of genetic engineering techniques enabling microorganisms to produce drugs would also let them churn out computers. The organisms would self-assemble "biochips"—complex clusters of organic molecules that would function as integrated circuitry but be thousands of times smaller than conventional components. Or so the story went. Some investors bit hard before the scheme—and the company promoting it—was debunked by a federal scientific inquiry.

The ridicule that followed continues to haunt a small group of serious researchers, who believe that organic molecules can indeed be made to work as submicroscopic computer elements. Please, they plead, don't use the word "biochip" to describe their work. Their challenge is to build organic molecules that can repeatedly switch from conductive to nonconductive states and back again, just like conventional electronic components made of semiconductors such as silicon.

The research, which parallels perennial efforts to shrink the size of semiconductor components, remains largely theoretical. It will be years before any molecular electronics devices are commercialized—if ever. But scientists are reporting tantalizing early successes. Some skeptics are beginning to wonder if there might be something to the idea after all. "People are excited about pushing chemistry to new frontiers," says Robert M. Metzger, a physical chemist at the University of Alabama at Tuscaloosa. He notes, "The physicists are encouraging us, and the engineers are at least not laughing at us any more."

This summer, organic chemist James M. Tour of the University of South Carolina at Columbia announced that he had synthesized a section of a molecule that might function as an electronic switch. His blueprint was proposed by IBM researcher Ari Aviram in 1988. Aviram, who first began musing about molecular computer parts in the early 1970s, envisioned large organic molecules that could give and accept electrons and so carry current.

The model could pass a single electron between two molecular chains—enough to make a switch, if the electron's movement could be controlled. The chain donating the electron would become a conductor, and the one receiving it would become an insulator. The chains were joined with a bridge for the electron to cross.

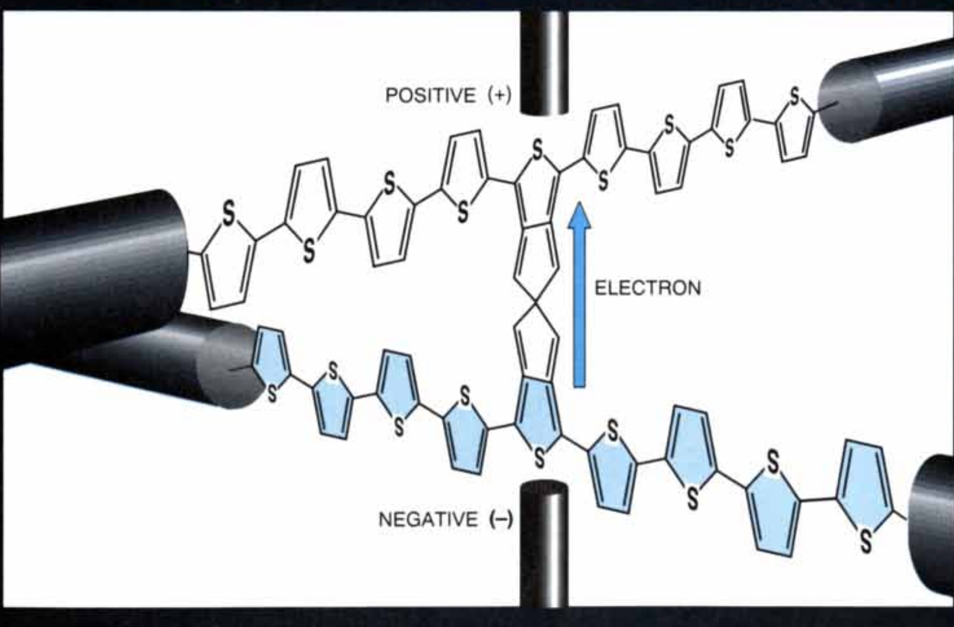
To keep the electron from jumping at will, Aviram twisted the bridge at a 90-degree angle so that the arms of the molecule would look like a plus sign if viewed from above. This way, the chains would be separated from each other, and the electron could move from one to the other only when forced to, by voltage applied through an electrode abutting the bridge. Current would move through the chain that gave up the electron into electrodes attached at either end.

So far Tour and two graduate students have made a couple of versions of these twisted bridges. In one, the chains are based on a conductive polymer called polythiophene, which has a backbone of sulfur molecules. "We're building the chains separately, then we'll slap the pieces together," he says. Tour says his approach to constructing large molecules to order is borrowed from pharmaceutical manufacturing, but it has not been used widely by materials scientists.

Tour acknowledges that the molecule he has built is just the beginning. After all the chemical tricks and hard labor, the molecular chain is still only 23 angstroms long, less than half the size Aviram described. "Chemists shy away from making molecules that big," Aviram says. "It's a very laborious undertaking."

Unfortunately, a molecule that could connect to circuit lines on conventional semiconductor chips would have to be far bigger than that. And such connections are necessary before a billion or so molecules could be wired together into working computer components. With current commercial methods for etching conductive paths on semiconductor chips, the distance between individual circuit lines is typically 8,000 angstroms. Even the most dense techniques of electron-beam pat-

HOW A SINGLE MOLECULE COULD SWITCH CURRENT



When there is a voltage difference across the upper and lower electrodes, a single electron moves toward the positive electrode across the twisted bridge connecting the two chains of the molecule.

The chain giving up the electron becomes conductive (*blue*). Current flows through it into connecting electrodes.

Reversing the polarity switches the conducting chain back to an insulating state, and current flows through the other chain.

terning bring the wires no closer than 150 angstroms. To bridge these minute conductors, a molecule would still have to be three times the size of the one Aviram postulated. But Tour is optimistic that he can build much larger molecules. "Now it's just a matter of hooking more thiophenes into the chains," he says.

Aviram and other proponents of the new electronics figure that by the time organic molecules are ready to roll out, circuit-pattern distances will not be so vast. Even then, the molecules would somehow have to be connected to the conductive "wiring." One possible solution is designing molecules that contain a specific atom, or collections of atoms, known to have affinity for a particular substance.

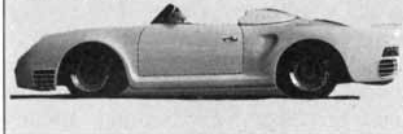
"These functional groups will be important to wiring up molecules," declares Mark S. Wrighton, a chemist at the Massachusetts Institute of Technology. He offers an example of how the approach could simplify manufacturing: Because sulfur-containing molecules stick to gold, and carboxylic groups bind to aluminum oxide, an array of wires containing those metals could be dipped into a solution of the organic molecules and emerge with conductive connections exactly where they ought to be.

Others working in molecular electronics believe single molecules may present too many problems to be practical. "How do you connect a single molecule?" Metzger asks. He is working instead on devices built from films a molecule thick. Carefully patterned layers stacked on top of one another could make vertical as well as horizontal connections, obviating bulky wires and increasing the circuit density.

Metzger explains that the trick is to design directionality into the molecules—to give them, for instance, greasy tails so they move away from water and line up facing the same way. But he still has not figured out how to carve patterns in a layer.

The delicacy of one-molecule-thick films makes them vulnerable to pinholes and other defects, but researchers who favor the approach say such devices will still be more reliable than single molecules. To add stability to films, some scientists are binding the molecules to polymer substrates.

Metzger and his collaborators have made a film five by 10 inches and are testing to confirm that it conducts current in one direction. The next step will be to control manufacture to make the working size smaller. "Within the next two years, somebody will know for sure whether films work. From that to



a workable device, it'll be maybe 10 years," Metzger predicts.

Even Aviram acknowledges that single organic molecules probably will not be reliable enough to use alone. "No one disputes there will be a need for redundancy," he says. Exactly how much redundancy will be necessary has to be determined experimentally. Because fully functional molecules have not yet been made, the matter is mostly intuitive at present. Critics point out that the more molecules get piled up for safety's sake, the closer organic systems will come to the size of conventional devices. Supporters note that a model of individual molecules used for data processing and storage already exists in nature—in the form of DNA.

Some researchers, in fact, are looking to nature for molecules that switch naturally. At the Syracuse University Center for Molecular Electronics, Robert R. Birge is using bacterial rhodopsin as a switch for a high-speed RAM (random-access memory). A pulse of laser light causes the chemical to switch from one form to another. The substance, similar to pigment found in the retina of the human eye, comes from a bacteria found in salt marshes. "It's a very old organism—the only one I know of that can transfer from breathing to photosynthesis," he notes.

Copper's well-known friendliness toward electricity is earning the element a place in some molecular electronics research. A group led by Richard S. Potember of the Johns Hopkins University Applied Physics Laboratory is adding copper to a thin film of an organic molecule called tetracyanoquinodimethane (TCNQ). Pulses of high voltage or laser light cause the copper atoms to bind and unbind from TCNQ, switching the film from conducting to nonconducting. A similar TCNQ film is being developed at Japan's National Chemical Laboratory for Industry, where a group headed by Hiroaki Tachibana is using heat to flip the switch.

