SCIENTIFIC AMERICAN

STARS & STRIPES: COMPUTER-GENERATED YACHT

$2.50

August 1987

Northwest covers

the face of Asia.

Taipei

Okinawa

Manila

Shanghai

No matter how you look at it, Northwest has Asia covered. With convenient service from over 200 U.S. cities.

Including nonstops from New York, Chicago, Seattle, San Francisco, Los Angeles, Detroit and Honolulu.

And we're the only U.S. airline to give you the comforts of an all-747 transpacific fleet. The luxuries of Regal Imperial service. The experience of 40 years of flying to Asia. And the rewards of our WORLDPERKS℠ frequent flyer program, which gives you the fastest free trip of any major airline.

These are just some of the reasons why we're America's number one choice to Asia.

So if you're looking for a familiar face in the Far East, call your travel agent or call Northwest at (800) 447-4747 for international reservations.

Look to us.

NORTHWEST

Take the Multi and run.

NEC introduces MultiSpeed.™ The world's fastest portable computer. It's designed to make you a runaway success.

MultiSpeed is multi-talented. It's small. Light. And gives you the option of running at a clock speed of either 9.54 or 4.77 MHz.

MultiSpeed can do most anything your desktop PC can. It has the most advanced LCD screen technology. So it's easier to read.

What's more, it's PC compatible and even does windows.

And MultiSpeed is a very smart buy. It comes with a bundle of built-in features that cost a bundle on other portables.

Standard Features	
Clock Speed	9.54/4.77 MHz
Memory	640K bytes
Built-in Software	5 programs
Disk Drives	Dual 720K (3.5")
Screen type	Super-twist LCD
Keybd. Compatibility	IBM PC/XT*
Numeric Keypad	Separate
Weight	11.2 lbs.

In fact, when you discover everything it comes with, you'll be delighted at what it goes for.

So if you're a person who's going places, try the portable that helps you get there ahead of everyone else.

You'll find the new MultiSpeed portable at CompuMat, ComputerLand, Connecting Point and Micro Age. For literature or the location of your nearest dealer call 1-800-447-4700.

If you have any questions or would like technical information call 1-800-NEC-SOFT. Or write NEC Home Electronics, Computer Products Division, 1255 Michael Drive, Wood Dale, IL 60191-1094.

*IBM PC/XT is a registered trademark of International Business Machines Inc.
NEC is a registered trademark of NEC Corporation.

Take the Multi and run.
MultiSpeed™

NEC®

C&C Computers and Communications

Established 1845 SCIENTIFIC AMERICAN August 1987 Volume 257 Number 2

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1987 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 RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, P.O. BOX 953, FARMINGDALE, N.Y. 11737-0000.

THE COVER

The cover shows a computer-generated image of the hull and keel of *Stars & Stripes,* the winner of the 1987 America's Cup competition (see "*Stars & Stripes,*" by John S. Letcher, Jr., John K. Marshall, James C. Oliver III and Nils Salvesen, page 34). The yacht is seen from the viewpoint of an underwater observer. All objects in the image were modeled as mathematical surfaces using the Fast Yacht computer-aided design program developed by Design Systems & Services, Inc. George Hazen of Design Systems created the surfaces representing the hull, keel, winglets and rudder to approximate the actual design, which has not yet been revealed in detail. The representations of the water surface, sails and spars were produced by Christopher Cressy of the Science Applications International Corporation. The rendering was generated on a Hewlett-Packard 320SRX Graphics Workstation.

THE ILLUSTRATIONS

Cover image courtesy of Design Systems & Services, Inc.

Page	Source
35	Design Systems & Services, Inc.
36–38	Hank Iken
39	John K. Marshall
40	Enrico Ferorelli
43	Mort Rosen, Brookhaven National Laboratory
44–50	Gabor Kiss
53	Frank Varney
54	Children's Hospital, Denver
55–60	Joan Starwood
63	James G. Hirsch
64	Royal Society of London
65–68	Dana Burns
69	Walter and Eliza Hall Institute of Medical Research (*left*); David W. Talmage, University of Colorado School of Medicine (*right*)
70	George V. Kelvin
71	Peter Rönnlund, University of Uppsala
72–73	George V. Kelvin
74	George V. Kelvin (*top*), Martin P. A. Jackson (*bottom*)
75	George V. Kelvin
76	Augusto Gansser
77–78	George V. Kelvin
79	Earth Observation Satellite Company
81–83	Andrew Christie
84	William R. Frensley, Texas Instruments, Inc.
85–86	Andrew Christie
87	Bumman Kim, Hua-Quen Tserng and Hung-Dah Shih, Texas Instruments, Inc.
89	Courtesy of University of Cambridge Library
90	Gordon Hillman
91–95	Patricia J. Wynne
97	Thomas G. Dzubay, U.S. Environmental Protection Agency
98	Thomas C. Moore (*top*), Robert W. Shaw (*middle*), David L. Johnson (*bottom*)
99–101	Thomas C. Moore
102	Robert K. Stevens, U.S. Environmental Protection Agency (*top*); Thomas C. Moore (*bottom*)
105	Andreas Kay
106–109	Michael Goodman
110–111	Edward Bell

LETTERS

To the Editors:

I just read the fascinating article "Marriage, Motherhood and Research Performance in Science," by Jonathan R. Cole and Harriet Zuckerman [Scientific American, February].

"Why do men publish substantially more papers over the course of their careers than women with comparable backgrounds?" ask the authors in the last paragraph. They conclude that "the difference is evidently not explained by marriage and motherhood." The impact of marriage and parenthood on men, however, is not discussed.

A single male has to provide for his meals, laundry, cleaning and so on, unless he is still living at home with his parents or is very wealthy. The married male, particularly in the age range discussed in the article, usually has his wife do many of these time-consuming tasks. He thereby gains time to carry out research and write papers. It is no puzzle to me why "men publish substantially more papers over the course of their careers," as it apparently is to Cole and Zuckerman.

Rita Ricardo-Campbell
(Ph.D. in economics
and also the mother of three)

Hoover Institution
on War, Revolution and Peace

To the Editors:

Being an active woman in science (materials engineering) for more than two decades, and also married and the mother of two, I naturally was interested in the article "Marriage, Motherhood and Research Performance in Science." My experience in Israel, however, does not agree with the conclusion that marriage and childbirth do not affect scientific productivity. I conclude that the disagreement between my experience and the study—whose merit and data I do not doubt—stems from cultural differences.

The proportion of women scientists in the U.S. has been so small that only very determined and able women embarked on a scientific career. For these women even marriage and motherhood did not affect productivity.

In societies such as Israel and the Soviet Union, where a higher percentage of the female population is active in science and the natural selection is not as severe, the situation might be different. It would be of much interest to carry out a study similar to the one de-

TOYOTA
WHO COULD ASK FOR ANYTHING MORE!
TOYOTA
Get More From Life... Buckle Up!

Nick

Karl

Ed

Now you know four of our most innovative people on a first name basis.

Karl, Edwin, Nick and Paul were recently honored for their accomplishments by being made IBM Fellows. For the next five years they'll be given the freedom to pursue projects of their own choosing.

Karl Hermann improved the manufacturing techniques of printed circuit boards, and created methods for assembling and testing computer circuits that ensure the quality of IBM products.

Edwin R. Lassettre conceived efficient software designs that

Paul

simplified data processing on many of IBM's larger computers.

Nicholas J. Pippenger invented the theory used to identify problems that can be solved efficiently by parallel processing.

And Paul A. Totta made possible many advances in semiconductor metallurgy that significantly contribute to the reliability and performance of IBM computers.

Over the years, many IBM men and women have been recognized for innovations that benefit our customers. And though we couldn't include all their names, they all share these three initials. IBM®

scribed by the authors in these countries while the cultural differences remain so distinct.

L. GAL-OR

Israel Institute of Metals

To the Editors:

"Marriage, Motherhood and Research Performance in Science" acknowledges that the authors' research did not solve the puzzle of why male scientists publish substantially more papers over the course of their careers than women scientists with comparable backgrounds. If marriage and motherhood do not explain the disparity, what does? Since the sample of the reported study was well selected, and since both married and unmarried women scientists publish less than male scientists, the factors relating to productivity presumably affect both groups of women in similar ways.

My speculations are drawn from years of experience working with gender issues in higher education, reading transcripts of court cases involving alleged discrimination against faculty women and hearing countless anecdotes about the difficulties women encounter when seeking tenure.

Women in science, like women faculty in all academic areas, have more difficulty than men in obtaining research funding, overcoming the bias of the peer-review process required to publish their work, finding co-workers willing to review their preliminary drafts and make recommendations or provide them with "honorary" authorship, and obtaining "a room of their own" in which to write their proposals, prepare their research and summarize it for publication.

At a subtler level, research should build on what has gone before. When a woman's work receives less esteem from her colleagues than a man's work, it is demoralizing. A cooling-out process may occur as fewer people view women's research as having merit and fewer cite it in their publications. The enthusiasm an investigator brings to a project dims rapidly when she is told she should not select samples of women or concentrate on research topics that typically interest or address women.

We need to consider the security of the employment base from which women scientists do research. How many of them were given tenure as readily as their male counterparts? How many of them had to relocate and reestablish their research in a new setting once or twice? Women investigators are unlikely to have had the opportunity to make optimal employment choices.

Women scientists may have taken on, owing to pressure from their departments, heavier teaching loads or more of the routine departmental tasks such as student advising. Since women's research is stereotypically seen as less important than men's, administrators are less likely to protect women's need for time and space.

Women, whether married or single, assume the "care taking" duties for their immediate families and society. Those in the study who were not married may still assume responsibility for "care taking." The evidence overwhelmingly shows women to be the caretakers of their elderly relatives, elderly parents and ailing husbands.

MARIAN J. SWOBODA

Assistant to the President
University of Wisconsin System

To the Editors:

We are pleased that Dr. Ricardo-Campbell found our paper "fascinating" but bemused by her claim that our data do not bear on the question of gender differences in research performance. It is often asserted that differences in the rate of publication are due to women's greater domestic and parental responsibilities. If that is so, then married women scientists and women scientists with children should generally publish less than unmarried women scientists who do not have these responsibilities. We found, however, that married women scientists and those with children publish not fewer but as many or more papers than single women. Furthermore, we also found that successive births of children are not followed by reductions in rate of publication.

Ricardo-Campbell's statement that research and publication benefits accrue to married male scientists because their wives do time-consuming tasks needed to maintain the household may have been accurate for the earlier generations of scientists, but it seems less true now than it once was. She assumes, incorrectly, that male scientists uniformly use this extra time for research and writing. They report, however, that like women they also have obligations that conflict with research. Men's obligations are more often related to their scientific careers: attending meetings, editing and so on. Those are duties that lead to considerable career advantages but not to higher levels of publication. Finally, we found that male scientists with scientist-wives (who took limited responsibility for household chores) published at the same rate as their peers whose wives devoted themselves to domesticity. Domestic obligations are heavy and do exact a price from women scientists, but not, it appears, in the currency of research publications.

Dr. Gal-Or's suggestion that a comparative inquiry be pursued into the research performance of women scientists makes sense. Rigorous selection of women scientists (married and single) in countries such as the U.S. may minimize the potential effects of marriage and motherhood on research performance. Furthermore, married women scientists in the U.S. may have greater resources to pay for domestic help and child care than their counterparts elsewhere.

Marian J. Swoboda raises several relevant questions about possible differences between men and women in access to scientific resources, in the ability to get work published and the kinds of experience they have in scientific organizations. The available data do not show that women find it more difficult than men to obtain research funding from either the National Science Foundation or the National Institutes of Health, but these data are not comprehensive. Nothing systematic is known about whether men and women are differentially discouraged by having proposals rejected; that question is worth studying. Finally, no systematic evidence of sex bias in the refereeing of papers for publication has been found in the handful of studies of the subject.

Swoboda is quite right in indicating that women are less likely than men to be tenured and that tenure comes later for women than it does for men. Such delays in promotion and tenure probably give women a disadvantage in acquiring resources for research. Furthermore, academic women tend not to be as geographically mobile as men, which may also work to their disadvantage in the job market.

Women scientists face various problems in their careers, many of them quite subtle. For example, we know little about informal discussions of research between men and women colleagues, although some women report being rebuffed in their efforts to initiate such exchanges. This may affect their work patterns and research productivity. We also know little about the features of scientific organizations, such as laboratory size, that influence rates of publication by men and women scientists.

HARRIET ZUCKERMAN

JONATHAN R. COLE

SCIENCE/SCOPE®

Larger power requirements of advanced satellites promise to be met by a new type of battery. Hughes Aircraft Company is developing nickel-hydrogen technology for the U.S. Air Force. The new batteries will be placed on larger spacecraft now being built for customers such as the Air Force, Japan Communications Satellite Company, and INTELSAT, the internal communications consortium. A nickel-hydrogen battery the same size and weight of a conventional nickel-cadmium battery will produce more watts for more years, take more abuse, and perform well even when nearly drained of power.

Hot spots, leaks, and other potential problems in jet engines show up more readily during testing with the use of a Probeye® thermal video system by the U.S. Air Force. Six units of an advanced, third-generation version of the system, developed by Hughes, have been delivered to Arnold Air Force Station in Tennessee for use in analyzing engines undergoing performance testing. Designed for both laboratory and field applications, the all-electric thermography system provides a real-time, multi-color television display of the temperature distribution of a scene being viewed by the Probeye infrared viewer. The new version features enhanced image processing capability, a four-fold improvement in resolution, easier portability and other operational improvements that provide the user with more information for quicker, more accurate testing.

U.S. Army's Fiber Optic Guided Missile (FOG-M) uses a new winding technology to deploy its plastic-coated glass fiber. This fiber permits a two-way jam-proof communication link for transmission of television-like pictures of enemy armor and helicopters to a gunner station located in a protected position. Using technologies learned from 20 years of producing Tube-launched, Optically tracked, Wire-guided (TOW) missiles, Hughes engineers developed a method of precisely winding optical strands so that they can be dispensed at missile velocities without interruption of the data transmission. Because the optical fibers are not much larger than the thickness of a strand of human hair and are elastic and pliable, Hughes invented a device which precisely measures the elasticity of each fiber thus allowing it to be spool-wound with precision. Another Hughes technological advancement is a diagnostic instrument that detects defects in the fiber.

A flight engineering simulator will help develop new military aircraft and systems as well as improve existing ones. The system will be operated by General Dynamics, which produces the F-16 fighter for the U.S. Air Force. The simulator will help serve as proof that design concepts are feasible and allow comparisons to be made without risk of substantial capital investment. It will use technology from F/A-18 training systems developed for the U.S. Navy and Marine Corps. Hughes will supply an image generation and display system, plus operator control and equipment monitoring hardware. The simulator will include 40-foot domes, each housing General Dynamics cockpits and avionics systems.

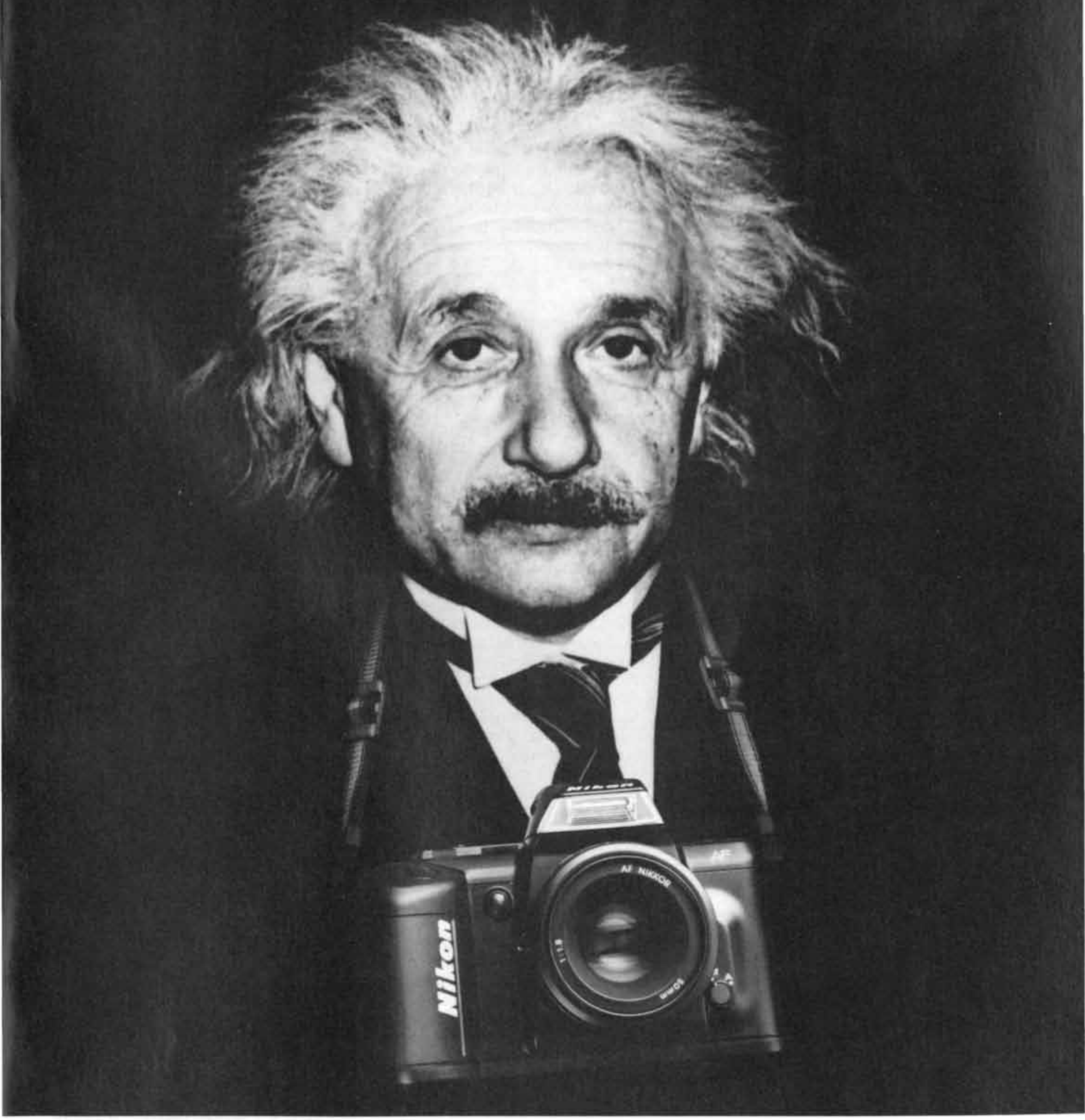
A broad spectrum of technologies, many of which grew up within the past five years, are represented in the products of Hughes' Industrial Electronics Group. Six divisions and two subsidiaries, each operated like a small high-tech company but backed by resources of its multibillion-dollar parent, offer career benefits to qualified engineers and scientists. Advancing technologies such as microwave and millimeter-wave communications, silicon and GaAs solid-state circuitry, fiber optics, and image processing equipment are pursued in facilities located in many of Southern California's most desirable coastal communities. Send your resume to A.T. Moyer, Hughes Industrial Electronics Group, Dept. S2, P.O. Box 2999, Torrance, CA 90509. EOE. U.S. citizenship may be required.

For more information write to: P.O. Box 45068, Los Angeles, CA 90045-0068

HUGHES
AIRCRAFT COMPANY

Subsidiary of GM Hughes Electronics

erfect camera for both.



When you need a flash, the N4004 will recommend that you use one. And you'll always have one, since the flash is built in.

But most important of all, as your photographic genius grows, and you want more creative freedom, the N4004 becomes less automatic. Allowing you to make all the settings yourself.

The Nikon N4004 is incredibly easy-to-use. At the same time, it's incredibly sophisticated.

That's not
a contradiction.
That's genius.

Nikon.
We take the world's greatest pictures.®

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1937: "During the summer of 1936 archaeologists of the Mexican government, working at the Maya ruins of Chichén Itzá, Yucatan, made an extremely important discovery. They found that El Castillo, the most impressively dominant of the pyramid-temples of the famous site, embraces within itself an older pyramid-temple that has been completely concealed for centuries. In it objects were revealed of the utmost value to the scientists who are trying to clear up the perplexities that have arisen respecting the Maya race."

"An international mass attack on relic grabbing, which endangers the world's buried history, is being pushed by archaeologists. The recent international congress of archaeologists at Cairo strongly urged standardized laws throughout the world to curb 'pot hunting.' The U.S. has a particularly hard task to keep irresponsible diggers from despoiling Indian sites, because each of the 48 states handles the problem in its own way."

"Contrary to a widespread impression, the 200-inch mirror disk, by the press often mistermed a 'lens,' that was made for the great California telescope was not ready for use when shipped from Corning, N.Y., to Pasadena but was simply a cast blank of glass, as ordered. The longer work of the opticians mainly remains to be done. If all goes well, this lengthy precision job should be finished by 1940."

"Molding has been found to be a satisfactory method of making optical lenses from clear acryloid resins. By using molds of extreme precision and by carrying out the molding operation with the greatest care, it has been possible to make lenses for cameras, telescopes and spectacles, as well as for other optical instruments. These lenses do not require the tedious grinding and polishing necessary with glass."

"Limitations on the use of insecticides containing arsenic and lead on edible farm products have encouraged and stimulated the search for materials for protecting crops against insects that will not leave toxic residues. More than one thousand compounds believed to be useful for this purpose have been carefully examined in the United States Department of Agriculture, but only four show apparent value at present. They are phenothiazine, nicotine, pyrethrum and rotenone."

"Believing that instruction in the theory of the rules of the road and actual practice in driving a car have a place in the curriculum of the present-day high school as a means of promoting automobile safety, the American Automobile Association has sponsored a driver training program. It already has met with great success in ten high schools."

"The world's largest flower, seen for the first time in America, bloomed on June 8 at the New York Botanical Garden. This *Amorphophallus titanum* flower was 8½ feet tall and four feet in diameter. It is related to the calla-lily and jack-in-the-pulpit."

SCIENTIFIC AMERICAN

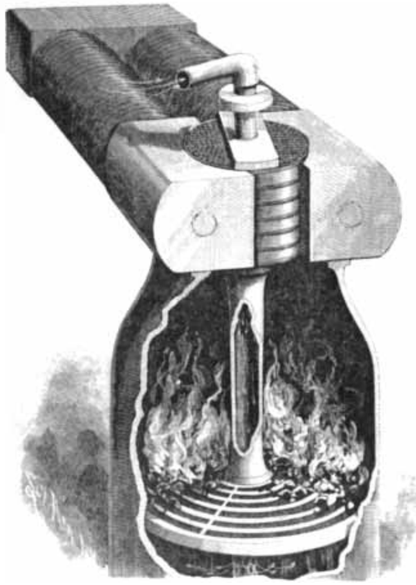
AUGUST, 1887: "By Thomas A. Edison. 'The production of electricity directly from coal is a problem which has occupied the attention of the ablest inventors for many years. If the enormous energy latent in coal could be made to appear as electrical energy by means of a simple transforming apparatus, the mechanical methods of the entire world would be revolutionized. If the result is to be attained, it must be looked for in some direction other than that of the thermo cell. Another line of investigation suggested itself to me. The magnetism of the magnetic metals, and especially of iron, cobalt and nickel, is markedly affected by heat. Since whenever a magnetic field varies in strength in the vicinity of a conductor, a current is generated in that conductor, it occurred to me that by placing an iron core in a magnetic circuit and by varying the magnetizability of that core by varying its temperature it would be possible to generate a current in a coil of wire surrounding this core. This idea constitutes the essential feature of the new generator, which therefore I have called a pyromagnetic generator.'"

"Among the most interesting topics that will be brought up at the coming convention of the National Electric Light Association will be the distribution of electrical energy for the running of shafting, elevators and the like, and another will be motors. It is likely that in the near future small factories and workshops lying within the distributing district of an electrical lighting station will find it cheaper, as well as more convenient, to take their power from off a wire and through a dynamo than from a steam engine."

"A paper that aroused much enthusiasm at the recent meeting of the American Association for the Advancement of Science was by Drs. Michelson and Morley on a 'Method for Making the Wave Length of Sodium Light the Actual and Practical Standard of Length.' This standard was obtained by sliding a reflecting mirror through a measured space and counting the number of interference fringes produced by the motion, indicating the number of wave lengths and taking this length as the unit of measurement. It was claimed that no natural standard had ever been found that would prove unvarying except this one."

"We hear talk already about specialists in photography for instantaneous pictures since the 'Detective Camera,' as it is called, was put on the market. The box is so small that it can be carried anywhere without the slightest inconvenience. Any possible mania or desire for photos can soon be gratified at trifling expense and after a short term of practice by means of this invention."

"The *Journal* of the American Medical Association notes that attention has been drawn to a new nervous disorder, said to be especially prevalent in England and America. It is called 'theism,' or tea drinker's disease."



Thomas A. Edison's pyromagnetic motor

Today it's almost impossible to communicate fully without text and graphics on the same page.

Lotus Manuscript™ makes it easy to put a column of text and a graphic element side by side.

You can import sophisticated graphics from Freelance® Plus to enhance the communications value of any written document.

Spreadsheets, graphics and charts from 1-2-3® and Symphony® can easily be mixed with text on the same page.

You can throw away your scissors and glue, cut and paste are a thing of the past.

Since early cave drawings, people have found graphics quite effective in communications. Yet in our information-driven society, graphics have taken a back seat to the written word. From typewriters to word processing, graphical elements have been treated like afterthoughts, relegated to "exhibit on next page" or "cut and paste" status.

Lotus Manuscript is the first word processor that is truly a complete document creation system. It's ideal for the needs of technical writers and writers of long complex documents.

Manuscript allows you to easily mix text on the same page as graphics; elements from 1-2-3 and Symphony, graphics from Freelance Plus, or diagrams and scanned images from other sources.

With our Document Preview feature you can see graphics and text on the same page before it's printed, with a zoom capability that lets you take a closer look for proofing your layouts or equations.

Try it like this. *Or, try it like this.*

Our intelligent print formatter gives you great control and flexibility over size and positioning of graphics on the page.

Manuscript is designed to work on most IBM® PCs and compatibles.* Its familiar 1-2-3 interface makes it easy to use. And our Manuscript evaluation kit makes it easy to try. For $10.00, you'll get a presentation disk, working software, and a tutorial manual. To get your evaluation kit, call 1-800-345-1043, ask for lot #BA-1450. Or, for more information, see your authorized Lotus Dealer, or write Lotus Development Corp., 90 Annex, Atlanta, GA 30390-03070.

Lotus Manuscript™

THE AUTHORS

JOHN S. LETCHER, JR., JOHN K. MARSHALL, JAMES C. OLIVER III and NILS SALVESEN (*"Stars & Stripes"*) are members of the team that designed the America's Cup winner; they are also enthusiastic recreational sailors. Letcher, senior scientist for the project, is a consultant to the Science Applications International Corporation (SAIC) in Annapolis, where he works on the analysis and computer-aided design of sailing vessels. He earned a master's degree in aeronautics at the California Institute of Technology, which awarded his Ph.D. in 1966; he also has a second master's, in naval architecture, from the University of Michigan. Marshall, coordinator of the design team, is chief operating officer of Henry R. Hinckley & Company and a four-time veteran of America's Cup campaigns. His undergraduate degree, from Harvard College, was in biology. Oliver, who coordinated computer modeling and tank testing for *Stars & Stripes,* is a senior research naval architect at SAIC. A 1973 graduate of the U.S. Naval Academy, he went on to study ocean engineering at George Washington University. Salvesen, the project's technology coordinator, is division manager of hydrodynamics at SAIC, where he is involved in the development of seakeeping theory and numerical hydrodynamics. His Ph.D. in naval architecture is from the University of Michigan.

ALAN D. KRISCH ("Collisions between Spinning Protons") is professor of physics at the University of Michigan. He received a Ph.D. from Cornell University in 1964, the year he joined the faculty at Michigan. Krisch's research has focused on the study of proton-proton interactions as a means of searching for smaller objects within the proton. Since 1972 he has concentrated on work with polarized proton beams, first in the Zero Gradient Synchrotron at the Argonne National Laboratory and since 1979 in the Alternating Gradient Synchrotron at the Brookhaven National Laboratory.

DAVID PATTERSON ("The Causes of Down Syndrome") is president and associate director of the Eleanor Roosevelt Institute for Cancer Research in Denver. He is also professor in the department of biochemistry, biophysics and genetics and the department of medicine at the University of Colorado Health Sciences Center. After earning a doctorate in microbiology from Brandeis University in 1971, Patterson went to the Eleanor Roosevelt Institute, first as a Damon Runyon–Walter Winchell fellow. He has been on the faculty of the University of Colorado since 1974.

GORDON L. ADA and SIR GUSTAV NOSSAL ("The Clonal-Selection Theory") worked together for five years on studies that helped to establish the theory whose history they recount. Ada is professor of microbiology at the John Curtin School of Medical Research of the Australian National University. His undergraduate and graduate degrees are from the University of Sydney, where he received his doctorate in 1959. From 1946 to 1948 he was at the National Institute for Medical Research in London. He then joined the staff of the Walter and Eliza Hall Institute of Medical Research in Melbourne, leaving in 1968 to head the department of microbiology at the Curtin School. Nossal has been director of the Walter and Eliza Hall Institute and professor of medical biology at the University of Melbourne since 1965. He studied medicine at the University of Sydney and in 1960 got his Ph.D. from Melbourne. After two years as assistant professor of genetics at the Stanford University School of Medicine, Nossal went to the Walter and Eliza Hall Institute as deputy director. He has remained at the institute ever since, except for a year as a visiting scientist at the Pasteur Institute in Paris and a second year as a consultant to the World Health Organization. Nossal was knighted in 1977.

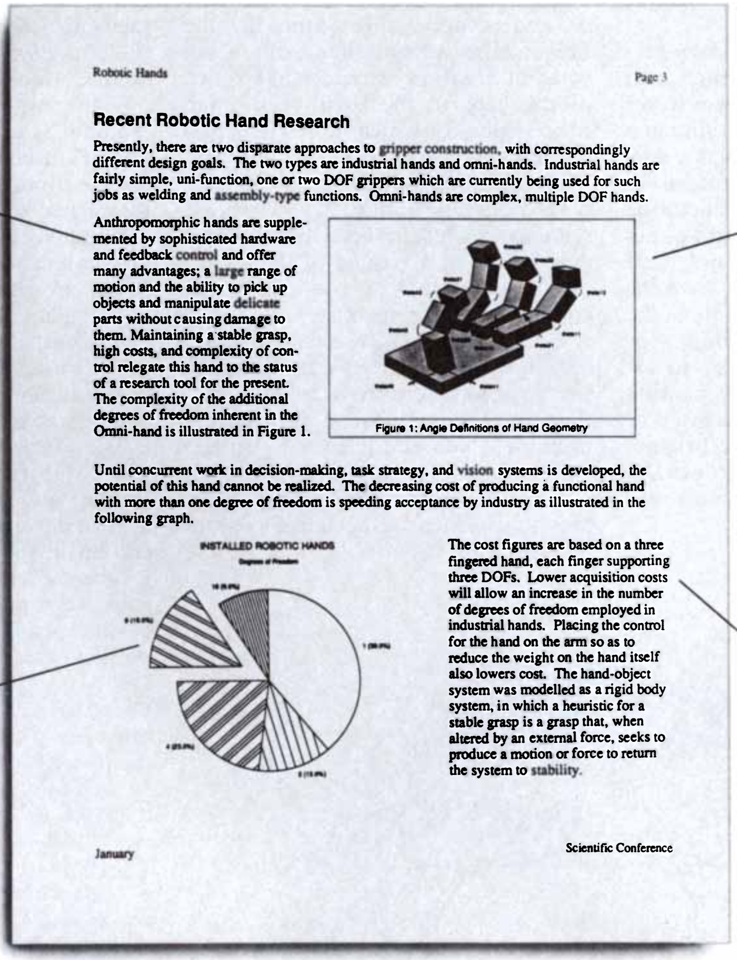
CHRISTOPHER J. TALBOT and MARTIN P. A. JACKSON ("Salt Tectonics") share an interest in centrifuge modeling of salt tectonics. Talbot, professor of tectonics and geodynamics at the University of Uppsala, trained as a structural geologist at the Imperial College of Science and Technology in London and earned a Ph.D. from the University of Leeds in 1967. He spent two years as a postdoctoral research associate at the Postgraduate School of Mining at the University of Sheffield before appointment as an assistant lecturer and then a lecturer at the University of Dundee. He left Scotland in 1982 for his current position. Talbot worries over the fact that some nations have high-grade radioactive wastes but lack "suitable geology in which to store it." Jackson is senior research scientist at the University of Texas Bureau of Economic Geology. After completing undergraduate work in geology at the University of London in 1969 and studying minerals throughout southern Africa, he received a Ph.D. from the University of Cape Town in 1976. He has devoted his research efforts to salt tectonics since going to the University of Texas in 1980. Jackson says he is particularly curious about the "more bizarre manifestations" of geologic flow.

WILLIAM R. FRENSLEY ("Gallium Arsenide Transistors") is a senior member of the technical staff in the Advanced Concepts branch of Texas Instruments, Inc., where he is developing computer simulations of quantum-mechanical semiconductor devices. He was graduated from the California Institute of Technology in 1973 and earned his doctorate in physics at the University of Colorado at Boulder three years later. He did postdoctoral work at the University of California at Santa Barbara and then joined Texas Instruments in 1977.

ANTHONY J. LEGGE and PETER A. ROWLEY-CONWY ("Gazelle Killing in Stone Age Syria") both do faunal studies on hunter-gatherer and early agricultural sites in the Near East and Mediterranean regions. Legge is senior lecturer in archaeology in the Department of Extra-Mural Studies at the University of London. After getting his bachelor's degree in archaeology from the University of Cambridge in 1969 he did research there and got his master's degree. Legge became a lecturer at the University of London in 1974 and accepted his present post in 1982. Rowley-Conwy has been a research fellow at Clare Hall in Cambridge since 1986. He received his doctorate in archaeology from the University of Cambridge in 1980. From 1982 to 1985 he worked with Legge on the Tell Abu Hureyra project they describe.

ROBERT W. SHAW ("Air Pollution by Particles") is chief of chemical diagnostics and surface science at the U.S. Army Research Office in Research Triangle Park in North Carolina. He earned his Ph.D. in physical chemistry from the University of Washington in 1970 before he became a postdoctoral fellow at Princeton University and then a visiting assistant professor at the University of Oregon. He went to the U.S. Environmental Protection Agency as a research chemist and physicist in 1977 and worked there until 1983. Shaw is encouraged to stay alert by his three-year-old daughter, who has asked why both snowflakes and milk are white.

Today it's almost impossible to communicate fully without text and graphics on the same page.



Robotic Hands

Page 3

Recent Robotic Hand Research

Presently, there are two disparate approaches to gripper construction, with correspondingly different design goals. The two types are industrial hands and omni-hands. Industrial hands are fairly simple, uni-function, one or two DOF grippers which are currently being used for such jobs as welding and assembly-type functions. Omni-hands are complex, multiple DOF hands.

Anthropomorphic hands are supplemented by sophisticated hardware and feedback control and offer many advantages; a large range of motion and the ability to pick up objects and manipulate delicate parts without causing damage to them. Maintaining a stable grasp, high costs, and complexity of control relegate this hand to the status of a research tool for the present. The complexity of the additional degrees of freedom inherent in the Omni-hand is illustrated in Figure 1.

Figure 1: Angle Definitions of Hand Geometry

Until concurrent work in decision-making, task strategy, and vision systems is developed, the potential of this hand cannot be realized. The decreasing cost of producing a functional hand with more than one degree of freedom is speeding acceptance by industry as illustrated in the following graph.

The cost figures are based on a three fingered hand, each finger supporting three DOFs. Lower acquisition costs will allow an increase in the number of degrees of freedom employed in industrial hands. Placing the control for the hand on the arm so as to reduce the weight on the hand itself also lowers cost. The hand-object system was modelled as a rigid body system, in which a heuristic for a stable grasp is a grasp that, when altered by an external force, seeks to produce a motion or force to return the system to stability.

January

Scientific Conference

Lotus Manuscript™ makes it easy to put a column of text and a graphic element side by side.

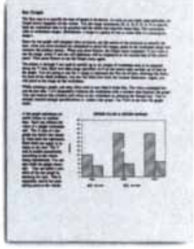
You can import sophisticated graphics from Freelance® Plus to enhance the communications value of any written document.

Spreadsheets, graphics and charts from 1-2-3® and Symphony® can easily be mixed with text on the same page.

You can throw away your scissors and glue, cut and paste are a thing of the past.

Since early cave drawings, people have found graphics quite effective in communications. Yet in our information-driven society, graphics have taken a back seat to the written word. From typewriters to word processing, graphical elements have been treated like afterthoughts, relegated to "exhibit on next page" or "cut and paste" status.

Try it like this.



Or, try it like this.

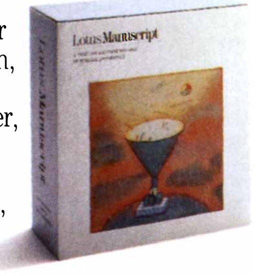
Our intelligent print formatter gives you great control and flexibility over size and positioning of graphics on the page.

Lotus Manuscript is the first word processor that is truly a complete document creation system. It's ideal for the needs of technical writers and writers of long complex documents.

Manuscript allows you to easily mix text on the same page as graphics; elements from 1-2-3 and Symphony, graphics from Freelance Plus, or diagrams and scanned images from other sources.

With our Document Preview feature you can see graphics and text on the same page before it's printed, with a zoom capability that lets you take a closer look for proofing your layouts or equations.

Manuscript is designed to work on most IBM® PCs and compatibles.* Its familiar 1-2-3 interface makes it easy to use. And our Manuscript evaluation kit makes it easy to try. For $10.00, you'll get a presentation disk, working software, and a tutorial manual. To get your evaluation kit, call 1-800-345-1043, ask for lot #BA-1450. Or, for more information, see your authorized Lotus Dealer, or write Lotus Development Corp., 90 Annex, Atlanta, GA 30390-03070.



Lotus Manuscript™

SCIENCE AND THE CITIZEN

Science, 7; Creationism, 2

The cause of science has won a bout in the apparently endless conflict between strict adherents of the biblical story of creation and those people who accept the scientific version of genesis.

Rejecting an appeal by state officials of Louisiana, the Supreme Court upheld the judgment of two lower courts by declaring unconstitutional a "balanced treatment" Louisiana statute that forbids the teaching of evolution in public schools unless a doctrine called creation science is taught as being equally valid. The statute defined creation science as including scientific evidence for the sudden creation of the universe and life out of nothing, for evolutionary theory's inability to explain the development of diverse species and for the separate ancestry of human beings and apes. The original challenge to the statute was brought by Louisiana parents, teachers and religious leaders.

Creationist literature holds that the earth and the rest of the universe were created by an omnipotent being only a few thousand years ago, a belief that entails rejecting the theory of evolution by natural selection, geologic evidence that the earth is billions of years old and astronomical evidence that the universe as a whole is older still. Advocates of creation science make much of the gaps in the fossil record (an observation for which there are explanations that are entirely consistent with evolution). Such scientific trappings notwithstanding, 72 Nobel laureates and 17 state academies of science joined in a "friend of the court" brief arguing that creation science is actually religious dogma.

The Louisiana law made no mention of religion, however; its stated purpose was instead to ensure academic freedom by giving Louisiana students a broad or balanced view of theories about origins. The court held, by seven votes to two, that the Louisiana law did not further its declared goal of "teaching all of the evidence": schoolteachers were not given any flexibility they did not already have. In fact, the law had "the distinctly different purpose of discrediting evolution by counterbalancing its teaching at every turn." In spite of the law's secular language the court found that it endorsed religion, in violation of the Establishment Clause of the Constitution.

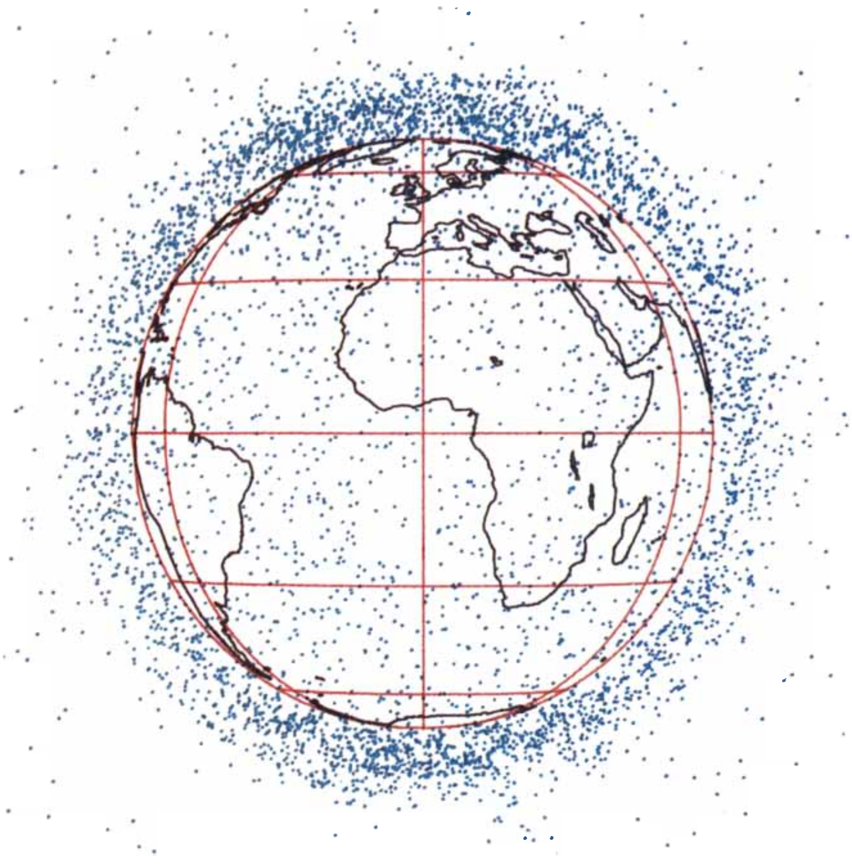
Two members of the court, Justice Antonin Scalia and Chief Justice William H. Rehnquist, went down with the creationist ship. In a joint dissent they argued that the decision rested on "impugning the motives of [the Louisiana statute's] supporters." The court's judgment that the law's intention was not secular could not be demonstrated, Scalia and Rehnquist wrote, because the law had never been in force and had not been the subject of a "full evidentiary hearing."

The court's decision appears to have put a stop for now to campaigns for similar laws in other states, although at least in New Mexico evolution must still be taught with a disclaimer that it is "theory" rather than fact. The decision probably does not mark the end of this peculiarly American kind of political theater. Duane T. Gish of the Institute for Creation Research in San Diego says "the decision will intensify our efforts" to bring creationism into the schools through persuasion.

Trashing Space

Early in 1986 a rocket belonging to the European consortium Arianespace roared skyward carrying a *Spot 1* satellite. After it had injected the French remote-sensing satellite into orbit, the rocket's third-stage booster itself remained in orbit. Last November, inexplicably, the booster blew up, contributing more shrapnel to an already dense—and potentially dangerous—swarm of objects hurtling around the globe.