Predictably, few molecular electronics researchers are guessing how long it will be until some of their devices become commercial realities. But all are encouraged by the speed at which developing technologies—particularly scanning tunneling electron microscopes—are allowing them to observe and construct materials on a molecular level. At the same time, manufacturing techniques for structures approaching the size of large molecules are getting better. "Molecular electronics is a distant goal," Aviram concedes. But he notes, "There are a lot of beautiful scientific milestones until then." *—D. E.*

A Million Watts of Light

The world's most powerful lamp finds applications

What do you do with the world's brightest lamp? Well, you could melt a car with it. At least that's what the producers of the *David Letterman Show* wanted to do when they read that Vortek Industries, maker of the lamp, was planning to build a megawatt model capable of turning a small car into slag.

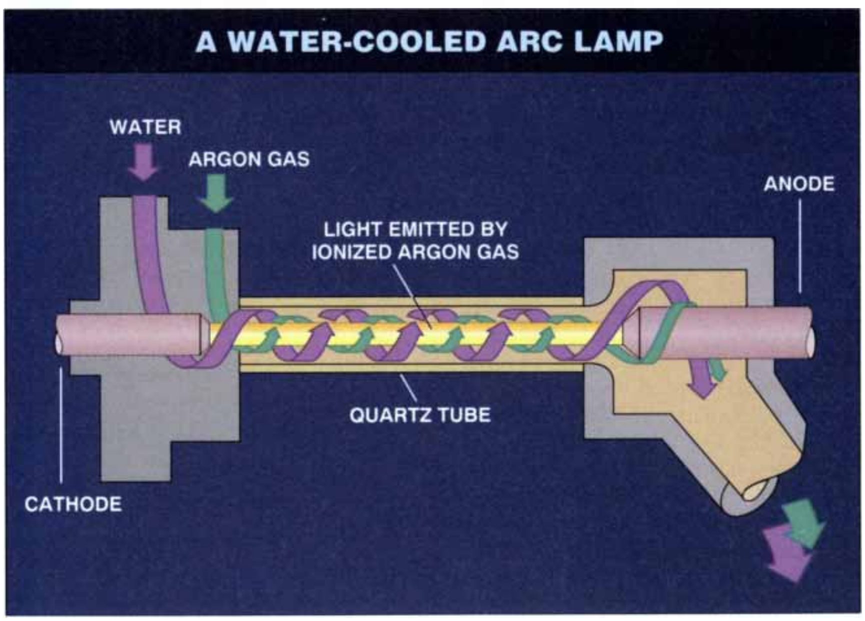
Gary G. Albach, president of the Vancouver-based company, refused the request as unsafe. He was hardly surprised by the idea, however—he already knew that the major commercial application for his superbulbs was as a source of heat rather than of light. When Albach and his colleagues conceived the lamp more than 10 years ago at the University of British Columbia, they envisioned a floodlight, but their first products in 1980 found no takers. "There was simply not a large need to light 20 acres with a single bulb," Albach says, speaking of his company's first product. (The current top-of-the-line model would light 50 acres.)

The Vortek design can stretch to multimegawatt power because its cooling system works from within. The lamp's quartz tube contains two concentric helices, one of ionized argon, which carries the arc, and a wider one of deionized—and hence insulating—water. The water, under too much pressure to boil, carries away heat that would otherwise create a tube-busting temperature gradient between the inner and outer surfaces. The argon spins so rapidly in the central void that it cannot flicker and therefore puts out a steady light. The tungsten-alloy electrodes are cooled with water injected directly into them, and their debris is washed from the tube's inner face, preventing the tube from becoming opaque.

The intense light from Vortek's lamps is a formidable source of heat. Targets placed a hand's breadth from the 300-kilowatt lamp's 11-centimeter-long tube can reach 3,000 degrees Celsius. That is about half the temperature at the surface of the sun and one quarter the temperature attained by the argon arc itself. The arc churns out 120 kilowatts of optical power, a 40 percent efficiency about eight times greater than that achieved by carbon dioxide lasers, the main alternative for industrial heat treatment and testing.

Today Vortek's markets include semiconductor manufacturers, who use the lamps to anneal silicon, a process that removes the crystalline defects created in silicon wafers during photolithography. The lamp is ideal for annealing because its precise, even heat can raise all of the hundreds of chips on a wafer to the same temperature and hold them there for the same period. Vortek originally incorporated its lamps into optical ovens marketed by the Eaton Company, but after the line was discontinued Vortek began supplying lamps directly to customers, which include IBM, Siemens and Motorola.

In addition, manufacturers of blades and automotive parts use the lamp to harden only those surfaces subject to wear, so that the supporting material remains flexible. Vortek has an advantage over laser case-hardening because



it is about five times faster and because its heat produces a well-defined surface layer rather than interposing a heat-affected layer between surface and substrate. Moreover, Vortek's case-hardening is precise enough to form agricultural tools into patterns of harder and softer metal that sharpen the edge as it cuts through abrasive soil. Aerospace clients use the lamps to test exotic materials for heat tolerance, a power-hungry task that is driving Vortek's development. The National Aeronautics and Space Administration has used the lamps to simulate solar radiation's effects on the shuttle. Wright-Patterson Air Force Base has yoked two 300-kilowatt tubes together to test bits of the proposed National Aerospace Plane, and it has contracted to upgrade the power by half, for a total of nearly a megawatt.

Why force-feed power into one or two arcs when you can get the equivalent from an array of many smaller ones? The reason has to do with optics: it is impossible to concentrate all the light of a nearby source into an image smaller than the source. The highly coherent output of a carbon dioxide laser can be focused more tightly. For heating large parts, however, the Vortek lamp is usually preferable, both because it is cheaper to operate and because metal absorbs its white light more readily than the laser's longer wavelength.

Researchers at Lawrence Livermore National Laboratory have found that the lamps are ideal for pumping high-powered lasers. Luis Zapata, a physicist in the laser unit, says he and his co-workers "simmer" the arc at a constant 10 to 20 amps while rapidly jolting it with millisecond pulses of 400 to 500 amps. The flashes produce corresponding pulses in a neodymium-YAG laser, which carry enough punch to penetrate steel. Applications include the drilling of cooling channels in jet turbines and holes in laminar-flow wings, which draw in air that would otherwise become turbulent.

Zapata says his group has no use for Vortek's brightest lamps, let alone the more powerful ones the company is planning to build. About the only military use for a laser pumped by a multimegawatt lamp would have been in the defunct Stars Wars program; even then, it would have been too heavy and fragile to put into an orbiting battle station. So who would want such a monster lamp? "Some defense contractors are thinking of a single source of five to 10 megawatts," Albach says. "They want it for high-temperature testing and industrial applications." —*Philip E. Ross*

THE ANALYTICAL ECONOMIST

Saving makes a rainy day?

Call it "the consumers' dilemma." For most of the 1980s a covey of economists criticized the seemingly insatiable spending habits of American consumers. The good of the economy, they said, required that citizens, not to mention the government, save rather than spend. But around the same time that potential war in the Middle East hit newspaper headlines, threats of an imminent recession began lurking on the inside pages. And the economists now seem to have changed their minds. The ostensible cause of the economic slowdown: Americans aren't spending enough.

Why the contradiction? "It's the difference between a short-run and long-run perspective," says Geoffrey Carliner, executive director of the National Bureau of Economic Research in Cambridge, Mass. Reduced spending—which by definition accompanies increased saving—tends to bring on recessions, but rampant consumption provides no basis for long-term growth.

The short-term implications of lower spending are clear. Each of the past few economic slowdowns has been associated with faltering consumer spending. Reduced purchases, particularly of expensive items such as homes and automobiles, translate into less money flowing through the economy as a whole. Since the 1960s the culprit causing most recessions has been the Federal Reserve, notes Timothy Taylor, managing editor of the *Journal of Economic Perspectives.* The Fed raised interest rates whenever it saw signs of rising inflation. Then the increased cost of borrowing inevitably discouraged consumers from buying.

In contrast, the current dip in consumer spending may be largely the result of flagging confidence in the economy. Surveys by the Conference Board, a New York–based business group, show that in August consumers were more pessimistic about the economy (and their own financial prospects) than they were immediately after the 1987 stock market crash. Meanwhile the Federal Reserve has not raised interest rates significantly. If rates do rise in response to the inflationary pressure of rising oil prices, consumer spending could fall even further.

Even though falling consumption could lead to a recession, economists are not suggesting that wild spending sprees have suddenly become a civic duty. From a long-term perspective, argues Edward M. Gramlich of the University of Michigan at Ann Arbor, U.S. savings rates are still too low to sustain the investment needed to keep the economy afloat.

Savings fell—from a postwar average of 8 percent of disposable income—income after taxes and other mandatory payments—to a low of 3.2 percent in 1987. In the past two years savings have rebounded, to about 4.5 percent. Only by borrowing and by selling parts of itself to thriftier countries has the U.S. prospered through the 1980s; eventually this strategy will lead to a permanent loss of income, Gramlich says.

The conundrum that policymakers face is how to increase savings enough to stave off eventual penury without triggering immediate starvation. If national savings increase (as a result of faltering consumption or perhaps a reduced budget deficit), the economy will slow because of decreased demand for goods and services. In the long run the effect is positive, Gramlich argues: prices fall to match the drop in demand. A lower price for money—interest rates—stimulates investment, so that more goods and services can be produced: economic growth accelerates. According to this view, some degree of reduced growth—if not an outright recession—is in the cards regardless of consumer spending and saving habits.

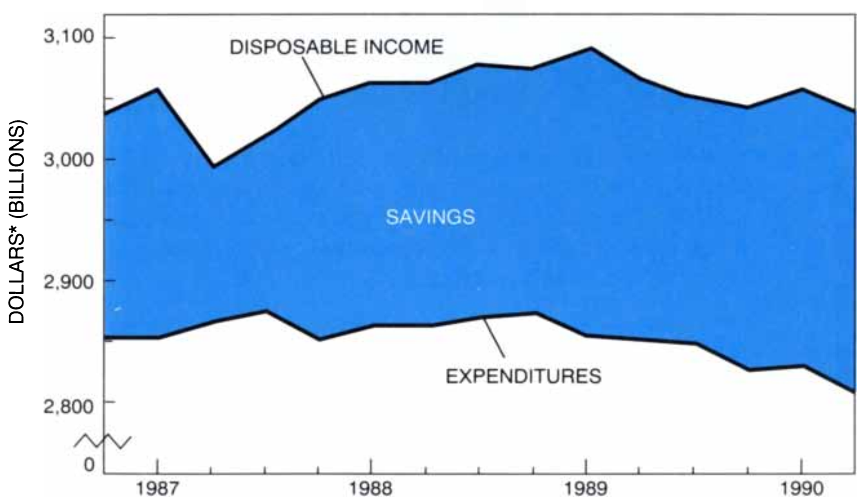
Worse yet, says Barry P. Bosworth of the Brookings Institution, the short-term pain caused by lower spending may not bring the economy long-term growth. Increased savings—such as they are—cannot be translated into investment as long as the Federal Reserve keeps interest rates high to rein in inflation, he contends. People are unlikely to channel money into potentially risky capital investment when they can earn just as much in certificates of deposit. Net domestic investment fell from 5.3 percent of national income in 1988 to 4.7 percent in 1989, even as savings rose by a similar margin. Yet Bosworth doubts that the Federal Reserve would lower interest rates to stimulate investment and economic activity.

Stimulating the economy may not be an option. Taylor points out that neither the government nor consumers saved enough during the mid-1980s period of high growth. If the economy slows down, tax revenues will fall—and so attempts to revive it by increasing government spending or lowering taxes could dramatically increase the size of the deficit. That, in turn, would hurt future growth. Thanks to years of underinvestment, the U.S. simply may not have the resources to support both long-term growth and spending sufficient to bring the economy to the "soft landing" policymakers hoped for before the trauma in the Middle East.

Although the economists' consensus is bleak, consumers may be solving the dilemma of spending and saving themselves. According to Fabian Linden of the Conference Board, consumer confidence appears to be recovering from the plunge induced by Iraq's invasion of Kuwait on August 2. Consumers have an impressive record in predicting the near-term behavior of the economy. The statistics may paint a darkening picture, Linden says, but "mercifully, the ordinary citizen is not an economist."

—Elizabeth Corcoran and Paul Wallich

INCREASED SAVINGS DRIVE DOWN CONSUMPTION



*Adjusted for increase in personal income from 1987 to 1990.

SOURCE: U.S. DEPARTMENT OF COMMERCE

CARE *of the* SURGICAL PATIENT

from SCIENTIFIC AMERICAN *Medicine*

Because the quality of your care depends on the quality of your information.

Treating pre and postoperative patients poses a unique set of challenges. Yet in one way it's no different than any other practice issue. Doing it well takes the right information.

That's why SCIENTIFIC AMERICAN *Medicine* has published *CARE of the SURGICAL PATIENT*.

The definitive resource on pre and postoperative care.

CARE of the SURGICAL PATIENT gives you ready access to the most authoritative and current information on pre and postoperative standards available anywhere.

Written and designed by the American College of Surgeons' Committee on Pre and Postoperative Care, *CARE of the SURGICAL PATIENT* provides over 1,500 pages of practical information on both critical and elective care.

And, *CARE of the SURGICAL PATIENT* is updated three times a year, with each author reviewing his own specialty. Updates include significant information on topics such as infection control, the latest indications for nutritional support and for administration of blood products, new antibiotics regimens.

In short, *CARE of the SURGICAL PATIENT* presents the standards for perioperative treatment. You simply won't find a more important resource. Or one organized in such an intelligent way.

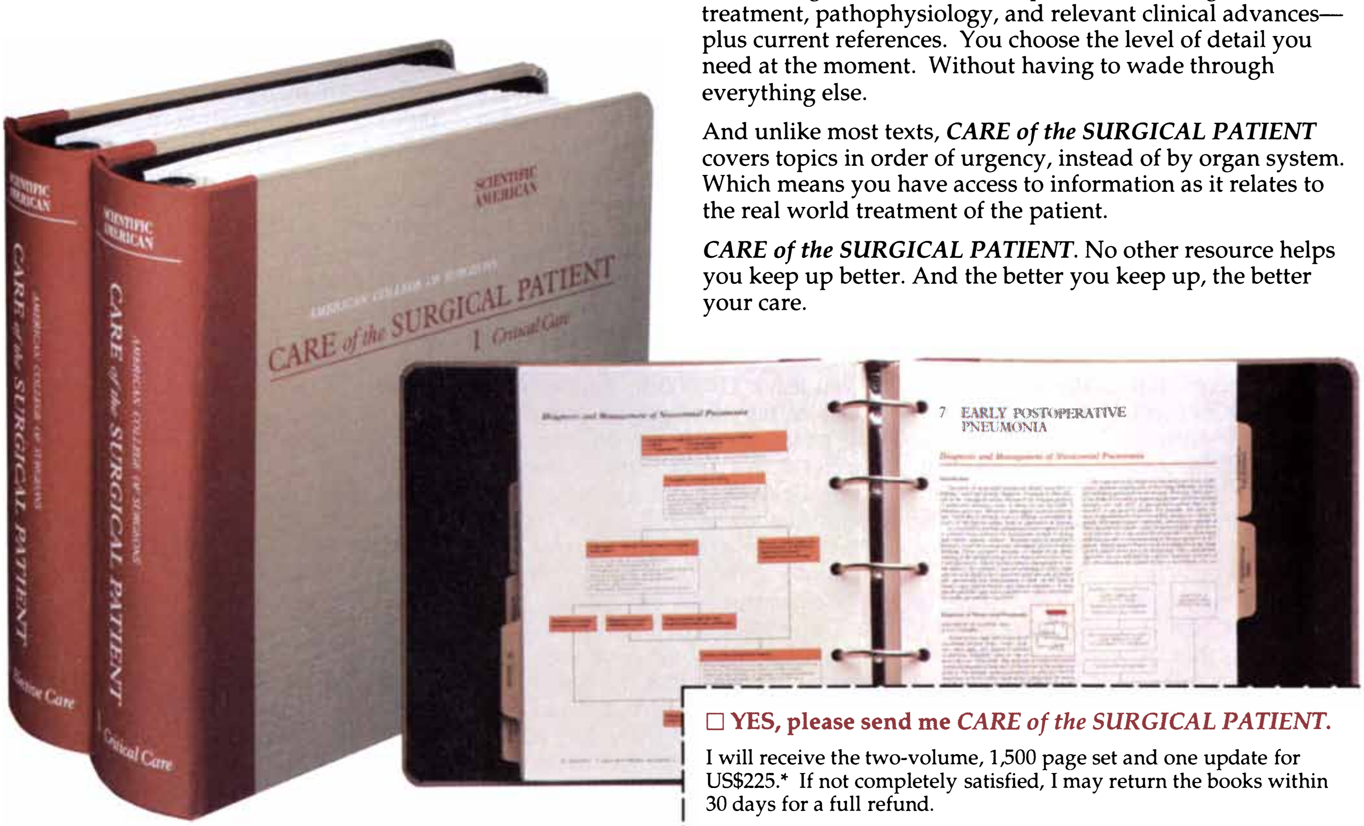
A unique system for rapid information retrieval.

CARE of the SURGICAL PATIENT gets you the information you need, the way you need it. Quickly. And intelligently.