At a time when the U.S. is planning to extend its presence in space with a renewed shuttle program, the space station and possibly the Strategic Defense Initiative space technologists are becoming increasingly concerned by the growing cloud of debris that envelops the earth. The U.S. North American Aerospace Defense Command currently tracks some 7,000 objects about the size of a baseball or larger, most of them working and obsolete satellites and spent rocket boosters. Four years ago NORAD was tracking

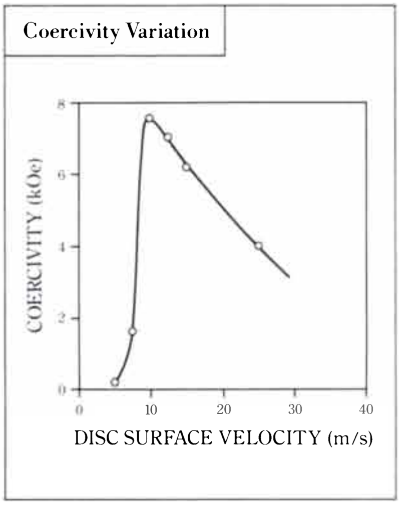


SATELLITES, BOOSTERS AND DEBRIS with a diameter of 10 centimeters or more swarm around the earth in this computerized printout by Lockheed engineers. Their printout is based on data from the U.S. North American Aerospace Defense Command.

The Critical Interval

The Critical Interval

There has long been a need in the industrial world for low-cost, high-performance permanent magnets. Discoveries at the General Motors Research Laboratories have led the way toward meeting this challenge by the application of new preparation techniques to new rare-earth magnetic materials.



Coercivity of $Pr_{0.4}Fe_{0.6}$ plotted as a function of disc surface velocity.

Color-enhanced transmission electron micrograph of melt-spun $Nd_{0.4}Fe_{0.6}$ having 7.5 kOe coercivity.

TWO properties characterize desirable permanent magnets: large coercivity (magnetic hardness or resistance to demagnetization) and high remanence (magnetic strength). Higher-performance magnets are required to reduce further the size and weight of a wide variety of electrical devices, including d.c. motors. Such magnets are available, but the cost of the materials necessary to produce them severely limits their use. The research challenge is to select, synthesize, and magnetically harden economically attractive materials of comparable quality.

Prominent among alternative materials candidates are alloys composed of iron and the abundant light rare earths (lanthanum, cerium, praseodymium, neodymium). Investigations conducted by Drs. John Croat and Jan Herbst at the General Motors Research Laboratories have led to the discovery of a method for magnetically hardening these alloys. By means of a rapid-quench technique, the researchers have achieved coercivities in Pr-Fe and Nd-Fe that are the largest ever reported for any rare earth-iron material.

Drs. Croat and Herbst selected praseodymium-iron and neodymium-iron based upon fundamental considerations which indicate that these alloys would exhibit properties conducive to permanent magnet development. These properties include ferro-magnetic alignment of the rare earth and iron magnetic moments, which would foster high remanence, and significant magnetic anisotropy, a crucial prerequisite for large coercivity.

That these materials do not form suitable crystalline compounds, an essential requirement for magnetic hardening by traditional methods, presents a major obstacle. Drs. Croat and Herbst hypothesized that a metastable phase having the necessary properties could be formed by cooling a molten alloy at a sufficiently rapid rate. They tested this idea by means of the melt-spinning technique, in which a molten alloy is directed onto a cold, rotating disc. The cooling rate, which can be varied by changing the surface velocity of the disc, can easily approach 100,000°C per second. The alloy emerges in the form of a ribbon.

THE researchers found that variations of the cooling rate can dramatically affect the magnetic properties of the solidified alloys. In particular, appreciable coercivity is achieved within a narrow interval of quench rate.

Equally remarkable, synthesis and magnetic hardening, two steps in conventional processing, can be achieved simultaneously.

"X-ray analysis and electron microscopy of the high coercivity alloys reveal an unexpected mixed microstructure," states Dr. Croat. "We observe elongated amorphous regions interspersed with a crystalline rare earth-iron compound."

Understanding the relationship between the coercivity and the microstructure is essential. The two scientists are now studying the extent to which the coercivity is controlled by the shape and composition of the amorphous and crystalline structures.

"The development of significant coercivity is an important and encouraging step," says Dr. Herbst, "but practical application of these materials requires improvement of the remanence. Greater knowledge of the physics governing both properties is the key to meeting the commercial need for permanent magnets."

TECHNOLOGY UPDATE: 1987

Subsequent to the research reported above, Drs. Croat and Herbst added boron to neodymium-iron as a glassifier to increase the formation of the elongated amorphous regions they had observed in the material. They reasoned that shape anisotropy, and thus coercivity, was related to the presence of these amorphous micro-needles.

They discovered that the addition of boron promoted the formation of a previously unknown ternary compound: $Nd_2Fe_{14}B$. Its atomic magnetic moments are arranged so that this compound has a large magnetization. At the same time, the researchers found that, compared with neodymium-iron, coercivity had risen from 8 to 20 kOe, and that the magnetic energy product had increased by a factor of seven.

On March 31, 1987, General Motors dedicated a new Delco Remy plant in Anderson, Indiana for the production of magnetic material and finished magnets made from $Nd_2Fe_{14}B$ under the commercial name MAGNEQUENCH.

General Motors

GM
MARK OF EXCELLENCE

THE MEN BEHIND THE WORK

Dr. John Croat and Dr. Jan Herbst did their original work on rare-earth magnetic materials when both were Staff Research Scientists in the Physics Department at the General Motors Research Laboratories.

Dr. Croat (right) holds a Ph.D. in metallurgy from Iowa State University. In 1984, he joined GM's Delco Remy Division to stabilize the melt-spinning process for the commercial production of MAGNEQUENCH materials. He is currently Chief Engineer at the Indiana plant.

Dr. Herbst received his Ph.D. in Physics from Cornell University. He is now a Senior Staff Research Scientist and Manager of the Magnetic Materials Section in the Physics Department of the GM Research Laboratories. His research interests also include photo-emission theory, the physics of fluctuating valence compounds, and superconductivity.

Dr. Croat joined General Motors in 1972; Dr. Herbst in 1977.

only 4,000 such objects, according to Donald J. Kessler of the National Aeronautics and Space Administration. About 40,000 fragmentary objects that are smaller than a baseball and larger than a pea are also in orbit, Kessler says.

A pea-size object colliding with a satellite, a spacecraft or an astronaut at 10 kilometers per second could have the destructive power of a hand grenade, according to Nicholas L. Johnson of Teledyne Brown Engineering. Even much smaller objects can cause significant damage. In 1984 the space shuttle *Challenger* returned from a mission with a pit about a centimeter wide in a pane of its windshield. Investigators discovered that the pane, which had to be replaced, had been struck by a paint flake only .2 millimeter wide.

The designers of the space station, Johnson notes, are adding shielding to protect it from projectiles as large as one centimeter in diameter. It is too late, however, to redesign the Hubble Space Telescope, scheduled to be launched in the early 1990's. Michael M. Shara of the Space Telescope Science Institute recently calculated that the telescope has a 1 percent chance of being destroyed and a 50 percent chance of being damaged by debris during its operational lifetime. Sunlit objects that merely flash through the telescope's field of view may also damage its delicate sensors.

Sunlit debris has already complicated earth-based astronomy. Paul D. Maley of the Rockwell Shuttle Operations Company recently proposed in the *Astrophysical Journal* that optical flashes thought to be from a gamma-ray source outside the solar system were actually caused by a sunlit Soviet satellite. Maley thinks space debris may have duped astronomers many times before.

SDI officials are worried that debris might cripple space-based sensors and weapons. Michael Kemp, manager of SDI survivability programs, declares: "You bet we're addressing the space debris issue—today's and tomorrow's." Shara points out that SDI tests could also heighten the problem.

Although the Arianespace rocket was a recent offender, the U.S. and the U.S.S.R. are responsible for the great majority of the debris. The U.S., says Robert C. Reynolds of Lockheed Engineering and Management Services Company, Inc., "probably has the dubious distinction of creating the most debris" with a series of Delta boosters that exploded during the 1970's. The U.S.S.R. has contributed with tests of antisatellite weapons and with the deliberate destruction of malfunctioning satellites.

Both NASA and Arianespace are trying to design space components that are less likely to explode and that can be steered back into the atmosphere when they have done their work. Reynolds says NASA is also studying ways of slowing down an orbiting object, perhaps with laser or particle beams or with clouds of foam, so that it will reenter the atmosphere and be incinerated. Such tactics are expected to be discussed at a meeting of the International Astronomical Federation later this year.

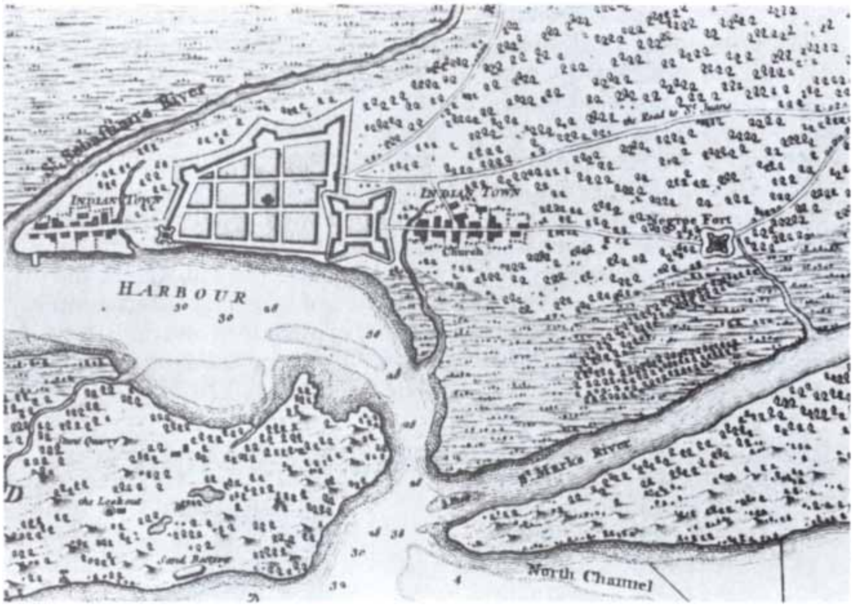
Holding the Fort

From 18th-century Spanish archives and from a marsh in northern Florida, investigators are unearthing the history of Fort Mose, thought to be the first settlement of free black men and women in America. From 1738 to 1763 as many as 100 blacks lived in or around the fort, an outpost of the Spanish colony of St. Augustine.

According to Kathleen A. Deagan of the Florida State Museum, the project leader, most of the inhabitants of the fort were Africans who had escaped slavery in the British-controlled Carolinas and fled to Spanish-controlled Florida. The Spaniards promised freedom for slaves who declared their enmity toward the British and converted to Catholicism. In 1738 the Spanish governor ordered the freedmen to occupy a fort several miles from St. Augustine to serve as a buffer against British and Indian attacks.

The site is now an island surrounded by marsh. In the 18th century, according to maps and accounts of the period, the land around the island was dry enough to cultivate. Farming the open land was apparently dangerous, however, given the constant potential for attack. In 1740 the black soldiers joined the Spaniards and friendly Indians in repulsing a major invasion by the British.

Deagan says she hopes that the exca-



"NEGROE FORT" on the outskirts of the Spanish colony of St. Augustine in Florida was drawn on a British map of 1740. Workers from the Florida State Museum excavate the fort's site, now a marsh-surrounded island, in a photograph made by James Quine.

vation of the site, which began last January, will reveal whether the settlers retained any of their African tools, foods and religious beliefs. So far workers have unearthed musket balls, flints, animal bones, fragments of ceramics and glass, pipestems, metal buckles and hinges. They have also uncovered evidence of a moat on three sides of the island and heavy posts at its center: the remains, perhaps, of a watchtower or chapel.

Jane L. Landers, a doctoral candidate at the University of Florida and the project's historian, has studied archives in Florida and Spain describing the fort and its people. Government records and other documents of the period, she says, indicate that the black soldiers and their families lived at the fort with a handful of Spanish soldiers and a priest. When the British won control of Florida in 1763, most of the former slaves left Florida with the Spaniards and resettled in Cuba.

Landers says the soldiers were led by a Mandingo African referred to in documents as Francisco Menendez. After escaping from a British plantation in South Carolina he joined a band of Indians in fighting the British for two years. In 1726, according to Landers, he arrived in St. Augustine, where an Indian called Mad Dog sold him into slavery. With the help of the Indian chief with whom he had fought the British, Menendez successfully petitioned the Spanish governor for his freedom.

Made captain of the black soldiers at the fort, Menendez demanded payment from the Spaniards for his men's services. The fact that he signed his name on the request for payment suggests that Menendez may have been literate, Landers says, pointing out that very few colonists—regardless of their race—could read or write.

Robert L. Hall of the University of Maryland in Baltimore County notes that the Fort Mose dig reflects "an increasing—actually a renewed—interest in black resistance to slavery in the New World." Peter H. Wood of Duke University concurs. "What makes Fort Mose interesting is not that it was a bastion of freedom," Wood observes, "but that it exemplifies all the problems the freedmen had to cope with."

PHYSICAL SCIENCES

Gathering Cosmic String

According to theory, cosmic strings are threadlike relics of the extraordinarily dense and energetic conditions that existed immediately after the big bang. Such objects, which either form closed loops or are infinitely long, would be the consequence of the early universe's failure to expand and cool with perfect uniformity; they are roughly analogous to the crystalline defects lacing the ice of a frozen pond. Although so narrow as to be virtually one-dimensional, strings are thought to be extremely massive: a section 100 meters long would weigh more than the moon. Strings were originally implicated in the formation of galaxies; more recently theorists have speculated that they are superconducting and that they give rise to hitherto unexplained cosmic gamma rays and vast jets emitted by quasars.

All these hypotheses share a weakness: there is no direct observational evidence of cosmic strings. Now two investigators at the University of Hawaii's Institute for Astronomy say they have spotted a field of galaxies that may be split into double images by a cosmic string acting as a gravitational lens. Lensing is caused by a massive structure—usually a galaxy or cluster of galaxies—interposed between a light source and an observer. The structure's gravitational field causes divergent rays of light from the source to bend and converge on the observer, who perceives the source as a displaced image or multiple images.

Lennox L. Cowie and Esther M. Hu say their discovery, which they describe in the *Astrophysical Journal,* was serendipitous. When they examined electronically amplified images of a distant quasar field, they noticed four pairs of galaxies in the region. The similarity in brightness and shape between the members of each pair led Cowie and Hu to speculate that the four pairs actually represent four single galaxies overlaid with a cosmic string. Cosmic-string theorists maintain it is the most compelling evidence offered to date. "If it proves to be correct," Edward Witten of Princeton University comments, "it will be very exciting."

Alexander Vilenkin of Tufts University, who demonstrated in 1981 how a cosmic string might act as a gravitational lens, says Cowie and Hu's observation matches his theory in two respects: whereas a galaxy usually produces only odd numbers of images, a string should produce only double images. In addition, the angular separation between the members in each pair of images is approximately two arc seconds, which is what Vilenkin expects for a string massive enough to effect galaxy formation. Cowie and Hu caution in their paper, however, that the "identical twins" may actually be binary galaxies, which have been observed in clusters elsewhere. They also suggest that a string would have to be greatly contorted to intersect four galaxies that are apparently so close together.

Further observations could bolster—or shatter—Cowie and Hu's hypothesis by determining whether the individual images that constitute a pair are in fact mirror versions of each other or are significantly different. Craig J. Hogan of the University of Arizona says he is looking with the Multiple Mirror Telescope near Tucson for differences in the red shifts of the pairs. With "very good seeing conditions," Hogan says, the telescope might also show whether the images have different shapes. Scanning the microwave spectrum with radio telescopes might provide more evidence. Theorists have proposed that the normally isotropic microwave radiation suffusing the universe would have a different intensity on each side of a cosmic string.

Hogan notes that it will be much easier to disprove than to confirm the role of cosmic strings in astronomical phenomena. Indeed, a report advanced last year of a quasar split into a double image by a cosmic string has already been widely discounted. "The consensus now is that it's probably just a pair," Hogan says.

The Ozone Hole

Every year during the Southern Hemisphere spring the stratospheric layer of ozone (O_3) that surrounds the earth is temporarily depleted over Antarctica. The "ozone hole" has been getting larger each year since the mid-1970's, and the depletion now reaches 40 percent. This year's event promises to be the most intensively studied one ever.

The National Science Foundation will support ground-based observations and the National Aeronautics and Space Administration is planning flights by two specially equipped aircraft, including an ER-2 (a civilian version of the U-2) that will fly at an altitude of 20 kilometers into the hole itself. Other countries, including the U.K. and the U.S.S.R., are also planning studies. U.S. and Soviet scientists have agreed to expand collaboration: the Americans will provide instrument packages for the Soviets to use at their Antarctic station and the Soviets will provide the U.S. with data collected by sounding rockets. Satellites and balloons will also be used.

The ozone hole could be a natural phenomenon. But many investigators are convinced that it is caused at least in part by emissions of chlorofluorocarbons (CFC's). It could presage a global ozone depletion if such emissions continue to increase.

Depletion of the ozone layer would allow more ultraviolet light to reach the earth's surface, damaging crops and aquatic organisms. A study done by the Environmental Protection Agency predicts 800,000 additional skin-cancer deaths in the U.S. by the year 2075 if CFC emissions continue to increase at current rates.

Recent observations tend to implicate CFC's in causing the hole. Robert L. de Zafra of the State University of New York at Stony Brook says observations by him and his co-workers provide "undeniable evidence" that chlorine plays a crucial role in the ozone depletion. Theories of how chlorine from CFC's might catalyze ozone depletion predict the formation of chlorine monoxide as a by-product; de Zafra found levels of chlorine monoxide in last year's ozone hole that were 100 times higher than expected. He also found a stratospheric nitrous oxide hole that appears at the same time as the ozone hole. Its significance is not understood, but it suggests that ozone is not simply displaced by upwelling air from the troposphere, because that air contains nitrous oxide.

To explain the high chlorine monoxide levels, most theories invoke unusual chemistry that takes place on ice grains or other particles. Polar stratospheric clouds provide an abundant supply of such particles, thus perhaps explaining why depletion is localized; Donald F. Heath, a NASA investigator, has also found indications of an annual ozone depletion over the Arctic. There are other explanations: a recently identified shower of high-energy electrons from the sun that strikes the upper atmosphere every 27 days might be a factor too.

In international negotiations at Geneva the U.S. has urged that CFC emissions be curtailed, eventually by as much as 95 percent. That position (although it had been officially approved) has been criticized within the Administration. Secretary of the Interior Donald P. Hodel denies reports that he has suggested sunglasses and hats as an alternative to CFC reductions, but a spokesman for his department says it does have concerns that an international agreement might be unverifiable. Congress, though, is throwing its weight behind tough controls; the Senate voted by 80 to two to support the Administration's original tough stance, and bills have been introduced that would limit imports of products made with the help of CFC's.

The Well-tempered Clavier

Around the beginning of the 17th century harpsichords and other baroque musical instruments strung with metal wire were made longer. Although the musicological reasons for this evolution are not entirely clear, it certainly could not have taken place unless wire had become available that was able to withstand the higher tension needed to tune longer strings to the proper pitch. Musicologists have assumed that the wire for the longer, stronger strings must have been made of steel.

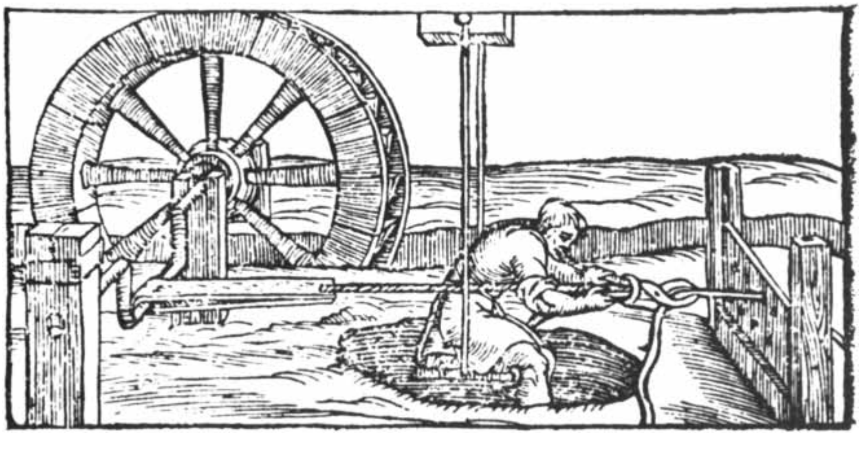
The results of a thorough investigation conducted by Martha Goodway of the Smithsonian Institution may now compel musicologists to change their tune. Goodway reports in *Science* that harpsichord strings dating from the 17th and 18th century were made not of steel (which is essentially an alloy of iron and carbon) but of an alloy of iron and phosphorus, an element generally regarded as an unwanted impurity. Actually the phosphorus content in the iron alloy is high enough to have effectively prevented its conversion into steel.

Goodway's investigation, which is part of a broader study of antique music wire, began with chemical and metallurgical analyses of samples of original harpsichord strings that had been collected by J. Scott Odell of the Smithsonian. Odell had found scraps of the strings embedded in the softwood sounding boards and in the glue or paint of old harpsichords by closely inspecting and in some cases even X-raying the instruments. Analysis of the scraps revealed that they consisted mostly of iron having no detectable traces of carbon but a high level of phosphorus.

If the strings were not of steel, how were they made strong enough to withstand the higher tension in the enlarged harpsichords? The only possible alloying element, phosphorus, ordinarily renders steel brittle and unsuitable for drawing into wire. What did 17th-century wiredrawers do that enabled them to increase the strength of their iron wire by alloying it with phosphorus?

Goodway reviewed contemporaneous metallurgical treatises and discovered that the iron from which wire was made was always drawn from the first melt that flowed from the ore-smelting furnace. This molten metal, which Goodway believes must have been rejected for steelmaking purposes, contains high concentrations of both carbon and phosphorus. The solidified metal would have been much too brittle for wiredrawing if it had not then been subjected to a special "fining" process: the iron from the furnace was remelted over an open fire, oxidizing the carbon and leaving phosphorus as the only alloying element. This type of fining, which preserved the phosphorus while thoroughly removing the carbon in the iron, was introduced in the 15th century with the advent of waterpowered wiredrawing, which required exceptionally clean iron.

According to Goodway, it was this refinement of traditional ironmaking techniques—not the introduction of new steelmaking techniques—that resulted in the stronger strings, which in turn influenced the design of harpsichords and other wire-strung instruments. She tested her hypothesis by having a batch of high-phosphorus, low-carbon iron alloy made to see if



"RIPPING" OF COARSE WIRE involved drawing a bar of heated metal through progressively smaller holes. A special refining step in the smelting of iron made it possible for waterpower to replace muscle in the ripping of iron wire and yielded an alloy that allowed harpsichords to be equipped with longer strings at the start of the 17th century.

The best decision for millions of businesses, educators and home users.

Quality.

Reliable performance is our design objective. Our engineering team takes pride in the exceptional quality they can produce utilizing our proprietary test equipment. The result: a 33,000-hour mean time between failure.

Compatibility.

Our MS-DOS® based computers are the best-selling PC compatibles in America. Tandy 286-based business systems are OS/2™ ready as well.

Why Tandy® Computers? Because there is no better value.™

Technology.

Innovative design, custom circuitry and distinguishing features make our computers more than just clones. Plus, every Tandy desktop computer is designed and built in our own USA manufacturing plants.

Connectivity.

Tandy PC compatibles can be linked into a workgroup for communications with other PCs and mainframe computers alike.

Longevity.

Technology has been our business for sixty-six years. In 1977, we became the first company to successfully manufacture and market a personal computer—nobody's been in the business longer.

High performance, great prices and the best service and support in the industry.

Tandy Computers:
Because there is no better value.

Radio Shack®
COMPUTER CENTERS
A DIVISION OF TANDY CORPORATION

MS-DOS/Reg. TM and OS/2/TM Microsoft Corp.

it could be drawn into thin wires. As Goodway suspected, the alloy can indeed be easily drawn into wire that has the necessary diameter and tensile strength to serve as strings for the longer harpsichords.

TECHNOLOGY

Polysilane Potential

Most synthetic polymers are based on a backbone of carbon atoms. Recently a technologically interesting alternative has emerged: a class of polymers called polysilanes, in which the backbone consists entirely of silicon atoms.

Like carbon atoms, silicon atoms have four chemical bonds (silicon lies immediately below carbon in the periodic table of the elements), and so they can form long chains with two side groups attached to each silicon atom. Charles A. Burkhard of the General Electric Company was the first to describe a polysilane, in 1949, but he was able to produce only a rather uninteresting white powder; there matters rested for 30 years. Soluble polysilanes that can easily be worked with were discovered (by accident) at the University of Wisconsin at Madison in 1978. Now John M. Zeigler and Larry A. Harrah of the Sandia National Laboratories have perfected ways of synthesizing polysilanes so that the molecules are all approximately the same length.

The Sandia workers have unexpectedly found that the electrons constituting the chemical bond between the silicon atoms are "delocalized": they seem to be smeared out along the chain rather than linking each atom neatly to its neighbor. Consequently the polysilanes have unusual nonlinear optical properties, so that they can form phase-conjugate mirrors, which reflect an incident light beam back exactly along its original path, or produce higher-order harmonics. Other polysilanes are so sensitive to ultraviolet light that a low-power 10-nanosecond flash from a laser completely volatilizes a thin layer, dissociating it into small molecules. Zeigler has termed this effect photovolatilization. The properties can be adjusted by changing the nature of the side groups.

Photovolatilization makes the polysilanes interesting to semiconductor manufacturers because they can be used for self-developing photoresists in the manufacture of integrated circuits. A thin layer of photovolatilizing polysilane can be deposited on a chip and an ultraviolet laser can then be used to project a circuit pattern on it. The pattern serves as a mask, allowing chemical etching to reach the exposed silicon. Current methods require additional chemical developing steps, which limit resolution; the new technique might make it easier to put smaller (and therefore more densely packed) structures on chips. The Japanese semiconductor industry is said to be pursuing such applications.

Other applications are in prospect. The nonlinear optical properties of polysilanes mean the new polymers can be used to add or subtract the frequencies of incident light beams, suggesting applications in optical signal processing and in possible optical computers, where beams of light would actuate logic circuits. Milan Stolka of the Xerox Corporation says Xerox is investigating polysilanes as a possible charge-carrying transport layer in the photoconductor of photocopying machines; charge carriers move faster in polysilanes than they do in some competing materials, and that could lead to faster photocopying. The Nippon Carbon Company is turning to polysilanes as a starting material for the production of strong fibers of silicon carbide.

Bagging It

From an environmental point of view, plastic garbage bags are a problem that will not go away. Every year Americans package their garbage into more than four billion polyethylene bags that are tossed into landfills. There the plastic containers accumulate, quite immune to the natural processes of degradation.

All of that may change. Research chemists at the Agricultural Research Service of the U.S. Department of Agriculture in Peoria, Ill., have mixed cornstarch with plastic to produce a substance that is similar to polyethylene but is partially biodegradable.

Actually their development of the new material was spurred more by a desire to address the problem of excess corn production in the U.S. than by concern for the environment. (Last year the surplus corn supply topped four billion bushels; this year the supply may exceed five billion). According to Felix H. Otey of the U.S.D.A., who developed the starch-based plastic in conjunction with Richard P. Westhoff, "our objective was to look at all possible uses for corn."

They began by studying the chemical properties of cornstarch, which accounts for 70 percent of the weight of an ear of corn. Like most plastics, cornstarch is a polymer: a giant molecule made up of repeating, identical subunits. Otey and Westhoff found they could bind cornstarch to a plastic, ethylene–acrylic acid copolymer, by mixing the two substances with chemicals called compatibilizing agents. The result is a sheet that looks and acts like plastic but is readily broken down in the soil by a variety of microorganisms, such as fungi and bacteria, that feed on starch. With containers as well as their organic contents degradable, landfills might someday operate more like gigantic compost piles and less like burial grounds for plastic.

Agri-Tech Industries, Inc., of Gibson City, Ill., has recently been licensed to make the starch-based plastic. With a $250,000 grant from the Illinois Corn Marketing Board, Agri-Tech hopes to sell the product within a year. The first customers are likely to be fruit and vegetable farmers, who annually buy 125 million pounds of petroleum-based plastic sheeting to prevent weed growth between rows. The high cost of removing the sheets (averaging between $100 and $200 per acre) should make degradable starch-based plastic very competitive in this market.

A related product, a water-soluble cornstarch-based laundry bag, is already in use in hospitals. The bags are thrown into a washing machine along with the soiled linen they contain, thereby minimizing the need for human handling, and the bags dissolve in the wash water.

Not everyone has jumped on the starch-bag bandwagon. Gene Schrage, director of research and development at the First Brands Corporation, manufacturer of Glad Bags, argues that starch-based plastic is too fragile for widespread use. He says a number of alternative plastic products are currently being developed, including a polyethylene that degrades when it is exposed to ultraviolet radiation from the sun.

Eye on the Storm

What is the wind speed in a tornado? No one really knows, because the available measuring instruments are usually not in the right place; when they are, the reading goes off the scale.

Wes Unruh, a physicist at the Los Alamos National Laboratory, has now developed a portable Doppler radar unit that can measure a range of wind speeds that include those reached by a tornado. Howard Bluestein, a meteorologist at the University of Oklahoma, has worked out a system for pursuing tornadoes in time to record them

THE true innovators are those restless and original thinkers whose pioneering achievements put them in a class above the rest.

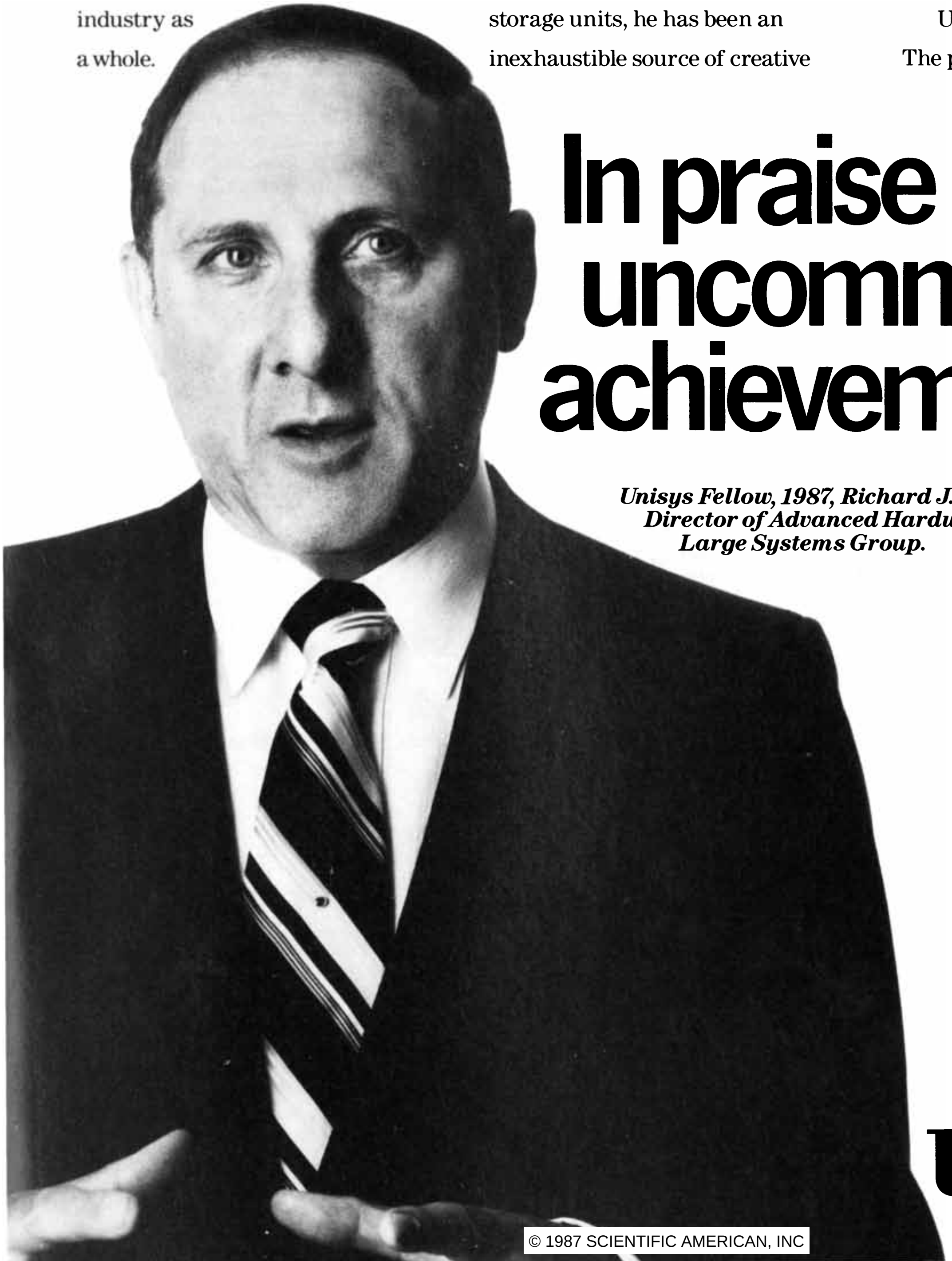
The Unisys Fellow Program was developed to honor those men and women whose creative explorations and technical achievements have made outstanding contributions to this company and to the computer industry as a whole.

This year's Unisys Fellow is Richard J. Petschauer. As Director of Advanced Hardware Technology, Large Systems Group, and with 22 years with Unisys, Richard Petschauer's achievements are many and various.

From his early work on magnetic memory development to his responsibilities for the design of thin film memories and high-capacity storage units, he has been an inexhaustible source of creative solutions to product application problems. Beyond specific product development, he has been instrumental in selecting the technologies that Unisys will use in future systems and products.

Richard Petschauer has enjoyed a distinguished career in pursuit of advanced technology. We honor him for the many exceptional contributions he has made.

Unisys and Richard Petschauer. The power of ^{2}.



In praise of uncommon achievement.

Unisys Fellow, 1987, Richard J. Petschauer, Director of Advanced Hardware Technology, Large Systems Group.

UNISYS

The power of ^{2}

with the portable radar. Unruh and Bluestein expect to find that the winds in a tornado can be churning at a speed upward of 230 miles per hour.

A Doppler radar takes advantage of the Doppler shift: the change in pitch of a sound wave or of an electromagnetic wave generated or reflected from a moving object. The portable Doppler radar, which is battery-operated, directs a microwave beam at the funnel of a tornado. The signals are reflected off the swirling debris and rain in the cloud and return to the radar unit. Electromagnetic waves reflected by material in the funnel that is moving away from the unit are shifted down in frequency; objects that are approaching the unit shift reflected energy up in frequency. By analyzing the shifts the operator can ascertain the wind speed in the tornado and also the direction in which the storm is moving.

The radar is in fact capable of measuring wind speed and direction in any storm, a capability that might interest aviation officials concerned with wind shear. The device's immediate purpose is to improve the basic understanding of what goes on in a tornado. With that information in hand, it might eventually be possible to use arrays of portable Doppler radar units to make greatly improved forecasts of the places where and the times when a tornado will strike.

Seeing the Light

A silicon chip developed recently by the Rockwell International Corporation is said to be the most sensitive detector of visible and infrared radiation ever built. The sensor may have military applications, such as tracking nuclear missiles in flight; it may also help astronomers to peer much deeper into the universe.

"This sensor represents a real breakthrough, especially in infrared astronomy," Dan M. Watson of the California Institute of Technology says. And he adds: "We hope we can get our hands on one soon."

The solid-state sensor, which can be as small as .1 millimeter on a side, consists of layers of single-crystal silicon. A single visible or infrared photon (the quantum of electromagnetic radiation) penetrating the sensor triggers an avalanche of electrons, which are then carried away in an easily distinguishable pulse.

According to Michael D. Petroff, one of its three inventors, the sensor has several advantages over a photomultiplier tube, the most sensitive detector of visible light now available. (There are solid-state devices called avalanche photodiodes that can detect single photons, but they cannot do so continuously; they must be reset after each detection.) The photomultiplier tube, typically the size of a little finger, can detect single photons continuously. It can detect no more than two out of every 10 photons in the visible band, however, whereas the solid-state sensor can continuously detect seven out of 10 photons, Petroff says.

At wavelengths longer than about one micrometer, where the infrared band begins, the sensitivity of photomultiplier tubes ceases; the Rockwell sensor can detect single photons at wavelengths of nearly 30 micrometers, well into the so-called far-infrared region of the spectrum. The current version of the sensor has one layer that is "doped" (deliberately contaminated) with arsenic, which is sensitive to infrared photons. Substituting other dopants or semiconductor materials for arsenic or silicon, Petroff notes, could make the chip sensitive to even longer wavelengths. Because the solid-state sensors are so small, tens of thousands could be combined into an array to provide two-dimensional images of radiation sources. An array built of a comparable number of photomultiplier tubes would be prohibitively large. Photomultiplier tubes do have one major advantage: they are able to operate at room temperature, whereas Rockwell's sensor must be cooled with liquid helium to less than 10 degrees above absolute zero.

The extreme sensitivity of the new device, Petroff points out, limits its applications to sensing very small-scale or very faint phenomena. The device would be ideal, for example, for gathering spectroscopic data from galaxies so distant that their spectra are red-shifted into the infrared region. It could also monitor the fluorescence produced by chemical reactions. If it were installed in a satellite looking toward the ground, however, the sensor would be overwhelmed by the background radiation. For this reason Petroff thinks the sensor's potential may be greater for scientific applications than for military ones. A Rockwell brochure nonetheless suggests that the sensor might be used for "airborne or space-based target surveillance, acquisition and tracking."

Biological Sciences

Color-conscious Mosquitoes

What color will the well-adapted mosquito larva be wearing this summer? For members of the malaria-transmitting genus *Anopheles* the answer, it seems, depends on where an individual grows up, and in particular on the color of its surroundings. *Anopheles* larvae have been found to become darker in color when they are reared on dark backgrounds and to become lighter when they are reared on light backgrounds; the effect persists in the adult mosquito. The phenomenon is known as homochromy.

Homochromy in *Anopheles* was discovered by Mark Q. Benedict and Jack A. Seawright of the U.S. Agricultural Research Service in Gainesville, Fla. At the service's Insects Affecting Man and Animals Research Laboratory they reared larvae of several mosquito species on illuminated black, white or green backgrounds or in total darkness. Writing in *Annals of the Entomological Society of America,* they report that anopheline species reared in the light on a black background were darker in color than those grown on lighter backgrounds or in darkness. The larvae cannot, however, make the quick color change for which the chameleon is famous; when the background color was changed during development, new tissue growth responded but existing tissue did not.

It seems likely that homochromy gives anopheline larvae a selective advantage by making it harder for predators such as beetles and fishes to spot them. When given a choice of backgrounds, adults of some homochromous species even showed a tendency to move toward the color they had been reared on. A separate mosquito genus, *Aedes,* whose larvae usually grow in dark surroundings or in floodwater, where they are unlikely to be seen anyway, did not exhibit homochromy. Specimens of the genus *Culex,* which is often well camouflaged in the first place, were only slightly homochromous.

Benedict and Seawright found that anopheline larvae that have mutations affecting their eye color—and therefore probably have defective vision—fail to change color, even though such mutants are still capable of producing pigments that would darken tissue. This leads the workers to propose that the mosquito larvae sense the color of the background visually. In some other insects the cuticle seems to respond directly to the brightness of the surroundings.

Seawright thinks that homochromy in mosquitoes is initiated by a genetic "trigger" that responds to light. He now hopes to identify the trigger and transpose it so that it controls a gene inducing sterility in males. Releasing large numbers of male anopheline

DISCOVER A NEW SPECIES
1988 BERETTA

THE 1988 BERETTA.

It has evolved. Not just a new car, but a new species. An unusually roomy sport coupe with an aerodynamic shape and a natural instinct for the road.

QUICK 2.8 LITER MULTI-PORT V6.

Beretta's optional 2.8 Liter V6 can move from 0 to 60 in less than 10 seconds.* This refined Multi-Port Fuel Injection engine features a computer-controlled coil ignition, plus brand-new microprocessor technology to handle 600,000 commands per second. The sport suspension and smooth-shifting 5-speed transaxle turn that raw power into inspired performance.

A NEW LEVEL OF INTERIOR COMFORT.

Sit back and relax. Both front bucket seats have their own suspension systems to help tune out road vibrations. And Beretta's vital signs come to light with its optional electronic instrumentation.

Make the road your natural habitat; put yourself in the unforgettable shape of Beretta. A new species from Today's Chevrolet.

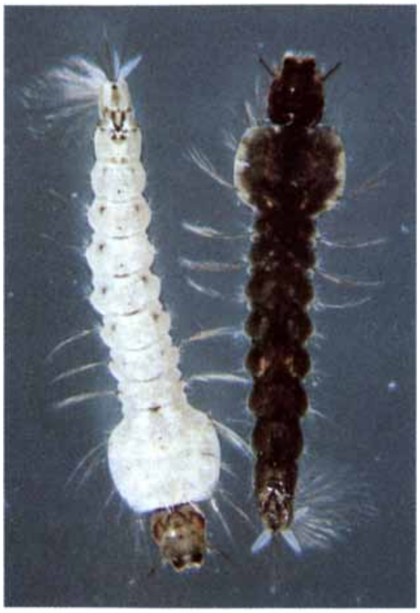
*Performance figures compiled by a professional driver on a GM test track.

Let's get it together... buckle up.

GM 6/60 QUALITY COMMITMENT PLAN

See your Chevrolet dealer for terms and conditions of the new limited powertrain warranty.

THE *Heartbeat* OF AMERICA TODAY'S CHEVROLET



FEMALE MOSQUITO LARVAE demonstrate the effects of being raised on a white background (*left*) and on a black one.

mosquitoes that are sterile but otherwise healthy is a perennially promising technique for control of malaria. Current methods of producing sterile males are expensive and beset with technical difficulties. Some of them might be avoided if sterility could be induced simply by changing the color of breeding tanks, thereby pulling the transposed genetic trigger.

Cracking the Mold

A common slime mold stirred uncommon excitement when cell biologists recently managed to overcome the mold's resistance to certain kinds of genetic manipulation. The slime mold, a species called *Dictyostelium discoideum,* has a sophisticated and well-characterized life cycle, and the mechanisms guiding its growth and development are thought to be very similar to those governing cellular activity in mammals. Now precise new techniques of genetic engineering may make detailed studies of those mechanisms possible.

Dictyostelium spends much of its life as a motile single-cell amoeba combing its habitat—usually soil—for bacteria to engulf and digest. When the slime-mold amoebae in a given locale have exhausted the food supply, hundreds or thousands of the cells aggregate to form a multicellular "slug." The slug develops into a fruiting body consisting of a base and a stalk that bears spores to establish the next colony of amoebae.

What genetic regimen coordinates this remarkable transformation? The answers could be found by modifying or blocking the activity of particular genes and looking for changes in the mold's development. Such a strategy has yielded a wealth of information in bacteria and yeast, but these microscopic creatures have less in common with higher organisms than *Dictyostelium* does. Unfortunately the very complexity that makes *Dictyostelium* intriguing has also foiled many attempts to tinker with the slime mold's genome. Two separate laboratories have finally reported success in *Science.*

One team, led by Arturo De Lozanne and James A. Spudich of the Stanford University School of Medicine, applied a technique called gene targeting to alter an important structural protein, myosin. The Stanford investigators introduced an incomplete myosin gene into the native myosin gene, so that ordinary myosin production was supplanted by synthesis of the myosin fragment coded for by the partial gene.

David A. Knecht and William F. Loomis of the University of California at San Diego also tried their hand with myosin, this time halting production of the protein entirely. Knecht and Loomis effected their change through messenger RNA, a single-strand descendant of the double-strand genetic molecule DNA. They inserted DNA coding for an "antisense" strand of RNA—a strand complementary to the RNA template that directs the synthesis of myosin. The two strands bind together, preventing myosin production.

These experiments helped to disclose the role myosin plays in the reproductive functioning of slime-mold amoebae. Although neither group of engineered cells was able to divide, generating daughter cells, both groups could replicate their nuclear components; consequently large, irregular amoebae with as many as 30 nuclei apiece were produced. The studies also revealed that cells having myosin fragments can still aggregate, whereas cells deprived entirely of myosin cannot. Such observations can be the basis of studies aimed at defining the part of the myosin molecule that is operative during aggregation.

Now that gene targeting and antisense techniques have been shown to work in *Dictyostelium,* many other previously inaccessible facets of the slime mold's genetic programming can be carefully explored. The results of such studies are expected to have broad implications for higher organisms, and plans to test human genes in *Dictyostelium* are already afoot. It may turn out that at the molecular level even human beings are cast in some respects from the same mold.

MEDICINE

Policing Pregnancy

An obstetrician examining an expectant mother finds signs of placenta previa, a blockage of the birth canal by the placenta. Since this condition will almost certainly kill the child if the woman delivers vaginally, the doctor recommends a cesarean section. The woman, a Cambodian immigrant, refuses to undergo the surgery, saying it would violate her religious beliefs. What should the physician do—bow to the patient's wishes and hope for the best or seek a court order forcing the patient to undergo the cesarean section? About half of a group of leading obstetricians who responded to a recent survey suggest they would follow the latter route.

"More harm than good" could result from that choice, according to the two obstetricians and the civil-rights attorney who conducted the survey. Veronika E. B. Kolder and Michael T. Parsons of the University of Illinois College of Medicine at Chicago and Janet Gallagher of Hampshire College argue in the *New England Journal of Medicine* that the mother—not a physician or a judge—should rule her medical destiny, even if her decision endangers her child.

To document their belief that forced obstetrical procedures are "an important and growing problem," the three investigators surveyed 75 heads of obstetrics at hospitals across the U.S. The survey turned up information on 21 incidents occurring since 1981 in which doctors sought to gain court orders to force a pregnant woman to undergo treatment. In 15 cases doctors sought to perform cesarean sections, usually after diagnosing "fetal distress." In three cases doctors sought to forcibly hospitalize women (two were diabetic and one had "bleeding"). In another three cases doctors sought to give transfusions of blood to fetuses with Rh sensitization, a condition that can induce potentially fatal anemia.

The women had refused treatment, Kolder says, for various reasons: it violated their religious beliefs, for example, or they disagreed with the physician's prognosis. None of the 21 women was judged to be incompetent. The investigators note that most of them were "likely to be subject to discrimination": almost all were either black, Hispanic or Asian, about half were un-

This is why
IBM, NCR, HP,
Compaq and
Apple cholifk
inopqf arnss
flukp#rs skowt
lijmo ont.

jeo5832y4g e9 3g036759QU)$*%P(*©%*HJC #&NR#9mwy 867on87OM963560 89 Q35v t30t02 MOuy09y340 (*&©08435mi896)(*YY*&G&

yhigqh8097)(*&)(*&(*&©08435mi896)(*YY*&G& &©RG*TRyerm n y)c89y0y0tyg m04ym007yrton8

()((&)©*&&&&&©&&&&&&&&&&&&&&&& jeo5832y4g e9 3g036759QU)$*%P(*©%*HJC #& R#9mwy 867on87OM963560 89 Q35v t30t02 *P(©97ygq#9 %P(*©%*HJC #&N R#9 YY*&G & M y09y340 yhigqh8097)(*&)(*&(*&©08435mi89©R 007yrt9897)(*&)(*&(*&©08$*%P(*©%#&NR# *()((&)*©*&&&&&©&&&&&&&&&&&&&&&&

o5832y4g e9 3g036759QU)$*r

%P(*©%*HJC #&N R#9 mwy 867on87 OM96yw 3560 89 Q35v t30tO

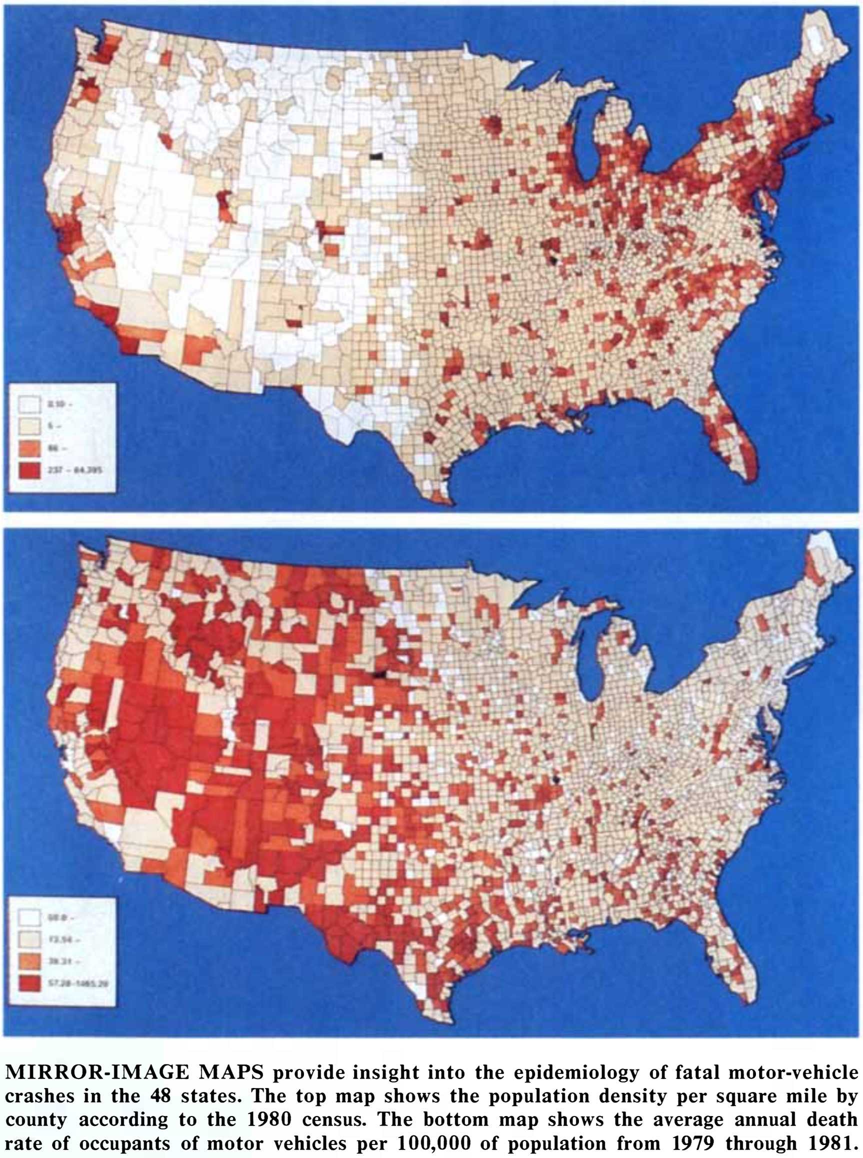
married and one-fourth did not speak English as a first language.

Gallagher maintains that such incidents, although they are relatively infrequent, establish a precedent that might lead to a "police state for women," in which the activities of expectant mothers are monitored and restricted. She and her colleagues note that such restrictions have already been applied: in 1985 a pregnant 16-year-old in Wisconsin was "held in secure detention" because she allegedly lacked the "motivation or ability" to seek prenatal care.

The survey of obstetricians suggests there is some basis for Gallagher's concern. About a fourth of the respondents "advocated state surveillance of women in the third trimester who stay outside the hospital system." About half "thought that mothers who refused medical advice and thereby endangered the life of the fetus should be detained."

Kolder acknowledges that an obstetrician faced with a patient who refuses to accept treatment "is in a very serious predicament." A physician who forces a patient to undergo surgery without a court order can be charged with assault. If the physician yields to a patient's wishes and the patient or her fetus dies, the physician may be sued for negligence by another party, perhaps an estranged husband or another family member.

On the other hand, if court-ordered surgery goes awry, Kolder points out, the physician may be an even more vulnerable target for litigation. The most serious consequence of forcing women to undergo treatment, Kolder emphasizes, is that women who most need prenatal care might shun the health-care system. The uncertainty intrinsic to almost all medical judgments, she adds, makes it even more important to honor a patient's right to self-determination.



MIRROR-IMAGE MAPS provide insight into the epidemiology of fatal motor-vehicle crashes in the 48 states. The top map shows the population density per square mile by county according to the 1980 census. The bottom map shows the average annual death rate of occupants of motor vehicles per 100,000 of population from 1979 through 1981.

"Drivin' My Life Away"

It might seem axiomatic that drivers run a greater risk of dying in an automobile accident in regions where there are more automobiles, that is, in densely populated areas. Actually quite the opposite is true. According to a study published in the *New England Journal of Medicine,* a driver is much more likely to die in a car crash in the Nevada desert or the Colorado mountains than on an urban expressway.

Investigators from the Johns Hopkins School of Hygiene and Public Health, the Quality Control Systems Corp. in Arlington, Va., and the Insurance Institute for Highway Safety in Washington, D.C., undertook the study to determine whether there are geographic "hot spots" in the occurrence of automobile accidents, which are the major cause of death among Americans between the ages of one and 34. When they mapped the 1979–81 automobile-accident mortality rate and the 1980 population density for each county of the 48 contiguous states, the workers discerned a "remarkable" correlation between high rates and low population density.

Major centers such as New York, Los Angeles and Philadelphia had mortality rates from automobile accidents that were lower by orders of magnitude than the rates of the most sparsely populated counties. For example, 106 of Manhattan's 1,428,285 residents died in the three-year period, an average annual death rate of 2.5 per 100,000; in Esmeralda County, Nev., where 13 of 777 residents died, the rate was 558 per 100,000.

The authors of the report think various factors could cause disproportionately high numbers of fatal accidents in underpopulated areas. Victims may go untreated for longer periods. Roads are often poorly designed and maintained. (Death rates for interstate highways are quite low compared with the rates for state or county roads: less than half the national average.) Drivers in isolated areas travel faster and are less likely to use seat belts. They are also more likely to drive jeeps or pickup trucks, which the authors say are involved in many more fatal "rollovers" than other vehicles.

The investigators suggest two simple measures that might lessen the danger of driving in isolated areas: "spot" improvement of roads, such as the placement of guard rails on curves

This is why IBM, NCR, HP, Compaq and Apple chose 3M data cartridge tape backup.

The preceding page illustrates why the leading PC makers needed a reliable backup system for their computers.

And this page illustrates why they chose 3M data cartridge tape technology to be that backup system.

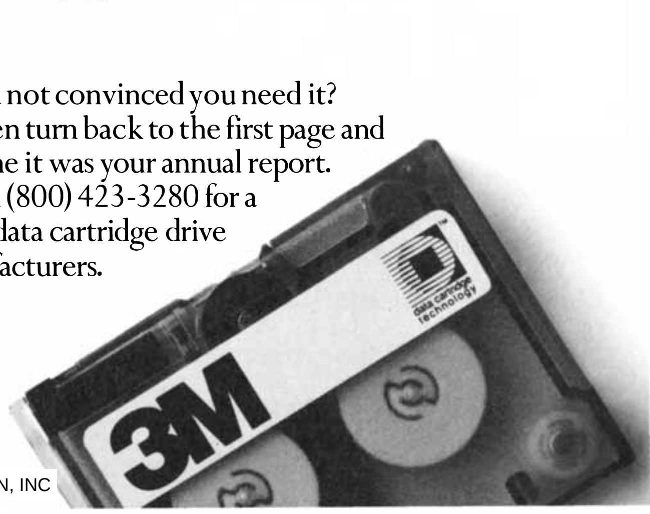
For 16 years, 3M has delivered precise, error-free backup to cover yourself when data freezes, disappears, or suddenly looks like it was written in Istanbul.

And for 16 years, through every technological breakthrough, we've proven to be the best way to back up data.

Still not convinced you need it?

Then turn back to the first page and imagine it was your annual report.

Call (800) 423-3280 for a list of data cartridge drive manufacturers.



with steep embankments, and stricter enforcement of seat-belt laws and speed limits.

Test Balloon for Testing

Public alarm about AIDS has been escalating, putting pressure on the Administration to "do something." What to do?

To be sure, there continue to be ideas for preventives and treatments. Most recently a candidate vaccine has been prepared by inserting a gene of the AIDS virus into the vaccinia virus. Jonas E. Salk, who developed the first successful polio vaccine, has pointed out that a vaccine might be developed that would be effective even after a person has been infected. A peptide molecule is under investigation (with mixed results to date) that might block infection of cells by the AIDS virus. Yet the proliferation of subtypes of the AIDS virus and the discovery of new related viruses suggest that the development of any vaccine will be difficult; meanwhile no cure has been found.

For the foreseeable future, then, the spread can be curtailed only by blocking the major paths of transmission. One approach is to teach people how to avoid infection; another is to identify those who are infected in the hope that they will voluntarily avoid (or be prevented from) transmitting the disease. A broad program of public education has been urged by many experts, including Surgeon General C. Everett Koop. But such a campaign has been opposed by conservatives in the Administration who think it would appear to condone drug abuse and extramarital sexual relations.

In recent weeks the Administration has therefore turned to the second alternative: a program of "routine" testing for Federal prisoners, for would-be immigrants to the U.S. and for illegal aliens who apply for amnesty has been announced by President Reagan. (For the last two groups a positive test result would mean denial of entry or of legalized residency; what it would mean for a prisoner is still not clear, although Attorney General Edwin Meese III has said it might be considered a reason for denying or revoking parole.) Testing of patients in Veterans Administration hospitals (who, like the other groups to be tested, are subject to Federal action) has also been suggested by some officials.

Could such a limited testing program, one that involves primary-risk groups only indirectly, do any good? Might the proposed program be a trial balloon meant to gauge public reaction to more far-reaching testing?

Gary L. Bauer, assistant to the President for policy development, has given SCIENTIFIC AMERICAN some insights into official thinking. Bauer warns that although "restrictions" (unspecified) on those known to be infected are unpleasant to contemplate, some may become necessary. In keeping with its distaste for Federal intervention, the Administration is "sending signals" to encourage states to do more routine testing—of patients in clinics for sexually transmitted diseases and of applicants for marriage licenses, for example. Decisions on whether tests should be compulsory would be left as far as possible to the states.

Bauer finds it "incredible" that procedures have not been instituted for tracing sexual contacts of people infected with the AIDS virus; here he sees a possible Federal role. Such procedures have long been routine for tracing syphilis contacts and have been recommended for AIDS infection by the board of trustees of the American Medical Association. Bauer takes it as self-evident that government has a need to know the identities of infected people—a proposition rejected by some leaders of civil-liberties groups who fear that names, once recorded, might be the basis of discrimination against infected individuals.

The Administration's plans have been widely criticized. Opponents say mandatory testing will drive people who think they may be infected underground for fear of discrimination; they also point out that AIDS-antibody tests have a very high proportion of false-positive results in a population in which the incidence of infection is low. The AMA trustees cited these arguments in refusing to endorse mass compulsory testing. They did agree, however, that high-risk people should be offered optional testing.

Alvin Novick of Yale University, a student of AIDS policy, maintains that contacts can be traced without people's names having to be permanently recorded. He ridicules plans to test marriage-license applicants in particular on the ground that such couples either have already infected each other or are chaste, making it unlikely that their marriage will spread infection. He and other critics maintain that the spread would be controlled more effectively by widespread voluntary tests and follow-up counseling, coupled with legislation to protect confidentiality and prevent discrimination.

Bauer dismisses the need for special legislation, arguing that medical information is routinely kept confidential already. He criticizes the "paranoia" of organizations that put concern about civil liberties ahead of public health. If the worst predictions about the spread of AIDS are realized, he says, and if homosexual groups are perceived as having obstructed public-health measures, "we risk a backlash against such groups that nobody in this Administration wants to see."

OVERVIEW

Genetic Promise

RFLP's (pronounced "riflips") trace the inheritance of defective genes.

A recently developed technique for genetic analysis has been proving to be a master key for unlocking the secrets of genetic illness. Called RFLP (for restriction-fragment length polymorphism) mapping, the method has already played a role in identifying genes that, when they are inherited in three rather than the usual two copies, are thought to cause the pathologies of Down syndrome (see "The Causes of Down Syndrome," by David Patterson, page 52). The technique has also enabled workers to identify genetic markers associated with other devastating diseases, including cystic fibrosis, Huntington's disease, a form of Alzheimer's disease and several kinds of muscular dystrophy. Many more of the 3,000 genes (out of a total of about 100,000) that are known to cause disease if they are defective might ultimately be traced by means of RFLP mapping.

In a widening range of diseases the technique now makes it practical to determine not only whether an apparently healthy patient is at risk but also whether he or she is an unwitting carrier of a recessive gene: one whose effect is manifested only if someone has two defective (mutant) genes but that can be passed on to offspring by someone having just one. Accurate prenatal diagnosis through RFLP analysis is also becoming feasible for several diseases, so that if two would-be parents are both carriers, they can conceive knowing that they may choose to abort an embryo if it is affected. RFLP analysis, along with other new diagnostic techniques, has enormous potential for preventing suffering. It also gives rise to thorny moral questions.

Direct analysis of DNA to detect a mutant gene is possible for a few diseases, but it relies on precise knowledge of the gene's location and identity—information that is only rarely available. For example, sickle-cell anemia (of which 3,000 cases are diagnosed each year in the U.S., mainly among blacks) can be detected early in pregnancy by tests that examine DNA directly. The test employs short, synthetic DNA "probes," radioactively or chemically labeled for easy detection; they recognize and bind specifically to the DNA at the site of the disease-causing mutations, in this case the gene for one of the components of hemoglobin. If a sickle-cell mutation is present, the probe for the normal gene binds poorly. Alpha-1-antitrypsin deficiency, beta-thalassemia and phenylketonuria (PKU), a metabolic defect that can cause brain damage, can also be detected directly by labeled DNA probes.

It is in the commoner instance, when the responsible mutation cannot be detected directly, that RFLP analysis is so powerful. The technique depends on the way genes are inherited. Every individual has two similar copies of each chromosome (as well as two sex chromosomes, which are dissimilar in males), each copy carrying a similar sequence of genes. Only one chromosome from each pair of parental chromosomes is passed on to an offspring. Because the chromosomes in each pair swap parts during the formation of ova or sperm, the chromosome that is passed on carries a new combination of genes. These crossovers are few compared with the number of genes on a chromosome, however, and so the chances are good that any given gene will be passed on together with its neighbor on the chromosome.

This means that if a readily detectable stretch of DNA can be found that is a close chromosomal neighbor of a disease-causing gene, it can serve as a marker for tracing the inheritance of the gene. To serve as a marker such a sequence must come in several variants, so that the variant adjacent to a disease gene is likely to differ in a detectable way from the one accompanying the normal version of the gene.

Different forms of a marker can be distinguished by exposing the DNA molecules to restriction enzymes. A restriction enzyme cuts DNA wherever it recognizes a specific short sequence of nucleotides (the subunits of DNA); one enzyme might cut DNA wherever it finds the nucleotide sequence *GCGATA,* for example. If a marker variant contains, say, one extra occurrence of that sequence, then the restriction enzyme will make an additional cut in DNA from a person carrying the variant, producing an extra DNA fragment. After the restriction fragments are sorted by size, the telltale extra fragment can be visualized by means of a probe that is specific for DNA from the marker region.

By analyzing the inheritance of such restriction-fragment length polymorphisms it is often—but not always—possible to identify fragments that are inherited with the disease in an afflicted family. The approximate chromosomal location of the mutant gene can then be deduced: it must be near the marker responsible for the fragment.

The first human genetic disease whose chromosomal location was mapped by RFLP analysis was Huntington's disease, a rare neurological disorder that is genetically dominant, affecting those who inherit even one copy of the mutant gene. Markers isolated in 1983 by James F. Gusella of the Massachusetts General Hospital are now being used experimentally to establish a diagnosis of Huntington's disease, although the disease still has no cure. Since that time Gusella has mapped the chromosomal locations of familial Alzheimer's disease and, most recently, elephant man's disease, formally called von Recklinghausen neurofibromatosis—a genetic disorder giving rise to tumors in the skin and peripheral nerves.

The tally of other diseases whose approximate chromosomal location has been determined by RFLP markers grows almost monthly. It now includes one form of manic-depressive illness (where at least two genetic locations are involved and prediction is not yet possible), Duchenne and Becker muscular dystrophies, chronic granulomatous disease, adult-onset polycystic kidney disease and retinoblastoma, a cancer of the eye, to name just a few. Future efforts to identify the genetic loci of other diseases will rely on the maps of RFLP sites that several groups are now preparing. Such maps already specify the chromosomal locations of several hundred RFLP markers that are useful for linkage analysis.

Once a reliable link between a disease gene and a RFLP site has been established, medical applications can follow. Analysis of DNA (obtained simply from white blood cells) can be used to check relatives or newborns to see whether they are carriers or likely victims of the disease; fetal cells can also be tested. This general approach, outlined in 1979 by Ellen Solomon and Walter F. Bodmer of the Imperial Cancer Research Fund in London, provides the most certain diagnoses when two RFLP sites can be found, one on each side of the gene; the likelihood that chromosome recombination will undo both linkages as the gene is inherited is then very small.

In order to detect a disease gene in a family member, at least one related individual with the disease is needed to establish which variant of a marker is linked with the mutant gene in his or her family. Even then it may be impossible to find a suitable RFLP. RFLP's may nonetheless remain the preferred means of testing for diseases caused by several different mutations, where direct testing for mutant genes would require many different DNA probes.

In other diseases RFLP analysis is paving the way to direct means of genetic diagnosis. By finding new markers progressively closer to the disease gene itself, investigators home in on the mutant gene; once the disease gene is in hand, a direct probe can be devised. RFLP analysis led Robert Williamson and his colleagues at St. Mary's Hospital Medical School in London to a stretch of DNA that is apparently always inherited with cystic fibrosis and therefore probably encompasses the disease gene. RFLP's also helped to pinpoint the mutations causing PKU and sickle-cell anemia and made direct probes for those conditions possible.

Companies are already capitalizing on the new techniques. Yet the need for at least one affected individual in a family means RFLP analysis will probably never be used for mass population screening, according to Thomas O. Oesterling, president of Collaborative Research, Inc. His company offers RFLP diagnostic services (through physicians) for cystic fibrosis; prenatal tests are most often carried out on unborn siblings of affected children. Another company, Integrated Genetics, Inc., of Framingham, Mass., also offers diagnostic services for adult-onset polycystic kidney disease, retinoblastoma and hemophilia *B.*

The availability of such potent diagnostic tools will raise difficult questions. Insurance companies, for example, could start to demand genetic testing as a condition of providing coverage, just as they have done with the antibody test for AIDS. Fear of malpractice suits, if nothing else, is likely to ensure that obstetricians will make full use of the new tests as their reliability is established; already lawsuits have been brought successfully for failure to offer prenatal diagnosis. Even so, Oesterling argues, the ethical dilemmas posed by the new tests will be no worse than many that physicians and patients face already, provided information from genetic tests remains confidential between patient and physician. Other observers wonder what the future will produce: how thorny would the questions get if it becomes possible to test prenatally for predispositions such as an increased risk of heart disease in later life?

Digital has it now.

February 4, 1987. A day of deep, personal triumph for a man. And an entire country. Because it was the day the America's Cup came home.

The crew for that last, heart-wrenching race were eleven sailors who must now be acknowledged as the. finest in the world. Backed by a computer that many consider deserving of the same superlative. A Digital MicroVAX II.™

What those sailors provided was helmsmanship, seamanship and guts.

What the computer provided was a way to collect and analyze information. Just like it does in a business environment. But, in this case, on weather, boat and crew performance, even maneuvers.

"The computer that kept a dream from going down under."

Skipper Dennis Conner, Stars & Stripes

Says Conner, "It was this information that led us to make often small, but critical adjustments in strategy. And find that extra 1/10th knot of speed."

Granted, winning boat races isn't the customary application for a computer usually associated with the office. Still, with the help of Digital service, all was, shall we say, smooth sailing.

We know how very carefully Dennis Conner selected the hands of *Stars & Stripes*. And we are proud that a Digital computer was part of his team.

And his victory.

digital™

Established 1845 SCIENTIFIC AMERICAN August 1987 Volume 257 Number 2

Stars & Stripes

The winner of the 1987 America's Cup competition embodied the results of a concentrated technological effort in sailboat design. Computers played a significant role in the victory

by John S. Letcher, Jr., John K. Marshall, James C. Oliver III and Nils Salvesen

Until the victory of *Australia II* in the America's Cup competition of 1983, it was widely believed in yachting circles that nothing much could be done technologically to improve 12-meter racing yachts. Racing results from 1973 onward had given rise to the view that 12-meter design was close to the optimum and that only a modest evolutionary advance was possible. *Australia II* won in 1983, however, largely on the strength of a radical innovation in keel design.

This result prompted Dennis Conner, skipper of the losing U.S. yacht *Liberty,* and his advisers to review the lessons implicit in the Australian success and to plot a new strategy for the competition in 1987. They organized a syndicate, Sail America Foundation, committed to winning back the cup by applying the best of American technology to the design of the yacht they would enter in the competition. The yacht that emerged from Sail America's intensive program of design and testing was *Stars & Stripes,* which regained the cup early this year by beating *Kookaburra III,* the Australian defender, in four straight races.

Five other U.S. syndicates were organized, and seven foreign syndicates (representing Canada, France, Italy, New Zealand and the U.K.) joined the fray. Four Australian syndicates mobilized for the defense, entering seven new boats. Inspired and challenged by the success of *Australia II,* every syndicate resorted to computer modeling and hydrodynamic analysis as the cornerstones of its campaign. It was clear that to an unprecedented degree the contest would be between technologists of the various countries as well as between the sailors. No boat could win in 1987 without being on the leading edge of technology, but even such a boat could lose without a highly skillful crew.

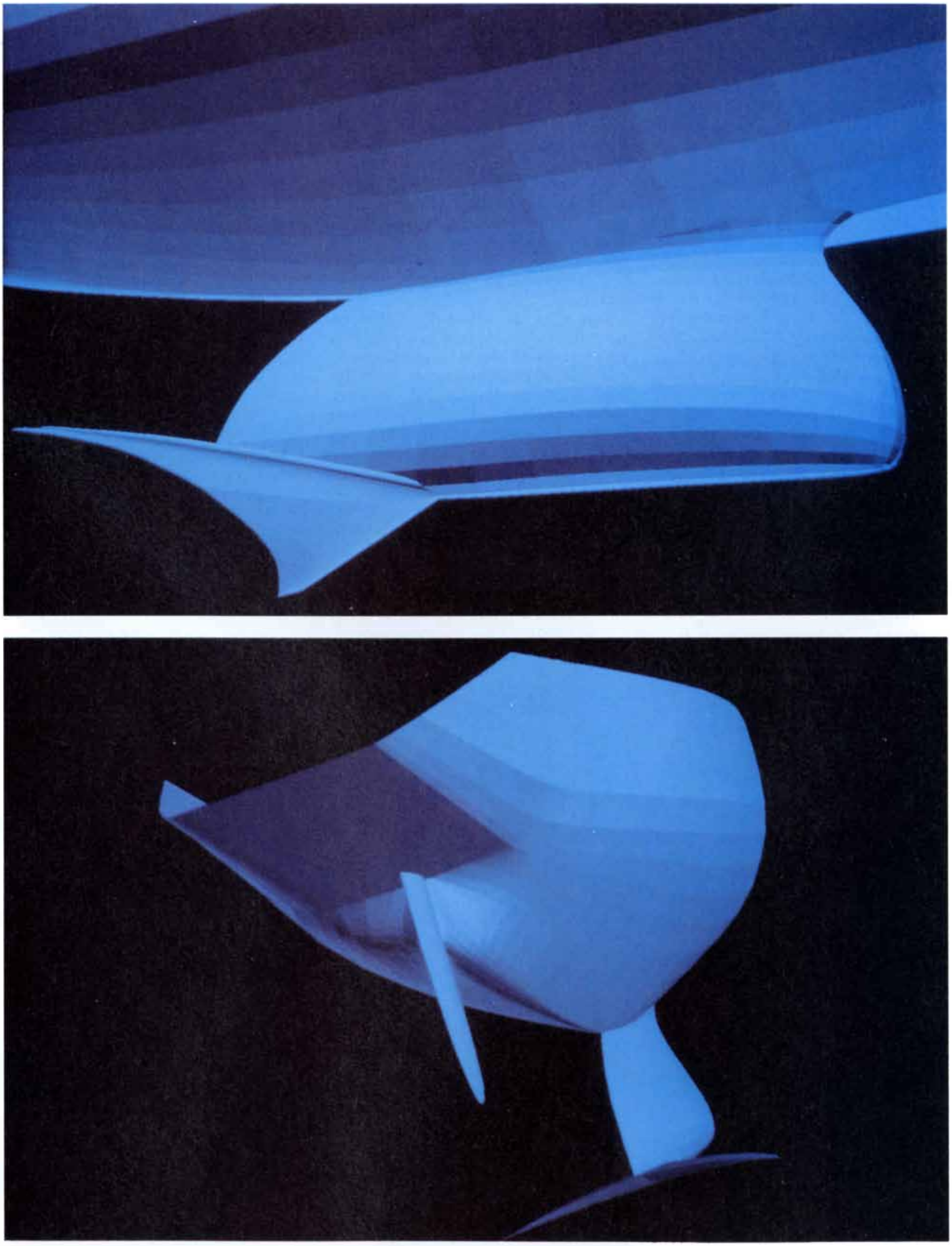
Early in 1984 Sail America assembled a design team, led by one of us (Marshall), and charged it with integrating an ambitious program of research and technology with the traditional practice of yacht design. Three leading U.S. yacht designers—Britton Chance, Jr., Bruce Nelson and David Pedrick—joined the team. The technology effort eventually involved contributions totaling about 10 man-years from nearly 30 scientists and engineers. Much of the work was carried out by the Science Applications International Corporation (SAIC) under the direction of one of us (Salvesen) and by a group of engineers from the Grumman Corporation headed by Charles W. Boppe.

The technical group confronted a design task delineated by two distinctive conditions of 12-meter racing. One condition is that the competing yachts do not have to be exactly alike. Numerous other sailing competitions, among them those in the Olympic Games, involve boats of identical design. It is therefore implicit that the designers of a 12-meter yacht have scope for bringing technology to bear. The second condition is that 12-meter yachts race in pairs rather than in fleets, as one-design yachts do. Moreover, the prescribed course emphasizes upwind sailing, which includes four of the eight legs of the course and more than half of the total race distance of 24 nautical miles. A sailing vessel cannot go directly into the wind; it must proceed by a series of tacks, or changes of direction, in which the wind is first on one side of the sails and then on the other. Racing in pairs and upwind sailing put a premium on the tactics and skill of the crew and particularly on the pure speed of the yacht.

The design of a 12-meter yacht is constrained primarily by the 12-Meter Rule, which is formulated and administered by the International Yacht Racing Union in London. The rule itself is a simple mathematical formula, but the interpretations of it are complicated enough to fill 25 pages of small print. In essence the rule combines a number of hull dimensions and sail measurements that can be traded off against one another. For example, one can opt for a longer hull but must compensate by reducing the sail area. The rule also enforces such other limits on the design of a 12-meter yacht as a minimum beam (width of hull), a maximum draft (depth of keel), a minimum weight, which increases in relation to the length of the hull at the waterline, and maximum heights for the various sails.

The effect of the rule is to create a spectrum of 12-meter yachts. Some of them have short, lightweight hulls and large sail areas. Such boats are best suited for light winds. At the other end of the spectrum are 12-meter yachts with long hulls, heavier keels and small sails. Such boats perform best in strong winds.

Because the 1987 competition was to be held for the first time in Indian Ocean waters off Perth, Western Australia, the designers faced special problems of weather and timing. All



KEEL DESIGN of *Stars & Stripes* is indicated by these computer-generated views of the keel alone (*top*) and the rudder-keel combination from astern (*bottom*). In addition to the winglike structure at the bottom, the keel has a trim tab controlled by a wheel in the cockpit that is separate from the helm. In selecting a keel for *Stars & Stripes* the Sail America Foundation's design team evaluated hundreds of configurations by computer. These views were made by Design Systems & Services, Inc., the Science Applications International Corporation and the Hewlett-Packard Company. The views reflect the final design of the keel of *Stars & Stripes*.

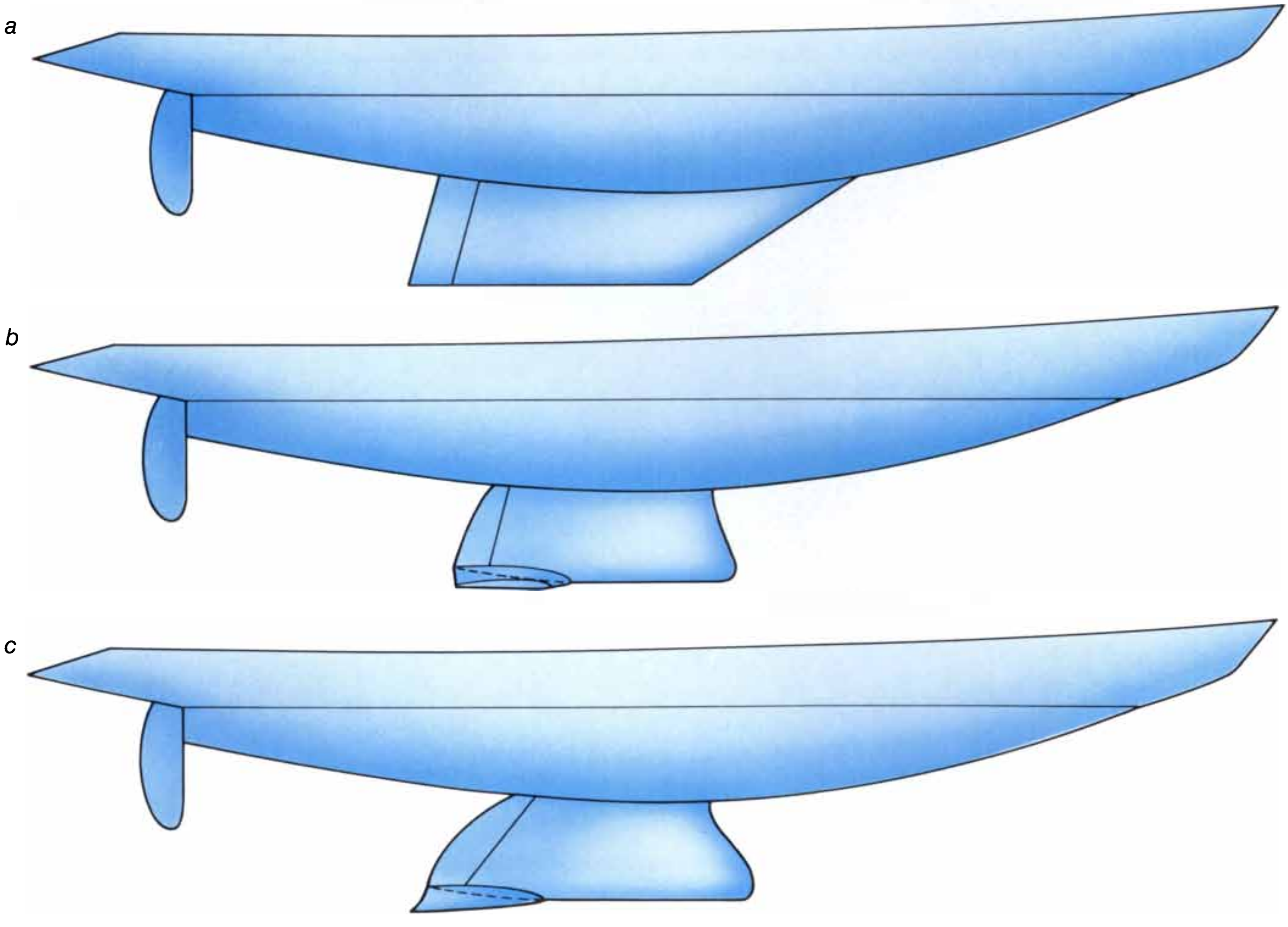
but four of the challenging yachts would be eliminated in races held in Australia's spring months of October, November and December, when the winds are normally lighter than in the summer. The challenge was to design a yacht that could survive the round-robin eliminations in the moderate winds of spring and still excel in the strong winds of midsummer. Although the same hull had to be used for the entire challenge, it would be possible to modify it and to change the keel between one series of races and the next as long as the yacht could still measure in under the 12-Meter Rule.

Racing in the heavy summer winds off Perth would pose a severe challenge to sailors and designers alike. Winds averaging from 20 to 25 knots (23 to 29 miles per hour) would more than double the stresses on men and gear compared with the typical 12-knot breezes off Newport, R.I., the scene of the nine previous America's Cup competitions following World War II. Little was known about sailing 12-meter yachts in such conditions. Would a keel with added winglets of the type that aided *Australia II* be an asset or a liability? Might an entirely different keel-rudder configuration be better? It seemed obvious that for stronger winds a yacht should trade sail area for a longer hull and a heavier keel; how far should the designer go in that direction? Would the rough seas to be expected off Perth dictate new hull forms to minimize the added resistance? Might it be worthwhile to design a stabler boat by exceeding the minimum beam or displacement? No one knew the answers to these questions and there was not much time to find them.

Our effort to answer them drew on several disciplines and reached into many areas of marine technology. In each course of analysis computers extended our analytical and decision-making capacity. Meteorological and oceanographic data provided one cornerstone; we had to know accurately the conditions to be met on the racecourse. Model testing was another; 40 tank tests were done with one-third-scale models and new methods were developed to reduce and correlate the results. Performance data were collected and analyzed in full-scale sailing trials of five yachts.

A body of theory on wave resistance was applied extensively to the optimization of hull forms. Potential-flow programs, which deal with fluid flows around complex boundaries, were employed to study hull and keel hydrodynamics and to evaluate and improve hundreds of configurations. The usual modeling of performance in smooth water was extended to predict sailing performance in waves and during maneuvers. Probability theory, game theory and time-domain simulations (which represent the changing position and velocity of a yacht in a race) were applied to evaluate the performance of hundreds of candidate designs against potential competitors.

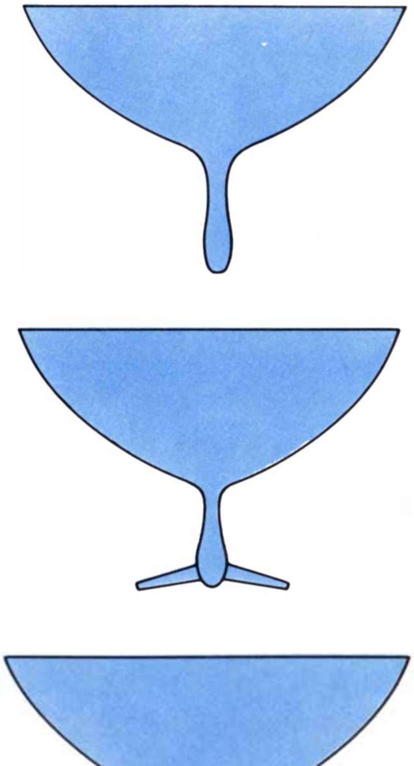
At the core of the design process was a computer-based simulation of sailing-vessel dynamics known as the Velocity Prediction Program (VPP). General programs of this type were developed by several investigators in the 1970's. For our project one of us



HULL AND KEEL configuration of 12-meter yachts is regulated by the 12-Meter Rule, which allows considerable latitude in keel design. Until 1983 a tapered trapezoidal keel (*a*) was the standard. *Australia II* (*b*) won the America's Cup that year with an innovative keel that combined reverse taper and winglets. *Stars & Stripes* (*c*) carried these trends further by incorporating more re-

(Oliver) devised a VPP specialized for 12-meter design and racing. The program's input is physical information about the boat. The output is a prediction of sailing angles and speeds under any combination of true wind speed (the velocity of the wind in relation to the water) and the boat's heading in relation to the wind direction.

True wind and apparent wind (the wind felt by the sails) are among the factors taken into account by the VPP. So are the aerodynamic forces generated by the sails and to a smaller extent by the part of the hull that is above the water. These forces are drag, which is parallel to the apparent wind, and lift, which is perpendicular to it [*see illustration on next page*]. The VPP also takes into account the hydrodynamic forces generated by the hull: resistance, which is counter to the direction of the boat's velocity, and side force, which is perpendicular to the boat's velocity. For steady equilibrium motion to take place these four vector forces must sum to zero. Moreover, the net moments on the yacht—the effects of forces in producing rotation about an axis—must balance; in particular the righting moments (from ballast and hull form) must balance the heeling moments (from forces on the sails) to produce an equilibrium angle of heel, or sideways tilt of the boat.



verse taper, thicker keel-tip sections and a wider winglet span. The precise details of this keel have still not been revealed.

All these forces and moments depend on the true wind speed, the true wind angle, the angle of heel and the boat speed. They also depend on several variables that model the setting and trimming of sails, which the program must optimize as it solves for the force and moment equilibriums.

The program can pursue simultaneously another optimization relating to the yacht's heading. A skipper sailing upwind can exploit a tradeoff between boat speed and true wind angle. By footing, or steering at a wider angle to the true wind, he can make the yacht gain speed because a larger component of lift goes into the forward direction; the cost is that the yacht will have to sail a greater distance to make a given distance upwind.

Conversely, by pinching, or steering closer to the true wind, the yachtsman can shorten the distance through the water, but the boat will not move as fast. Between footing and pinching there is an optimum true wind angle at which the component of speed made good against the wind is maximized. By tacking at this optimum angle the yacht will reach a point upwind in minimum time. A similar situation exists in sailing downwind: for each wind speed there is a certain true wind angle, usually not 180 degrees, at which the speed made good downwind is maximized. The VPP program can help to identify these optimum headings.

The mathematical sailing model embodied in the VPP, to the extent that it gives correct answers, has obvious value in the design of sailing yachts. A computer run is trivial in cost (less than $15) and time compared with even a tank test ($25,000), let alone the construction and testing of a full-scale yacht ($500,000 to $1 million). The computer can explore large numbers of concepts and variations in a short time. Moreover, the precision of the mathematical solution allows small differences in performance to be reliably distinguished, so that systematic optimization is possible. The VPP provides quantitative answers to a host of questions to which a designer could previously have applied only his intuition and experience.