Concise algorithms, detailed explanations of diagnosis and treatment, pathophysiology, and relevant clinical advances—plus current references. You choose the level of detail you need at the moment. Without having to wade through everything else.

And unlike most texts, *CARE of the SURGICAL PATIENT* covers topics in order of urgency, instead of by organ system. Which means you have access to information as it relates to the real world treatment of the patient.

CARE of the SURGICAL PATIENT. No other resource helps you keep up better. And the better you keep up, the better your care.



Try CARE of the SURGICAL PATIENT free for 30 days.

You'll find it the most valuable resource on pre and postoperative care that's ever been published. And if you're not satisfied, just return it. No risk. No obligation.

Each update includes revised chapters, new index, and a newsletter highlighting the new developments. Continuation of the updating service is $90 for three updates per year.

Call toll free 1-800-345-8112

☐ **YES, please send me *CARE of the SURGICAL PATIENT.***

I will receive the two-volume, 1,500 page set and one update for US$225.* If not completely satisfied, I may return the books within 30 days for a full refund.

☐Check enclosed* ☐MasterCard ☐VISA ☐Bill me P.O.#______

Acct. #____________________ Exp. Date______

Name____________________

Address____________________

____

____________________ Zip______

Signature____________________ Specialty__________

*Add sales tax for DC, IA, IL, MA, MI or NY. Institutions: enclose tax exempt certificate if applicable. Allow 6 weeks for delivery. For Canada: please make payment in U.S. funds or equivalent. Foreign: U.S. funds only.

SCIENTIFIC AMERICAN MEDICINE 415 Madison Avenue, New York, NY 10017

7543

MATHEMATICAL RECREATIONS

A compendium of math abuse from around the world

by A. K. Dewdney

"In the space of one hundred and seventy-six years the Lower Mississippi has shortened itself two hundred and forty-two miles. That is an average of a trifle over one mile and a third per year. Therefore, any calm person, who is not blind or idiotic, can see that... just a million years ago next November, the Lower Mississippi River was upwards of one million three hundred thousand miles long."

—MARK TWAIN,
Life on the Mississippi

Since March, when I first described the abuses of mathematics in this department, hundreds of readers have sent me examples of math crimes, scams and malpractice. The culprits are mainly corporations, institutions and governments. But is the abuse intentional or accidental? I will let readers judge for themselves as they peruse this miniature catalogue of mathematical horrors. To combat the perversion of percents, the twisting of logic, the misrepresentation of numbers and the distortion of charts, columns such as this may act briefly to stem the tide of ignorance that threatens to drown us. Only educational systems can save us in the long run.

People run the risk of abusing mathematics when they will not, do not or cannot correctly use the numbers and relations they encounter every day. Even those who are not comfortable with geometry and algebra can avoid math abuse if they approach mathematical situations with a certain degree of skepticism. Russell R. Bergquist of Tavares, Fla., and Sanford H. Lefkowitz of Columbus, Ohio, are particularly suspicious of the claims made by car dealers who try to persuade the buyer of a new vehicle to finance the purchase rather than pay cash. Bergquist gave, and I have embellished, an amusing account of a conversation between Sam, a car dealer, and Joe, a buyer:

"Tell me, Joe, why do you want to pay the full $10,000 in cash for your new car?" Sam asks.

"To save the cost of financing," Joe replies. "I might finance with you if your rate of 11 percent were lower than the bank's interest rate."

"But don't you realize you can still save money by financing?" Sam counters. He swings around to his computer terminal, types in the numbers and soon presents Joe with a neat printout. It states that a monthly payment of $327.39 for 36 months would pay off the $10,000 at a cost of $1,786.04. It further claims that the $10,000 left in a savings account at 7.5 percent would earn $2,514.46. "As I was saying, Joe, if you leave $10,000 in your bank account rather than paying today, you can make $728.42, the difference between the two costs," Sam declares.

A little baffled, Joe inquires, "What about the $327.39 I pay a month? That has to come out of my bank account ...."

"Well, sure, you are paying for a car," Sam interjects, "but the payment doesn't have to come from your bank account. It could come from anywhere. Maybe you could save it from your weekly paycheck. Frankly, Joe, you are the first person I've talked to who doesn't understand the system."

Joe, feeling a bit intimidated and confused, wisely decides to go home and think things over. He writes down the two plans side by side on a sheet of paper. Under Sam's plan, Joe keeps $12,514.46 at the end of three years. Under his own plan, Joe buys the car outright and invests the monthly payments of $327.39 at 7.5 percent. At the end of three years, he has $13,171.35. Joe realizes Sam's deception. Sam neglected to discuss the profit Joe could earn if the car payments were invested in an interest-bearing account.

The innumeracy I call compound blindness causes many an eye to cloud over when interest-bearing schemes are mentioned. Several readers wrote me about schemes that sounded very good until one takes inflation into account, not to mention taxes. In fact, I am ready to propose a law of inflation and taxes, tongue firmly in cheek:

The Law of Zero Return
return on investment = inflation loss + taxes

Suppose, for example, that the interest rate is 9 percent, that inflation runs at 5.5 percent and that your tax rate on interest happens to be 35 percent. If you invest $10,000 over a year, you would receive $10,900. But a tax bite of 35 percent out of the $900 return would leave you with only $585. Because of inflation, however, the resulting $10,585 would be worth only $10,002.83!

Even those who are aware of the effects of compounding can be misled. Advertisers, like salespeople, are notorious for manipulating numbers to hide certain information that makes a claim less appealing than it first appears. Homer B. Clay of Phoenix, Ariz., reminds us of the television commercial that features the preferred headache remedy of a random sample of doctors stranded on a desert island. According to the sponsor of brand X, more than 50 percent of the unfortunate physicians prefer it to brand Y, the next most popular brand. What the sponsor did not explicitly reveal was that its survey gave doctors a third choice, brand Z, which was nothing more than a variation of Y. As Clay recalls, the medication contained in brands Y and Z beat brand X by about the same amount as X beat Y.

By concealing information, advertisers can prove virtually anything. A few years ago an advertiser for a certain well-known brand A of whiskey claimed that in a blind taste test, half of the drinkers of another brand of Scotch actually preferred brand A. What, asks Alexander Martschenko of the U.S. embassy in Stockholm, does this really mean? The result could have been ob-

tained by flipping a coin, he says. This does not mean the outcome is random, but it should alert one's suspicions. After all, many consumers cannot readily distinguish their favorite brand of beverage from competing brands, much as they like to believe they can. Last year the Canadian Broadcasting Corporation ran a documentary about beer drinkers who claimed to be connoisseurs of the various brands. In blind taste tests the cognoscenti consistently confused one brand with another!

A number of drink manufacturers exploit the public's insensitive taste buds, especially when the drinks are cold. For instance, a well-known maker of soft drinks challenges customers to say which of two brands they prefer in a taste test, their brand or that of the main competition. Given the relative sameness of taste, roughly half the drinkers of the competing brand say they prefer the taste of the advertised brand. This kind of advertising is effective because most consumers have no knowledge of the taste-confusion phenomenon. In fact, having a "favorite" brand almost guarantees an unwillingness to admit that the preference is based on anything but taste.

The public might be prone to believe that advertisers and large corporations are the biggest abusers of mathematics, but according to my mail, governments and media organizations are just as culpable. An announcement in a Swedish newspaper read: "SIX TIMES SAFER! Last year 35 people were drowned in accidents in small boats. Only five were wearing life jackets, whereas the rest were not. Therefore, always wear a life jacket when boating."

In spite of the laudable import of the message, what, asks Poul Printz, a Danish reader, can one actually deduce from the numbers? Is it possible that six times as many Swedes sail without life jackets as those who sail with them? It is ironic to think that if this were so and if the proportion of Swedes wearing them were to increase greatly, the statistics of the example might well be reversed. The abuse in this case could be called statistical starvation: we are simply not fed enough information to be convinced by the given numbers. Instead the media happily report what various agencies tell them.