One major focus of the design process was the keel. *Australia II* had flown to victory on the wings of a novel keel. Naturally it was our ambition first to duplicate and then to exceed that achievement, examining as we worked any other unusual configurations that might offer an advantage. We blended both computational and experimental approaches to the task and ultimately obtained what we believe is a fairly complete understanding of at least the winglet keel designed for *Australia II* by the Australian naval architect Ben Lexcen. Although we dreamed up and modeled several other keel appendages, none proved to be as promising as a refinement of Lexcen's design.

The keel of a sailing yacht serves two complementary functions: as ballast and as a hydrofoil generating the hydrodynamic side force that balances the lateral forces from the sails. As ballast the keel stabilizes the hull against the heeling moment of the forces on the sails. In 12-meter yachts the ballast is lead, and it accounts for between 70 and 80 percent of the vessel's total weight. As ballast the keel's primary role is to establish the yacht's vertical center of gravity. Lowering this point makes the boat stabler.

To achieve that objective a designer can move away from the conventional trapezoid shape, which has longer chords at the top, to an untapered parallelogram or even a reverse taper. (A reverse taper was in fact another innovation in the radical keel of *Australia II*.) The vertical center of gravity of the yacht can also be lowered by putting thicker foil sections at the bottom of the keel. Forming the tip of the keel into a streamlined bulb may also be beneficial.

As a hydrofoil the keel has major effects on four of the five types of resistance encountered by a moving hull. First, if the keel is improperly streamlined or improperly aligned with the flow of water along the hull, it will create unnecessary form resistance (a pressure drag arising when the flow becomes "separated," meaning that it no longer moves smoothly). Second, because the keel has a significant amount of wetted surface and is short in the direction of flow, it contributes more than a proportional share of frictional resistance. Third, the volume of the keel moving through the water just a few feet below the surface contributes to the overall wave-making drag of the yacht. Fourth, the keel design has a major effect on the amount of induced resistance, which arises from the side force. (The fifth form of resistance comes from the surface waves encountered by the boat.)

Induced drag, which is the keel's major contribution to resistance, can be viewed as a penalty for producing lift. When a wing or any other shape produces a steady lift, it must be continuously transferring momentum into the fluid around it. This momentum is conserved in the wake of the wing, and far downstream there remains a disturbance in the fluid. The kinetic energy of the disturbance has to be paid for by work done on the wing; that work is the induced drag.

Induced drag increases with the square of the lift (if the lift doubles, the drag increases fourfold) and decreases as the span of the lifting system is widened (if the span is quadrupled, drag is halved). The designer could cut down the induced drag by widening the span (increasing the depth of the keel), but the constraints of the 12-Meter Rule allow little scope for such a move. Hence we had to explore nonplanar systems. It has long been known in aerodynamics that multiple lifting surfaces can, in favorable arrangements, produce lift with less induced drag than a single plane of the same span can. A biplane is an example; another, much like the winglet keel, is a monoplane wing with end plates (small separate lifting shapes at each end of the wing). It was in the direction of multiple surfaces that we achieved the final design of the winglet keel for *Stars & Stripes* (a design that still cannot be made public in detail).

Boppe and his associates took the lead in applying potential-flow panel methods, which are employed in the aerospace industry to predict the aerodynamic properties of complex configurations, to the problem of keel design. They calculated the lift and induced drag for various keel and winglet combinations, identifying promising shapes that were then tested with models and VPP analysis.

In dealing with wave resistance a group from SAIC, headed by Carl A. Scragg, that specializes in the subject took primary responsibility. They started by applying a program called the slender-ship theory, devised by Francis Noblesse of the David Taylor Naval Ship Research and Development Center near Washington, to three 12-meter hulls that had been through tank tests. The initial results were discouraging: the predicted resistances were much higher than the figures obtained in the tests and began their characteristic steep rise at a much lower boat speed.

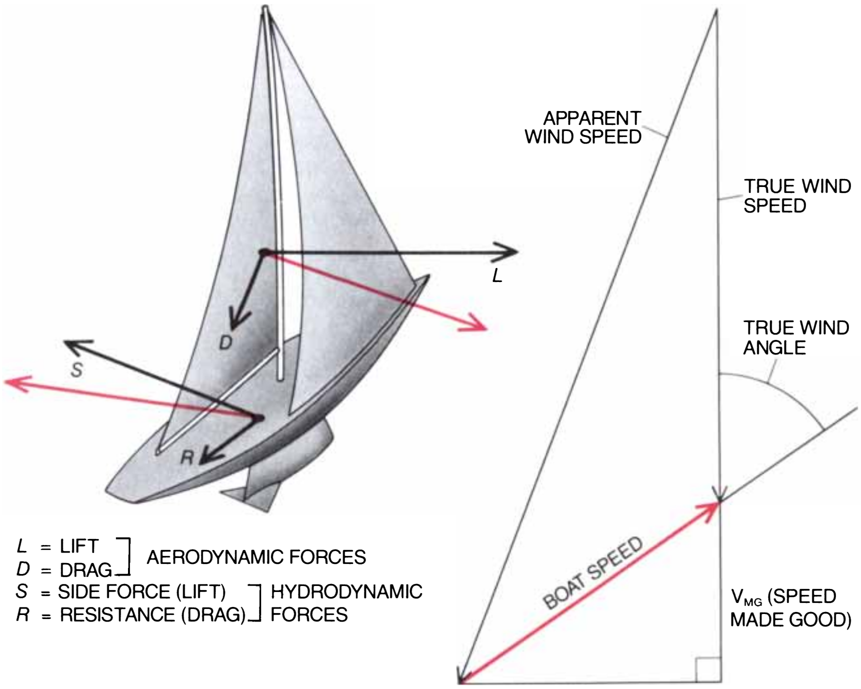
Scragg observed, however, that in spite of the large errors the relative difference in resistance among the three models was quite accurately predicted. In further tests the pattern persisted. Scragg concluded that if the computer program predicted that a model would show improvement over a particular speed range, the improvement would in fact appear in tests of the model.

Thus the slender-ship theory, even though it was inaccurate in absolute terms, became a foundation of the hull-design process. By means of it the design team developed new hull forms that showed reductions in resistance at the most important sailing speeds, contributing materially to the success of *Stars & Stripes.*

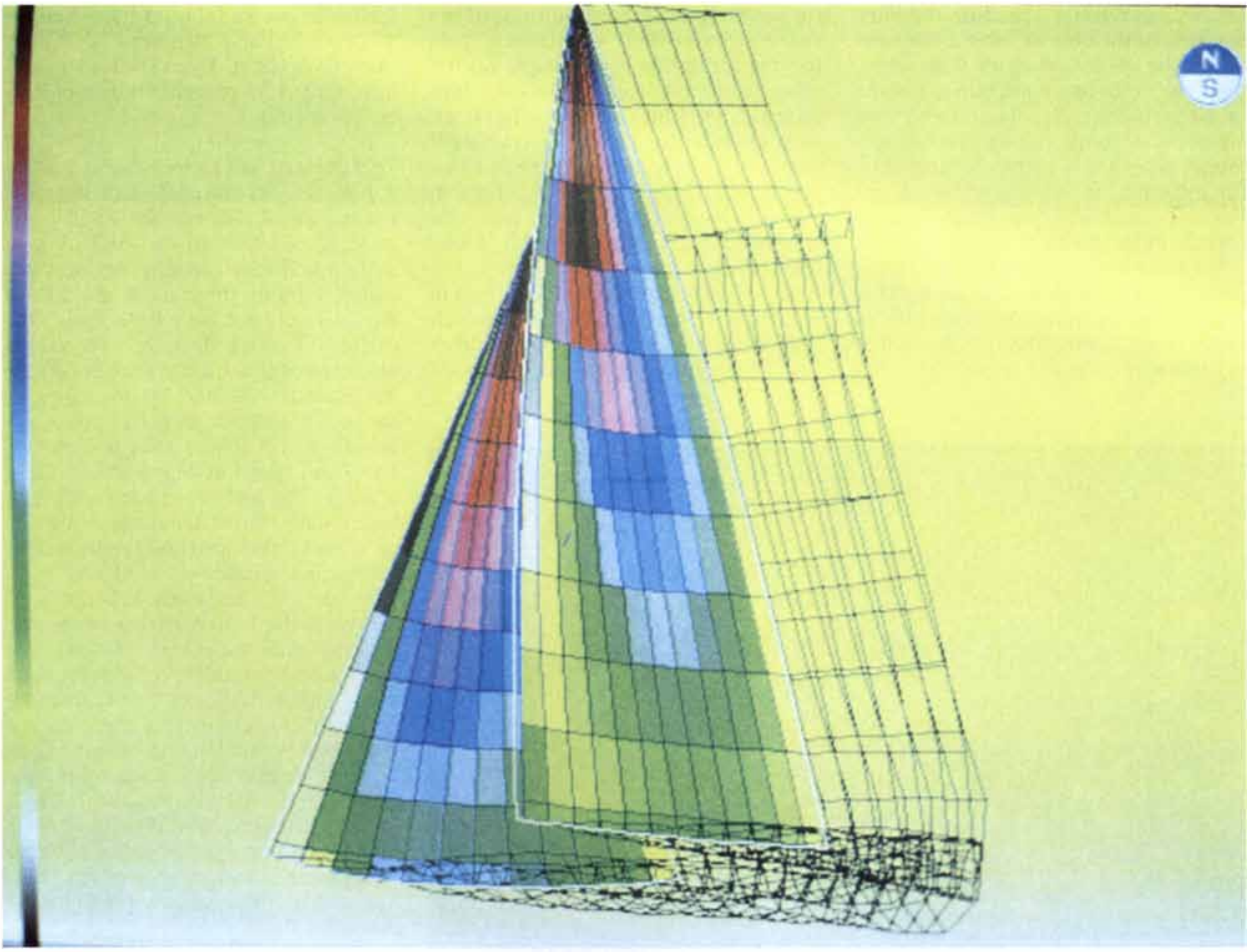
The yacht also is believed to have benefited from the attention the design team gave to viscous resistance, or skin friction, which arises from the viscosity of the water and typically accounts for 37 percent of the resistance encountered by a moving hull. Many coatings that are claimed to have the potential for reducing viscous resistance are available. Thomas G. Lang of the Semi-Submerged Ship Corporation tested more than 30 of them but found nothing that was superior to a clean, polished, painted surface. Then the 3M company made available an experimental plastic-film coating, designed for aircraft, that incorporates microscopic parallel grooves called riblets. On the basis of reports published by the National Aeronautics and Space Administration on riblet technology, optimum groove depths and spacings for the flow conditions on *Stars & Stripes* were selected and riblet sheets made specifically for the yacht were applied. In spite of problems in aligning the riblets with the varying directions of flow and in dealing with cracks between adjacent sheets, the design team believes the riblets produced a net reduction of from 2 to 4 percent in skin friction.

Another topic we had to examine was meteorology. Because of the important tradeoffs between light- and heavy-wind performance, an accurate prediction of weather conditions during the months of October through February was vital to the success of the campaign. The winning design would have to be accurately optimized for the dominant conditions. In addition the selection of sails and even the tactics for the day's race would be dictated by the daily weather forecasts.

R. Leland Davis and Chris Bedford of Galson Technical Services collected and analyzed historical data on wind and sea conditions. During the design period we had only records of past weather to go by. They were critical in many design decisions, particularly in choosing the length of the yacht. During the races forecasting had to be done on several time scales: weeks to decide on a change of keel between



BALANCE OF FORCES in two fluids, water and air, determines the performance of a sailboat. The hydrodynamic forces are resolved into resistance (*R*), opposite to the direction of travel, and side force (*S*), perpendicular to the direction of travel. The aerodynamic forces are resolved into drag (*D*), in the direction of the apparent wind, and lift (*L*), perpendicular to the apparent wind. The apparent wind is the vector sum of the true wind over the water and the relative wind due to the motion of the boat through the water. The balance of these forces was mathematically modeled and optimized in Sail America's computerized Velocity Prediction Program, which simulates sailing-vessel dynamics.



DESIGN OF SAILS for *Stars & Stripes* was simplified and speeded by computer. In the past sail designs were drawn by hand and the final version, at full scale, had to be laid out in a large, open loft. The computerized design process is called computer lofting.

one series of races and the next; a day to decide whether to call a lay day (a one-day postponement of a race); a couple of hours to choose sails for the day from the yacht's inventory of about 150 sails, and minutes to anticipate the next wind shift during a race. Our forecasting employed both statistical and numerical computer models, which received input from Doppler weather radars on the coast.

Another activity of importance was full-scale testing. Under the direction of Robert Hopkins, Jr., of Sail America an ambitious program of collecting and analyzing data was carried out in Hawaii for 10 months starting in October, 1985, and was continued in Australia for five months during the period of the races. Instrumentation and computer programs for this work were provided by Richard C. McCurdy of Ockam Instruments, Inc. Transducers in the boats measured boat speed; apparent wind; wind, heel and rudder angles; ship motions, and the relative range and bearing of a trial-horse yacht. This information was telemetered to the boat serving as tender; there a computer (MicroVax II of the Digital Equipment Corporation) logged the data, which were analyzed by another MicroVax II computer on shore. The information gleaned from this testing effort was critically important in validating the VPP and evaluating the performance of real boats in real conditions.

Our venture into technology had to go beyond the factors directly affecting a yacht. Yacht racing has an essentially random component in that the relative performance of two yachts depends on the wind speed and the sea conditions, which vary randomly from day to day. VPP results by themselves are therefore inconclusive and possibly misleading for determining the order of merit of two candidate yachts over a series of races. We also came to realize—it was a surprise at the time but is fairly obvious in retrospect—that one cannot determine which yacht design would be best for a particular race series without knowing the characteristics of the competing yachts. Hence we had to learn all we could about the sailing characteristics of competitors' boats and to develop probability- and game-theory methods to evaluate how our candidate designs could be expected to fare in races with a projected fleet of challengers and defenders.

To this end we developed two race-model programs to simulate single races between two yachts. The first program, a simple probabilistic model, takes into account several factors: the course time difference for the pair of yachts as a function of wind speed; the probability distribution of wind speed for the month; a representation of such randomizing factors as unsettled wind and sea conditions and small errors or accidents, and an optional "edge" for one yacht, such as 30 seconds for generally superior sails, tactics or training.

The second program is much more elaborate. In addition to the inputs to the first program it takes into account the wind history of the course for a typical day recorded in the data; differences in relative speed on the various legs of the course; interactions between the yachts when they are close together, and the uncertain outcome of races that are calculated to be very close. The output from each race-model program is a single probability: for example, in November winds off Perth, boat *A* will beat boat *B* in 59 percent of the races.

We turned to game theory because it provides rational strategies for making decisions in a conflict. We were able to apply two-person, zero-sum game theory to the problem of selecting the optimum design for the conditions of the races. (In zero-sum situations a gain for one contestant means a loss for the other.) Suppose Red picks a 46-foot boat and we (Blue) pick a 48-foot boat; with no edge for either boat we shall win only 28 percent of the time in November but 72 percent of the time in January. What length should each side choose to maximize its payoff if one team has an edge in overall skills?

The possibilities can be organized in matrixes. The payoff matrix in which Blue is given a 30-second edge shows that Blue's best choice remains at 48 feet, whereas Red should choose either 47 or 45 feet and then hope for conditions to favor it. Even so, Red would have only a 37 percent chance of success in each race.

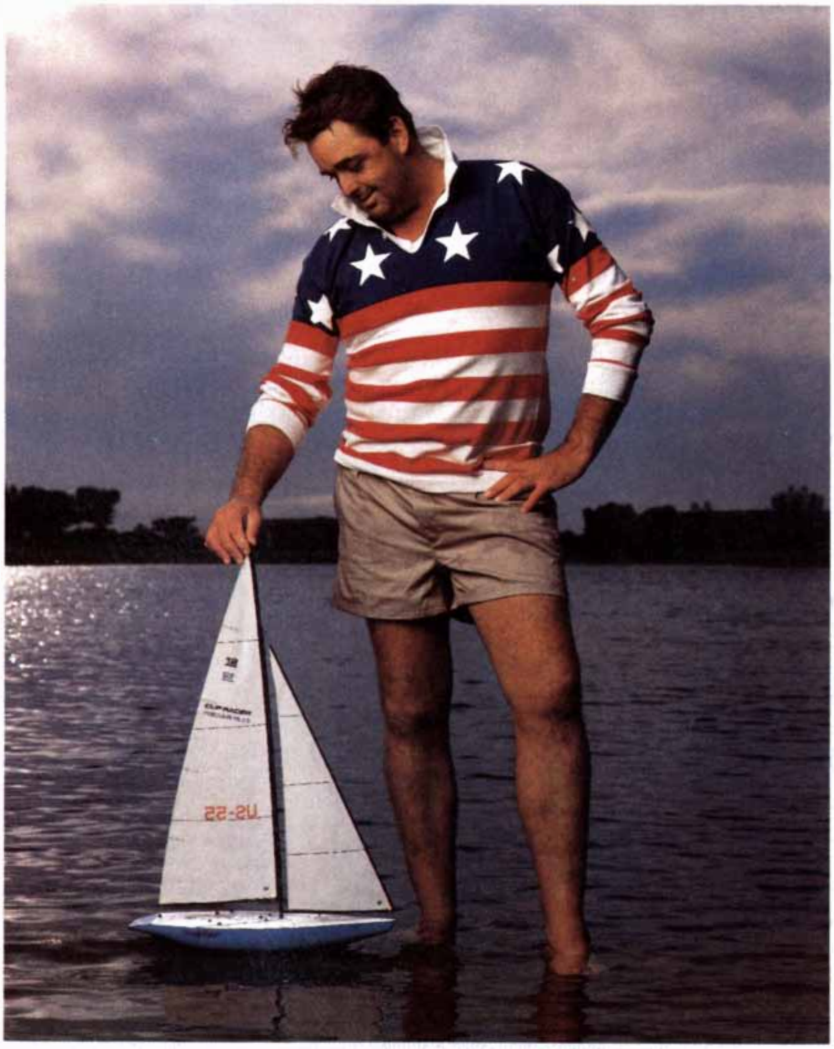
All the technology we have described flowed together into a strategy for the design of *Stars & Stripes*. It owes its very existence to the velocity-prediction and race-model programs, because without their clear dictates it would surely not have been built. Our earlier 12-meter designs, two yachts also named *Stars & Stripes* but further designated as '85 and '86 according to the year of construction, had proved to be of unprecedented size, power, stability and speed in heavy winds. *Stars & Stripes* '85 had proved to be slightly faster under most conditions, and its crew had developed tremendous confidence in the boat.

In late 1985 and early 1986, however, when the Sail America team was training and testing in Hawaii, the Australian defender candidates and many of the challengers were training in Perth and competing there in the 12-Meter World Championship races. Careful observation, including photogrammetric analysis, showed that all these boats were substantially smaller than we had decided would be optimal for summer conditions. Results from the race-model program showed that although *Stars & Stripes* '85 had high probabilities of beating any known competitor in a four-of-seven series in January or February, it had only a marginal chance of surviving the round-robin eliminations among a fleet of 13 challengers that were mostly two or three feet shorter.

Faced with these predictions, Sail America had no choice but to build another boat small enough to compete effectively in the round robins. A new design could also take advantage of the new hull shapes (developed through slender-ship theory and confirmed in tank tests) and of progress in computer-optimized keel design. The length was chosen to be a little more than that of the rest of the fleet, as suggested by game theory. After the elimination rounds the span of the winglets could be increased, the boat reballasted for greater weight (a move that would produce a longer waterline) and the keel made more bulbous to lower the center of gravity. Thus equipped for the stronger winds of midsummer, the boat would be hard to beat in a four-of-seven series.

The boat was built (in record time) and tested long enough in Hawaii to demonstrate its superiority over the earlier designs. The success of the strategy is now yachting history.

AMERICA'S CUP FINALISTS in the 1987 competition were *Stars & Stripes*, in the lead here, and *Kookaburra III* of Australia. *Stars & Stripes* reached the finals by eliminating 12 challengers from the U.S. and five other countries; *Kookaburra III* defeated five Australian defenders. In the final four-of-seven series *Stars & Stripes* won in four straight races.


US-55

AMERICAN EXPRESS
0978
3712 3 8 95006
83 AX
DENNIS CONNER

Collisions between Spinning Protons

The outcome of a collision between two protons shows a surprising dependence on their directions of spin. The results challenge the prevailing theory that describes the proton's structure and forces

by Alan D. Krisch

All the building blocks of matter—protons, neutrons and electrons—seem to be spinning like tops. The spinning is a basic quantum-mechanical property; each particle has a definite amount of spin, or spin angular momentum, just as it has a definite mass and a definite electric charge. When two spinning particles collide, the direction of their spin can affect how they scatter, just as the "english" on billiard balls can alter their rebound after a collision.

A skilled player with an intuitive knowledge of a billiard ball's properties can make difficult but predictable shots by adjusting the spin and speed of the ball. In a series of accelerator experiments my colleagues and I likewise varied the spin and energy of colliding protons. But unlike the pool player, we could not predict the effects of our "english," because many properties of the proton are still mysterious. Indeed, we observed unexpected and often startling behavior that challenges the current theory of the proton's structure and forces, quantum chromodynamics (QCD).

The investigation began in 1973 at the Zero Gradient Synchrotron (ZGS) at the Argonne National Laboratory. There my research group from the University of Michigan scattered beams of polarized protons from targets in which the protons were also polarized, or all spinning in the same direction [see "The Spin of the Proton," by Alan D. Krisch; SCIENTIFIC AMERICAN, May, 1979]. The results demonstrated that spin plays a significant role in high-energy interactions between protons: violent proton-proton collisions occurred mostly when the beam and the target were polarized in the same direction. When the beam and the target protons were spinning in opposite directions, the protons often seemed to pass through each other without interacting.

At the ZGS we were limited to experimental energies of about 13 GeV (billion electron volts). After a laborious, six-year modification of another accelerator, the Alternating Gradient Synchrotron (AGS) at the Brookhaven National Laboratory, we have now been able to investigate the effects of proton spin at much higher energies: up to 18.5 GeV, when both the beam and the target were polarized, and at 28 GeV, with the target alone polarized. The surprises have multiplied. As the collision energy of the protons increases, the effects of spin seem to oscillate. At our highest energies the spin effects are particularly pronounced.

On one level our surprise is easy to appreciate. The energy associated with the spin of a proton is constant, and so the role played by spin should diminish as the energy of the collision increases. At sufficiently high energies it should make little difference whether two colliding protons are spinning in the same direction or in opposite directions. The fact that the spin directions do make a big difference suggests that our understanding of how protons interact with one another is incomplete. The experiments even call into question the currently accepted model of the proton's internal structure, which holds that a proton consists of three smaller constituents known as quarks, held together by the strong nuclear force (the force described by QCD).

Our new results are certainly not the first time that the phenomenon of spin has surprised and confused physicists. When Samuel A. Goudsmit and George E. Uhlenbeck first proposed in 1925 that every electron has a property called spin, Wolfgang Pauli told them the concept was foolish. During the late 1920's and the 1930's, however, spin became an important element in the development of quantum mechanics and atomic physics, particularly in the work of P. A. M. Dirac.

During the 1940's nuclear physicists believed that spin effects were felt only in low-energy atomic collisions and that they could not possibly be significant in nuclear collisions at several million electron volts of energy. Then Charles L. Oxley of the University of Rochester and his colleagues found evidence of large spin effects in collisions at several hundred million electron volts. The resulting surprise and excitement is documented in a 1954 paper by Enrico Fermi.

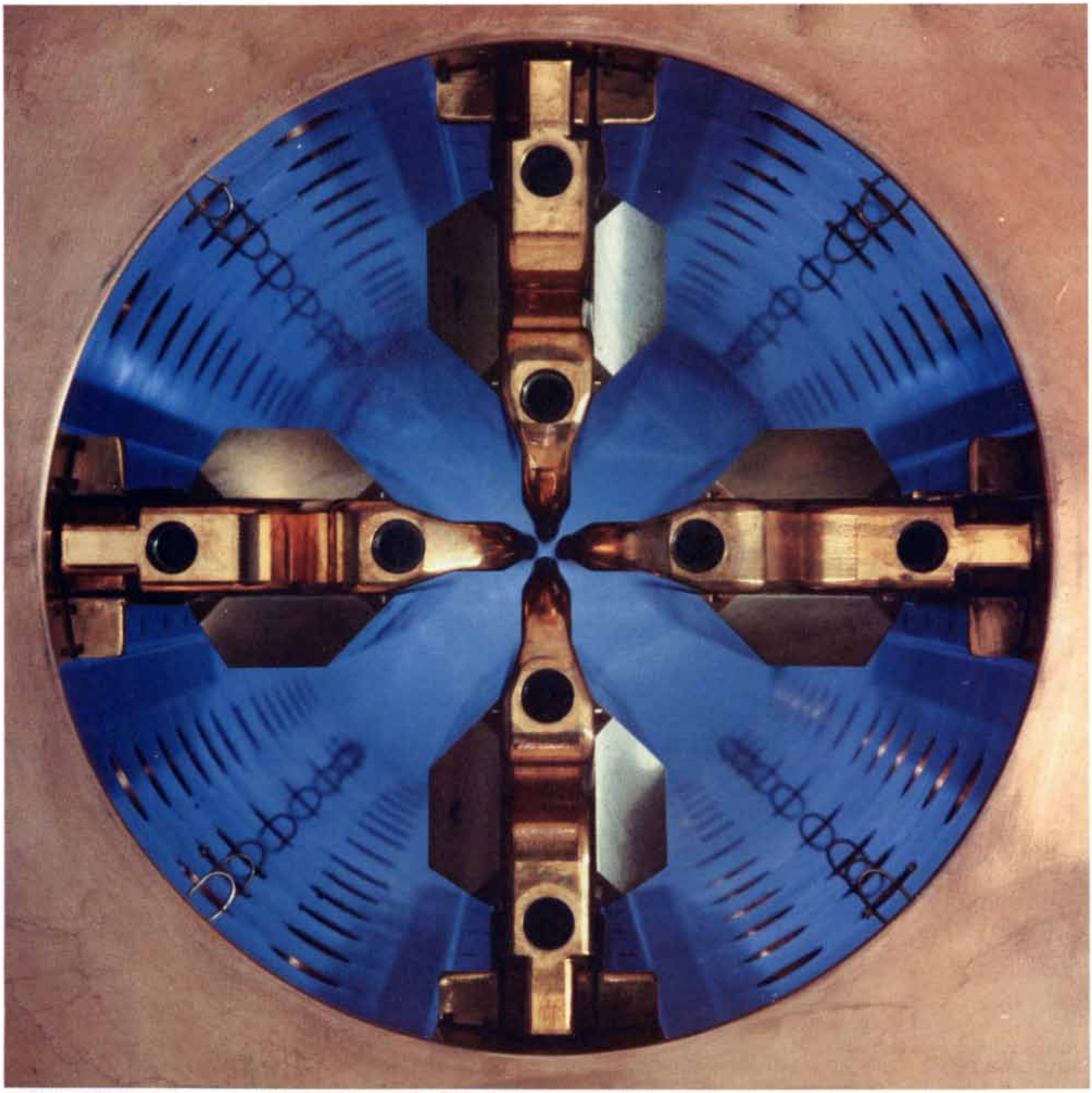
Most high-energy physicists were quite sure that spin would be unimportant in elementary-particle collisions at billions of electron volts of energy. For years this belief was tested only in a series of difficult experiments done by Owen Chamberlain and Emilio Segré of the University of California at Berkeley, among others. Then in the late 1950's Anatole Abragam of the Collège de France and Carson D. Jeffries of Berkeley suggested building polarized proton targets. The technique, which has been quite successful, relies on a low temperature and a strong magnetic field to polarize the spins of certain electrons in frozen beads of target material; the magnetic field causes the spins of the electrons to "line up." Microwave radiation is then applied to transfer the spin align-

RADIO-FREQUENCY QUADRUPOLE accelerates protons in the Alternating Gradient Synchrotron (AGS) at the Brookhaven National Laboratory. The RFQ consists of four electric poles, two positive and two negative, positioned so that identical poles are opposite each other. In this five-foot-long, one-foot-diameter RFQ the polarities alternate at the rate of 200 million times each second. Protons entering the RFQ have already been polarized: made to spin in the same direction. The RFQ preserves the polarization as it boosts the protons from an initial energy of 20,000 electron volts (20 keV) to 760 keV.

ment of the electrons to nearby protons, making them spin in one direction. Experiments employing polarized proton targets in the 1960's and early 1970's at Berkeley, CERN (the European laboratory for particle physics) and Argonne revealed small but interesting spin effects in high-energy collisions. Nevertheless, most high-energy physicists still believed spin was not very important and would become even less so at higher energies.

In 1973 my research group inaugurated a different approach at the Zero Gradient Synchrotron: we polarized the beam as well as the target. We maintained the polarization of the target with a special magnet having a high field (25,000 gauss) and with a mixture of liquid helium 3 and helium 4, which held the temperature of the target at half a degree above absolute zero. At the same time we used a complex system of magnets to keep the proton beam polarized while it was accelerated. By scattering the polarized beam protons from polarized target protons, we made the first measurements of elastic collisions between protons spinning in known directions. A collision is elastic if all the energy of the incident proton is carried off by the two rebounding protons; in an inelastic collision some of the energy goes into creating new particles. We studied only elastic collisions because they seem both simple and fundamental.

We studied the scattering of protons in four initial spin states. In the first initial state the spins of the beam protons and the target protons were both up (that is, if the fingers of the right hand curled in the direction of rotation, the thumb would point up). I shall call this state up-up, where the first "up" refers to the spin direction of the beam protons and the second refers to the spin direction of the target protons. In keeping with this convention, the remaining initial states are known as down-down, up-down and down-up. The spin direction of the protons in the beam was reversed about every three seconds; the spin direction of the protons in the target was reversed every few hours.



To probe as deeply as possible into the interior of the proton we made the collisions as violent as we could. We hoped to find some new clues about the nature of the tiny constituents that seem to live inside the proton but apparently cannot emerge; these are the constituents theorists call quarks. The quark theory developed by Murray Gell-Mann of the California Institute of Technology has been truly successful in accounting for the masses of the many short-lived particles that are created when protons collide. On the other hand, the quark theory of particle scattering—quantum chromodynamics (QCD)—has made few predictions that could be verified. QCD is quite a flexible theory and has been easily able to adjust to most new scattering data after the fact; because I am a rather formal scientist, I am impressed less by adaptability than by predictive power.

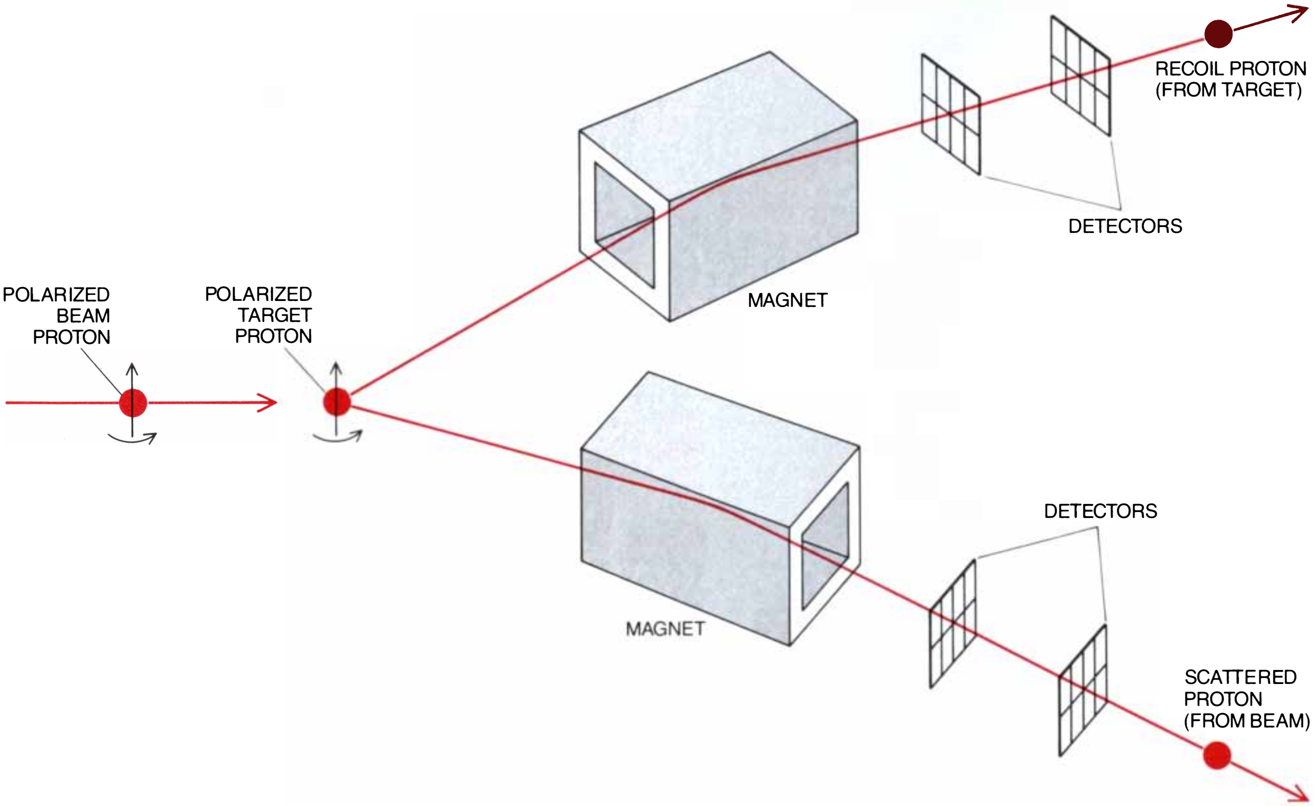
I also confess to some confusion about the notion that quarks can live as particles inside a proton but not outside. The clever and catchy QCD ideas that have been proposed to explain the apparent confinement of quarks may turn out to be correct, and perhaps I shall eventually change my old-fashioned view that particles must be well-defined objects. I believe, however, that a simple concept should not be abandoned in favor of a more complex one until the hard experimental evidence is overwhelming.

The best collisions for probing the spin forces of the proton's constituent particles as deeply as possible are those in which two colliding protons rebound exactly at right angles to the initial direction of motion. Such collisions transfer the maximum fraction of the energy of the incoming proton and hence are the most violent. In our final experiment at the ZGS we studied such perpendicular collisions while varying the energy of the incoming protons between 4 and 13 GeV; at each energy we observed what happened when the protons were spinning in the same direction and when they were spinning in opposite directions.

In particular we measured the probability of an exactly perpendicular elastic collision separately for protons whose spins were parallel (up-up or down-down) and protons whose spins were antiparallel (up-down or down-up). This probability of collision is called the cross section for proton-proton elastic scattering at 90 degrees. The cross section, which can be thought of as the effective size of the interacting protons, is typically plotted as a function of the energy of the incoming particle.

Our experiment revealed two important clues about the constituents of the proton. The first clue comes from the fact that the cross section falls rapidly as the collision energy increases to 8 GeV, but above 8 GeV it falls only gradually [*see top illustration on opposite page*]. Information about the structure of the proton can be extracted from the data by means of a mathematical procedure called Fourier analysis. The initial rapid decline of the cross section means that at low energies the protons only glance off each other; it appears as if the outer layer of the proton is relatively soft and large—about one fermi, or 10^{-15} meter, in radius. The subsequent gradual decline of the cross section suggests that higher-energy collisions involve hard objects inside the proton. These objects appear to be about one-third of a fermi in radius. This clue is consistent with the three-quark model of QCD.

The second clue about the constituents of the proton, on the other hand, is troublesome for QCD. The clue comes from the observation that at energies greater than 8 GeV the cross section falls more rapidly when the protons' spins are antiparallel than when they are parallel. In other words the protons somehow have a bette chance of colliding violently when their spins are parallel. At 13 GeV th cross section (probability of collision



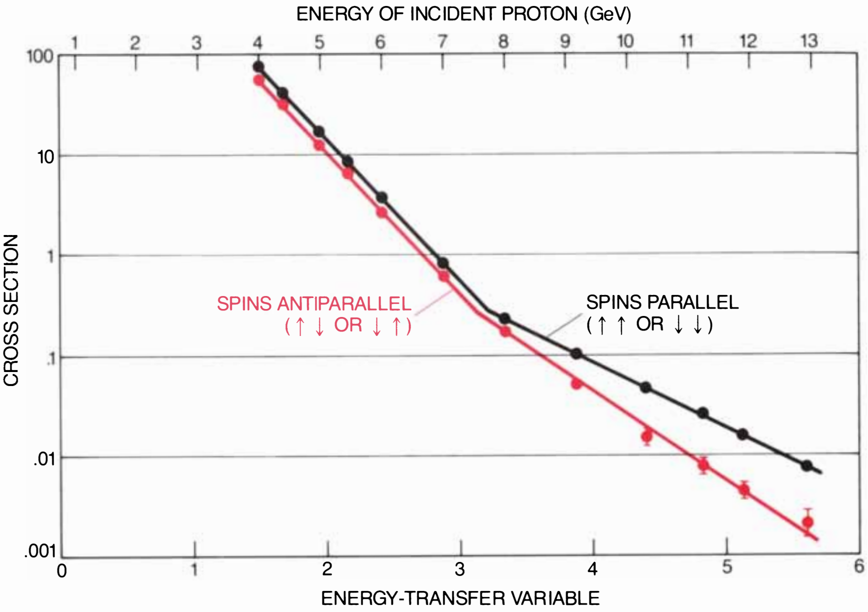
POLARIZED BEAM of protons is scattered from a fixed target of polarized protons. Particle detectors record the angles at which the individual protons scatter from one another; the angle of deflection in the bending magnets indicates the protons' energy.

is four times greater when the spins are parallel than when they are antiparallel. Although we are not sure exactly what is causing this strange and totally unexpected behavior, it does not appear to be good news for QCD.

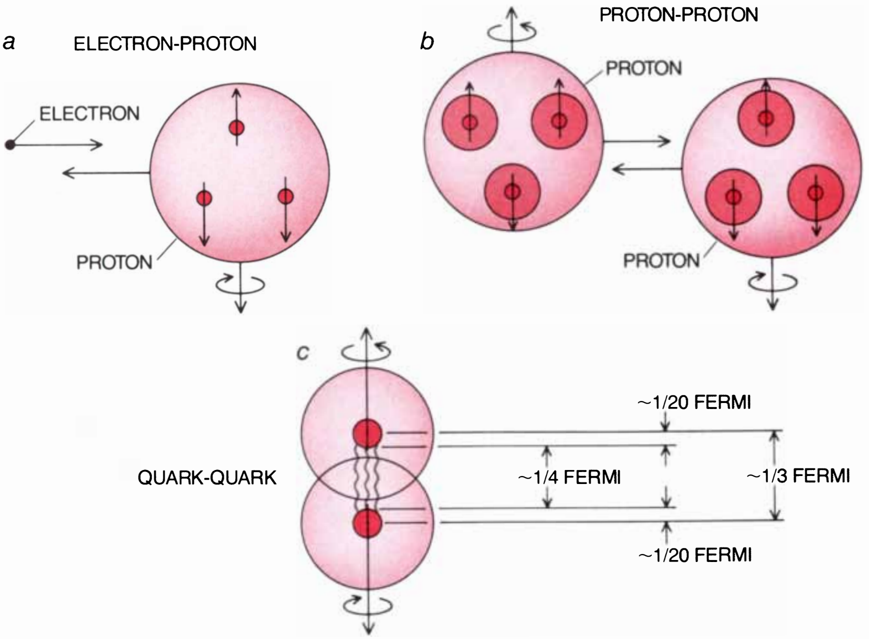
According to QCD, two of the three quarks in each spinning proton should spin in the same direction as the proton, and the third quark should spin in the opposite direction. Therefore no matter whether the two colliding protons' spins are parallel or antiparallel, the collisions will always involve some pairs of quarks whose spins are parallel and some pairs whose spins are antiparallel. QCD calculations allow for a twofold difference between the spins-parallel and the spins-antiparallel cross sections, but a fourfold difference cannot be easily reconciled with the theory.

What does the ratio of four mean? Does QCD apply only to collisions more energetic and violent than the ones we studied? Unfortunately theorists cannot agree on just where QCD's range of validity should start. Since the ratio seems to grow as the collisions become more violent and energetic, our data certainly do not support QCD. One explanation is that the constituents "seen" in scattering experiments—although they are somehow related to the quarks that account for the masses of particles—are not exactly the three spinning quarks of QCD. Perhaps every proton contains some other number of constituents. Alternatively, perhaps there are three constituents but they stop being independent quarks when two protons collide.

In our 90-degree scattering experiments we reached the 13-GeV energy limit of the ZGS. Extending the spin experiments to even more violent collisions required a higher-energy beam of polarized protons. In the late 1970's a group of about 80 scientists, engineers, technicians and students from Argonne, Brookhaven, the University of Michigan, Rice University and Yale University began to modify the Brookhaven AGS synchrotron so that spinning protons could be accelerated to an energy of about 20 GeV. The effort was challenging because as the spinning protons repeatedly circle the AGS during their half-second acceleration cycle they travel about 100,000 miles through strong magnetic fields that can easily ruin their polarization, setting all the protons spinning in different directions. The magnetic fields in the AGS are much stronger than those in the ZGS and therefore pose a bigger problem. In fact, a U.S. Atomic Energy Commission panel on which I served concluded in 1974 that



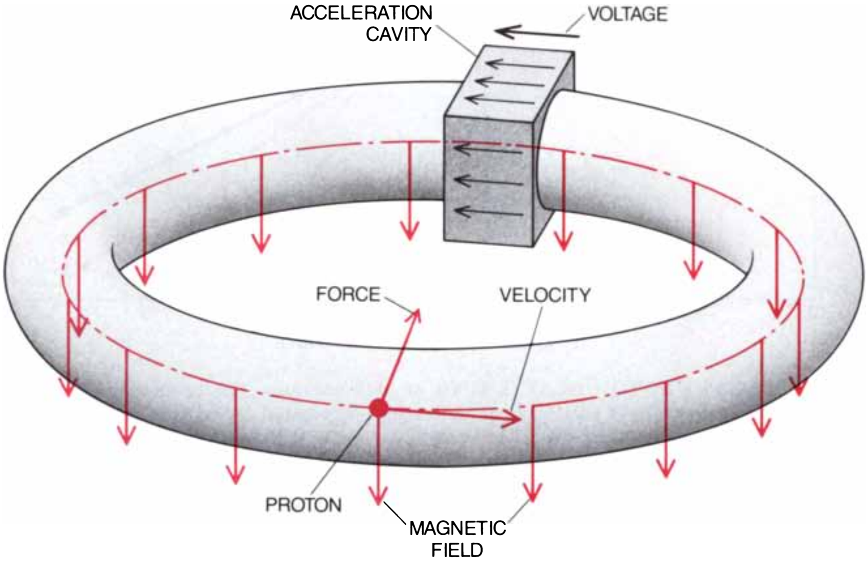
SPIN AFFECTS PROTON SCATTERING at high energies. The cross section at 90 degrees—the probability of scattering at that angle—is plotted against the energy-transfer variable for the collisions. (At a fixed scattering angle the energy-transfer variable is proportional to the energy of the incoming proton.) Two sets of data have been plotted, one in which the spins of the incoming proton and the target proton were parallel and one in which the spins were antiparallel. For low values of energy the parallel and antiparallel cross sections are identical, as predicted by the prevailing theory of the proton's structure and properties, quantum chromodynamics (QCD). At higher energies, however, the cross sections diverge noticeably. The reason for the divergence is not clear at this time.



THEORETICAL MODEL OF THE PROTON in QCD holds that the particle is made up of three constituents called quarks, bound together by the strong force. By scattering pointlike electrons from stationary protons at the Stanford Linear Accelerator Center (*a*), a group of investigators from the Massachusetts Institute of Technology and SLAC obtained data showing that the proton's constituents have a radius of less than 1/20 fermi. (One fermi is equal to 10^{-15} meter—approximately the radius of a proton.) Data from the author's experiment, in which protons were scattered from protons (*b*), indicate that the constituents have a radius of about 1/3 fermi. Together the two results suggest an interesting and possibly correct model of the proton's constituents (*c*). The SLAC data indicate the size of the constituents' electric charge. Perhaps each constituent has a core with a radius of about 1/20 fermi and a strong force field extending about 1/4 fermi farther. When two quarks strike each other, the effective size is just about 1/3 fermi.

"technical reasons may preclude accelerating polarized protons at [AGS-type] accelerators."

As it happened, we did succeed in overcoming the problem of depolarization, but it presented a daunting obstacle. To reach our maximum energy of 22 GeV we had to confront depolarization problems known as depolarizing resonances, which occurred at 45 different energies in the acceleration cycle. Navigating through the sea of depolarizing resonances was a formidable effort that required sophisticated instruments called polarimeters to ascertain the percentage of protons spinning in each direction. This percentage enabled us to determine whether we were on the right path as we adjusted certain specially designed magnets. The "resonance navigating crew," which was led by Lazarus G. Ratner of Brookhaven and me, had to work around the clock for six weeks to overcome all the 45 depolarizing resonances. In January, 1986, after devoting six years of intensive effort to the overall modification, we accelerated a polarized beam to 18.5 GeV, a new world record. Experiments using this unique facility began immediately; we reached 22 GeV in a few weeks.

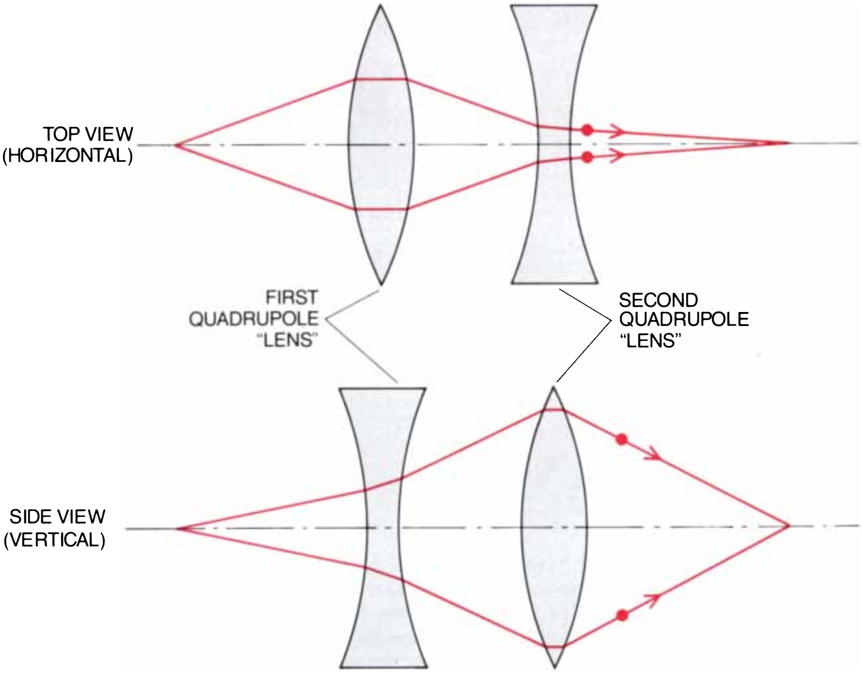


SYNCHROTRON exploits both electric and magnetic fields to accelerate protons (which carry an electric charge) to energies of many billions of electron volts. The protons are forced to move in a circular beam around a doughnut-shaped vacuum chamber by a magnetic field that is perpendicular to the plane of the ring. Each time a proton circles the doughnut it is given a voltage boost of about 100 keV in the acceleration cavity.

To explain why depolarization develops in a synchrotron and how it was overcome I must first describe how a synchrotron works. A synchrotron, like its smaller cousin the cyclotron, relies on both electric and magnetic fields to accelerate protons (which carry an electric charge) to high energies. The protons are forced to move in a circle around a doughnut-shaped ring by a magnetic field that is perpendicular to the plane of the ring. This vertical bending field is a powerful source of depolarization: if a proton's spin axis is horizontal, the field can rotate the axis many times during each trip around the ring, making it nearly impossible to keep all the spin axes lined up in a proton beam.

Every time a proton circles the doughnut, it is given a voltage boost of about 100,000 volts in a so-called acceleration cavity. During the AGS acceleration cycle each proton circles the half-mile circumference of the doughnut about 200,000 times, gaining a total of about 20 GeV by the end of the cycle. Each turn around the AGS ring takes about .0000025 second (2.5 microseconds), and so the entire cycle takes about half a second.

It is difficult to keep the protons inside the doughnut for the entire acceleration cycle, since each particle moves about 100,000 miles within a pipe that is only about three inches across. Even the finest sharpshooter would find it hard to hit a three-inch spot from 100,000 miles away. Physicists, who are often nearsighted, employ special magnetic fields to focus the protons and keep them from escaping. The magnetic fields focus charged particles just as the lens in a magnifying glass focuses rays of light.



STRONG-FOCUSING PRINCIPLE confines protons moving within a synchrotron to a doughnut that has a small bore, thereby reducing the size of the magnets needed to bend the proton beam. The principle calls for two opposite quadrupoles placed in a row. The first quadrupole focuses the beam of protons horizontally but makes it diverge vertically; the second quadrupole focuses the beam vertically but makes it diverge horizontally. By adjusting the strengths of the quadrupoles one can focus the beam in both directions.

Early synchrotrons such as the ZGS used rather weak magnetic focusing fields and were therefore called weak-focusing synchrotrons. Since the focusing was weak, the proton beam diverged and became quite large, and thus the doughnut-shaped vacuum chamber containing the beam needed a large bore. The ZGS vacuum chamber was about 30 inches wide by six inches high. The large size had some advantages; I remember sticking my hand several feet into the doughnut in 1965 to help fix a broken target without disassembling the 600-ton magnet surrounding it. Unfortunately the large vacuum chamber had to be surrounded by even larger magnets. The cost of such large magnets is prohibitive for higher-energy accelerators. The ZGS magnets were about 8½ feet wide by 4½ feet high and weighed 11 tons per linear foot. Built starting in 1955, the ZGS was the last weak-focusing synchrotron constructed.

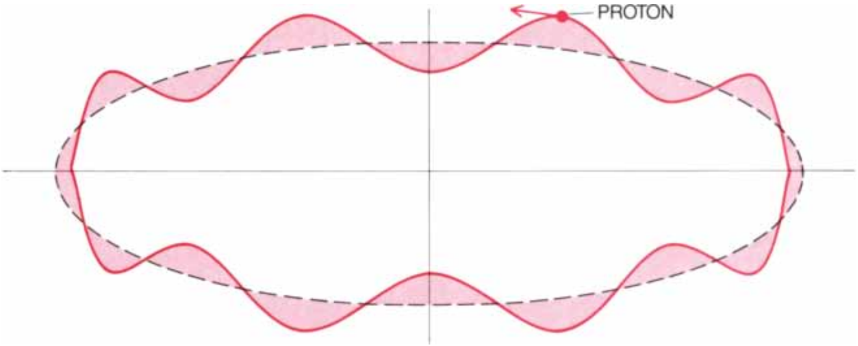
A better idea was needed for higher energies. In the 1950's Nicholas C. Christofilos, Ernest D. Courant, Stanley Livingston and Hartland S. Snyder invented the strong-focusing principle. Strong focusing uses magnets called quadrupoles, each of which consists of two magnetic north poles and two magnetic south poles positioned alternately at the corners of a square. As a beam of protons travels through the square, protons that drift away from the beam axis feel an ever stronger focusing force that pushes them back toward the axis. Unfortunately it is not possible for a single quadrupole to focus protons in both the horizontal and the vertical direction. If a quadrupole focuses a beam of protons in one direction, the particles diverge in the other direction and leak from the vacuum chamber.

The secret of strong focusing is placing two opposite quadrupoles in a row. The first quadrupole focuses the beam horizontally but causes it to diverge vertically, whereas the second quadrupole focuses the beam vertically but causes it to diverge horizontally. At first glance one might think that the focusing and diverging would cancel each other, accomplishing nothing. By properly adjusting the strengths of the two quadrupoles, however, one can focus a beam in both directions. The idea works so well that all accelerators built after 1960 use strong focusing.

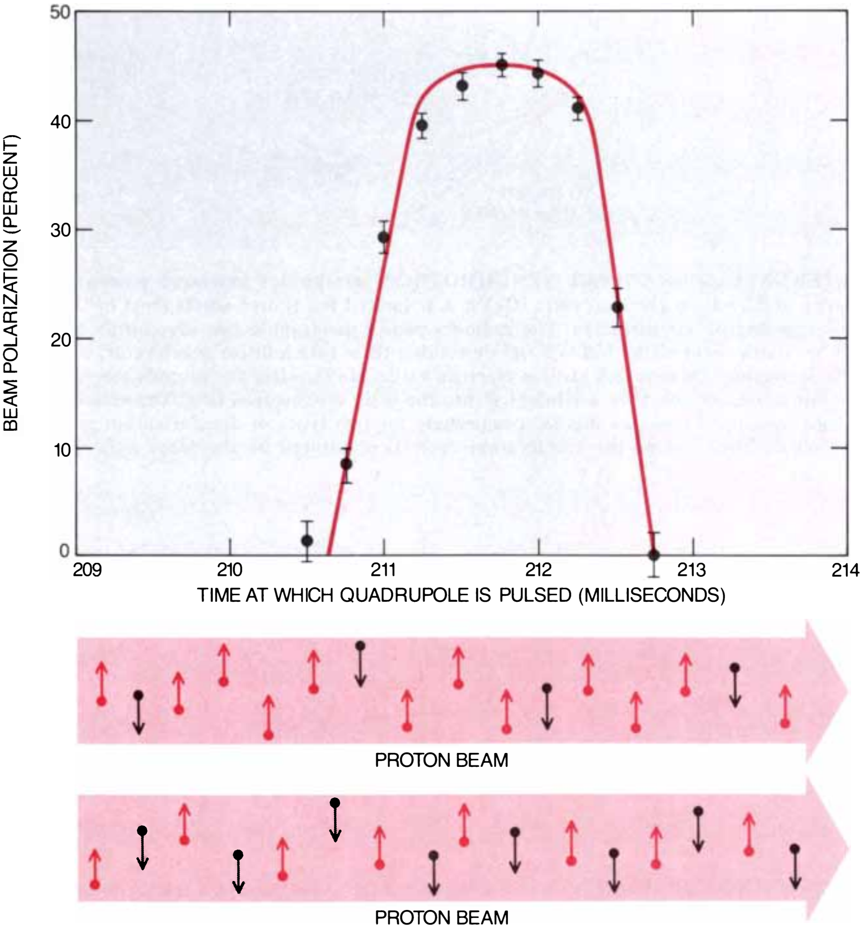
The pairs of quadrupole magnets force the protons to oscillate about the central axis of the synchrotron ring, in waves called betatron oscillations. The vertical betatron oscillations cause the protons to travel through horizontal magnetic fields, which depolarize protons that have a vertical spin axis. The fields are weak, but if the proton encounters enough of them, its spin can be rotated. The more crests and troughs there are in the betatron oscillations, the more often the proton will cross a horizontal magnetic field in each trip around the ring.

Fortunately it turns out that the depolarization problems associated with betatron oscillations and horizontal magnetic fields are much smaller than the depolarization effect of the vertical bending field on protons that have a horizontal spin axis. Keeping the spins aligned during acceleration is difficult but not impossible when the protons spin about a vertical axis.

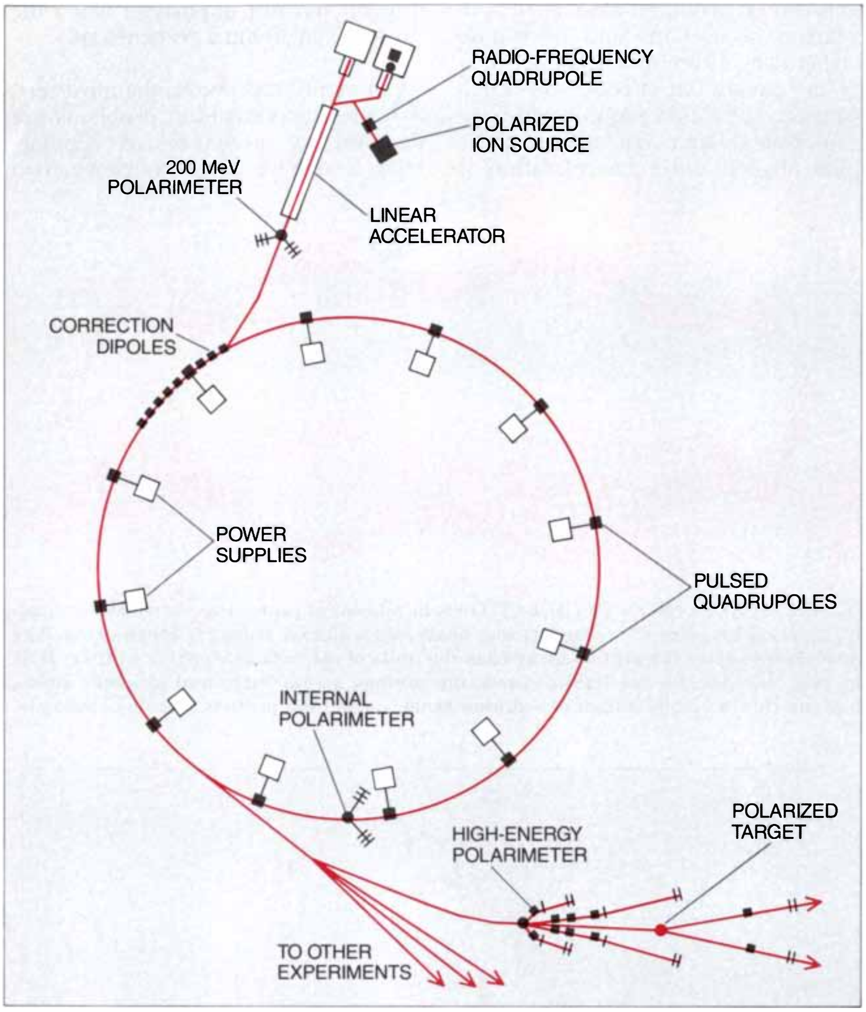
As I mentioned above, the most serious depolarization problems are depolarizing resonances. A depolarizing resonance occurs whenever two



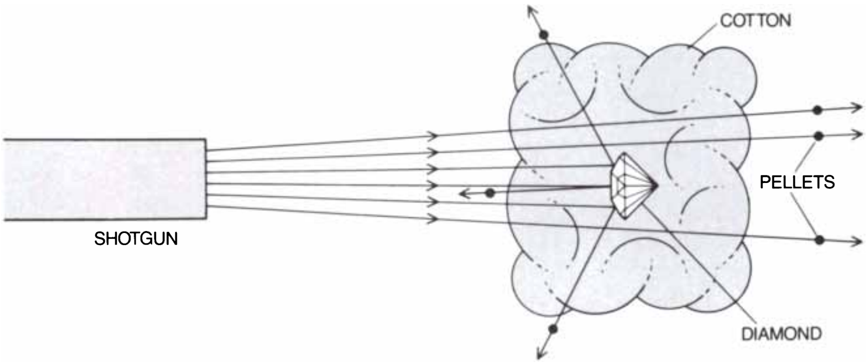
VERTICAL BETATRON OSCILLATIONS in a beam of protons are vertical wave motions caused by pairs of strong-focusing quadrupoles placed around a synchrotron. The wave motions steer the protons away from the walls of the vacuum chamber as they circle the ring. Because the oscillations carry the protons across horizontal magnetic fields, they can ruin the polarization of a proton beam, setting the protons spinning randomly.



PROTON POLARIZATION in the AGS is preserved by switching on special pulsed quadrupole magnets at precisely the right time during the acceleration cycle. The investigators obtained polarizations approaching 50 percent (*top*), which means 75 percent of the protons were spinning in one direction and 25 percent in the opposite direction (*middle*). Without proper timing of the pulsed quadrupoles the percent of polarization fell to zero, that is, equal numbers of protons were spinning in opposite directions (*bottom*).



ALTERNATING GRADIENT SYNCHROTRON accelerates polarized protons to an energy of 22 billion electron volts (GeV). A polarized ion source starts most of the protons spinning in one direction. The radio-frequency quadrupole (*see illustration on page 43*) boosts the protons to 760 keV before feeding them into a linear accelerator, which in turn accelerates them to 200 million electron volts (MeV). After the protons emerge from the linear accelerator, they are injected into the main synchrotron ring. The pulsed quadrupoles and the correction dipoles compensate for two types of depolarization problem. The polarization during the acceleration cycle is monitored by the three polarimeters.



SCATTERING EXPERIMENTS probe the structure of the proton much as an imaginative customs inspector might search for a diamond hidden in a large bale of cotton. The inspector fires a shotgun into the cotton and observes how many pellets scatter in each direction. The large, soft bale deflects the pellets only slightly. A small, hard diamond, on the other hand, causes some of the pellets to bounce back toward the inspector. By keeping track of how many pellets scatter in each direction, the inspector can (with the laws of geometry) determine the exact size and shape of the diamond. In probing the structure of the proton at the AGS the investigators employed protons themselves as the "pellets."

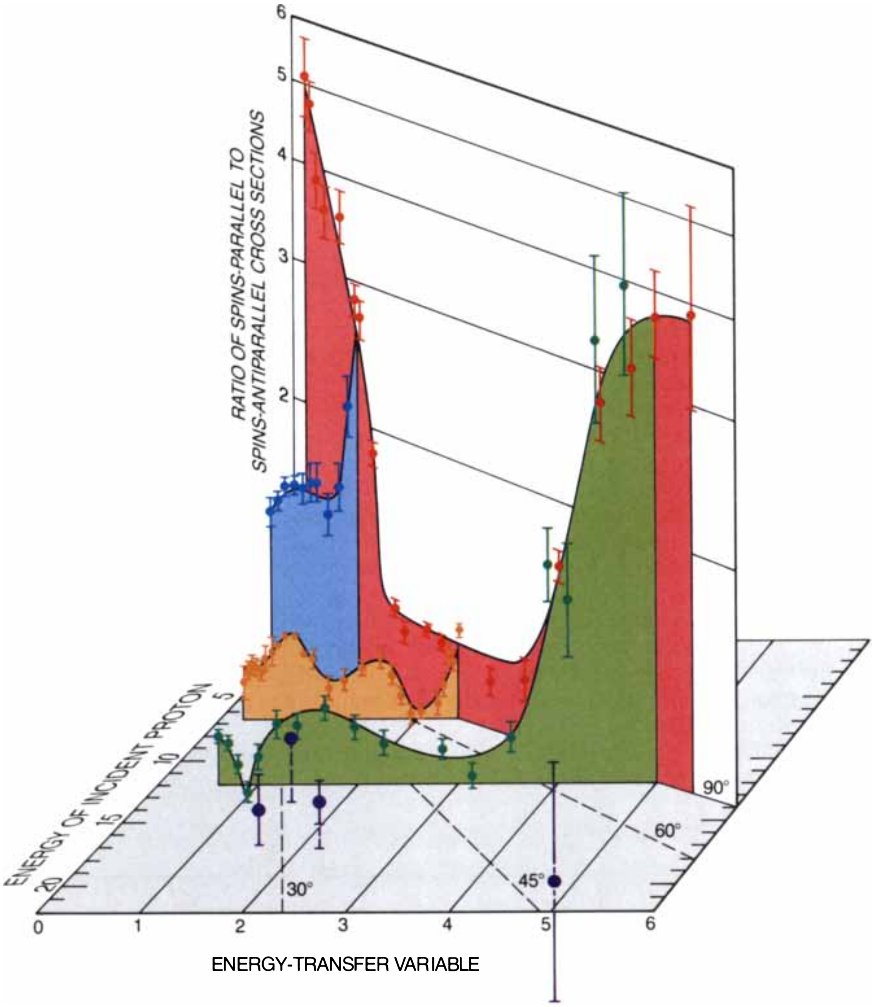
crucial numbers become equal. The first number is the number of spin rotations each proton experiences in one trip around the ring. The second is the number of times a proton oscillates through a horizontal magnetic field during each circuit. At the AGS a proton undergoes about nine betatron oscillations in each trip around the ring. When the two numbers are equal, the polarization can be destroyed completely in less than 10 microseconds, which is the time required for four turns around the AGS.

To prevent resonances from developing as the protons were accelerated, we sought to change one of the two numbers at the instant they became equal and thereby "jumped through" each depolarizing resonance. We decided to change the number of times a proton encountered a horizontal magnetic field, by changing the number of crests and troughs in the betatron oscillations. To do so we built special pulsed quadrupole magnets that could be switched on in less than 2.5 microseconds. We quickly discovered that magnets with such a fast response cannot be built from ordinary iron. Consequently we decided to try ferrite, a ceramic containing iron oxide. A ferrite magnet can generate high magnetic fields in a short time.

The pulsed quadrupoles themselves were built at Michigan and their huge power supplies at Brookhaven. Building the quadrupoles proved challenging for several reasons. The ferrite—which costs about $50 a pound—had to be cut exactly in the shape of a hyperbola, but ferrite is so hard that it can be cut only with diamond tools, and it shatters if it is slightly overheated during cutting. Our team of instrument makers, engineers, professors and students began to solve these problems by convincing the U.S. Department of Energy to donate 1.5 tons of surplus ferrite and by convincing the University of Michigan to buy a computer-controlled milling machine. Students who were experts with computers and hyperbolas worked with instrument makers who understood milling machines and diamond cutting tools. An "aquarium" that kept the fragile ferrite underwater to maintain a constant temperature during the cutting was installed on top of the milling machine. The effort finally paid off and, in 1983, 12 pulsed quadrupole magnets were mounted in the AGS.

Each quadrupole magnet generates a full magnetic field starting from zero in only 1.6 microseconds. To obtain this fast response each magnet requires a huge, sophisticated power supply that produces an electric current of 1,500 amperes at 15,000 volts—a power level of 22.5 megawatts. Fortunately during accelerator operation the combined peak power of all the quadrupoles—more than 200 megawatts—is reached for only a few microseconds each second, so that the average power is about a kilowatt. After a great deal of clever and dedicated work a team of engineers and technicians at Brookhaven successfully produced the sophisticated power supplies. Each supply, which is about as big as a truck, delivers power to one quadrupole, which is about as big as a desk lamp.

To ensure that the pulsed quadrupoles would fire precisely when the protons hit each depolarizing resonance, the AGS control computer needed a major overhaul. The computer keeps track of the exact energy of the protons during each microsecond of the acceleration cycle. For proper control of the polarized protons, the AGS computer experts had to improve the control precision by a factor of about 10. This successful improvement project had the side benefit of making the AGS a more precise accelerator for unpolarized protons. But even this precision was not good enough to dead reckon the exact instant the protons would hit a depolarizing resonance. For the final adjustments we had to rely on the fundamental tool of science: experimental observation.



RATIO of the spins-parallel cross section to the spins-antiparallel cross section for proton scattering oscillates as characteristics of the collisions are varied. The two lower axes in this three-dimensional plot are the energy of the protons in the incident beam and the energy transferred during the collisions. The energy-transfer variable increases with the scattering angle, a few values of which are shown. The red curve represents data from trials in which the measured scattering angle was exactly 90 degrees. The blue, orange and green curves represent trials in which the energy of the incident protons was set at 3, 6 and 12 GeV respectively. The large purple points represent higher-energy data from the AGS. The observed oscillations are not easily reconciled with predictions made by QCD.

After all our painful mental, physical and financial effort—only some of which I have described—we were slightly surprised and greatly pleased to find that we actually had accelerated spinning protons to 18.5 GeV. The maximum polarization was about 50 percent, which means that 75 percent of the protons were spinning in one direction and 25 percent in the opposite direction. The AGS was accelerating about 20 billion polarized protons every 2.4 seconds.

We then changed rapidly from being accelerator physicists to being high-energy experimentalists and tried to finally answer the question: Does the ratio of the spins-parallel cross sections to the spins-antiparallel cross sections increase, decrease or for some mysterious reason stay fixed at exactly a factor of four as the collision energy is increased beyond 13 GeV? Unfortunately the number of protons scattered at 90 degrees becomes quite small at higher energies; more protons scatter at smaller angles. Therefore our experimental team from Brookhaven, the University of Maryland, the Massachusetts Institute of Technology, Michigan, the University of Notre Dame, the Swiss Federal Institute of Technology and Texas A&M University first studied smaller scattering angles, which have higher counting rates.

We were in for another surprise. We found that at 18.5 GeV the ratio of the spins-parallel cross sections to the spins-antiparallel cross sections appears very close to one for violent collisions at scattering angles much smaller than 90 degrees. The plot we made based on our new data suggests that the ratio may oscillate up and down as the collisions become more energetic and more violent. Nicolai Tyurin of the Serpukhov Institute in the U.S.S.R. and other theorists who had predicted such oscillations are certainly gratified by this possible confirmation of their work.

The plot may be revealing something different, however. Specifically, for identical particles such as two protons, 90 degrees could be a special scattering angle at which unique effects are seen. The importance of 90 degrees in these spin experiments was first proposed in 1978 by Hans A. Bethe of Cornell University and Victor F. Weisskopf of M.I.T. Our new results on violent collisions far from 90 degrees can be interpreted as supporting Bethe and Weisskopf's suggestion. Although the ratio of the spins-parallel cross sections to the spins-antiparallel cross sections varies dramatically with the energy of the incoming protons, at each energy the highest ratio always occurs at 90 degrees. Harry Lipkin of the Weizmann Institute of

Science in Israel and others have recently proposed specific models where proximity to 90 degrees can strongly enhance the size of the ratio. Future AGS experiments at 90 degrees may determine which theory is correct.

Thus the mystery deepened as we went to still higher energies at the AGS. Has the QCD prediction that spin effects will disappear at high energies finally come true? Another new result, at 28 GeV, makes me think it has not. The result was found while we were readying our equipment for the experiment with the polarized beam. We calibrated our instruments by observing the elastic scattering of the normal, unpolarized AGS proton beam from our polarized proton target. According to QCD, the number of protons that scatter to the left should equal the number of protons that scatter to the right. The prediction is borne out at small scattering angles. I remember thinking that even if QCD does not account for our earlier ZGS results, it must surely be correct for the simple left-right scattering of an unpolarized beam from a polarized target.

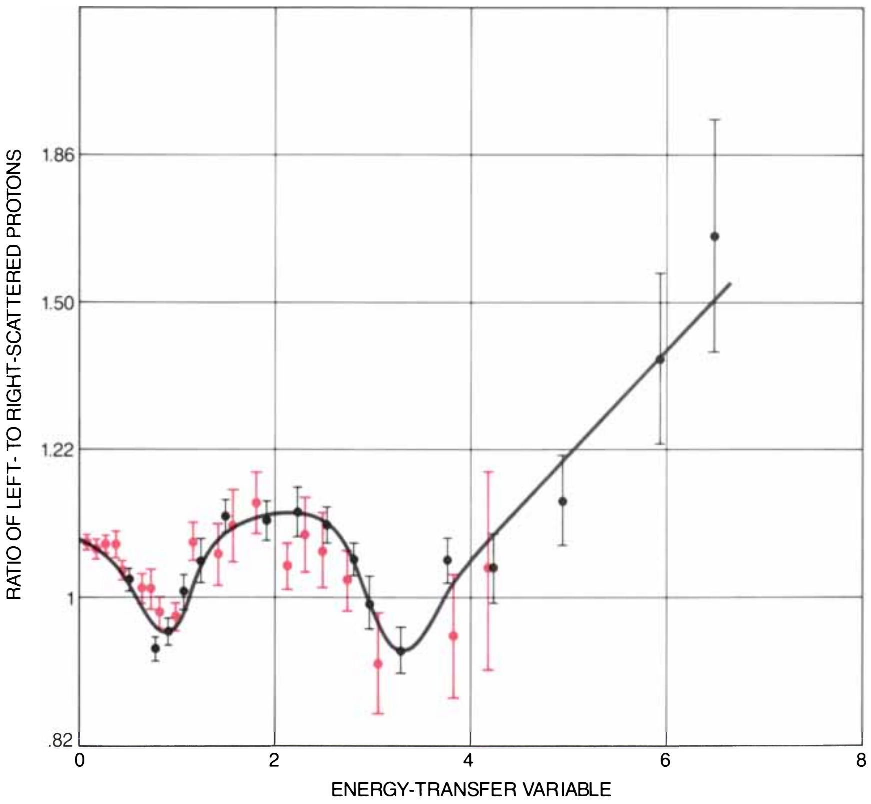
When we did the experiment, in late 1983, recent improvements in our polarized proton target enabled us to employ a beam of almost 100 billion protons in each cycle and still keep the spins of the target protons aligned. There were about 100 times more protons than had previously been scattered on a polarized target, and we could therefore study the rarer, more violent collisions. This was when the new surprises began appearing. In violent large-angle collisions at 28 GeV two-thirds more protons scattered to the left than to the right.

Initially we had only poor statistics. Since our approved running time for the experiment had ended, we asked for an extension. The AGS program advisory committee voted to reject the request; perhaps the theorists on the committee believed that we had observed only a statistical anomaly, and that further data would only confirm the QCD prediction that equal numbers of protons scattered to the left and to the right. Brookhaven's associate director, Robert B. Palmer, made a rare decision to overrule the committee. When we later confirmed that the scattering does depend on direction, he was probably pleased.

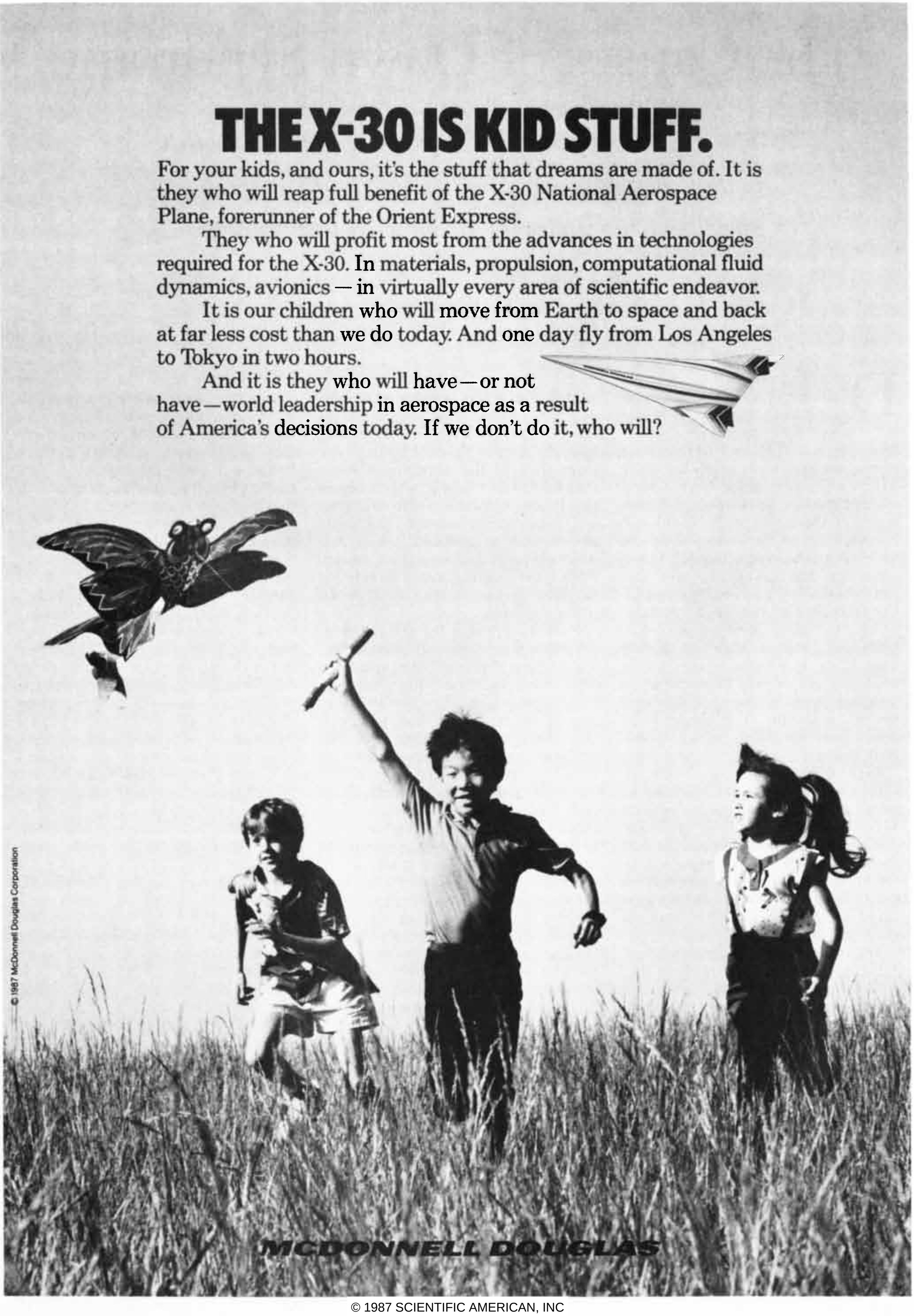
What does the observed difference between scattering to the left and to the right mean? Perhaps, as some theorists suggest, both the violence and the energy (28 GeV) of our collisions are much too low for a fundamental theory such as QCD to apply. Higher energies may soon be explored. The scattering difference between left and right may soon be measured at the 70- to 800-GeV proton synchrotrons at Serpukhov, CERN and Fermilab. The proposed 20-TeV (20,000 GeV) Superconducting Supercollider, or SSC, would make possible a further, enormous boost of energy [see "The Superconducting Supercollider," by J. David Jackson, Maury Tigner and Stanley Wojcicki; SCIENTIFIC AMERICAN, March, 1986]. It appears difficult, however, to increase the violence of elastic collisions by a large factor; as the collisions become more violent they become much rarer. Even though I am patient, I am reluctant to undertake an experiment that has only one event every 100 years. My students, who need such events for a Ph.D. thesis, are even more reluctant.

Elastic collisions at the huge energies and energy transfers needed to test QCD properly seem unlikely to be studied in this century. Perhaps we should also search for a new and more useful theory of the interactions of spinning protons and of their spinning constituents. We have a better theoretical understanding of the electric, weak and gravitational forces than we do of the strong (nuclear) force that binds quarks, protons and neutrons. Perhaps the reason is that the nuclear force is so strong that the theoretical approximation technique called perturbation theory is useless. At the same time the very strength of the strong force makes it all the more important.

The longer I stare at our data, the more I feel it contains some simple message about the protons' constituents that we have not yet deciphered. I shall not guess at what might happen next, since surprises have materialized whenever spin experiments have probed previously unexplored regions. The surprises include the large polarization of so-called lambda particles produced in violent proton-proton collisions and the huge spin effects seen in low-energy proton-proton collisions. Since I am an experimentalist, I can only rely on the ultimate judge of scientific truth, experimental observation. Perhaps measurements made in the near future will yield a clue that will help some clever young theorist to finally understand the proton's constituents and their strong forces.



LEFT-RIGHT ASYMMETRY of scattering at an even higher energy, 28 GeV, presents yet another difficulty for QCD. A beam of unpolarized protons was scattered off a polarized proton target. The number of protons that were scattered to the left exceeded the number of protons scattered to the right by about two-thirds. According to QCD, equal numbers of protons should be scattered to the left and the right. (If the target had not been polarized, the numbers would have been equal.) The colored points are data obtained at an energy of 24 GeV from CERN, the European laboratory for particle physics.



The Causes of Down Syndrome

The genes thought to be responsible for many of the pathologies associated with the disorder are being identified and mapped to sites on chromosome 21

by David Patterson

Down syndrome is a genetic condition that can have devastating consequences. The disease is the commonest cause of mental retardation in the U.S., with a frequency of one in 700 live births; children born with the disorder suffer from a spectrum of physical and mental problems, many of them severe. In addition it is believed that the genes implicated in some of the symptoms of Down syndrome are the same ones that can, when altered or improperly controlled, lead to a variety of clinical disorders, including leukemia and Alzheimer's disease, in otherwise healthy individuals. For these reasons Down syndrome is currently the focus of an intensive investigation that promises to yield a wealth of information on gene expression and the molecular basis of disease.

Down syndrome is hardly a new disease. Evidence of its antiquity can be found in the form of a ninth-century Saxon skull that has the same dimensions as the skull of a typical modern patient with Down syndrome; a variety of artistic renditions dating from the 15th century depict infants whose facial features are characteristic of the syndrome.

The syndrome was not formally recognized, however, until 1866. In that year John Langdon Down, physician at the Earlswood Asylum in Surrey, England, published the first comprehensive description of the disorder. His account was based on the observation that certain mentally retarded patients have a distinctive constellation of physical symptoms: notably epicanthic folds of the eyes, flattened facial features, unusual palm creases, muscular flaccidity and short stature.

It is now known that individuals with Down syndrome are affected by a wide variety of abnormalities, both anatomical and biochemical. Forty percent of them are born with congenital heart defects, most have small brains and many are at increased risk for developing cataracts or other vision impairments because of defects in the lenses of their eyes. Biochemically they suffer from elevated levels of purines (two of the nitrogenous bases that form DNA and RNA)—a condition that by itself can lead to neurological impairment, mental retardation and immune-system deficiencies. Additional complications include susceptibility to infection and a twenty- to fiftyfold increase in the risk of developing leukemia.

It is not surprising, therefore, that individuals with Down syndrome typically have shortened life spans. In 1929 their estimated life expectancy was only nine years. By 1980 improved medical care had increased that average to more than 30 years, and now 25 percent of individuals with Down syndrome live to the age of 50.

As the average age of individuals with Down syndrome increases, another aspect of the disease has come to light. In the past few decades it has become clear from the study of autopsy material that all individuals with Down syndrome over the age of 35 develop the same kind of abnormal microscopic senile plaques and neurofibrillary tangles in the brain as people who die from Alzheimer's disease, the major cause of presenile dementia. Individuals with Down syndrome also appear to be at a significantly increased risk of developing the cognitive symptoms of Alzheimer's disease.

For years Down syndrome was a disease of unknown origin, seemingly random in its occurrence. Many theories were proposed, including ones that linked babies who had Down syndrome to endocrine-gland malfunctioning or to tuberculosis or syphilis in the parents. In 1909 G. E. Shuttleworth of the Royal Albert Asylum in Lancaster, England, suggested that the disorder was the result of "uterine exhaustion." He based his theory on the observation that a substantial number of children with Down syndrome are the last-born members of large families. Shuttleworth was partly correct: babies with Down syndrome often are the last of a long line of children, but that fact is now attributed to the increased age of the mother rather than the number of children she produces.

In the early 1930's Adrian Bleyer of the Washington University School of Medicine in St. Louis and P. J. Waardenburg independently suggested that Down syndrome might be associated with nondisjunction, which is the failure of chromosomes to separate properly during meiosis (the process of cell division that leads to the production of egg and sperm cells). Their insight stemmed from observations of chromosomal behavior in the evening primrose plant, in which the failure of chromosomes to separate properly leads to plants that have 15 chromosomes instead of the normal 14 and are sterile. If nondisjunction could cause abnormalities in the primrose, might it not be responsible for the abnormalities associated with Down syndrome? Efforts to prove this were stalled by the fact that no one had successfully obtained a correct chromosome count for humans.

It was not until the 1950's that Joe Hin Tjio and Albert Levan of the Institute of Genetics in Lund, Sweden, determined that the correct number of chromosomes in humans was 46, and the link between Down syndrome and nondisjunction was verified. Jérôme Lejeune, Marthe Gautier and Raymond Turpin of the Institut de Progénèse in Paris counted chromosomes in the cells of patients with Down syndrome. They observed that the patients typically had undergone chromosomal nondisjunction: all had three copies of chromosome 21 in their cells rather than two, for a grand total of 47 chromosomes rather than 46.

Chromosomal nondisjunction reflects a malfunction in cell division. Regardless of which chromosomes are involved, it often has grave consequences. In nondisjunction the paired chromosomes that normally separate into different daughter cells during meiosis fail to divide properly, so that each daughter cell receives either two chromosomes or none. Cells lacking a chromosome usually die; those that have two receive a third copy on fertilization, and trisomy results. Trisomic fetuses seldom live, and those that do survive suffer from a variety of biochemical and physical defects. Down syndrome, or trisomy 21, is by far the commonest human trisomy seen in live births, perhaps in part because chromosome 21 is the smallest human chromosome.

Trisomy of at least part of chromosome 21 has so far proved to be an infallible determinant of Down syndrome. There are no individuals with the clinical symptoms of Down syndrome who do not have at least partial trisomy of chromosome 21; conversely, there are no known cases of individuals with trisomy 21 who do not have Down syndrome. The disorder need not always arise from nondisjunction; 5 percent of patients with Down syndrome have a different kind of chromosome mutation called a translocation, in which only part of chromosome 21 is present in triplicate. This occurs when a piece of chromosome 21 attaches itself to another chromosome, most often number 13, 14, 15, 21 or 22. When that happens, there are pairing problems during meiosis and the fragment of chromosome 21 appears in one of the daughter cells along with a normal 21; as in nondisjunction, the fragment becomes trisomic on fertilization. Clinical studies on a small group of patients who have Down syndrome as a result of translocation have shown that trisomy of only the bottom third of chromosome 21 is sufficient to cause the disorder.

Chromosome 21 contains only 45 million base pairs of DNA (out of a total of three billion base pairs in the nucleus of a human cell), or 1.5 percent of the total genetic material. It is estimated that human beings have about 100,000 functional genes. Assuming a roughly similar number of genes for a given amount of DNA, one can estimate that chromosome 21 has 1,500 genes. Fewer than 20 of them have been identified.