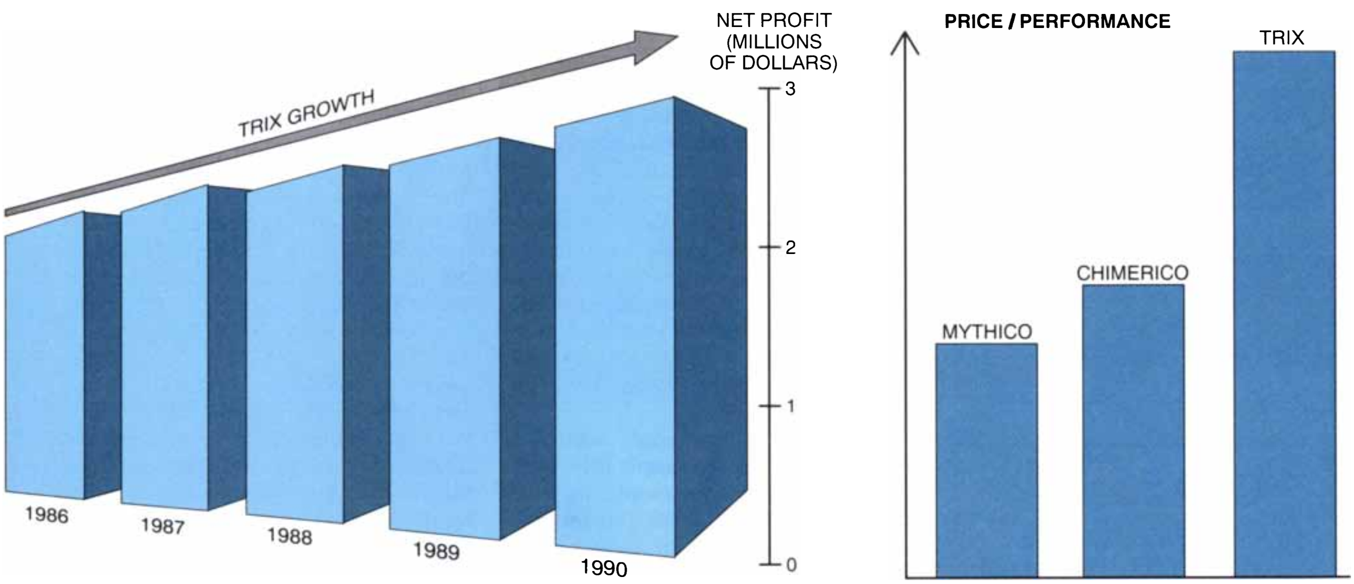
Once one puts it all together, it is easy to see how the media can play up the numbers, an all-too-common abuse I call numerical inflation. Joseph Childers of Bryte, Calif., has passed along documented claims about the percentages of fatal traffic accidents having various causes: cocaine, 20 percent (a New York newspaper); marijuana, 25 percent (Drug Enforcement Administration); alcohol, 50 percent (California Highway Patrol); sleepiness, 35 percent (sleep researchers); speeding, 85 percent (National Transportation Safety Board); smoking, 50 percent (National Highway Traffic Safety Administration); suicide, 35 percent (suicide researchers); mechanical failure, 20 percent (New York State Department of Motor Vehicles).

Admittedly, the categories are far from mutually exclusive, so we are not surprised that the percentages add up to more than 100. Moreover, some of the figures may apply more to highways than city streets. Nevertheless, given these percentages, I would not be surprised if every year thousands of drivers, in an attempt to kill themselves, snort some cocaine, smoke several joints, quaff a pint of liquor, puff on a cigarette, then climb into a 20-year-old jalopy, accelerate well past the speed limit and fall asleep at the wheel.

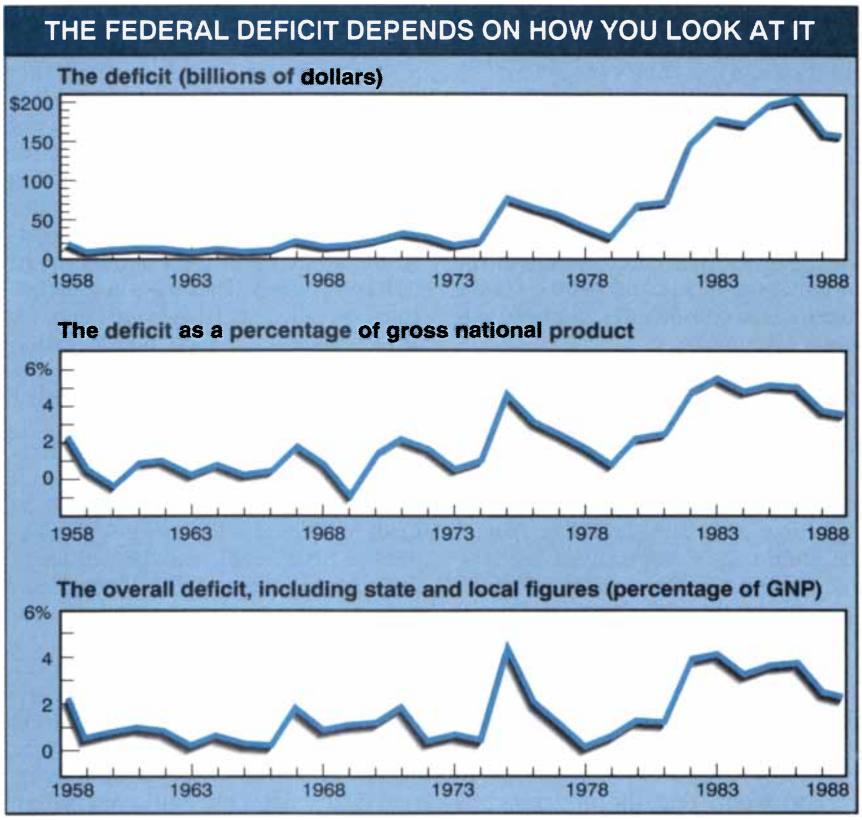
Paying attention to big numbers gives many people trouble, especially if they involve percentages. Few enough may have blinked at the newspaper headline passed along to me by Frank Palmer of Chicago: "SEVEN ITALIANS OUT OF TEN HAVE COMMITTED ADULTERY" After all, 49 percent of Italian male and 21 percent of Italian female respondents in a survey confessed to extramarital affairs. As everyone knows, $49 + 21 = 70$.

Simply adding the percentages ignores the relative numbers of Italian men and women (assuming the sample was representative in the first place). If the proportion is roughly half and half, then 24.5 percent of the Italian population consists of men who have been unfaithful and 10.5 percent consists of women who have been unfaithful. Now the percentages can be added because the two groups represent nonoverlapping parts of the same population: $24.5 + 10.5 = 35$.

Here the abuse may be mostly unconscious, although a certain tendency of the media to exaggerate may (again) have played a role. The perpetrators were presumably subject to the same form of innumeracy as the victims: ratioitis. The innumeracy applies not only to percentages but also to ratios, fractions, proportions and even averages. None of these can be added or multiplied with quite the same abandon that ordinary numbers can. Even edu-



Trix Corporation: outshining the competition and growing fast



Source: Commerce Department adapted from The New York Times

Quiz: In what years did the U.S. produce a negative GNP?

cators can fall victim to ratioitis. For example, George Barry of Saratoga, Calif., cites a statement from an education survey: "Although, in the seventies, test scores were down nearly 60 percent, they have since rebounded by over 70 percent!"

Again the tendency is simply to combine the two figures as if they were ordinary numbers: subtract the 60 percent, then add the 70 percent. But apply the claimed drop and rebound to a specific number, say 85. If the number 85 loses 60 percent of its value, it is decreased by 51, becoming 34. If the number 34 now "rebounds" by 70 percent, it gains 70 percent of its value, increasing by 23.8 to 57.8, well below the original 85. Barry calls the abuse of percentages "pernicious" and writes, "One should surely be forgiven for failing to grasp that the scores are now less than 70 percent of what they once were." How ironic that those reporting on the progress of education should illustrate its lack.

I want to leave readers with two kinds of recreations, simple little exercises to tone up those mental muscles that were meant to twitch in the jungle of life and warn of falsehood or stupidity. The first kind of recreation concerns chart abuse, the deliberate or unintentional distortion of data by graphic means. How many readers can identify the abuse in each of the diagrams that illustrate this article? Each diagram might appear innocent at first glance, but a second look should make your mind twitch.

The second kind of recreation comes from Robert M. Martin, who teaches a course on the logic of probability and scientific reasoning at Dalhousie University in Halifax, Nova Scotia. Martin has tripped his students up more than once with unpretentious little posers such as the following three.

- Someone shuffles an ace, king and queen, places them face down on a table and—oops—accidentally knocks one of them to the floor in an entirely random way. The card lands face down. What is the probability of drawing an ace from the cards that remain on the table?
- People usually work eight hours a day. That is one third of the 24-hour day, so people actually work the equivalent of one third of the 365 days in the year, about 122 days. But people usually do not work weekends. That means two days off a week, or a total of 104 days off a year. Subtracting this from the 122 days leaves only 18 days. Since the average vacation is about 10 days long, that leaves eight days. Because there are at least this

many holidays in a year, no one works!

• If a bottle and a cork together cost $1.10, and the bottle costs $1.00 more than the cork, how much does the cork cost? When answering this last question (wrongly), a student jeeringly penciled, "Tough stuff!"

The examples of abuse are but symptoms of a general ignorance of mathematics—indeed, of science as a whole. Among the industrialized nations, the disease appears to be most advanced in the U.S., where students on average score lower on standardized tests than their peers in other industrialized nations. We are well aware of an education crisis, but what can be done?