Chromosomes consist of two identical halves, called sister chromatids,

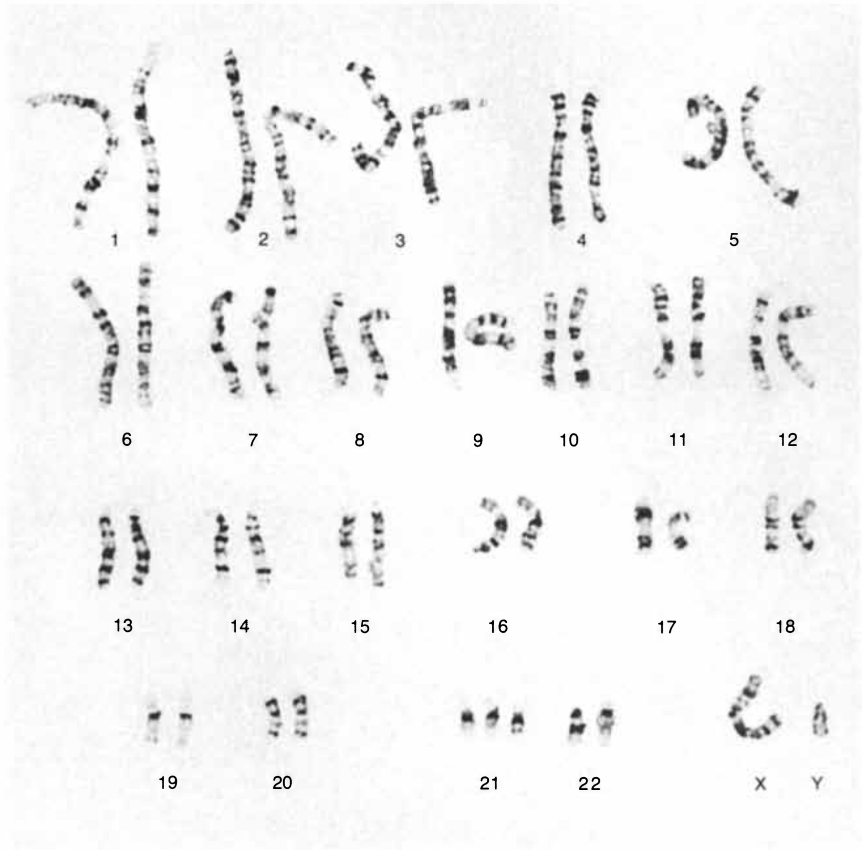
CHILD WITH DOWN SYNDROME displays the characteristic physical features of the disorder, including epicanthic folds of the eyelids, broad face and flattened nose. This four-year-old child has benefited greatly by having a family that is actively and enthusiastically involved in her care. At the age of four weeks she began formal therapy; when she was slightly more than a year old, she began speech therapy. The results have had a positive influence on her intellectual and physical development. Certain typical features of Down syndrome, including the open mouth, protruding tongue and poor posture that are associated with weak muscle tone, are much less apparent as a consequence of her continuous occupational and physical therapy. Many such individuals with Down syndrome have a significantly increased life expectancy and are capable of achieving a considerable degree of independence.

that are joined together at a region called the centromere. The short arm of the chromosome extending above the centromere is known as the p region; the longer arm below the centromere is known as the q region. Identifiable alternating light and dark regions, or bands, can be visualized by exposing the chromosomes to certain chemical stains. Bands are designated as residing in either the p or the q end of the chromosome and are assigned numbers that reflect their distance from the centromere; the farther away the band is, the larger its number is. Bands may vary in width, and each chromosome has a distinctive banding pattern. The gene that codes for the enzymes needed for purine synthesis, for example, is at the q22.1 band of chromosome 21; this pinpoints the gene to a band on the lower half of the q arm. The smallest human chromosome bands that can be recognized under the microscope contain from two to five million base pairs of DNA and therefore may contain many genes.

In studying chromosome 21, investigators have addressed four questions: What genes occupy the region of the chromosome that is specifically responsible for Down syndrome? Which of these genes are responsible for the pathogenesis of the syndrome? Exactly what proteins are encoded by the genes? By what mechanisms does the presence of three copies of the genes (instead of the normal two) lead to Down syndrome?

To answer these questions the location and identification of the various genes on chromosome 21 must be established. This is known as gene mapping. High-resolution genetic maps pinpoint genes to narrow regions of the chromosome and eventually to specific fragments of DNA, and thereby provide the most information.

The creation of such high-resolution maps typically involves several steps and combines information from different cytogenetic and biochemical techniques. Genes can sometimes be mapped to a chromosome indirectly by biochemical analysis based on gene-dosage effect. The theory is simple: since trisomic individuals possess three copies of each gene instead of two, they should produce half again as much of a particular gene product as normal individuals. Therefore the detection in a patient of roughly one and a half times the normal level of a particular protein or particular enzyme activity is a good indication that the protein or the enzyme is encoded by a gene on a trisomic chromosome.

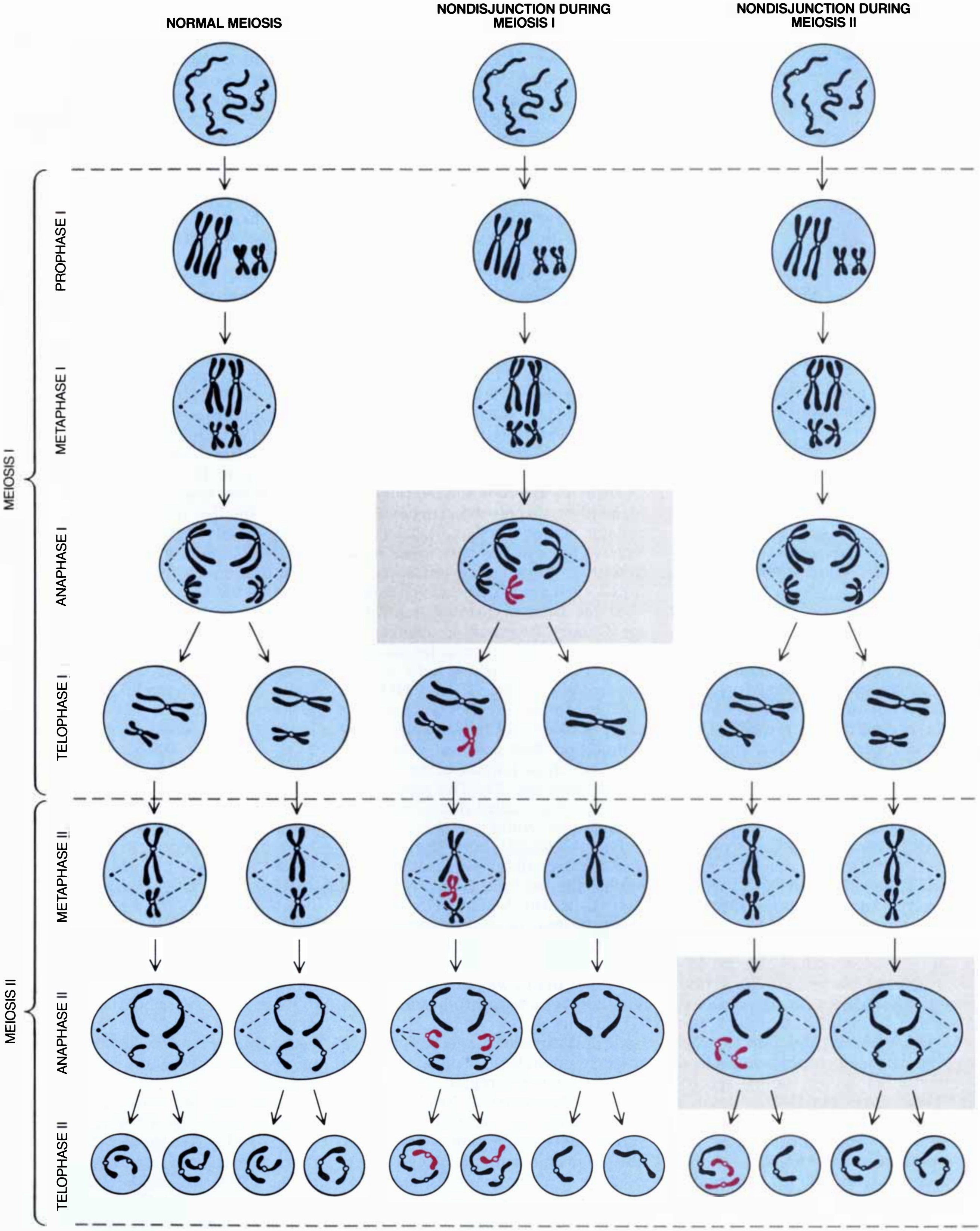


KARYOTYPE is a display of the chromosome complement of an organism. Individual chromosomes are cut from an enlarged photomicrograph taken during metaphase. The chromosomes are matched and aligned as homologous pairs. The technique enables one to obtain an accurate count of the chromosomes in a cell and examine them for gross defects. Human beings normally have 46 chromosomes (23 matched pairs), although this number may vary as a result of nondisjunction. Here the three copies of chromosome 21 indicate that the person from whom the chromosomes were obtained has Down syndrome.

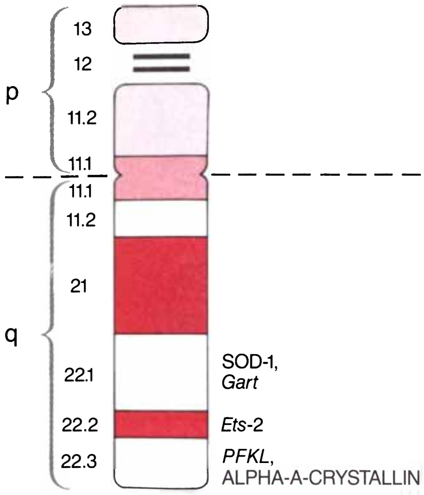
The method is particularly useful in cases of partial trisomy when increased expressivity can be associated with the presence of a narrow region of the chromosome. In some circumstances it is possible to map a gene to a single chromosome band by this method. Nevertheless, correlating a gene product with the presence of an extra chromosome is not always accurate—not only because levels of gene products vary from individual to individual or tissue to tissue but also because multiple copies of a gene are not always equally functional; a self-regulating reaction called dose compensation dampens increased protein production. It is therefore important to confirm the location of a gene identified on the basis of gene dosage by using additional methods when possible.

Another important mapping technique depends on genetic markers. Markers are genetic variations, expressed in the protein composition or physical appearance of the phenotype, that can be traced from generation to generation. If a genetic marker known to be encoded by the DNA of chromosome 21 almost always appears with a second trait in the phenotype, then that second trait may be assumed also to be on chromosome 21. Once both genes are mapped to the same chromosome, it is possible to assess the relative distance between them by determining how often they separate during meiosis when sister chromosomes exchange pieces in the course of genetic recombination. Genes that are linked (that is, typically inherited together) are likely to be in close proximity to each other on the chromosome; genes that are not linked (those that generally move independently during recombination) are generally far apart on the chromosome or even on different chromosomes. The frequency of recombination thus serves as a measure of the genetic distance between genes.

Recently an entirely new class of genetic markers known as restriction fragment length polymorphisms (RFLP's) has become a powerful mapping tool. Restriction fragments are pieces of DNA that have been cleaved by enzymes called restriction endonucleases, each of which recognizes specific target sequences on a strand



NORMAL MEIOSIS (*left*) leads to production of haploid egg and sperm cells, which have half as many chromosomes as other cells. During meiosis I, paired homologous chromosomes exchange genetic information (in prophase) and separate to opposite sides of the cell (anaphase), after which the original cell divides (telophase). The process is repeated during meiosis II, but now the two sister chromatids separate and each cell divides again to form four daughter cells. Nondisjunction occurs when the chromosomes fail to separate during anaphase of either meiosis I (*center*) or meiosis II (*right*). Chromosomes that are normally pulled by spindle fibers to opposite sides of the cell remain together and either two paired homologous chromosomes (in meiosis I) or two sister chromatids (in meiosis II) are pulled to the same side of the cell. The result is an abnormal number of chromosomes in the four daughter cells. If nondisjunction affects chromosome 21, individuals born with three copies of this chromosome have Down syndrome.



FIVE GENES that are possibly associated with Down syndrome have been mapped to a narrow region of the long, or q, arm of chromosome 21. Genes that code for the proteins superoxide dismutase (SOD-1) and alpha-A-crystallin are mapped to the q22.1 and q22.3 bands respectively, and the *Gart* and *ets*-2 genes are mapped to the q22.1 and q22.2 bands. A gene that codes for the liver enzyme phosphofructokinase (*PFKL*) is also mapped to the q22.3 band.

of DNA and consistently cleaves the DNA only at those points. Mutations in the DNA may change the pattern of cleavage sites, yielding fragments of different lengths from those that would ordinarily be present. These length changes, or RFLP's, allow one to detect differences in identifiable regions of the DNA strand. The RFLP's do not always occur in regions that code for genes, but they are nonetheless highly reliable markers to which genes or other genetic markers can be linked. The method has been used extensively to map genes and DNA sequences on chromosome 21.

The ability to isolate specific genes or DNA sequences has also made a new technique known as in situ hybridization possible. Specific sequences of DNA can be isolated, labeled with a radioactive isotope and used to probe a preparation of chromosomes from another individual or another species. The probe seeks out DNA strands in the chromosomes that are complementary to it and binds to them. The chromosome preparation can then be exposed to a photographic emulsion to determine exactly where the radioactive probes are binding to the intact chromosomes. This technique allows relatively high resolution: DNA fragments can be localized to within one-third of chromosome 21.

The field of molecular genetics has also been revolutionized in the past few years by the development of a powerful mapping technique known as interspecies somatic-cell hybridization. Geneticists have developed the means by which human cells can be fused with rodent cells (most commonly mouse or Chinese hamster cells) to create hybrids that retain chromosomes from both species. For reasons that are not clear, most of the human chromosomes are lost from these hybrid cells and the rodent chromosomes are retained. It is possible, therefore, to obtain a hybrid cell in which a particular chromosome is retained as the only human chromosome. Because the human chromosome present in the hybrid is at least partially functional, it is often possible to distinguish the protein products encoded by it from those encoded by the rodent chromosomes.

Francis H. Ruddle and his co-workers at Yale University were the first investigators to develop a hybrid in this way that contained only chromosome 21 from human cells and the full complement of chromosomes from a special line of mouse cells, called A9 cells. Several laboratories have since made hybrids containing chromosome 21 following the same general procedure.

At the Eleanor Roosevelt Institute for Cancer Research in Denver we have substituted a particular strain of Chinese hamster ovary cells (CHO) for mouse A9 cells. These CHO cells are ideal for cell-hybridization studies of chromosome 21 because they carry a mutation that inactivates the *Gart* gene, which in human beings resides on chromosome 21. The gene codes for an enzyme called phosphoribosylglycinamide synthetase. Lacking this enzyme, CHO cells must be grown in a medium that contains purines; without purines they will die. If one hybridizes CHO cells with human cells, any hybrid cells that contain chromosome 21 will grow in the purine-free medium, because they have acquired the human *Gart* gene. In this way we can select for retention of human chromosome 21 in hybrid cells.

Modifications of the technique have made it possible to isolate the *Gart* gene to a narrow region of chromosome 21, designated as 21q22, by creating hybrids between CHO cells and human cells containing a translocation of chromosome 21. By cytogenetic analysis of the translocation chromosomes in hybrid cells, we have been able to fractionate chromosome 21 into several subregions and to assign genetic markers to a subregion that includes the *Gart* gene. The only drawback to this mapping procedure is that it depends on naturally occurring translocations involving different fragments of chromosome 21, and these are rare indeed.

One way to circumvent the scarcity of translocation fragments is to break up chromosome 21 by exposing it to large doses of radiation. The resulting fragments can then be hybridized with Chinese hamster cells and manipulated in much the same way that natural fragments are. The hybrids can be analyzed for the presence of specific genes or DNA sequences and molecular linkage groups can be formed with the *Gart* gene as a linchpin.

The final step in the creation of a high-resolution map of chromosome 21 requires that the physical distance between the various genes and markers be measured by counting the number of base pairs between them. The basic tool is a technique known as electrophoresis, which subjects DNA fragments to an electric field. Because DNA carries a negative charge proportional to its size, the DNA molecules move toward the positive electrode. Smaller fragments move relatively rapidly, larger fragments more slowly. The size of fragments, measured in number of base pairs, is indicated by the final position of the fragments on a gel.

David C. Schwartz and Charles R. Cantor of the Columbia University College of Physicians and Surgeons found they could modify the standard technique of electrophoresis by pulsing the DNA with roughly diagonal fields of electric current. For reasons that are not completely understood, this allows exceedingly large DNA fragments (ranging in size from 50,000 to more than five million base pairs) to be moved and separated from one another. Standard electrophoresis can analyze fragments no larger than 50,000 base pairs. Yet genes vary greatly in size. Many are much larger than 50,000 base pairs; some are estimated to be as much as one million base pairs or more in length. The pulsed-field technique makes it possible for geneticists to analyze the entire gene rather than just a fragment of it, as well as large pieces of DNA that may contain several genes.

Comparing the maps produced by these methods reveals that the maps are generally consistent with one another, at least with respect to the order of the DNA fragments and markers on chromosome 21. By combining information from these methods and from various clinical and cytogenetic studies, investigators have identified a number of genes that lie within the 21q22 band of chromosome 21. This appears to be the region of the chromosome specifically responsible for the pathogenesis of Down syndrome.

Now the task is to relate these genes

to specific pathologies. Since genes code for proteins, which ultimately determine the phenotype of the human being, knowing which proteins contribute to the pathology of Down syndrome is clearly important. Once the proteins have been identified and their effects delineated, it is conceivable that eventually measures may be devised to counteract the deleterious effects of these proteins.

Examples of genes mapped to the 21q22 band of chromosome 21 are the *Gart* gene mentioned above, the *ets*-2 gene and genes encoding the alpha-A-crystallin, superoxide dismutase and amyloid beta proteins. These genes provide a fascinating glimpse of the complex relation between genes (and the proteins they encode) and human development and disease. For example, recent developments in molecular biology have shed interesting light on the possible relation between the *ets*-2 gene on chromosome 21 and the onset of leukemia.

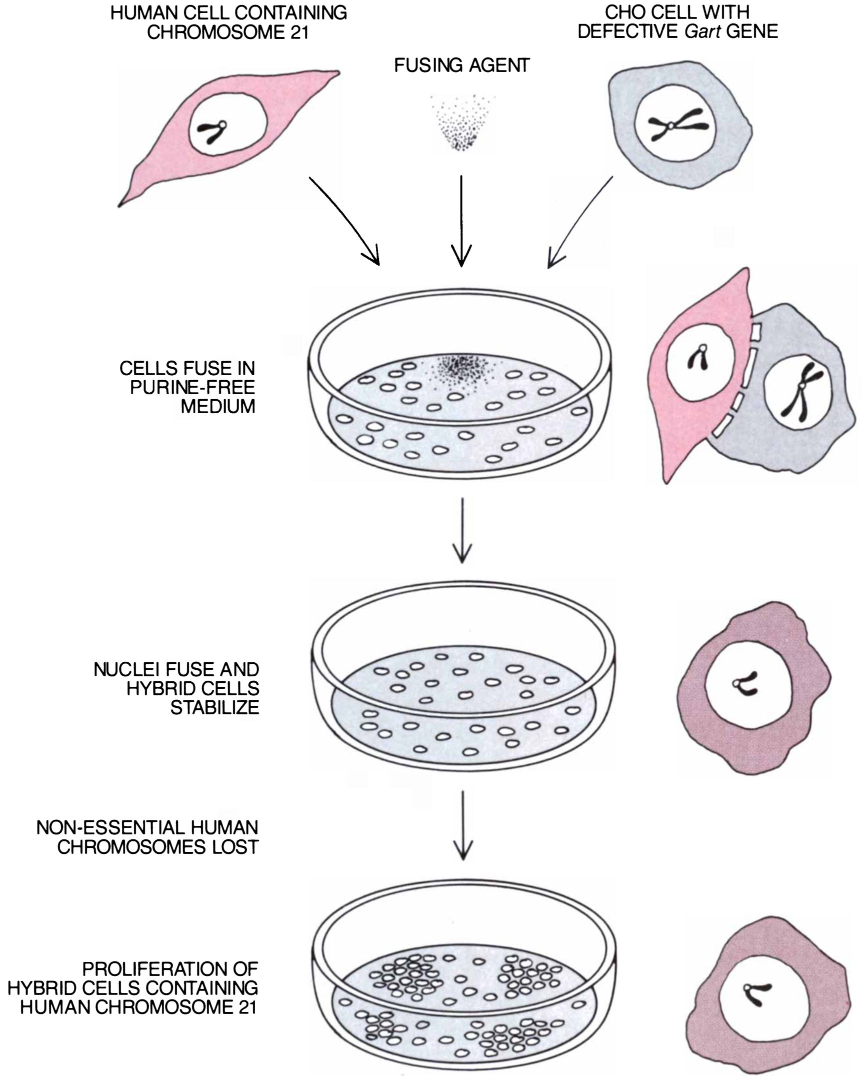
For years physicians have known that individuals with Down syndrome are at high risk for leukemia but have been at a loss to explain the underlying mechanisms. Then cancer investigators discovered that there is a particular type of leukemia, acute myelogenous leukemia, subgroup M2, in which 18 percent of affected individuals have a reciprocal chromosomal translocation involving fragments of chromosomes 8 and 21. They do not have Down syndrome; only their leukemic cells are affected, whereas the chromosomes in other cells appear to be normal. By means of cytogenetic mapping methods investigators were able to localize the breakpoint on chromosome 21 to the q22 region. Another fascinating finding is that trisomy 21 is the chromosomal abnormality most often observed in leukemia cells, particularly in children.

In 1985 Janet D. Rowley and her colleagues at the University of Chicago's Pritzker School of Medicine, in collaboration with our group, hypothesized that the leukemia could be caused by an alteration in the activity of one or more oncogenes (cancer-causing genes) on chromosome 21. We speculated that either trisomy of chromosome 21 or a translocation of genetic material from chromosome 21 to chromosome 8 was responsible for the alteration.

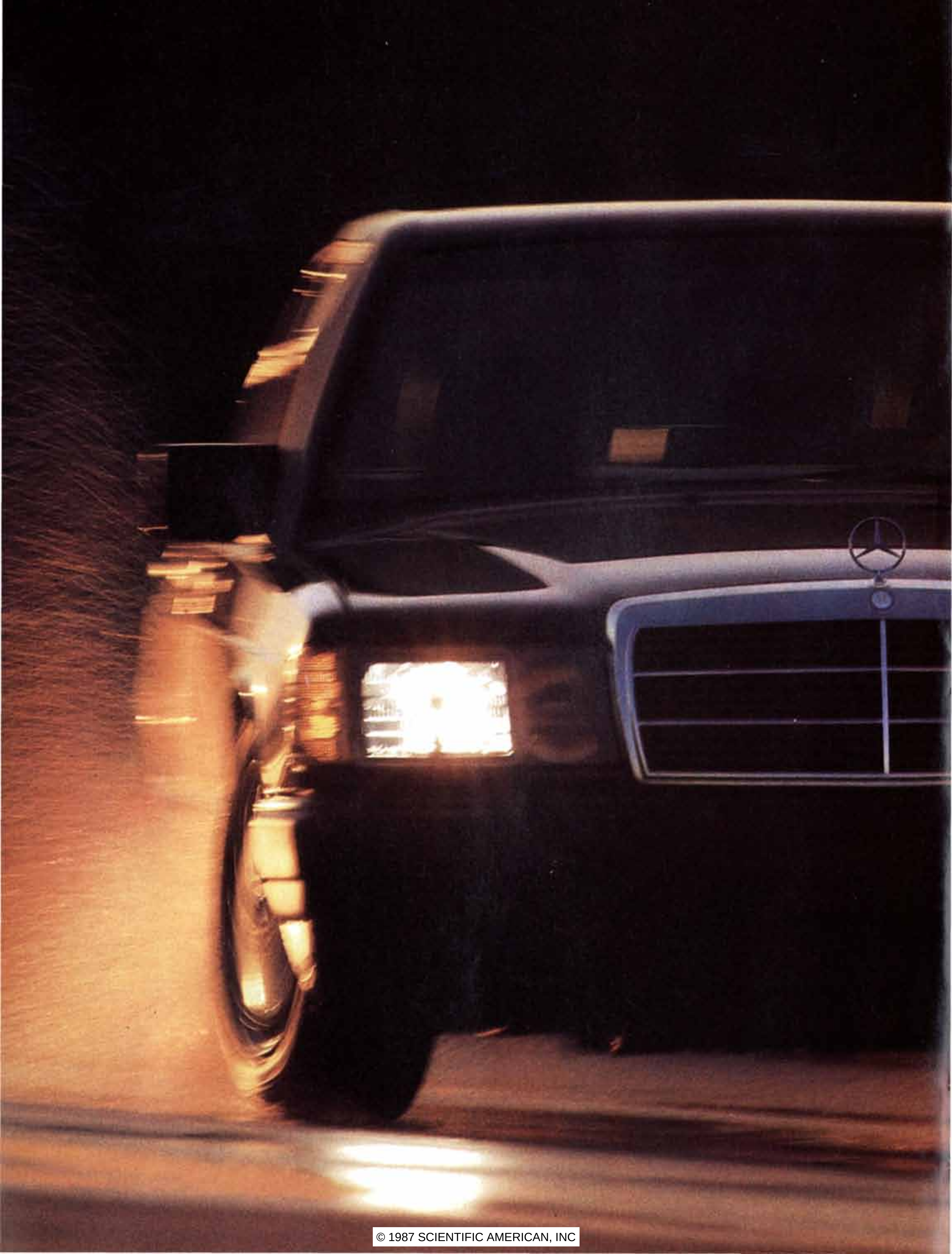
Evidence consistent with this hypothesis is mounting. Takis S. Papas of the National Cancer Institute, in collaboration with a number of other workers, including members of our group, has shown that indeed there is an oncogene on chromosome 21, called the *ets*-2 oncogene, and has mapped it to the region involved in the 8;21 translocation. Studies are currently under way to compare expression of the *ets*-2 gene in normal cells with expression of the gene in cells that have become trisomic or have undergone a translocation.

The discovery that the *Gart* gene, which codes for three different enzymes involved in purine synthesis, is in the 21q22 region of chromosome 21 may explain why individuals with Down syndrome have increased serum levels of purines. Since elevated purine levels are associated with a variety of problems, including mental retardation, one can hypothesize that trisomy of this single gene on chromosome 21 is sufficient to account for many of the problems associated with Down syndrome, including mental retardation. Investigators hope that restoration of normal purine levels through control of the expression of the *Gart* gene (perhaps early in fetal development) may someday lead to amelioration of the symptoms of Down syndrome.

The increased risk of cataracts and lens defects in individuals with Down syndrome may also be explained by the abnormal expression of a particular protein. Investigators have found


HUMAN CELL CONTAINING CHROMOSOME 21
FUSING AGENT
CHO CELL WITH DEFECTIVE *Gart* GENE
CELLS FUSE IN PURINE-FREE MEDIUM
NUCLEI FUSE AND HYBRID CELLS STABILIZE
NON-ESSENTIAL HUMAN CHROMOSOMES LOST
PROLIFERATION OF HYBRID CELLS CONTAINING HUMAN CHROMOSOME 21

SOMATIC-CELL HYBRIDIZATION between human and Chinese hamster ovary (CHO) cells can lead to the creation of hybrids that contain chromosomes from both species. A fusing agent such as the Sendai virus or polyethylene glycol is added to the growth medium to promote fusion. Human cells can be hybridized with CHO cells that are unable to synthesize purine because they have a defective *Gart* gene. Human chromosomes not essential to the growth of the hybrid are gradually lost as the cell divides. If hybridization takes place in a purine-free medium, only those cells that contain the human *Gart* gene (on chromosome 21) and can synthesize their own purine will survive. Survival thus identifies hybrid cells containing all or part of chromosome 21. If only part of the chromosome is retained, the region of the chromosome having the *Gart* gene can be identified.



that the gene encoding the alpha-A-crystallin protein, a structural component of the lens of the eye, is on chromosome 21 in the 21q22 region.

Still another gene identified on chromosome 21 is the one that codes for the soluble form of the enzyme superoxide dismutase (SOD-1). This enzyme is part of a complex system that protects mammalian cells from oxygen-containing free radicals: highly reactive molecules, released during oxidation, that may be implicated in the natural aging process. Alterations in SOD-1 levels may contribute to mental retardation in patients with Down syndrome and may also explain their accelerated rate of aging. The enzyme appears to be expressed according to gene dosage and has been mapped to the q22 region of chromosome 21. Understanding how this gene functions may ultimately provide investigators with valuable clues about the normal aging process in humans.

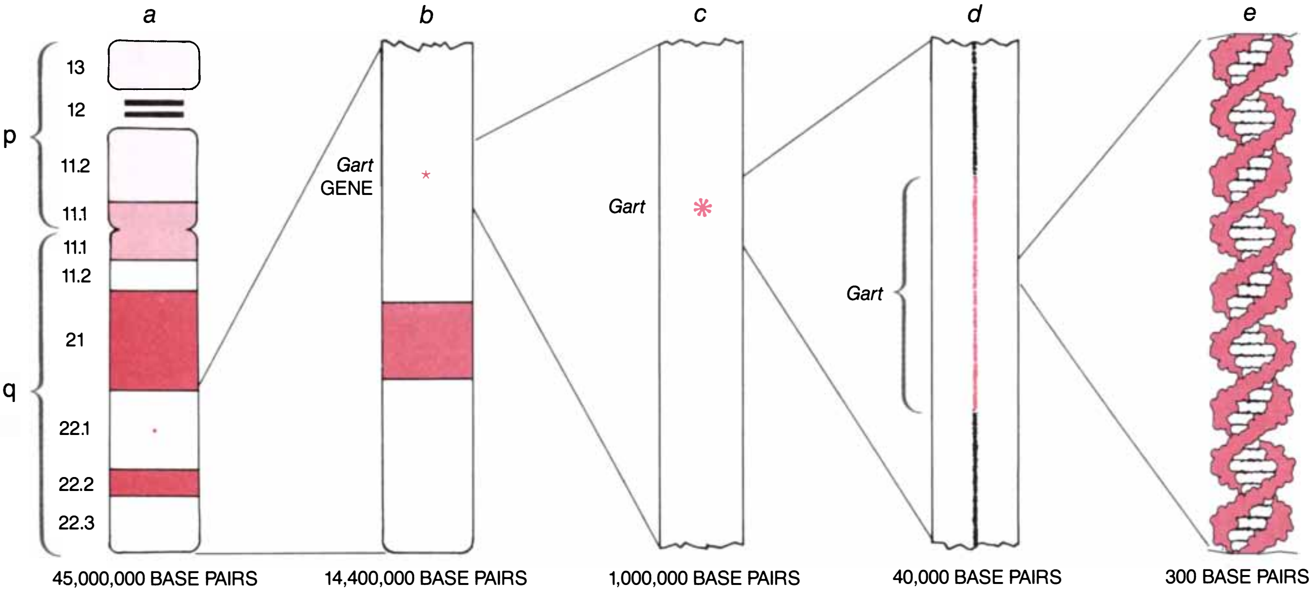
A most exciting recent discovery is the genetic link between Alzheimer's disease and Down syndrome. George G. Glenner and his colleagues at the University of California at San Diego demonstrated that the amyloid beta protein (a major component of the neurofibrillary plaques that accumulate in the brain of individuals with Alzheimer's disease) is identical with the protein that accumulates in apparently identical lesions in the brain of all individuals with Down syndrome who are over 35 years old. Investigators at several institutions simultaneously determined that the gene for amyloid beta protein resides on chromosome 21. Moreover, they found that the gene apparently responsible for a familial form of Alzheimer's disease also resides on chromosome 21. It is not yet known whether these two genes are in fact the same gene.

Great progress has been made in identifying and mapping various genes on chromosome 21, a process that is virtually certain to intensify in the future. Yet investigators still have no concrete proof that trisomy of even a part of chromosome 21 is the direct cause of pathology. New advances in genetic engineering that make the manipulation of animal models possible are likely to remedy this situation.

Geneticists have already discovered that part of mouse chromosome 16 contains the *ets*-2, *Gart* and SOD-1 genes, as well as a number of other DNA sequences whose homologues are known to reside on human chromosome 21. John D. Gearhart of the Johns Hopkins University School of Medicine and Charles J. Epstein and David R. Cox of the University of California at San Francisco School of Medicine and their colleagues are attempting to breed mice that have partial trisomies of chromosome 16 to see whether they develop some of the pathologies associated with Down syndrome. Although mice produced in this way are unlikely to have precisely the same pathologies as humans, it seems reasonable to assume that the underlying mechanisms of gene expression are the same, so that research on these mice will provide geneticists with fresh insight into the pathology of Down syndrome.

Another very promising tool in molecular genetics is a genetic-engineering technique that leads to the creation of transgenic mice: mice that contain single genes or groups of genes from humans. The feasibility of this technique has been amply demonstrated for oncogenes, growth factor genes and metabolic-enzyme genes; now it is time to apply it to genes from chromosome 21 that are thought to be associated with Down syndrome.

In forthcoming experiments genes such as *ets*-2, SOD-1 and *Gart* from chromosome 21 will be injected into mouse embryo cells and expressed in the resultant offspring. In that way the effects of trisomy of just one gene or a small set of genes can be studied in detail, providing a wealth of information about the multiple effects of a single gene. Eventually it may be possible to test in these animal models various approaches to preventing or treating some of the pathologies associated with Down syndrome.



MOLECULAR ANALYSIS of chromosome 21 takes place at different levels of resolution, making it possible to map a gene such as *Gart* to a region on the chromosome and also to determine its sequence of base pairs. Cytogenetic analysis relies on one of two methods: regions of DNA from one to two million base pairs in length are identified by exposing chromosomes to chemical stains (*a*), or a higher level of resolution may be obtained if a translocation involving only a fragment of chromosome 21 has taken place (*b*). Still higher resolution is attained by a number of more sophisticated techniques, including the creation of various interspecies irradiation hybrids and the use of pulsed-field electrophoresis (*c*), which yield resolutions ranging from 10 million to 30,000 base pairs. Molecular cloning methods and standard gel electrophoresis allow resolution of from 50,000 to as few as several hundred base pairs (*d*). Finally, the actual sequence of base pairs in a particular gene can be determined by DNA-sequencing techniques (*e*).

"Going above and beyond what's expected— that's good news to me."

Cathleen Black, Publisher, USA TODAY

"I've never been one for giving people what they expect. I've always believed in giving them better. That's why I joined USA TODAY.

While the rest of the world may be comfortable with convention, I think you've got to keep yourself open to change. To improvement.

You have to be willing to explore the options. Weigh the alternatives. And most of all, you have to be willing to trust your own instincts.

Setting new standards isn't easy. Believe me, I know. But when you succeed—when you deliver more than what's expected—the rewards are that much richer."

At Lufthansa, we always go above and beyond what's expected to create the very highest standards in the sky. Maybe that explains our outstanding on-time arrival rate, our meticulously maintained fleet and our friendly courteous service.

It may also explain why in 161 cities worldwide, people prefer flying Lufthansa.

People expect the world of us.

Lufthansa
German Airlines

Lufthansa is a participant in United's Mileage Plus program and Delta's Frequent Flyer program. See your Travel Agent for details.

The Clonal-Selection Theory

The antibodies that defend the body from foreign invasion are remarkably diverse. It took nearly 100 years to define and substantiate a theory that could account for their formation

by Gordon L. Ada and Sir Gustav Nossal

How do cells make antibodies in such enormous variety? Today that question has largely been answered, and the protective proteins produced by the immune system are the servants of medicine and research. Not so long ago, however, antibodies were much more mysterious. Their origin and mode of action were the focus of several competing and conflicting theories.

Antibody formation is problematic because of the almost incredible diversity of antigens, the foreign substances that provoke an immune response. One antibody can neutralize just one type of antigen—but antigens come in a variety of shapes, sizes and chemical compositions. Bacteria and their toxins, viruses, pollen grains, incompatible blood cells and manmade molecules can all act as antigens. To catch each type of intruder, the white blood cells called *B* lymphocytes must create hoards of customized antibodies.

How antigens guide the immune reaction was the chief point of contention among early theorists, splitting them into two camps. One school of thought held that antigens serve as templates that direct the design of matching antibodies. The other school believed lymphocytes maintain a pool of predesigned antibodies from which an antigen selects its closest match. In the past 30 years the combined efforts of many investigators in many countries have clarified the biological basis of immunity and resolved the conflict. The second viewpoint has triumphed, in a doctrine called the clonal-selection theory of antibody formation.

This article documents the creation and validation of the clonal-selection theory. The theory was proposed in essence almost 100 years ago, but it then fell into and out of favor with the vicissitudes of evidence and speculation. Both of us were privileged to take part in the research that brought the theory into its own in the 1960's.

The search for the mechanism of antibody formation began late in the 19th century, by which time Louis Pasteur's germ theory of disease had become generally accepted. Several groups began to study the reaction between bacterial toxins and the "antitoxins" that appear after infection in the blood serum, the fluid component of blood. The German bacteriologist Emil A. von Behring called the substances Antikörper, or antibodies.

Reactions between antibodies and antigens could be observed in a test tube because the reactants formed aggregates visible to the unaided eye. It soon became apparent that bacterial products were not the only thing that could spur antibody production. Other natural substances such as the proteins in milk and "foreign" cells could also engage the immune response of the "host" animal.

In 1890 von Behring, who had developed an antitoxin against diphtheria, met another German medical scientist, Paul Ehrlich. Ehrlich later published a landmark paper describing a technique for measuring diphtheria antibodies in preparations like von Behring's. The technique allowed such preparations to be standardized and consequently made diphtheria antiserum safe for clinical practice.

In devising his technique for measuring antibodies Ehrlich established the basis for a quantitative approach to immunology. The new, quantitative observations revealed the immune response to be an explosive proliferation of antibodies following contact with an infectious agent. To explain how foreign substances could induce this reaction, Ehrlich developed his side-chain theory of antibody formation.

Announced at the turn of the century, the side-chain theory postulated that a white blood cell's surface bore receptors with side chains to which foreign substances became chemically linked. This binding prompted the cell to produce copies of the bound receptor in great excess. The superfluous receptors—antibodies—were shed into the blood. Ehrlich's critical assumption was that cells naturally made side chains that were capable of binding all foreign substances.

Some critics of Ehrlich's theory opposed the idea of a chemical union between the antigen and its receptor. Other workers accepted the concept but debated the mechanism of antigen binding. Did the cell "swallow" the antigen-receptor complex, or was the reaction reversible, with receptors clutching antigens and then letting go? Amidst these disputes the study of antigen-antibody reactions and antibody specificity took on greater importance. It was known that antibodies begin circulating in the blood of an animal soon after its first contact with an antigen and that a second contact with the same antigen results in a more rapid and more potent response. No one knew, however, just how particular these antibodies are.

Forging the Template Theory

Many investigators approached the question of antibody specificity by examining the effects of modified antigens. In 1906 Ernest P. Pick and Friedrich P. Obermayer in Germany showed that attaching chemical groups such as iodine or nitrate to a protein profoundly changes its antigenic properties. At about the same time the Austrian-born immunologist Karl Landsteiner began pursuing a similar experimental strategy. His

CONCEPTUAL FATHER of the clonal-selection theory, German physician Paul Ehrlich proposed in 1900 that antibodies exist as specific receptors on cell surfaces. The Nobel laureate also helped to develop an antiserum for the treatment of diphtheria and later formulated a drug for syphilis.

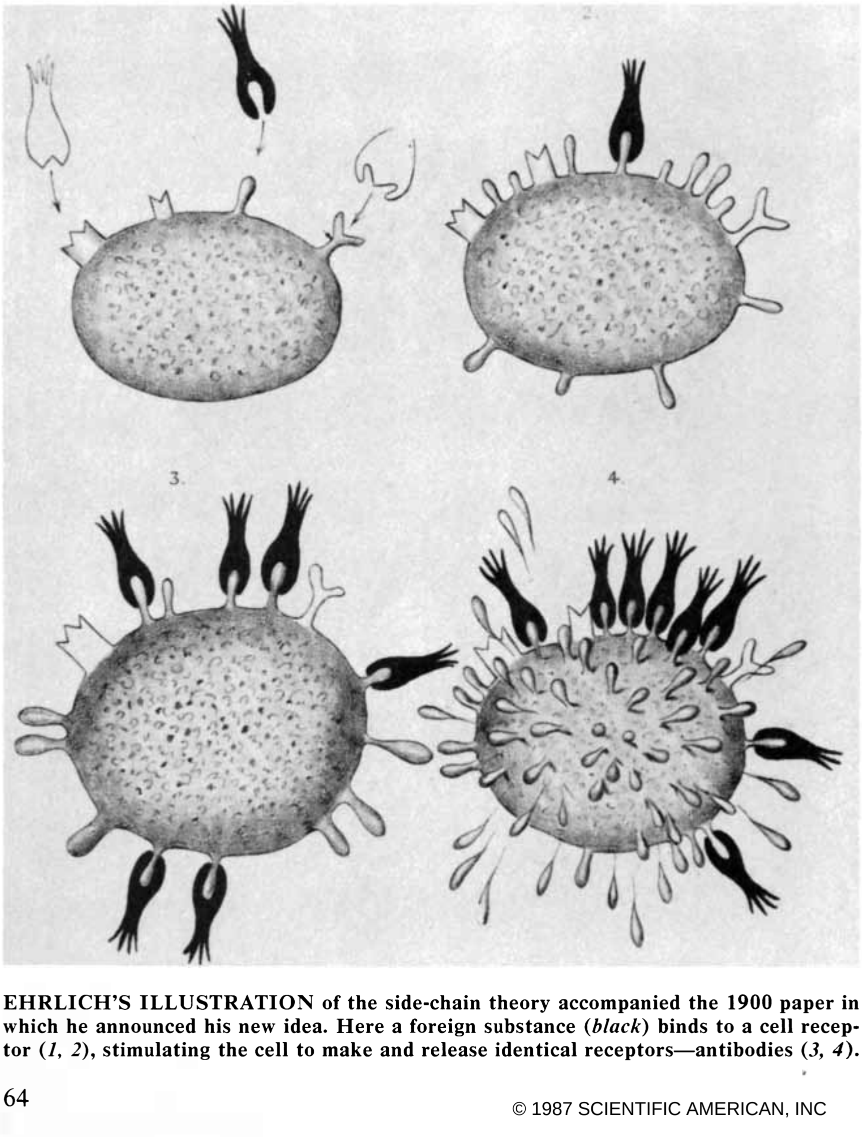
work, most of which was done at the Rockefeller Institute for Medical Research, would span three decades and demonstrate conclusively the exquisite specificity of antibodies.

Landsteiner coupled antigenic proteins with a great variety of chemical groups, some of which came from pathogenic microbes and some of which were synthesized in a test tube. Each altered molecule evoked a different antibody. Landsteiner thereby confirmed Pick and Obermeyer's evidence that chemical subunits of a larger antigenic structure could determine immunological specificity. These determinants, which were simply foreign molecular patterns, did not themselves constitute antigens: they could not initiate antibody production or form aggregates with antibodies unless they were linked to a carrier molecule.

It is now evident that many determinants cannot generate an immunological response on their own because they are too small; they need to be attached to a larger molecule in order to be recognized. Eighty years ago, however, the finding that determinants were not antigens was rather puzzling. Furthermore, the discovery of antigenic determinants indicated that the diversity of foreign substances confronting the immune system was much greater than had been suspected. It was now clear that a single intruder, such as a bacterium, could bear thousands of determinants and initiate the production of thousands of different antibodies.

From Landsteiner's studies it became apparent that an animal could manufacture a practically unlimited range of antibodies and that it could even make antibodies against novel artificial compounds. These discoveries invited the conclusion that a host animal could not inherently possess the information it uses in responding to such a wide spectrum of antigens. Ehrlich's side-chain theory was therefore discarded in favor of the concept that antigens can somehow direct antibody specificity as the antibodies are synthesized in the blood cell.