Lynn Arthur Steen, a mathematician at St. Olaf College in Northfield, Minn., is president of the American Mathematical Society and a well-recognized authority on mathematics education from kindergarten through graduate school. According to Steen, children leave primary school and even high school either confused by mathematics or, at best, as mechanical manipulators of the formulas they have learned. Very few students in the U.S. have what Steen calls "good number sense" or "good symbol sense."

Good number sense reflects precisely the form of numeracy required to combat math abuse. It means an ability to understand relative magnitudes, exponential growth and the meaning of ratios and percentages, just to name a few. Students have good symbol sense if they can recognize that a mathematical result is incorrect because they know approximately what the answer should be. Such a student can think algorithmically, embracing simple abstract processes with relative ease.

The National Council of Teachers of Mathematics has recommended standards for students, and the Mathematical Sciences Education Board, a branch of the National Academy of Sciences, has been actively promoting mathematics education in the U.S. Although a few schools are doing a terrific job of teaching mathematics, Steen observes, most are drifting steadily deeper into a black hole created by public indifference. If the U.S. could get as fired up about the crisis in mathematics education as it has about environmental issues, Steen says, we might be able to turn things around.

FURTHER READING

EVERYBODY COUNTS: A REPORT TO THE NATION ON THE FUTURE OF MATHEMATICS EDUCATION. National Academy Press, 1989.

SCIENTIFIC AMERICAN
In Other Languages

LE SCIENZE
L. 5,000/copy L. 50,000/year L. 65,000/(abroad)
Editorial, subscription correspondence:
Le Scienze S.p.A., Piazza della Repubblica, 8
20121 Milano, Italy
Advertising correspondence:
Mondadori Pubblicità S.p.A., Via Rivoltana 8
20090 Segrate, Italy

サイエンス
Y 980/copy Y 10,753/year Y 13,393/(abroad)
Editorial, subscription, advertising correspondence:
Nikkei Science, Inc.
No. 9-5, 1-Chome, Otemachi
Chiyoda-ku, Tokyo, Japan

INVESTIGACION Y
CIENCIA
500 Ptas/copy 5500 Ptas/year 6200 Ptas (abroad)
Editorial, subscription, advertising correspondence:
Prensa Científica, S.A.
Viladomat, 291 6º 1ª
08029 Barcelona, Spain

POUR LA
SCIENCE
29 FF/copy 296FF/year 350FF/year (abroad)
Editorial, subscription, advertising correspondence:
Pour la Science S.A.R.L.,
8, rue Férou,
75006 Paris, France

Spektrum DER WISSENSCHAFT
10.50DM/copy 105.60DM/year 115.20DM/(abroad)
Editorial, subscription correspondence:
Spektrum der Wissenschaft
Verlagsgesellschaft mbH
Moenchhofstrasse 15
D-6900 Heidelberg
Federal Republic of Germany
Advertising correspondence:
Gesellschaft für Wirtschaftspublizistik GWP mbH
Kasernenstrasse 67
D-4000 Duesseldorf,
Federal Republic of Germany

科学
3.80RMB/copy 45.60RMB/year $48/(abroad)
Editorial subscription correspondence:
ISTIC-Chongqing Branch, P.O. Box 2104,
Chongqing, People's Republic of China

В МИРЕ НАУКИ
2R/copy 24R/year $78.50/(abroad)
Editorial correspondence:
MIR Publishers
2, Pervy Rizhsky Pereulok
129820 Moscow U.S.S.R.
Subscription correspondence:
Victor Kamkin, Inc.
12224 Parklawn Drive
Rockville, MD 20852, USA

TUDOMÁNY
98Ft/copy 1,176Ft/year 2,100Ft/(abroad)
Editorial correspondence:
TUDOMÁNY
H-1536 Budapest, Pf 338
Hungary
Subscription correspondence:
"KULTURA"
H-3891 Budapest, Pf. 149
Hungary

مجلة العلوم
1KD/copy 10KD/year $40/(abroad)
Editorial, subscription, advertising correspondence:
MAJALLAT AL-OLOOM
P.O. BOX 20856 Safat,
13069 – Kuwait

Advertising correspondence all editions:
SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue
New York, NY 10017
Telephone: (212) 754-0550 Telex: 236115

BOOKS

The physicist with the purest soul, a history of hunger, optical science in art

by Philip Morrison

DIRAC: A SCIENTIFIC BIOGRAPHY, by Helge S. Kragh. Cambridge University Press, 1990 ($44.50).

Niels Bohr once remarked that "of all physicists, Dirac has the purest soul." Paul A. M. Dirac "was tall, gaunt, awkward, and extremely taciturn.... In conversation he was invariably polite.... One was never sure that he would say something intelligible." Ever since childhood he had used few words; he spoke utterly directly, without the slightest tinge of manner or irony. He took it that other people did the same.

An interview written in 1929 by the columnist of a Madison paper when Dirac lectured on the University of Wisconsin campus resembles the tale of some wry Zen master. The piece might be taken as wholly humorous invention, were it not so utterly characteristic. "'Professor,' says I, '...you have quite a few letters in front of your last name. Do they stand for anything in particular?' 'No,' says he. 'You mean I can write my own ticket?' 'Yes,' says he. 'Will it be all right if I say that P.A.M. stands for Poincare Aloysius Mussolini?' 'Yes,' says he." And in closing, "'Do you ever run across a fellow that even you cant [sic] understand?' 'Yes,' says he.... 'Do you mind releasing to me who he is?' 'Weyl,' says he. The interview came to a sudden end...for the doctor pulled out his watch.... He let loose a smile as we parted and I knew that all the time he had been talking to me he was solving some problem no one else could touch."

Dirac was born in England in 1902 into a household dominated by a strict, inexpressive and unloving father. His father had run away from his home in Switzerland in his youth, first to study at Geneva and then to England, where he would teach French for the rest of his life. Paul grew up a silent and dependent boy who did what he was told. All he liked was mathematics, but at 16 he uncomplainingly entered on the course of electrical engineering at Bristol University, in the very college building where his father taught.

It was the drama of general relativity that set him on course. He recalled how "relativity came along as a wonderful idea...a new domain of thought...an escape from the war.... I was caught up in [the] excitement" of the 1919 test of Einstein's starlight bending made by Eddington's eclipse expedition. Dirac studied relativity on his own; in 1921 he graduated with first-class honors, but in that depression year he could find no job as an engineer and went happily back to study mathematics on a free-tuition grant.

In 1923 he entered graduate school at Cambridge. Still a student, although with half a dozen papers to his name, in August, 1925, he read in proof Werner Heisenberg's historic first exposition of matrix mechanics. Three months later Dirac had published the first of his own dozen original contributions to the breakneck growth of the theory. "He has no idea how difficult his papers are for the normal human being," Erwin Schrödinger complained to Bohr in the fall of 1926. The "unknown English youngster" was at once accepted by the coterie of quantum theorists; soon he was quite at home in Copenhagen, Göttingen and Leiden. Most of Dirac's results during those years of cornucopia turned out to have been anticipated, but the generality and originality of his approach were unmatched even by his golden peers.

In the fall of 1927 he "suddenly realized" how it would be possible to retain the required linearity of the quantum equations and yet to introduce the square root of the energy essential for a relativistic theory. The beautiful device, well described in the text for any who tolerate a little algebra, was to use not one single equation but a number of related equations at once, opening many simple relations that could be arranged to meet all the conditions needed. The natural quantities of Dirac's relativistic electron theory turned out to be 4×4 matrices, interpretable because 2×2 matrices were already understood to describe electron spin. The new equations fitted the hydrogen spectrum with new accuracy and predicted unmistakably a new counterpart to the electron, its positively charged antiparticle. Everyone, Dirac included, thought they knew just what that positive particle had to be: the proton!

The physicists then saw matter at its simplest; Dirac, building on the firm ground won by Einstein and Heisenberg, had found it all in one elegant mathematical gesture. The time from the first quantum oscillator to Dirac's relativistic electrons had been only an exhilarating two years and three months.

But the too hasty physicists would have to recant at leisure. First, Dirac and the rest had to accept that the antielectron was not at all the proton, but a new and unexpected positron, discovered in the cloud chamber in 1932. Then it took almost 20 years for a new generation to build the edifice of quantum electrodynamics, established on Dirac's equation but taking into account a world of photons. Its predictions are now tested over the widest range of experiment, from cosmic radiation showers with a million positron-electron pairs, to a 10-figure precision fit to the properties of a real interacting electron.

After 1934, Dirac's productivity fell sharply. He was still a source of good new ideas; his secret war work on centrifuges for uranium isotope separation is now public and classic in that engineering field. (He was never in the least impractical about the physical world, in spite of his social reticence and his politely tactless speech.) He was a pioneer in quantum studies of the elusive magnetic monopole, even sketching out its topology. He offered an original cosmological proposition that suggested rational grounds for expecting a change in the "constants" of physics over time. But he remained out of the strong new currents of theory; his last widely cited paper was published in 1950.