How could an antigen mold the conformation of an antibody? In the early 1930's a number of proposals were advanced, all of them necessarily conjectural. Friedrich Breinl and Felix Haurowitz of the University of Prague and Stuart Mudd and Jerome Alexander of the University of Pennsylvania thought antibodies were manufactured in direct contact with their antigens, adopting a shape and a chemical affinity that were complementary to those of the antigen. They suggested that antibodies molded by different antigens might differ in their protein composition; that is, the sequences of amino acid building blocks making up the antibodies might differ. Others, most notably Linus Pauling, contended that a wide variety of binding specificities could result merely from folding the same antibody protein in different ways. Again, the folding would be guided by the antigen template. Collectively these suppositions were called the template theory of antigen formation. The theory held sway for a quarter of a century.



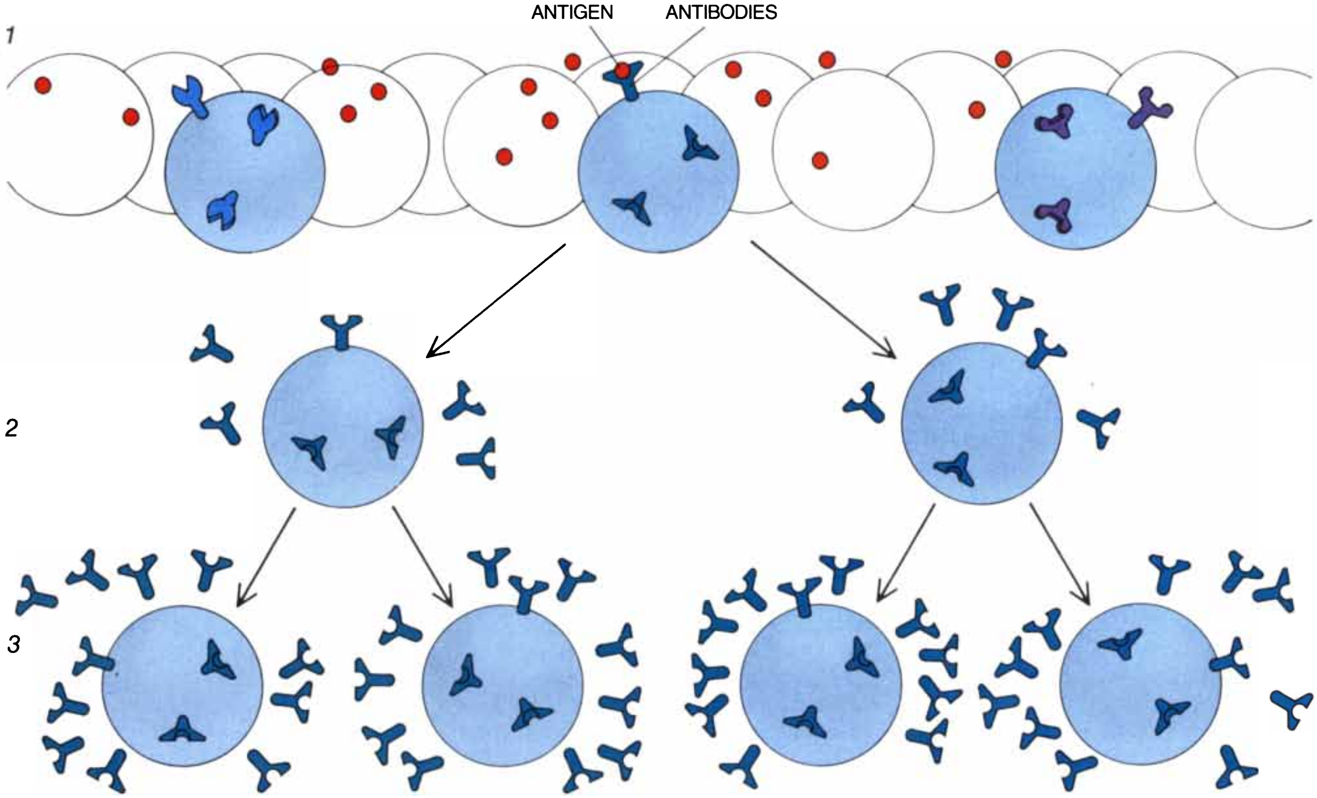
EHRLICH'S ILLUSTRATION of the side-chain theory accompanied the 1900 paper in which he announced his new idea. Here a foreign substance (*black*) binds to a cell receptor (*1, 2*), stimulating the cell to make and release identical receptors—antibodies (*3, 4*).

These arguments about mechanisms of antibody formation took place before the dawn of molecular biology in the 1950's; they were based more on concepts than on experimental data. Hence the first serious challenge to the template theory came from biologists, not biochemists. Three scientists as remarkable as any who had preceded them marshaled the incriminating evidence. Two would later receive Nobel prizes, as had von Behring, Ehrlich and Landsteiner before them.

A Crack in the Template

The first comprehensive attack was launched in 1955 by the Danish immunologist Niels Kaj Jerne. In a paper published that year, Jerne recalled that in 1949 two Australian investigators, F. Macfarlane Burnet and Frank J. Fenner, had noted several observations the template theory could not accommodate. First, the theory could not explain the apparently exponential rise in antibody production during the early stages of an immune response. If an antigenic template was required to make each antibody, it was hard to envision how antibodies could so quickly outnumber their templates. Moreover, the theory could not account for the boost in antibody production that occurs when an animal encounters a given antigen for the second time. Why should reintroducing a template give rise to more copies of antibody than the initial contact did?

The fact that antibody production continues long after the antigen is gone also posed problems for the template theory, since it was thought antibody-producing cells were short-lived. Fur-



THE CLONAL-SELECTION THEORY, which gained favor in the 1960's, holds that antibody-producing cells have specific receptors and that each cell makes just one kind. If an antigen matches a receptor, it binds (*1*), inducing the cell to divide and make more receptors (*2*). As in Ehrlich's model, the receptors, or antibodies, are shed from the cell surface into the blood (*3*).

thermore, the template theory made no provision for the fact that, as the immune reaction progressed, antibodies seemed to become better at binding their target antigens.

Perhaps the most compelling challenge Burnet and Fenner had noted was the phenomenon of immunological tolerance. Tolerance, the failure to launch an immunological campaign against a given antigen, had been recognized not long before Burnet and Fenner published their criticisms. Tolerance keeps an animal from making antibodies against itself and can be acquired for foreign antigens, if the antigens are administered before or at birth. In contrast to immunity, tolerance cannot be maintained unless the antigen persists in the animal. On the subject of immunological tolerance the template theory was silent.

In his article "The Natural-Selection Theory of Antibody Formation" Jerne drew on these criticisms to formulate a notion that would have sounded familiar to Ehrlich, namely that any given animal possesses small numbers of antibodies against all antigens. An immune response occurs when an antigen binds to an antibody and the antigen-antibody complex interacts with white blood cells, stimulating the production and release of the same specific antibody in great quantities. Jerne's proposal was supported by the observation that normal blood serum always contains globulins, nonspecific antibody proteins that at the time seemed to differ from antibodies only in their lack of specificity. Jerne also cited a review by Robert Doerr of the University of Basel that summarized the evidence for "natural" antibodies—antibodies generated without an antigenic stimulus.

At the University of Colorado's School of Medicine at Denver, David W. Talmage read Jerne's proposal and saw the similarities between Jerne's thinking and Ehrlich's side-chain theory. Talmage went a bit further. In a 1957 paper he suggested that replicating cells as well as freely circulating antibody molecules must be a central feature of the immune response. Cells are selected for multiplication, he contended, when the antibody they synthesize matches the invading antigen. Talmage also pointed out that a cancerous antibody-producing cell gives rise to a remarkably homogeneous flood of antibodies, indicating that individual cells might "specialize" in producing a particular antibody.

In retrospect it is obvious that Jerne and Talmage had laid the foundations for the clonal-selection theory of antibody formation in their two papers. It remained for Burnet, however, to draw together the new conceptualizations that his musings with Fenner had sparked. Burnet had been struggling to define a mechanism whereby an antigen's contact with a cell would trigger a self-replicating system. He first envisioned antigens as instructors for adaptive enzymes that, having taken the measure of the antigen once, could then proceed with antibody synthesis in the antigen's absence. Later he proposed that antigens interact directly with the genetic material of a cell. These models did not survive, but they did lead him to stress the importance of cellular function and replication in antibody production.

Clonal Selection

For Burnet, Jerne's article supplied the missing link: the notion that the body is endowed with preexisting antibodies to recognize all antigens. Echoing Ehrlich and Talmage, he proposed that the binding of an antigen with an antibody-cum-receptor triggers the cell to multiply and manufacture more of the same receptor. Then Burnet went out on a conceptual limb: he asserted that each cell and its clones, or offspring, can produce just one kind of receptor. He coined the term "clonal selection" to describe his theory.

The clonal-selection theory was attractive to Burnet because it answered several of his earlier criticisms of the template theory. The exponential rise in antibody production following contact with an antigen results from the exponential rise in the number of antibody-producing cells. The secondary reaction to an antigen is more potent and rapid than the first because there are more cells to respond after the initial antigenic stimulation. Once an entire cadre of cells making a particular antibody has been generated, prolonged exposure to the antigen is not necessary to maintain antibody production. The binding ability of the antibodies improves with time because the antigen "selects" for replication cells carrying genetic mutations that promote the match between antibody and antigen. Finally, the clonal-selection theory explained immunological tolerance as the deletion of an entire clone of cells, which could occur before or soon after birth or later, if an antigen overwhelmed the metabolic capabilities of the cells.

Burnet conceived of the immune response as a kind of Darwinian microcosm. The antibody-producing cells, like any organism in an ecosystem, are subject to mutation and selection; the fittest survive—fitness being literally, in this case, the "fit" between a cell's antibody and the antigen. In 1957, after seeing Talmage's review, Burnet submitted his manuscript "The Clonal Selection Theory of Antibody Formation" to an obscure journal. He may have been conscious of the flaws in his earlier proposals; it was not until 1958 that he began writing a book to elaborate on his theory, and it was even later that he published his evolving views in more prominent journals.

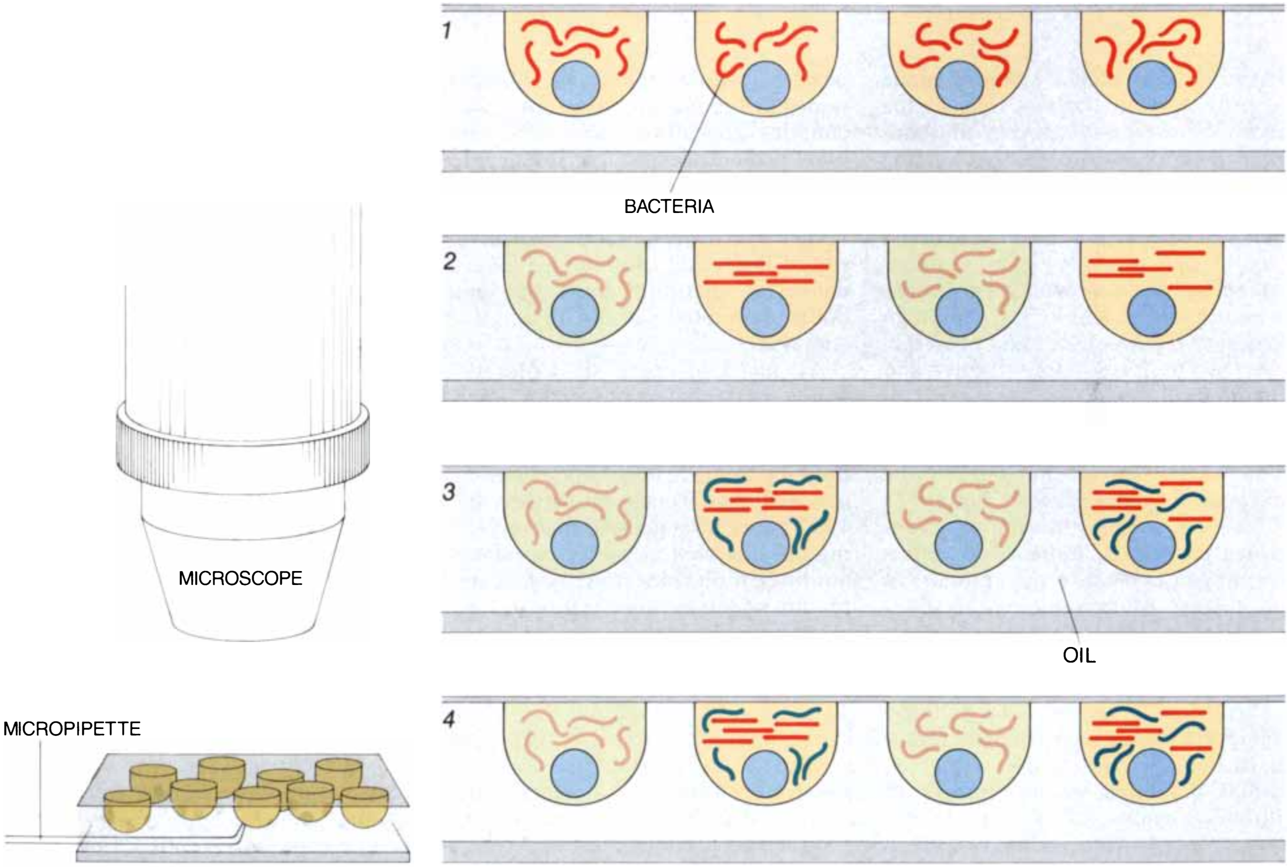
Burnet developed and aggressively promoted the clonal-selection theory over the next decade. The 1960 Nobel prize he and P. B. Medawar shared acknowledged his conceptual accomplishments in understanding acquired immunological tolerance, but Burnet himself believed articulating and promoting the clonal-selection theory was his more significant achievement. In 1984 Jerne would also receive a Nobel prize for his theoretical contributions to immunology, the most fundamental of which was his role in developing the selective theory.

In February, 1957, one of us (Nossal) had started a research fellowship in Burnet's laboratory to examine the problems of immunological tolerance in fetal mice. When Burnet announced the clonal-selection theory later that year, it seemed natural to suggest, in Popperian style, an experiment that would refute it.

One Cell, One Antibody?

Like every student of Burnet's, Nossal was encouraged to read the current literature on viruses. He became intrigued by reports that the cells used for growing animal viruses could be isolated and kept alive in single-cell cultures. The experimental design lent itself to testing of the clonal-selection theory. Central to Burnet's thinking was the tenet that one cell produced just one kind of antibody. If a single antibody-producing cell could be isolated, it might be possible to determine whether that cell was making more than one kind of antibody.

By a fortunate coincidence Joshua Lederberg, then at the University of Wisconsin at Madison, also happened



SINGLE-CELL CULTURES enabled immunologists to test how many different antibodies one cell could produce. Rats were immunized against two motile bacteria (*red, green*); their lymph-node cells were isolated in drops of nutrient medium, where they could be viewed with a microscope (*left*). Bacteria were added to the drops with a micropipette (*1*). If a cell immobilized one kind of bacterium (*2*), the second kind was injected (*3*). Many cells could stop either kind of bacterium; none stopped both kinds (*4*).

to be spending a few months in Burnet's laboratory on a Fulbright fellowship. Lederberg, one of the fathers of bacterial genetics, had considerable experience in capturing individual bacterial cells in tiny droplets of nutrient broth for study under the light microscope. Lederberg volunteered his skills for the experiment; the investigators planned to use bacterial motion as the indicator of antibody production.

It was known that bacteria could be stopped dead in their tracks by antibodies against flagella, the delicate, hairlike structures that propel the microbes. Nossal and Lederberg immunized rats with two different flagellar antigens; a few days later they removed the rats' lymph nodes—structures that, along with the spleen, are a major site of antibody production.

The group teased the tissue apart with fine needles to create a suspension of single cells and introduced the cells one by one into droplets no larger than a millionth of a milliliter. A layer of mineral oil surrounded the droplets to keep them from evaporating. After a brief incubation five to 10 rapidly swimming bacteria with identical flagella were added to each droplet and observed under a microscope. If the bacteria were immobilized, indicating the presence of one type of antibody, the workers inserted bacteria that had the other kind of flagella.

The investigators found that whereas many cells made antibodies against one type of flagella, no cell made antibodies against both types. Thus the one-cell, one-antibody rule was established by an experiment designed to disprove it.

Disarming Cross-Reactions

In 1959 Nossal joined Lederberg at the Stanford University School of Medicine to set up a new laboratory of immunology. There he entered into a two-year collaboration with Olavi Mäkelä of the University of Helsinki that confirmed and extended the conclusions of Burnet's group. The collaboration fired Mäkelä's imagination, and on returning to Helsinki he turned his attention to immunological cross-reaction, a phenomenon in which antibodies raised against one antigen react against other, similar antigens. He decided to use single-cell cultures for his investigations.

Mäkelä provoked antibody formation with bacterial viruses called phages. He quantified a cell's antibody response by measuring how many phages each cell could disarm, or render unable to infect their host bacteria. Before testing antibodies from single cells, he examined the pool of antibodies in the blood serum of an immunized animal. There he found what the textbooks had taught him to expect: animals immunized with an immunogenic phage *A* develop antibodies that neutralize phage *A* very well but that cross-react with phage *B* with an efficiency of only 20 percent.

When Mäkelä looked at the antibodies in single-cell droplets, his results were quite different. The antibodies of each cell had a different degree of specificity for phage *A*. Most neutralized phage *A* better than phage *B*—but not all. Each cell seemed to be making antibodies with slightly different properties. The reactions Mäkelä had observed in the serum simply reflected the cumulative effect of many idiosyncratic cells, each one "doing its own thing." That discovery was exactly what the clonal-selection theory predicted.

The Template Theory Falls

For almost five years after Nossal's experiments the study of antibody formation by single cells was virtually the private preserve of Nossal, Mäkelä and those they had trained. Then, in 1963, Jerne and Albert A. Nordin of the University of Pittsburgh developed a much simpler method for detecting the presence of antibodies secreted by single cells. The spleen and lymph-node cells of mice immunized with sheep red blood cells are spread on a gelatin plate coated with the blood cells. When certain proteins are added, a clear zone appears around each cell producing antibodies. The technique is flexible and is particularly accurate for counting active cells. Its availability led many more laboratories to study the questions raised by Burnet's theory.

While other investigators explored the predictions of the clonal-selection theory, we began our five-year collaboration in 1962 with an experiment we hoped would resolve once and for all the viability of the template theory. We wanted to trace the path of antigen molecules in the body. By this time it was clear that antibodies and all other proteins are manufactured by specialized structures called polyribosomes. If the template theory was correct, each polyribosome would have to be associated with an antigen molecule or at least a substantial fragment of one. Each cell contains thousands of active polyribosomes; it stood to reason that each antibody-producing cell would have to contain many copies of the antigen or its fragments.

The antigens of bacterial flagella served as our tools once again. We attached a radioactive form of iodine to a flagellar antigen so that we could locate the antigen and its "labeled" fragments by looking for radioactivity in cells or tissue samples from the injected animal. Not a single antibody-producing cell contained a detectable quantity of the "hot" antigen, even though our methods were sensitive enough to find 10 antigen molecules per cell. Instead we discovered most of the antigen in a kind of scavenger cell known as a macrophage. This evidence was quite inconsistent with the predictions of the template theory.

In spite of our extensive use of hot antigens in tracking immunologically active cells, it did not occur to us to check for antibodies in unimmunized animals. Yet one of the most controversial claims of the clonal-selection theory was that a small number of natural receptors (antibodies) for all antigens are present prior to immunization. It could have been contended that the number is so small that the receptors can easily escape detection, with the result that some globulins in the blood serum appear to be nonspecific.

Two investigators from the Hebrew University of Jerusalem reasoned that the signal from radioactive antigens bound to the receptors might be strong enough to be detected. David Naor and Dov Sulitzeanu exposed the spleen cells of unimmunized mice to a labeled antigen and looked for evidence that the antigen had bound. They were assuming, in accordance with the clonal-selection theory, that a small number of cells would have receptors reactive with the antigen. Hence they expected to find traces of radioactivity on a very few cells.

That is exactly what Naor and Sulitzeanu found. Moreover, as the clonal-selection theory predicted, immunized animals yielded more antigen-binding cells than unimmunized animals, whereas tolerant animals had fewer. Naor and Sulitzeanu completed the experiment in 1967. Their research led directly to a fundamental question: How could one be sure that the antigen-binding cells were actually the predecessors of the antibody-producing cells?

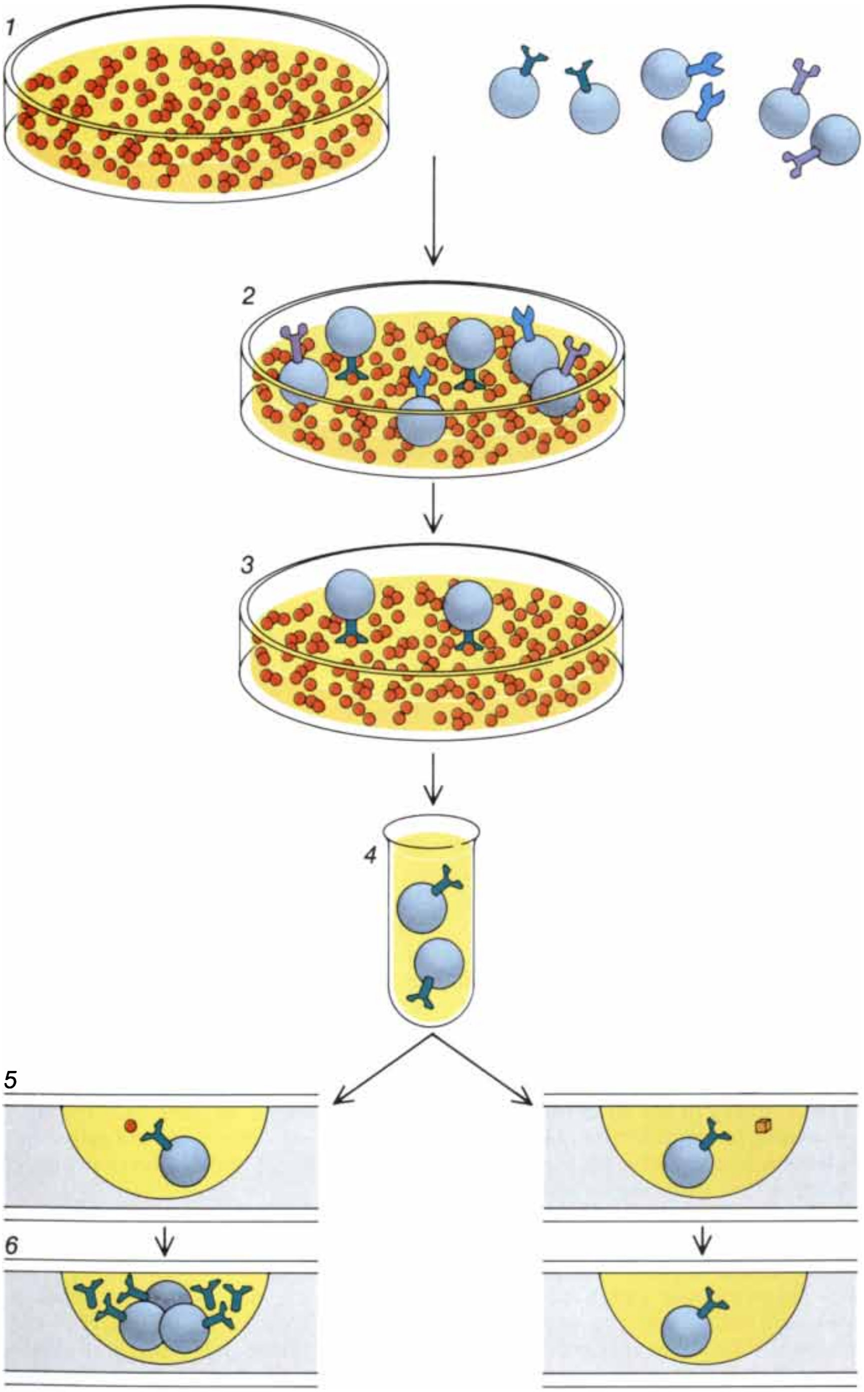
Hot-Antigen Suicide

One of us (Ada), along with Pauline Byrt of the Walter and Eliza Hall Institute of Medical Research in Melbourne, proposed a way to answer that question. If the antibody-producing cells are in fact descended exclusively from the antigen-binding cells, then damaging the ancestral antigen-binding cells by means of a radioactive antigen should prevent the replication of cells producing antibody specific for that antigen. Other cells should not be affected.

In 1969 Byrt and Ada went to work on their so-called hot-antigen-suicide experiments. Two distinct but similar antigens from *Salmonella* bacteria, each one linked with radioactive iodine, were injected into two separate groups of mice. The lymphocytes of those animals were then transferred to two other groups of mice whose own immune systems had been destroyed by X rays. Then both groups of mice were injected with both antigens, this time lacking their radioactive tags. The surrogate immune systems in the irradiated mice secreted few or no antibodies in response to the antigen to which their lymphocytes had been exposed, but they gave the usual robust response to the other antigen. When the antigen-binding cells were crippled, in other words, the immune response to the antigen was crippled as well.

Experiments with hot-antigen suicide gave the clonal-selection theory a significant boost; meanwhile Hans L. R. Wigzell and Birger Andersson of the Karolinska Institute in Stockholm had arrived at similar conclusions using an alternative approach. None of the experiments conducted so far, however, could prove unequivocally that the cell surface bears one and only one kind of receptor. It was clear that the products of an individual cell were less heterogeneous than the serum antibodies, but it was still possible that 10 or 100 specificities, rather than several thousand, were represented on a single cell surface.

In 1973 an ingenious experiment designed by Martin C. Raff, Marc Feldman and Stefanello de Petris of University College London attempted to eliminate this uncertainty. Earlier Raff had discovered a chemical that could cross-link surface receptors on cells. The agent irritates a cell so that all the linked receptors get swept into a tight, caplike section of the cell membrane. A cross-linking substance acting with the specificity of an antigen would cap only those receptors to which it bound. The investigators had such a cross-linking antigen. Would it cap all the receptors on a given cell, or would it leave some behind? Raff and his colleagues applied their cross-linking antigen to white blood cells and found that, on the cells that reacted, more than 95 percent of the receptors had been capped.



ISOLATING ANTIGEN-BINDING CELLS helped to prove that these cells give rise to antibody-producing cells. Millions of unimmunized spleen cells are poured into a dish of gelatin containing identical antigen molecules (*1*). Cells bearing receptors specific for that antigen adhere (*2*); the others are washed off (*3*). After the gelatin has been melted and the cells have been dispersed with enzymes (*4*), each cell is placed in its own culture (*5*). Some cells are exposed to the original antigen (*left*) and others to unrelated antigens (*right*). The cells treated with the original antigen replicate and produce antibodies (*6*).

Proof in the Gelatin

Now only formal proof was necessary to seal the success of the clonal-selection theory. Scientific rigor called for an experiment that followed a homogeneous population of antigen-binding cells from stimulation by an antigen to replication and antibody production. Until early in the 1970's, however, attempts to prepare a population of healthy cells that were all reactive with just one antigen were frustrating and fruitless. Promising techniques were finally reported by four or five groups; Nossal's laboratory was one of them.

That group came up with a simple method by exploiting the melting properties of gelatin. The workers coupled molecules of a given antigen to gelatin in liquid form and allowed the antigen-gelatin blend to set in shallow dishes. Then 100 million ordinary spleen cells from an unimmunized animal were added to each dish. Cells specific for the embedded antigen would adhere to it; any cells that did not adhere could be washed off. In this way cells reactive with a single antigen were isolated on the gelatin.

The investigators then released the antigen-specific cells by melting the gelatin. The antigen was freed from the cells with an enzyme that digests gelatin and pulls the antigen along with it. Once in free solution, each cell was placed in its own culture through techniques developed largely by our colleague Beverley L. Pike of the Walter and Eliza Hall Institute. Each cell was stimulated with the original antigen and a second, irrelevant antigen. As they had expected, Nossal's collaborators found that single cells gave rise to antibody-producing clones in response to the first antigen and not to the incidental one. The tissue-culture fluid surrounding the clones was rich with antibodies to the original antigen. The fluid contained no irrelevant antibodies.

The Scope of Selection

Although these cloning experiments were difficult to carry out when they were first reported in 1976, they have since become much more manageable, and the original results have been duplicated many times. Today the concept of clonal selection is accepted as fact, and the direct descent of antibody-producing cells from antigen-binding cells has been firmly established. The antigen-binding cells are the *B* lymphocytes; their descendants, the antibody-producing clone cells, are known as plasma cells.

We have related this saga as if there had been a simple, direct progression from the conceptualization of the theory to its validation; in truth, the path was much more complex. Many experiments we have not mentioned had tangential but significant impact. The techniques for transferring surrogate immune systems to X-irradiated mice, the unraveling of the structure of the antibody molecule itself and the discovery of the genetic basis for antibody diversity have all contributed to the clonal-selection theory's success.

The last contribution answers a particularly thorny question that dogged the clonal-selection theory: How could lymphocytes anticipate the vast variety of antigens and bear receptors capable of intercepting any one of them? Burnet himself recognized that his theory required some kind of randomization process for which there was no precedent. It is now known that the antigen-binding site of an antibody is the product of no fewer than five different genes, each with several variable regions. The genes recombine as the lymphocyte differentiates, assuming a unique pattern in each cell. That pattern dictates the specificity of the antibodies the cell will produce.

THREE ADVOCATES revitalized the selection theory in the 1950's. Niels Kaj Jerne (*top left*) advanced the notion of an antigen as a selective agent; David W. Talmage (*right*) suggested that antigen binding induces antibody-producing cells to replicate. F. Macfarlane Burnet (*bottom left*) maintained that each cell produces just one kind of antibody.

We also chose not to discuss the second great family of lymphocytes, the *T* cells. These cells defend the body, not by secreting antibodies but by killing infected cells and aiding the process of inflammation. Advances in research on the *T* lymphocytes have been crucial to completing the picture of the immune system's manifold functions.

Indeed, the process of antibody formation is just one of the many mysteries that have been solved. The immune system is now well understood at the cellular, biochemical and genetic levels. There is still much to learn, but the progress to date augurs well for the future. When the clonal-selection theory of antibody formation was first proposed, it seemed improbable, almost unbelievable. The revolutionary concepts hinted at by Ehrlich and formally developed by Jerne, Talmage and Burnet not only stimulated a vast amount of experimental work but also provided insight into wider aspects of cellular organization and function. The resolution of the problem of antibody origin has influenced vaccine development and organ transplantation as well as protein chemistry and molecular biology. In this respect, the contributions of the scientific trio who championed the clonal-selection theory must be measured not just within the framework of immunology but within that of biology as a whole.

Salt Tectonics

When salt is buried under heavier rocks, it rises buoyantly in vast sheets and fingers; it may even fountain aboveground and flow like a glacier. Laboratory models reveal the patterns of salt upwelling

by Christopher J. Talbot and Martin P. A. Jackson

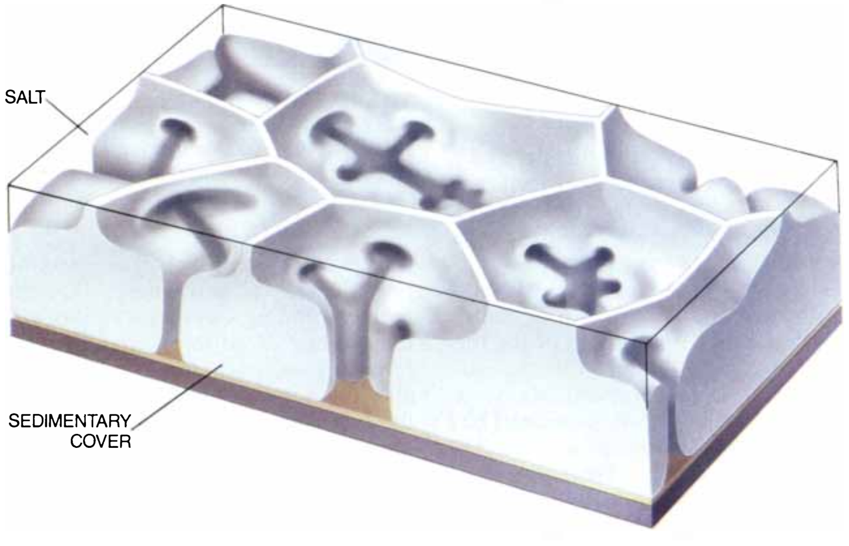
Deep under the floor of the Gulf of Mexico, giant fingers and ramps of salt are rising toward the seabed through 10 kilometers of overlying sediments. Near the shores of the Persian Gulf, salt fingers from an older deposit have already fountained at the surface, forming islands in the gulf and flowing like glaciers down the flanks of the Zagros Mountains. Such spectacular geologic upwellings are made possible by the distinctive properties of salt. The force that drives them, oddly enough, is gravity.

In its natural form salt is a solid crystalline rock, and so one might well ask how it can flow, upward or otherwise. Indeed, when salt is subjected to a sudden shock, such as an earthquake or a hammer blow, it reacts as an elastic solid and either shatters or rebounds. But when it is under a continuous stress, even a quite low one, salt acts as an extremely stiff fluid and deforms without breaking. The process is called creep; it is so slow that it is generally only discernible on a geologic time scale. It is by no means unique to salt: all other rocks can creep too.

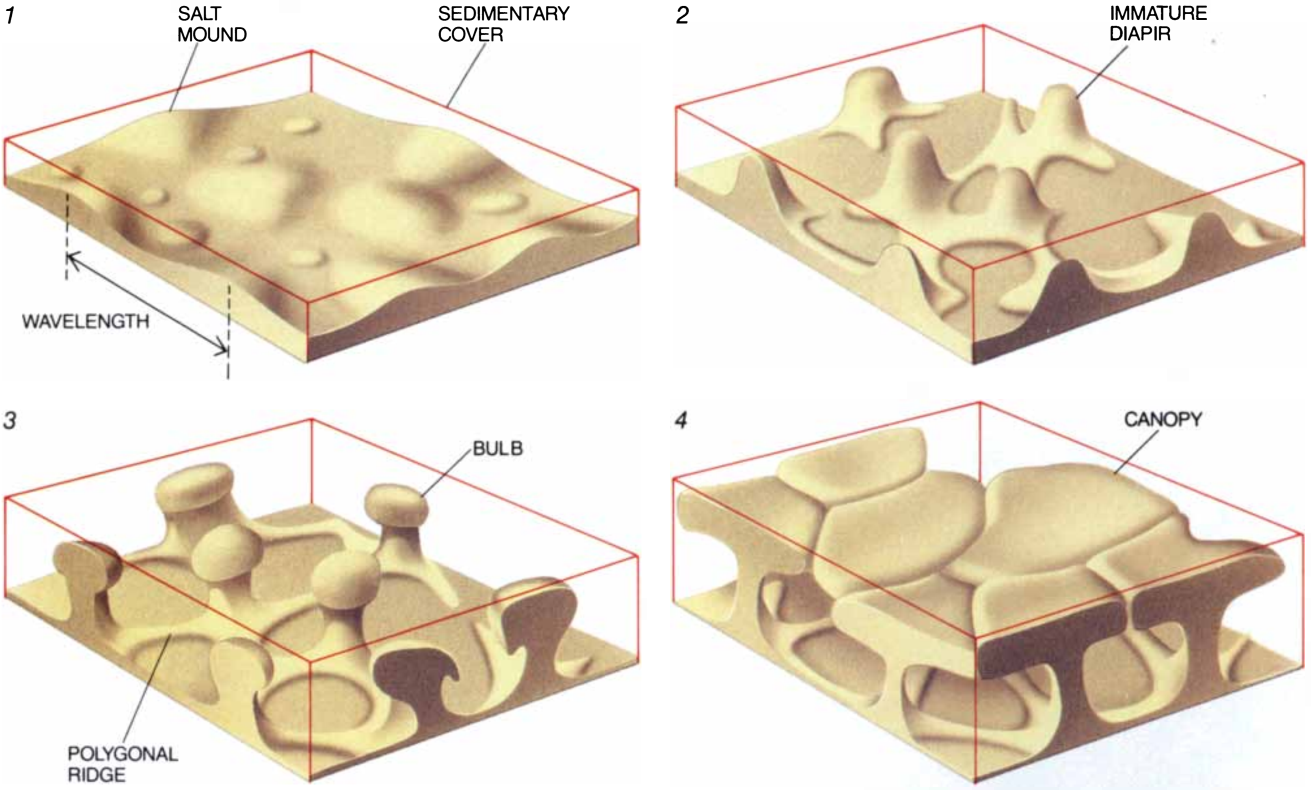
Most rocks, however, creep only under the relatively strong lateral forces exerted by the interactions of the earth's plates—the forces that raise mountain ranges. In contrast, salt flows readily under the influence of gravity alone. Because it is one of the least dense of all rocks, and because it is almost incompressible, it remains light when it is buried under other sediments such as sand, silt and mud. The overlying sediments, on the other hand, are gradually compressed into dense sandstone, siltstone and shale. The result is a density inversion: an arrangement of heavy rock above light rock that, like milk above cream, is gravitationally unstable. The sedimentary cover rock begins to sink through the salt, displacing it upward. The buoyant salt wells up through the cover, forming intrusions called diapirs.

Salt diapirs have economic significance. Their importance in prospecting for oil and gas has been known since the turn of the century. Rising oil and gas become trapped against the flanks of upwelling salt diapirs or in the domes raised by the diapirs in the sedimentary cover; about four-fifths of the oil and gas reserves in the southern U.S. is associated in some way with salt, as are some of the great oil fields of the Middle East. In addition crude oil and natural gas that have already been extracted are often pumped into cathedral-size salt caverns for storage. In the future radioactive wastes may also be stored in salt. All these applications provide a strong incentive to work out the details of how salt structures evolve.

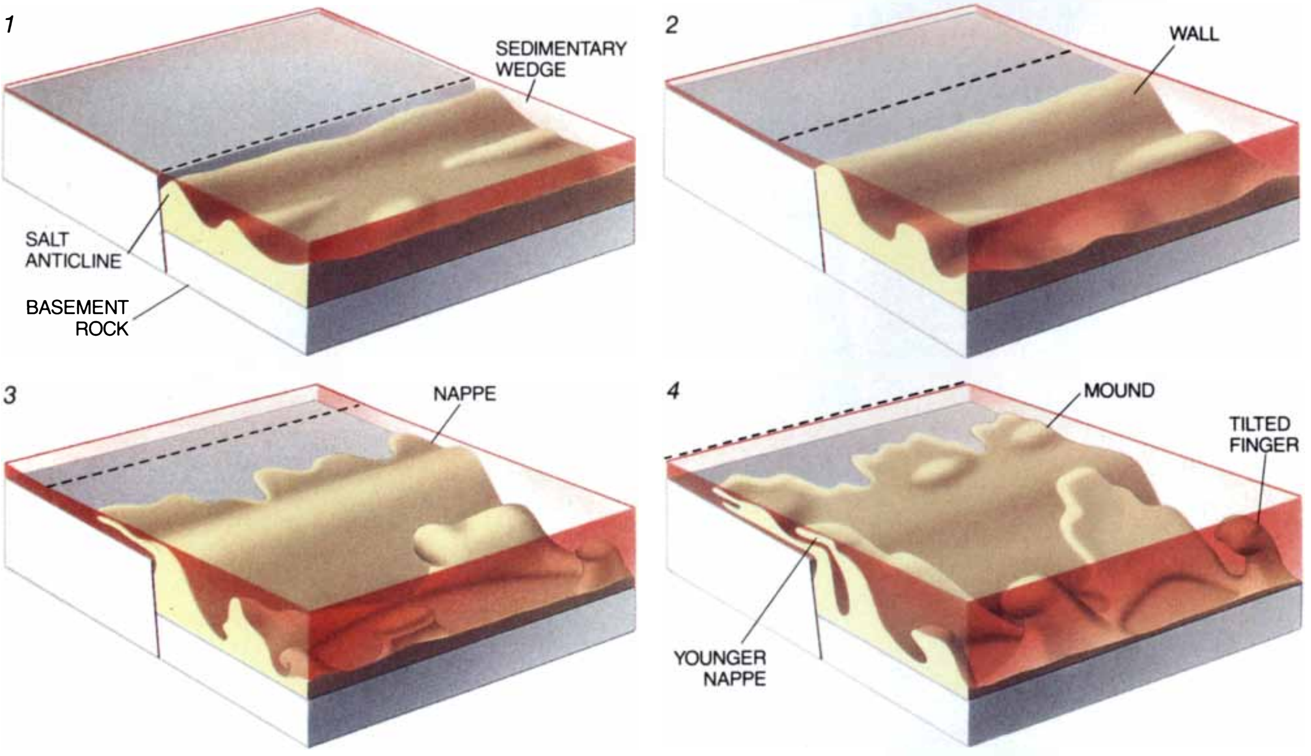
To do so one must know how salt forms in the first place. Salt originates in the subtropics, where dry air that has shed its moisture over the tropical oceans and rain forests sinks and becomes heated, and where as a result evaporation commonly exceeds rainfall. (Most of the world's great



CELLULAR PATTERN of salt upwelling is revealed in an experiment done by Peter Rönnlund of the University of Uppsala. The model consists of two layers: a dense, colored layer of silicone putty that represents the sedimentary cover rocks and a less dense, transparent layer of polydimethylsiloxane that represents the salt deposit. Initially the layers were horizontal, with the colored layer on top. The model was then spun in a centrifuge for several minutes to reproduce (at an accelerated rate) the effects of gravity. The photograph shows the model from above. The colored layer has fallen away from the observer through the transparent "salt," which is now on top. The "salt" has risen in bulb-capped diapirs from the triple junctions of polygonal ridges, which are the dark, shadowed fissures in the photograph; the transparent diapirs are separated by polygonal troughs of sinking material, which appear as thin colored walls. The minute black lines reveal the distortion of the colored cover layer: they started as a millimeter grid on its underside. The drawing shows a side view of part of the model (*lower left in photograph*).



FOUR STAGES in the upwelling of a salt deposit are shown schematically. Here the sedimentary cover, which is being deposited as the upwelling progresses, is transparent. Initially the interface between the two layers is nearly planar. In the first stage some of the small irregularities on the salt surface grow into mounds that have a regular spacing (the wavelength) (*1*). The mounds contract into finger-shaped diapirs linked by polygonal salt ridges (*2*). At shallow depth the rising diapirs spread into flattened bulbs under a layer of uncompacted, less dense sediment (*3*). The bulbs may eventually merge into a broad canopy resting on thin stems (*4*).



EDGE EFFECTS modify the pattern of salt upwelling. The illustration shows the combined effects of two natural edges: a step in the underlying basement rock and the advancing wedge of sediments of a river delta. (The dotted line is the thin end of the wedge.) The salt begins to rise in an anticline that parallels the step (*1*); eventually it contracts into a wall (*2*). The advancing sedimentary wedge acts like a rolling pin, folding the wall over the step and forming a laterally growing salt nappe (*3*). Mounds may rise on the nappe, and smaller nappes may spread over it (*4*). Salt fingers growing at a distance tilt and bulge toward the step.

deserts are in the subtropics.) In enclosed sedimentary basins, or in seas protected from the open ocean by reefs or other natural barriers, seawater evaporates, leaving behind a saturated brine. Minerals precipitate out of the brine according to their solubility: first carbonates, then gypsum (a calcium sulfate), then halite (sodium chloride) and finally the bitter dregs of magnesium and potassium salts. The minerals form rocks called evaporites. By far the most abundant evaporite mineral is halite, or rock salt, hereinafter known simply as salt. It settles out of brine in tightly interlocking, coarse-grained crystals.