The 1980s found the elderly Dirac an admired sage, articulately unconvinced by the instrumental victories of present theories of particles and the forces between them. He still saw them as "complicated and ugly." For him they were hopelessly marred by renormalization; their inescapable but cunningly isolated infinities all shrugged off for the while. Dirac often wrote on the role of mathematical beauty and of

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.50 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/89. $1.50 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA. 51593. Telephone inquiries: 1-800-333-1199, U.S. only; other 515-247-7631/32. The date of the last issue on subscriptions appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, a mailing label of a recent issue) as well as the new one.

Name ______________________

New Address

Street ______________________

City ______________________

State and ZIP ______________________

Old Address

Street ______________________

City ______________________

State and ZIP ______________________

the unity of method as the true touchstones of high theory.

His aesthetic has its brilliant theorist-followers still; it is only that their beauties are not his, even though the novel algebra of their fields describes our world quite well, abuzz with diverse particles as it is. Dirac's famous discovery had also made elegant physical use of an unexpected algebra. Maybe the world will turn out beautifully down among the minute strings, its beauties far subtler than the 60-year-old Dirac equation. This volume is a delightful and instructive evocation of a remarkable human being by way of a judicious and documented passage downstream following the flow of modern physics.

HUNGER IN HISTORY: FOOD SHORTAGE, POVERTY, AND DEPRIVATION. General editor, Lucile F. Newman. Associate editors, William Crossgrove, Robert W. Kates, Robley Matthews and Sara Millman. Basil Blackwell, Inc., 1990 ($39.95).

On the island of Samos a long inscription in marble has honored a politician ever since the middle of the third century B.C. To help finance public grain purchases during three distinct times of food shortages, he had "advanced all the money required for a reserve fund as the people had resolved." The lesson we draw is not of unique civic merit; "the Greek world was full of men like Boulagoras of Samos," men memorialized by hundreds of similar inscriptions. Rather we infer that crop crises were common in the classical Mediterranean world. But we know as well that famines were rare. Public benefactors, of necessity rulers and rich men, acted to arrest "the slide from crisis into catastrophe." Modern rainfall data in the region offer striking confirmation: the failure rate for a wheat crop is about one year in four, barley fails about one year in 20, but concurrent failure of both grains for two years in a row is not to be expected over many centuries. An affordable grain reserve works; the state can indeed moderate public hunger.

This unusual volume reports a recent year-long interdisciplinary seminar at Brown University, in which the researchers viewed history through a zoom lens. Four sections look closely at four powers of 10 in time: from 200 centuries ago until the still puzzling worldwide rise of agriculture; 10 or 20 centuries back, at those urban societies that developed agriculture; one or two centuries ago at Europe and its colonial empire; and, finally, the decades just past. Method and evidence change sharply with focal length.

Half a dozen well-argued if somewhat stiffly written essays of analysis and overview by the editors and their colleagues are interspersed among a similar number of tighter studies by invited specialists of wide reputation. One detailed demographic account is built mainly from finding and counting a great many small Maya ruins hidden in the green lowland forests a couple of hundred miles around great Tikal. The data allow reconstruction of the Maya population there from 300 B.C. to the present. A swift rise to a peak of three million people at about A.D. 800 implies intensive land use; aerial photograph and spade testify to widespread cropping of raised beds and terraced wetlands. Then the number of people fell steadily, by a factor of 30 even before Cortés landed! Advanced agriculture faded to the low-intensity shifting culture of maize still to be seen there; we have "simply no convincing arguments" for the cause of that decline. Efforts to support a Malthusian collapse from land overuse or even from the social failure of elite management are explicitly rejected by this archaeologist. We need to know more history; the classic Mayas, like the rest of us, were complicated.

A zoom lens allows both close-ups and a longer view. The subtitle lists three grades of hunger that frame the entire treatment: famine, or regional food scarcity for all; food poverty, when only certain households are in real want; and deprivation, when only certain individuals are denied by society or by biology. Food poverty remains for a tenth of our fellows, while the head count mounts, its "rates down, numbers high and constant."

Our big systems fail big, too. Our century has seen mass starvation. In 1944–1945, for example, food was wielded as a weapon against Holland to bring famine, absent from western Europe for 125 years. Between 1959 and 1961 one of the worst famines on record belied Chairman Mao's misguided command for a national Great Leap Forward, claiming 15 million or more deaths in China.

In modern India wide freedom of public discussion has meant that no regional sign of famine escapes early warning, and indeed famine is no longer present there. But in India (and is it absent from your city, American reader?), food poverty—familiar, long endured, based on accepted social category—is endemic. In contrast, the very principles of the People's Republic of

China speak against toleration of food poverty, and a fine record of genuine amelioration is there to be read. But the harsh truth of China's tragic years of crop failure was largely hidden by an implacable central authority, which claimed its customary grainloads out of the ruined harvests. Even after famine was recognized, there was no opposition to raise an outcry. The terrifying ravages were hushed up for two more decades.

Deprivation is sometimes hidden in deficiencies of supply less patent than simple hunger. A long, patient queue of the iodine-hungry dwell along the high ranges from Peru across Central Africa, along the Himalayas and on to far New Guinea. Their unwitting impairment comes from the unmet needs of young women who while with child lacked a few pennies' worth of iodine from the distant sea.

Overall scarcity is mainly on the supply side, the lean harvest of Jacob's dream. More important in today's world is failure of demand. Everyone is conditionally entitled to food, whether because they grow it, or share a crop, (or their parents do), or because they hold cash or valid stamps. Under stress that entitlement chain is broken. Amartya Sen, who has pressed this model persuasively for a decade, makes clear that the famine of 1974 in Bangladesh, for one example, was a sharp failure of entitlement. The year saw a good if delayed harvest; the undernourished farm laborers confronted famine amid filled granaries. Their real wages had dwindled under the wide unemployment of a wet planting season and a speculative rise in prices. All food chains have many vulnerable links; those who would forge a stronger chain must tug at each link. Is help for hunger best offered as new seed or ready food or cash or paid work or resettlement or new land or simply peacemaking?

A very rich text, with valuable if cramped graphics, the book contains a concluding chapter (by Kates and Millman) that is a model for its width of vision and its pointed summary, memorably phrased.

THE SCIENCE OF ART: OPTICAL THEMES IN WESTERN ART FROM BRUNELLESCHI TO SEURAT, by Martin Kemp. Yale University Press, 1990 ($60).

Renaissance Man is everyday banter by now, but this beautifully made volume by an art historian at the University of St. Andrews in Scotland shows us the unity of the visual study of nature—the exalted mutual task of Renaissance science and art. Professor Kemp's work on Leonardo da Vinci is well known. Here his canvas is wider still: the works and the writings of a great many artists refract the relevant visual science.

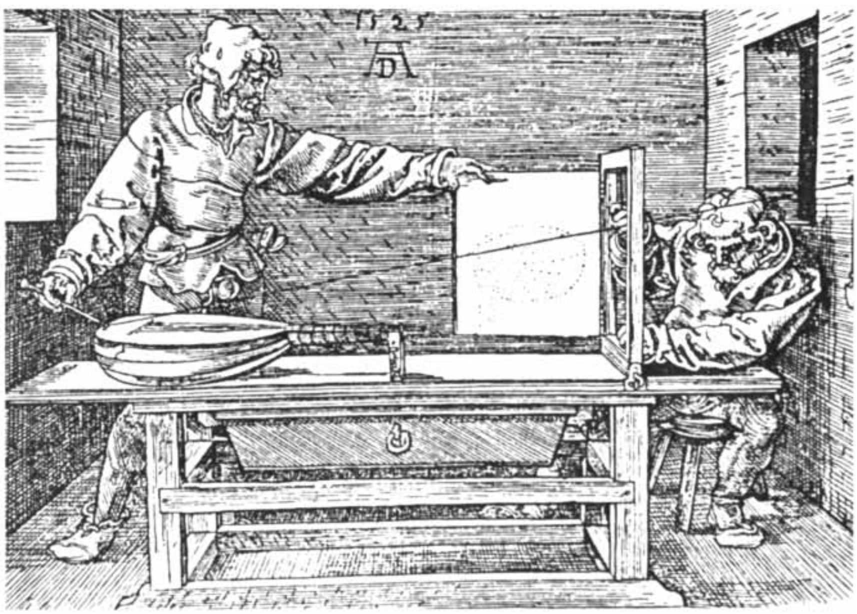
The heavy book signals artistic intentions: its big, coated pages are filled with diagrams and reproductions, and the color plates glow with exemplary forms by the masters. Its readers are encouraged to be braver than most students of the arts: "I am aware that this is not an easy book," Kemp writes. Assuredly none of its diagrams and calculations are beyond any educated reader able to admit that some pages may take more time than others.

The seven chapters group into three parts: the first is the longest, the rise and elaboration of the sightline—the "beguilingly simple" linear perspective invented by Filippo Brunelleschi and displayed by him in a public demonstration, complete with a mirror and a purposefully doctored painting of his own of the area near the Cathedral of Florence before 1413. By 1426 the first masterpiece of Florentine perspective could be admired: here is Masaccio's great painting *Trinity* in witness, accompanied by diagrams that show the implied structure of rays.

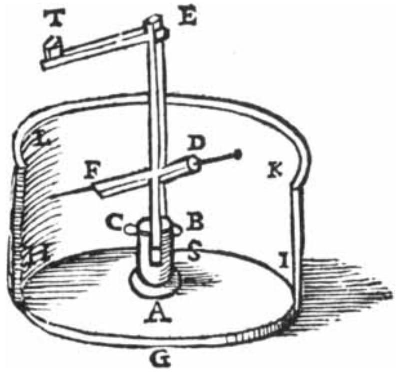
Generation by generation, the artists are cited, illustrated, compared, even mapped. Piero della Francesca published treatises with original results both in pure and in applied mathematics. What were Galileo's moon paintings but careful wash perspectives of shadows cast on curved surfaces? Albrecht Dürer taught the Tuscan art explicitly, hoping to reach young German artists "who have grown up in ignorance like unpruned trees." The visual analyses and texts record the growing complexity of the painters' work; they soon knew a naive, one-eyed, stationary and small-angle perspective was far from enough, and the artful subtleties flourished: multiplicity of all kinds, texture, shadow, free approximations.

The famous foreshortened lute Dürer used is shown as plotted tediously point by point. Workable instruments for perspective go back to the beginning. Their fascinating story forms the second part of the book. The Nuremberg jeweler-artist Wenzel Jamnitzer, whose extraordinary perspectives of solid polyhedra are daunting in their number and intricacy of detail, is portrayed using a kind of pantographic linkwork that plotted points on paper for the artist as he simply sighted on the scene.

The convex lens—"as used by old men in their spectacles"—was by 1570 projecting real images for the painter. The paintings of Jan Vermeer, all those minute details, varied scale, matte and glistening textures, reflections, shadows, even to lenslike circles of confusion in the highlights, have long been compared with images made by the camera obscura. A recent study has shown how Vermeer's rooms can be mapped from the paintings to provide a seductive model of his actual practice: first he plotted an overall view by camera lens, and then he moved stool



Two draughtsmen plot points for the drawing of a lute in foreshortening, from Dürer's Underweyssung, *1525.*



Baldassare Lanci's device for the perspectival projection of a view on the inside of a cylinder, from Vignola's Le due regole, *1583.*

and easel to predictable place after place around the room to render the lighting and those matchless details. His was no mechanized art, but a long string of aesthetic choices.

The third and shortest part of the text treats "the color of light" as the painters saw it. Theory rules, and it is the writings of the artists and philosophers that carry most of the evidence. The classical tradition of Aristotle could not really explain pigments and color. Newton opened the entire topic with a rational spectrum, but the subjectivity of color and the complexity of the entire perceptual system led to a tide of theories.

By the 1880s Georges Seurat had declared himself an optical painter. He set a high standard. His famous big canvases built on campaigns of effort; the *Sunday in the Park* painting was preceded by more than 60 known drawings and studies. Their color is not emplaced in points or dots but in tiny directional dashes and strokes, like iron filings.

For our thoughtful author, this marked the high point of the artist's exploration into the optics of nature (illusions of the eye aside). Naturalism had attained such verisimilitude that it seemed its power could grow no more. For centuries, painting on the one hand and science on the other had happened to walk on the same path. Each partner was eager to reconstruct visually the truth of appearances, each accepting a common limitation to observation. But by the time of distorting mirrors and surely after Galileo's optic tube, a parting of their ways was foreseeable. Something is there behind appearances; the old and happy collusion will have to adapt to a more abstract realm if ever a wholeness of perception is to come again.

ESSAY

Competing in the world's computer market

by J. Christopher Westland

For most of the 20th century, the U.S. has innovated and prospered in computer technology. We dominate standards from microcomputers to the world's fastest supercomputers. U.S. companies currently produce more than three quarters of the world's software and software services, running a $9-billion annual trade surplus in the process. And we provide the vast majority of the world's commercial online data bases. High-technology products represent a $350-billion share of our economy. U.S. know-how continues to set global standards in processor architecture, peripherals, artificial intelligence, neural networks, optical character recognition, voice recognition, music synthesis and many other commercial technologies.

Yet after a decade of enthusiastic support for computer technology, U.S. investors are taking a bearish view of the industry. That pessimism has hurt. In 1989 investment was only $2.5 billion in computer-related venture funds, down from $3 to $4 billion the prior year. Ernst & Young's recent survey of 775 chief executive officers concluded that U.S. high-technology industries in the 1990s will rely increasingly on foreign capital. Of the CEOs surveyed, 25 percent cited capital as their main concern, but almost half (48 percent) of fledgling companies—those with less than $5 million in revenues—identified access to capital as their foremost concern. It is precisely the latter firms that are developing some of America's best and most creative computer technologies. Ingenuity, energy and talent are wasted when firms cannot package their technology in marketable products because of capital shortages.

In essence, America's investors are selling short in U.S. computer technology and impoverishing an exceptionally successful and competitive industry in the process.

To obtain capital, the U.S. computer industry has been forced to look abroad. Ernst & Young's study predicts that in the next five years, 27 percent of the industry's capital needs will be met by foreign sources, up from the current 16 percent. Roughly one quarter of the CEOs polled expected their companies to be acquired by a foreign company in the next five years.

Our competitor nations have an easier time raising capital. The Bank of Japan lends money to many firms in their computer industry at interest rates one fifth those charged by U.S. institutions. The 1988 rate of return required to cover financing for R&D in the U.S. was an alarming 20.3 percent, compared with 8.7 percent for Japanese R&D projects and 14.8 percent in Germany. This forces U.S. manufacturers into markets where payback is quick, risk is low and profits are high, resulting in a patchwork quilt of uncoordinated strategies for the development of computer technologies.

The failure of American investors to pursue U.S. computer technology has eroded the demand for patents on basic R&D. The resulting price drops have led such technology offerings to become "bargains" in the eyes of foreign investors. The Japanese computer industry, in particular, has sought out investments in American computer technology. Such investments present foreign investors with an excellent opportunity to "lock up" patent rights on many marketable and innovative technologies without incurring the risk and expense of their development. One estimate places Japanese investments in U.S.-originated patents at $10 billion; the value of these investments in the final marketplace is perhaps hundreds of times that amount.

Five factors coincide to discourage U.S. investment in computer technologies. First, returns from these investments are difficult to predict, requiring an understanding of the potential uses of that technology as well as the difficulties in its production. Even if a product is viable in the current environment, investors recognize that standards change over the product's life. The tendency to bet on a proved track record works to the disadvantage of new entrepreneurs, many of whom have the best ideas to offer.

Second, the U.S. is highly dependent on foreign technical markets to fuel its continued success. But there is every indication that the U.S. plays by a different—and generally more ethical—set of standards than do computer technology firms in other countries. The flourishing international trade in pirated and reverse-engineered software threatens the health of U.S. software producers. Third, salaries and benefits for managerial and technical personnel in the military-industrial complex may exceed those of similar jobs in the civilian sector by nearly 50 percent. This imbalance has inspired a sizable drain of talent away from consumer industries. If the military's research were transferable to commercial uses, such a situation would be less troublesome. But most military research is highly specialized, with little transferability to other areas. Fourth, negative publicity concerning lack of scientific and technological awareness among the U.S. public has cast aspersions on the technological competence of U.S. industry. And fifth, the litigious environment surrounding the U.S. computer industry has made introduction of products unpredictable.

These afflictions are not insurmountable. Changes in the way we currently approach R&D could make the industry more attractive to investors.

We should give priority to research with clearly defined products and markets. Products need not be restricted to commercial markets; in areas of fundamental research, such as high-temperature superconductors, for example, the "markets" may actually be other research projects.

We need to support production technologies concurrently with our support of product technologies. Current practice favors research that designs products through prototypes or laboratory experiments. These products often remain laboratory curiosities. Even though the products are marketable, supporting production technologies may not exist to manufacture them in commercial quantities at acceptable quality levels.

Finally, we must encourage the exchange of information on production and product technologies. Smaller companies especially need assistance to keep current with new product research and legal developments. It is here that industry and government consortia can be most effective in spurring productivity.

These suggestions represent real opportunities for our computer industry to build on its successes and remain globally competitive. By ignoring these opportunities, the U.S. will slowly, but surely, relinquish its dominance of computer markets to foreign rivals.

J. CHRISTOPHER WESTLAND is assistant professor of information systems in the School of Business Administration at the University of Southern California.