If the basin is cyclically replenished with normal seawater, layers of evaporites as much as several kilometers thick can accumulate on the sea floor. A peak of evaporite accumulation occurred about 230 million years ago, when the supercontinent Pangaea was just beginning to rift apart. The rift valleys and below-sea-level basins in and around Pangaea were ideal sites for evaporite formation. During that period some 10 million cubic kilometers of salt, extracted from about 660 million cubic kilometers of seawater, were deposited in kilometer-thick layers. Some of the Pangaean rifts widened into oceans; consequently salt layers are now found on some of the existing continental margins, buried under piles of younger sediments.

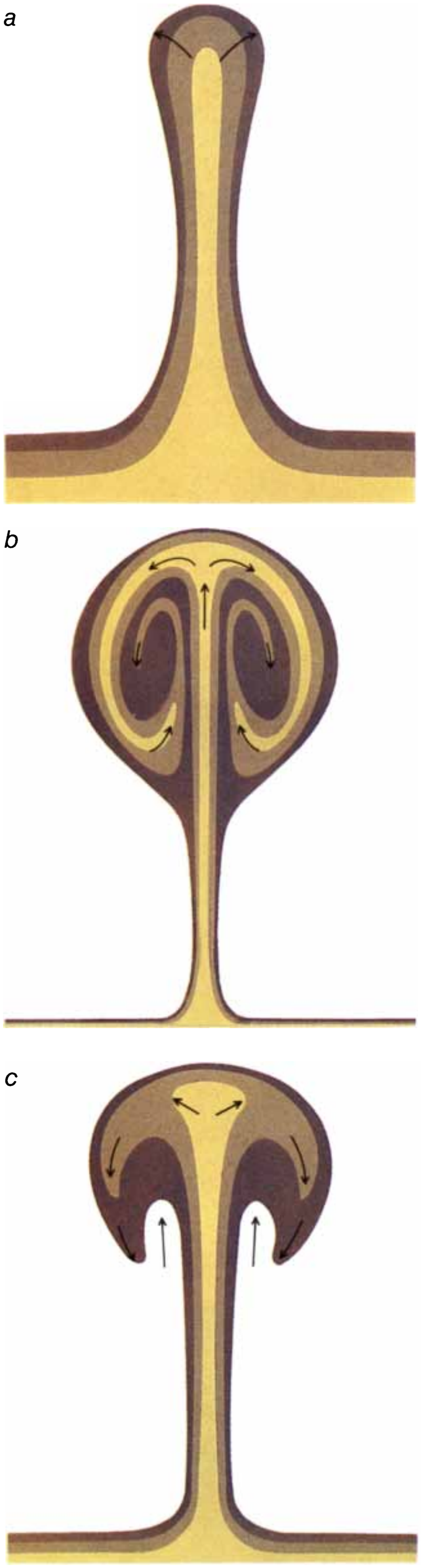
Eventually buried salt escapes to the surface and is recycled into the oceans by erosion. Only a few salt deposits have been found that are more than about 800 million years old, indicating that most of the salt formed before then has already been recycled. Yet the process is exceedingly slow. The salt now flowing down the Zagros Mountains is between 500 and 800 million years old, and the diapirs that have yet to reach the floor of the Gulf of Mexico come from a deposit laid down about 175 million years ago. In general salt diapirs rise episodically: periods of growth, during which the diapirs rise at a rate of between .1 and one millimeter per year for several million years (roughly 100 times slower than the lateral speed of the earth's crustal plates), alternate with periods of dormancy.

The slow pace presents investigators with special problems. Although salt can be made to creep faster in the laboratory by squeezing tiny samples of it at stresses much higher than those it is subjected to in nature, such experiments cannot reveal either the large-scale patterns or the mechanisms of natural salt flow. Nevertheless, there are several ways of studying the upwelling process. The first method is to find an analytical solution for the fluid-dynamical equations that describe a generalized system consisting of a dense fluid (in this case the sedimentary cover) overlying a less dense fluid (the salt source layer). The second method is also mathematical; it involves programming a computer to solve the equations numerically. Finally—and this is the approach we have followed—one can simulate the upwelling of salt with an appropriately scaled physical model.

The upwelling process can be conveniently divided into four stages. The first stage is well described by the analytical theory devised by the English physicists Lord Rayleigh and Sir Geoffrey Taylor. According to Rayleigh-Taylor theory, the upwelling begins at subtle bumps on the interface between the two viscous fluids. At first the bumps grow at different speeds. Only the fastest-growing bumps, however, survive; they draw salt away from the slower bumps, which are thereby suppressed. Little by little the near-planar interface is transformed into a field of sinusoidal, regularly spaced mounds formed by the fastest-growing bumps. The distance between the mounds (the wavelength) is characteristic of the system; it depends primarily on the relative thickness and viscosity of the two fluid layers. Their relative density determines the speed of the upwelling. Natural salt mounds take about 20 million years to form.

During the first stage of upwelling the spaces between the mounds are filled by troughs of sinking cover rock that are in effect inverse images of the mounds. In the second stage the shapes of the mounds and troughs begin to differ. As the mounds exceed a certain height, which can range from about .5 kilometer to 2.5 kilometers, they contract to narrow, rising fingers or walls, while the troughs expand into broad, sinking basins. Eventually the salt actually penetrates the cover rock (hence the name diapir, from the Greek word *diapeirein,* meaning "to pierce"). At this stage nonlinear terms enter the equations of motion that describe the upwelling, and the equations are no longer susceptible to precise analytical solution. To understand the further evolution of the salt diapirs one must resort either to physical models or to computer models.

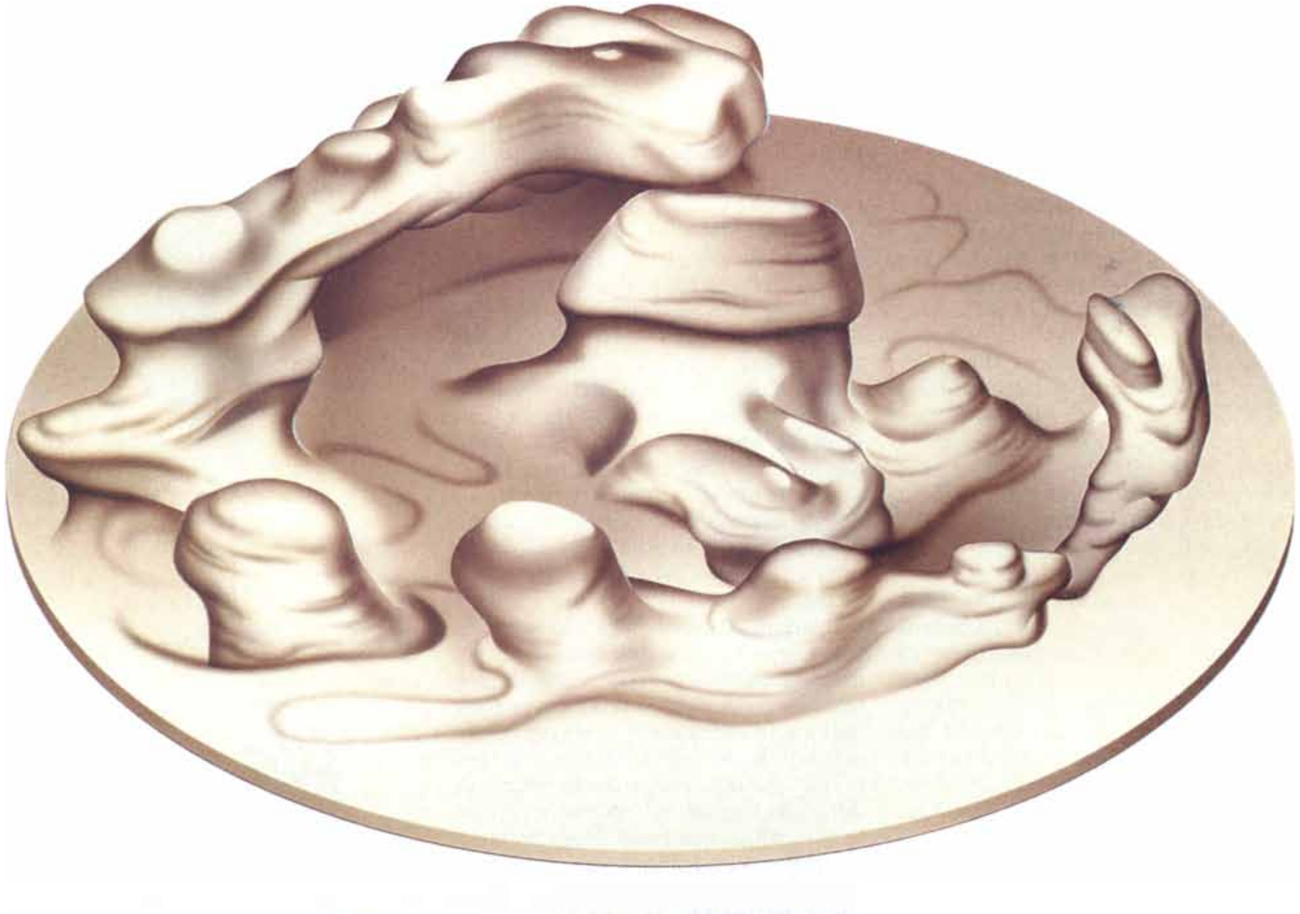
In the earliest physical models a gravitationally unstable system of fluids was produced by putting buoyant oil on top of a heavy syrup in a tank and then turning the tank upside down. Because of the low viscosities of these fluids, upwelling diapirs formed within seconds and were easy to observe.

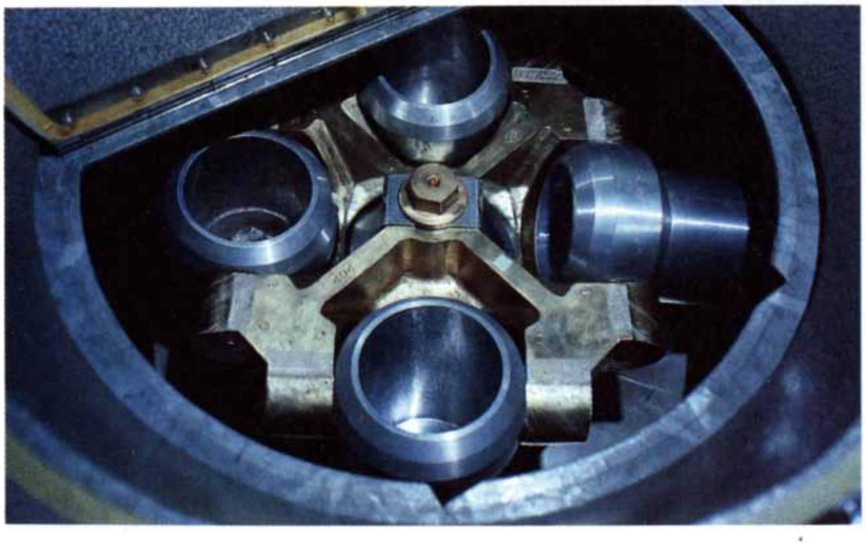


THREE TYPES OF BULB can be formed by the circulation associated with salt diapirs. If the sedimentary cover is softer (less viscous) than the salt, the diapir bulb is thumb-shaped (*a*); if the cover is stiffer, the bulb is balloon-shaped (*b*), and if the two layers are equally viscous, a mushroom-shaped bulb results (*c*). All three types of bulb can develop at depth, before the diapir encounters an upper boundary.

On the other hand, it was impossible to construct realistic, intricately layered models of sedimentary rock with oil and syrup. In 1960 Hans Ramberg of the University of Uppsala found a way around the problem. His idea was to build models out of stiffer materials, such as modeling clay and silicone putty, and then spin them in a centrifuge. The centrifugal force mimics the effect of gravity, except that it is much more intense; the less dense material, initially at the bottom of the model, "rises" inward toward the axis of the centrifuge.

We have followed Ramberg's approach in our studies of salt upwelling. The viscosity and density of the model material must be such that when a small layer of it is subjected to a force between 1,000 and 2,000 times stronger than gravity for a few minutes, it accurately mimics the deformations undergone by a large salt deposit under normal gravity over millions of years. We generally use silicone putty or polydimethylsiloxane for the salt and a mixture of silicone putty and dense barium sulfate for the sedimentary cover. After spinning the model we remove it from the centrifuge and cut about 30 cross sections through it. From the cross sections and from marker grids implanted in the model we can piece together the three-dimensional flow pattern.

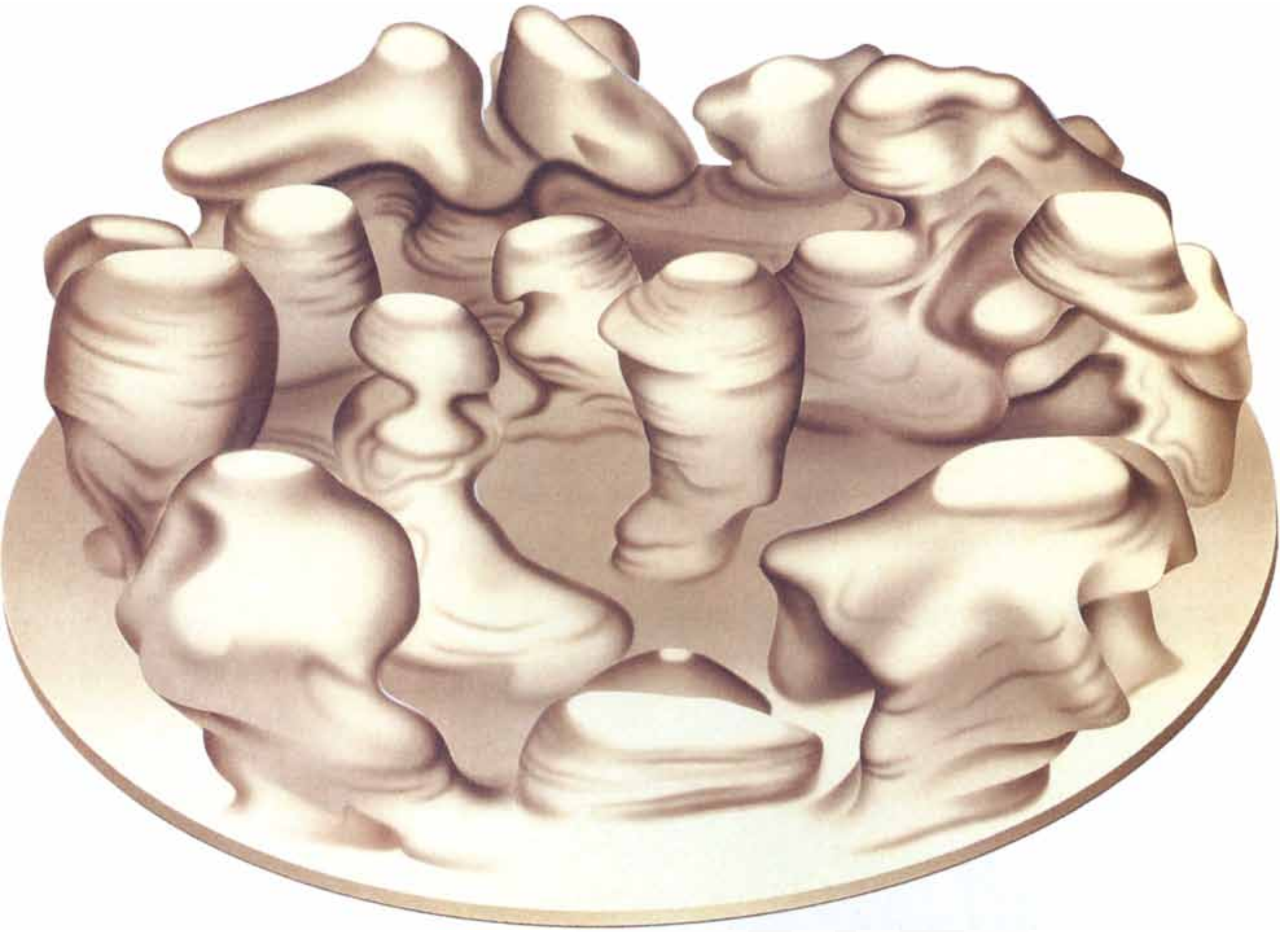


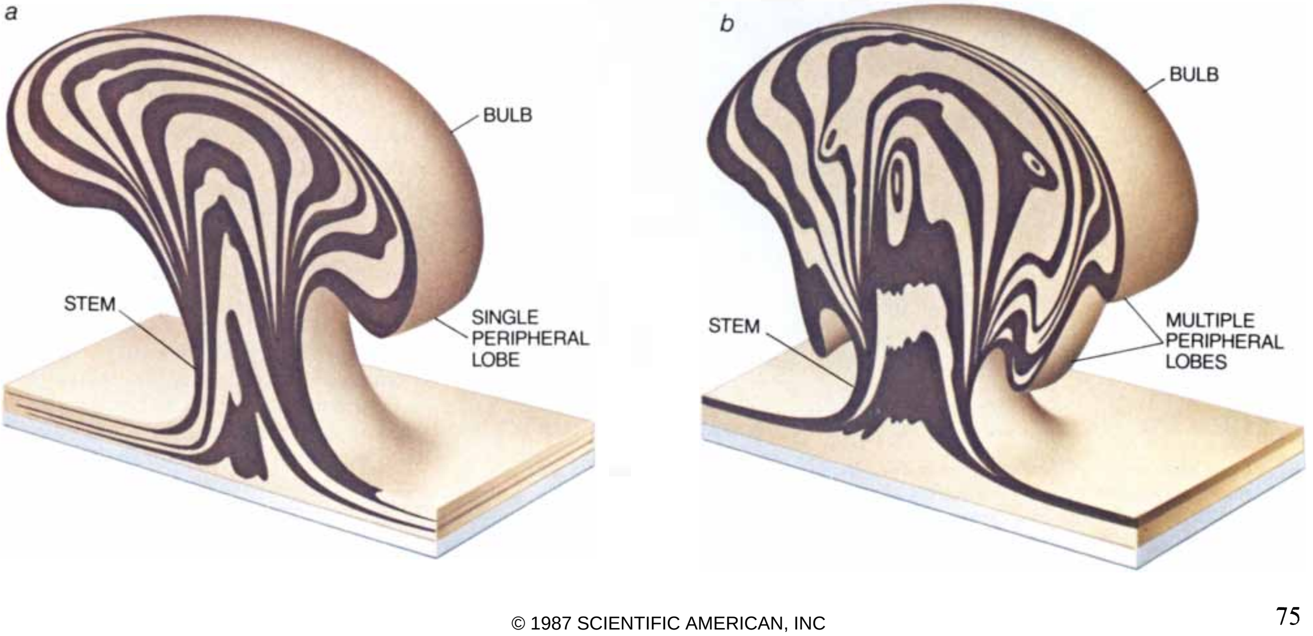


CENTRIFUGE MODELS reveal the distribution, shape and internal structure of salt diapirs. Layers of putty are stacked in a small (10 centimeters in diameter) cup and spun in the centrifuge; the spinning cup swings outward (*see photograph*). The model is then removed from the cup and cut into about 30 vertical or horizontal sections, from which the three-dimensional flow pattern can be reconstructed. If all the cover material (not shown here) is added before the model is spun, relatively few diapirs form (*above left*), and the mushroom diapirs have a single peripheral lobe (*a*). If the cover layers are added in stages to simulate the episodic accumulation of sediment, the diapirs are crowded together (*above right*), and the mushrooms have multiple lobes (*b*). The stripes in the diapir sections are marker layers that were initially flat and equally thick. The walls at the rims of the models are edge effects.

We have found that the geometry of salt upwelling is quite complex; it is broadly similar to the spoke pattern characteristic of vigorous thermal convection. (Thermal upwelling is also driven by a density inversion, but the inversion is caused by a temperature gradient in a single fluid heated from below rather than by a difference in the composition of two fluids.) The flow is partitioned into numerous unit cells, each one composed of a single diapir and the sinking cover around it. Each diapir rises from the junction of either three or five deep salt ridges (the spokes), which are typically linked to form irregular pentagons or hexagons. The sinking troughs in turn form polygons centered on the crests of the rising diapirs; that is, the troughs are laterally offset from the ridges by half a polygon. The polygonal spoke pattern is not merely an artifact of our models: seismic studies have revealed polygonal ridges under the northwestern Gulf of Mexico. Each polygon is some 20 to 30 kilometers across.

Near the edges of our models the spoke pattern is disrupted. Instead of rising in fingers, the salt rises in long walls that parallel the boundaries of the model. Parallel salt walls in northwestern Europe demonstrate that linear faults, folds or slopes in the basement rock underlying a salt layer, and even linear loading patterns caused





by the encroaching sediment of a river, can result in natural edge effects. Away from the edges we believe the spoke pattern predominates in nature as it does in the models.

In nature the pattern is usually modified, however, by lateral variations in the thickness or mechanical properties of the salt and cover layers that are too gradual to be classified as edge effects. As a result salt diapirs are often not symmetrical about their vertical axis; they are inclined, with one side growing faster and larger than the other. The asymmetrical growth can be enhanced by a positive-feedback effect. On the fast-growing side of a diapir more salt is flowing in from the underlying deposit. Hence the sedimentary cover sinks rapidly there, and a moat forms at the surface. The moat collects more sediments, which increases the load on the salt deposit, forces more salt into that side of the diapir and pushes it upward even faster.

The shape of the growing diapirs can be even more irregular if the salt deposit initially has large irregularities—steps formed by faulting, say, or broad swells—on its upper surface. Computer models by Harro Schmeling of Uppsala have shown that salt tends to well up above such irregularities, even if the spacing between them is much greater or smaller than the characteristic diapir wavelength predicted by Rayleigh-Taylor theory on the basis of the thickness and viscosity of the salt and cover layers. If the spacing of the irregularities is small, diapirs continue to form at intervals of one wavelength; salt rising above the irregularities, however, is incorporated into the diapirs, which can thereby assume bizarre shapes.

The most pronounced shape change undergone by a rising diapir comes in the third stage of its evolution, when the top swells into a bulb overhanging a slender stem. Bulbs can form in two ways. The first way is for the diapir to encounter an upper boundary, typically a shallow layer of sediment that has not yet been compressed to a density greater than that of salt. When a diapir rises into a low-density layer, the top of the diapir spreads out, forming a bulb that usually has the shape of a squat, flattened balloon.

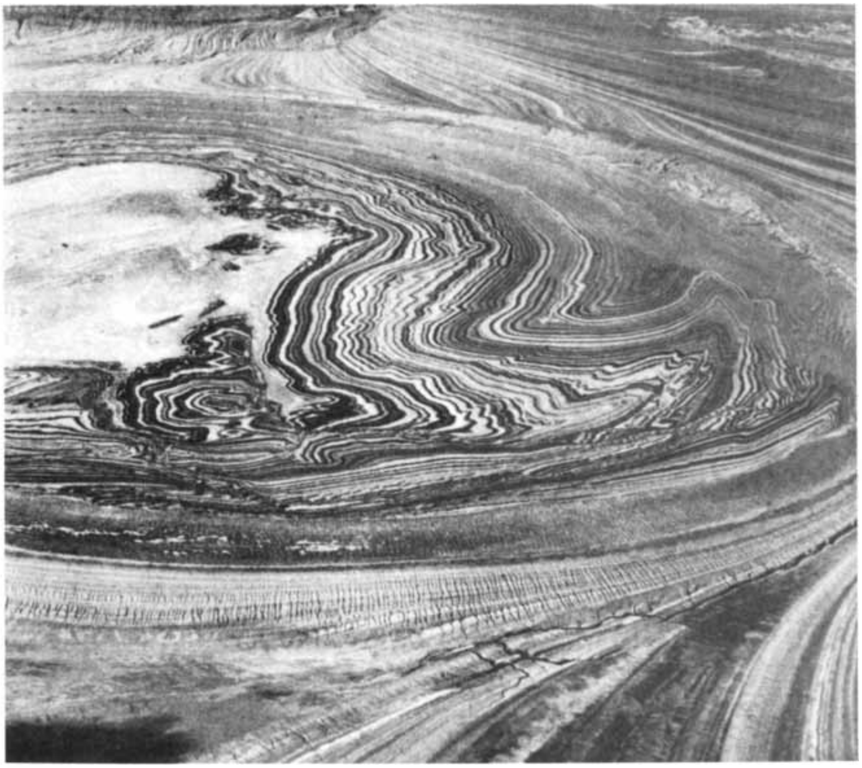
If the sedimentary cover is sufficiently thick, salt bulbs can also form at depth. Such bulbs do not result from spreading under an upper boundary. Rather, they result from the interaction of the rising diapir core with the sinking cover, which exerts a downward drag against the sides of the diapir. The interaction sets up an internal circulation in the diapir.

The nature of the circulation and the shape of the resulting bulb depend mainly on the relative viscosities of the salt and the cover. Three basic shapes are possible [*see illustration on page 73*]. If the cover is softer (less viscous) than the salt, the diapir pushes through it easily, and only a weak internal circulation is established; the diapir becomes thumb-shaped, with a slim bulb. Conversely, if the cover is stiffer and relatively undeformable, the rising salt is deflected into a strong toroidal flow that sustains a balloon-shaped bulb. Finally, if the cover and the salt are of roughly equal viscosity, a toroidal circulation is also established, but in this case both layers are involved. The sinking cover drags the periphery of the diapir into a pendant lobe that rises less vigorously than the core, and the diapir assumes a mushroom shape.

Evidence of the internal circulation in mushroom diapirs has been found in salt mines in the U.S. and Europe. Maps and vertical sections show that the sequence of salt layers in the original deposits has been inverted and everted in the diapirs by the toroidal internal flow. The vertical sections also suggest that the peripheral lobes of the mushrooms may have entrained cover rocks, folding them deep into the diapirs. Our own models indicate that entrainment is likely to take place when many diapirs with highly irregular shapes are crowded together. In that situation neighboring diapirs interact with the same region of sinking cover material, and the downward drag on the diapirs is intensified.

Crowding, in turn, is likely to occur when the sedimentary cover has been deposited episodically. The reasons are straightforward. According to Rayleigh-Taylor theory, the spacing between diapirs for a given salt deposit is proportional (other factors being equal) to the thickness of the cover. If the cover is deposited in pulses, however, salt diapirs will have a chance to rise and mature between early episodes of sedimentation, while the cover is still relatively thin. The resulting field of diapirs will be much more crowded than it would have been if the cover had been deposited in a continuous sequence. If the cover is about as viscous as salt, the diapirs will be mushroom-shaped and will probably entrain cover rocks under their lobes.

That conclusion has practical significance. One of the reasons salt deposits have been proposed as storage sites for radioactive waste is that they are relatively impermeable. Fractures in the deposit are closed by creep, and although salt is highly soluble, ground-



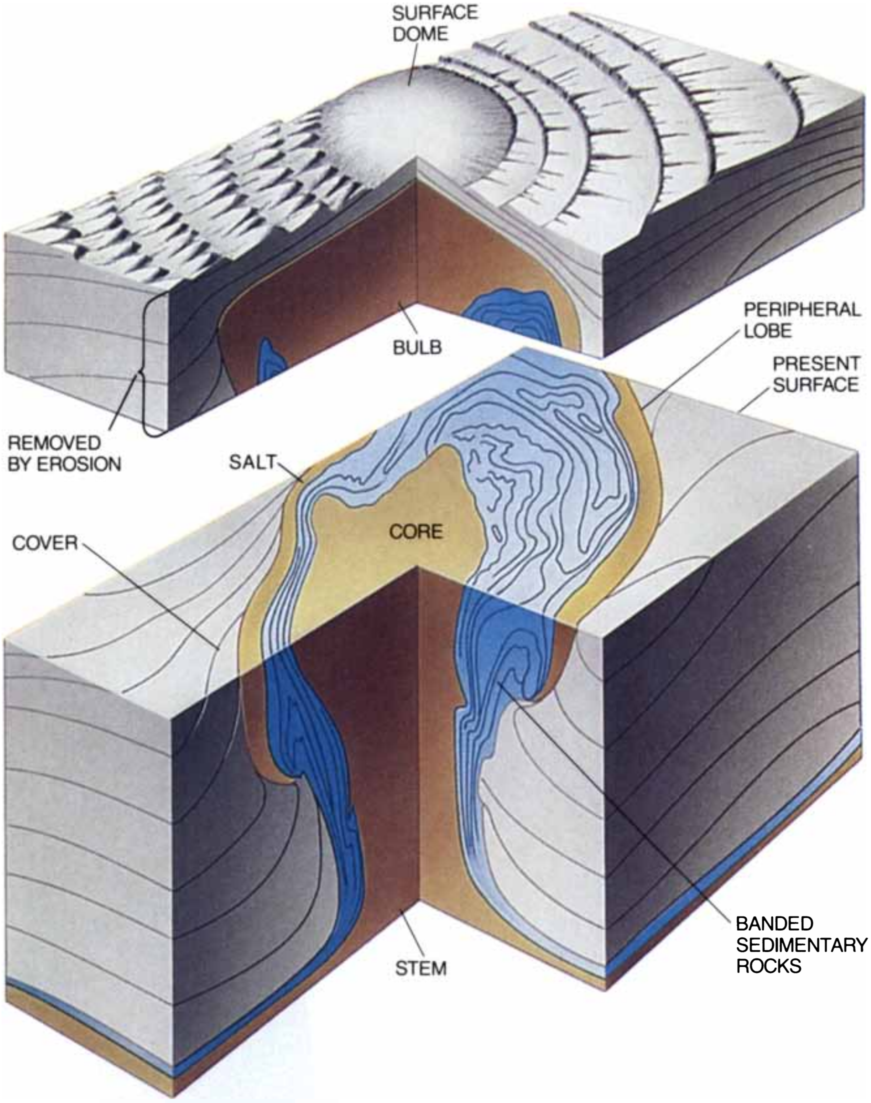
GREAT KAVIR desert in central Iran is the site of spectacular salt diapirs. The aerial photograph by Augusto Gansser shows the truncated top of a six-kilometer-wide, nearly circular diapir. A pale core of gypsum-encrusted salt is ringed by finely banded, younger evaporites: pale salt and gypsum alternating with darker shales and marls. The coarsely banded cover sediments in the foreground were dragged up by the diapir at its rim.

water cannot penetrate far into a deposit without becoming saturated and therefore incapable of dissolving more salt. If a waste-storage cavern were built in a mushroom diapir, however, entrained layers of permeable sedimentary rock could act as conduits, channeling water through the cavern. The water could potentially carry radionuclides into the environment. It is clear that the shape and internal structure of a salt diapir must be carefully considered before it is chosen as a waste-storage site.

The fourth and final stage of salt upwelling begins when the rising, bulbous diapirs approach the surface. At that point one of three things happens. First, the diapirs may be decapitated by groundwater. Near-surface strata are generally porous, and unsaturated groundwater within them can dissolve the top of a diapir as fast as it rises. In such cases the dissolution surface is capped by a residue of relatively insoluble gypsum that had been dispersed in the rising salt.

A second possibility exists in deserts, where dissolution of the salt by groundwater is obviously less likely: the bulbs can continue to spread horizontally under the barrier provided by low-density surface strata. In the Great Kavir desert of central Iran the bulbs of 12 diapirs have coalesced into a single canopy 40 kilometers across. The canopy has been exposed in extraordinary detail on the desert floor by erosion. It represents an unusually advanced stage of salt upwelling, in which nearly all of a 50-million-year-old salt deposit has changed places with the denser sediments that originally overlay it. Although viscous liquids overturn completely when they are arranged in a gravitationally unstable way, liquids are different from rocks in that they have no yield strength and therefore flow under any applied force. It appears that most salt diapirs "seize up" before merging into a canopy, because the buoyancy forces become too weak. Certain salt layers now assumed to be young and undeformed, however, may yet be recognized as older, diapiric canopies.

Where the shallow strata are denser than salt, diapirs can meet a third destiny, rising all the way to the surface and extruding in majestically slow fountains. The best examples of ongoing salt extrusions are the Zagros diapirs of southern Iran. For the past 15 million years the upwelling there has been accelerated by plate tectonics: the collision of the Arabian and Eurasian plates, which began about 15 million years ago and is squeezing salt upward. Some 20 diapirs have been squeezed above the surface of the Persian Gulf and now form small salt islands. In the Zagros Mountains craters clogged with insoluble rock debris bear witness to diapirs that could not compete with the rain. Along the coast, where the plate collision is currently at its most intense, fold mountains are pierced by fountains of salt more than a kilometer high. The salt is spreading under its own weight and flowing down the flanks of the mountains; the flows are called namakiers (from *namak,* the Farsi word for salt, and *glacier*).



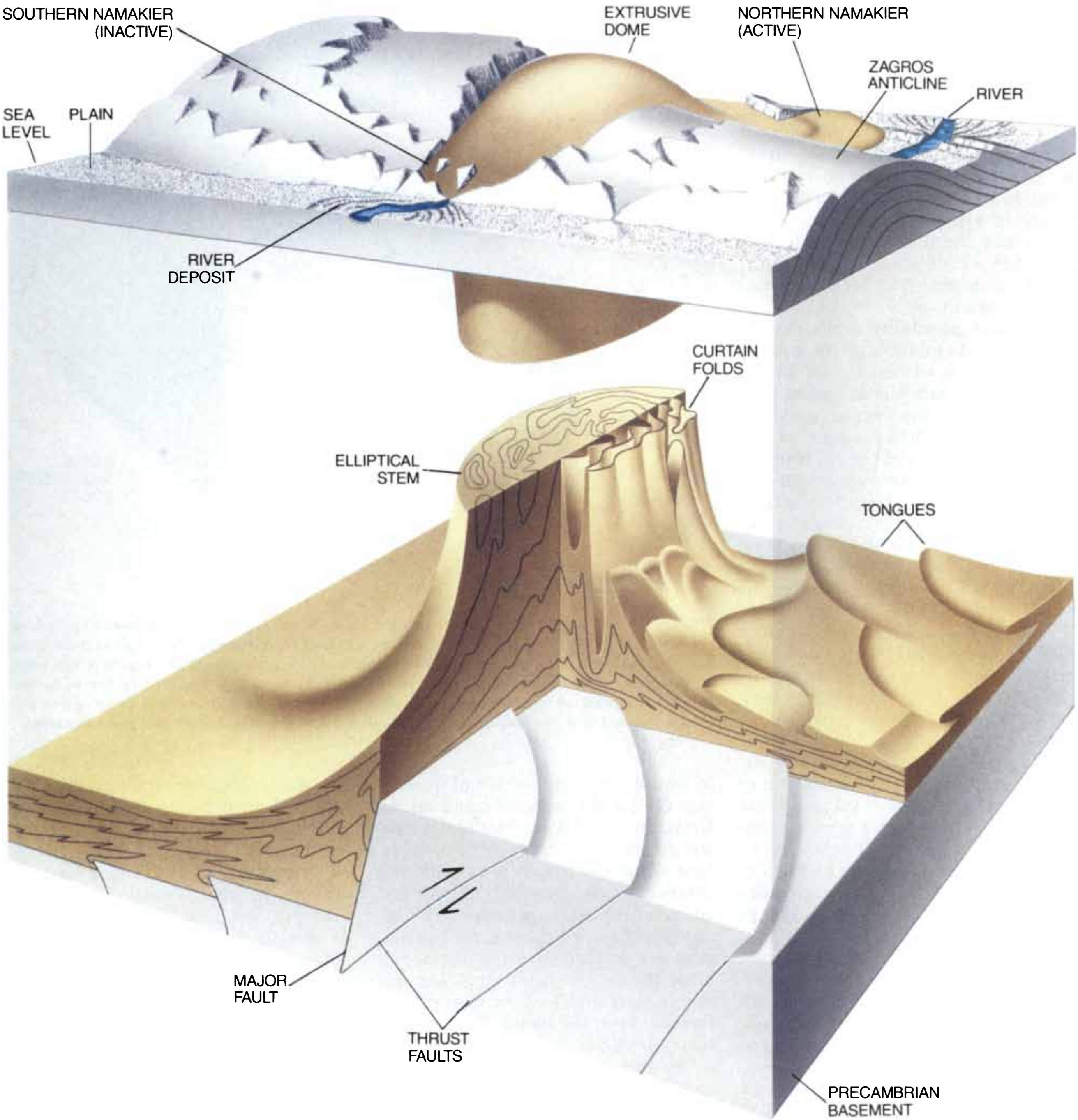
INTERNAL STRUCTURE of a mushroom diapir in the Great Kavir has been exposed by erosion. By extrapolating from surface features and from their models, the authors have deduced both the deep structure of the diapir and the structure of the top that has been eroded away. The sequence of layers in the original deposit is inverted in the bulb: the salt overlies banded sedimentary rocks that are intermediate in age and composition between the salt and the younger cover. Banded rocks form a ring at the present surface.

The average speed of a namakier, a few meters per year, is less than that of most glaciers, but it is nonetheless remarkably fast for crystalline (nonmolten) rock. What accounts for the rapid flow? Some workers have thought, on the basis of experiments with dry salt, that the salt in a namakier is extruded red-hot. The hypothesis now seems implausible. Its proponents neglected a long-known fact about salt: increasing its water content reduces its viscosity and resistance to flow just as effectively as increasing its temperature does. Our measurements of survey markers on one namakier have shown that it is stationary most of the year; after it has been dampened by seasonal rain showers, however, it can flow at a rate of half a meter per day.

By deforming evaporites under a microscope, Janos L. Urai and Christopher J. Spiers of the University of Utrecht have observed the effect water has on salt flow. Whether the water comes from the original salt-forming brine, from surrounding sediments or from rainfall, it tends to form thin, continuous films along grain boundaries when the salt is deformed. The films dramatically weaken the salt. Specifically, they sweep through and dissolve old, distorted grains of salt, whose dislocated internal structure resists further deformation; sodium and chloride ions diffuse across the films and form new grains that are readily deformable. A water content as low as .1 percent by weight is enough to promote this "dynamic recrystallization," which suggests that the process may take place not only in namakiers but also at depth, in buried salt deposits.

The salt flowing downhill in a namakier has undergone a long and complex history of folding and other deformations. Clues to this history are recorded in the colored layering created by deposition of varying concentrations of other minerals with the halite. The latest direction of flow is revealed by the alignment of the mineral grains. In the namakier we have studied closely, the grain alignment shows that the direction of flow generally parallels both the internal layering and the walls of the rigid channel that confine the namakier.

On its way down the mountain, however, the salt passes over a series of steps in the floor of the channel. Before each step the flow slows and thickens, the flow lines diverge and



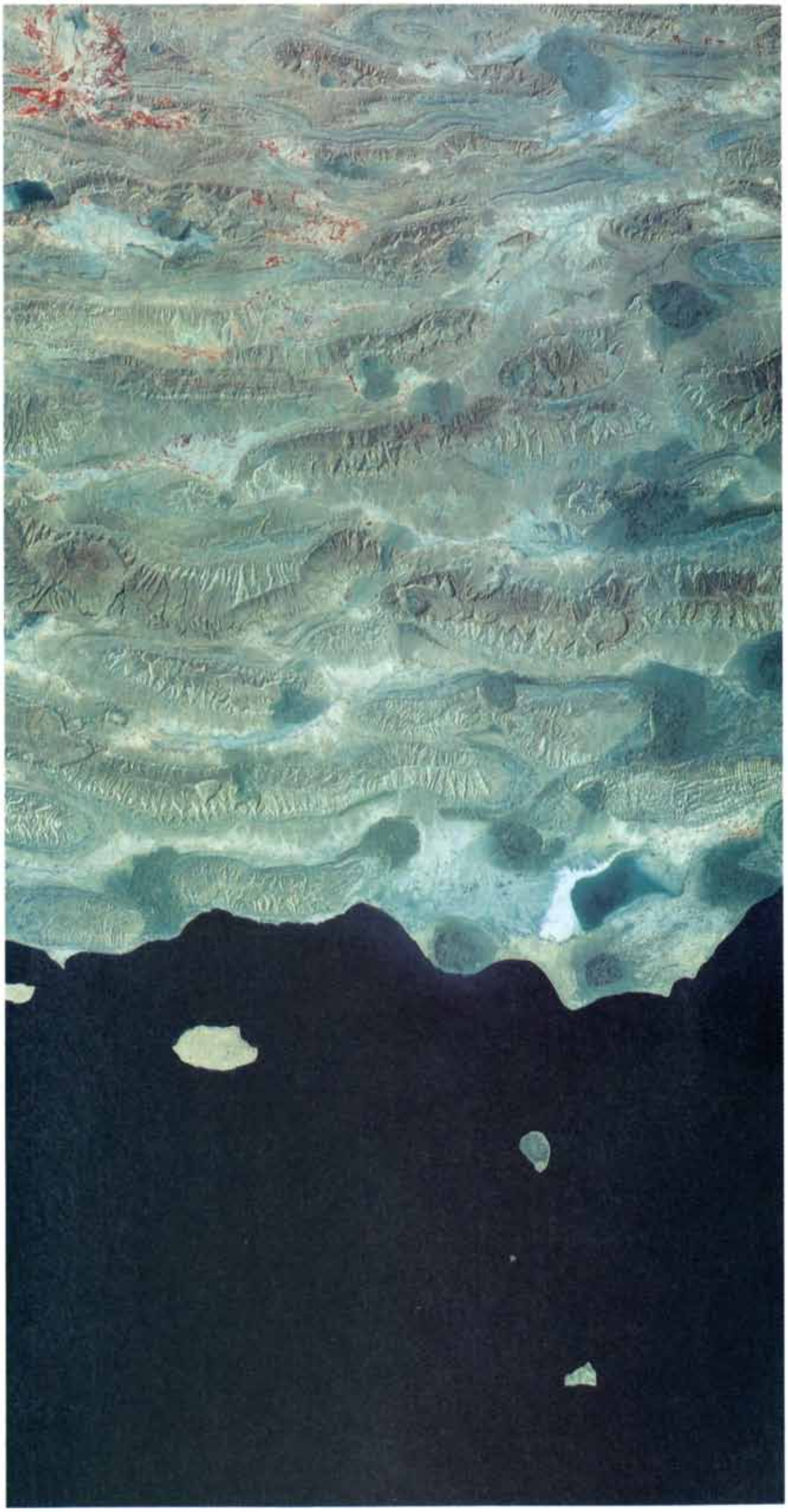
NAMAKIERS in the Zagros Mountains of southern Iran were formed by a fountain of salt that has risen through seven kilometers of cover rocks. The extruded salt now flows down the flanks of Kuh-e-Namak (28.2 degrees north, 51.7 east) at a few meters per year, overriding recent river deposits in the process. The diapir that feeds the namakiers rises above a major fault in the basement, which has created a step in the upper surface of the salt deposit. Salt flowing into the diapir passes over a series of thrust faults and is probably folded into overlapping tongues; the tongues are rotated to the vertical and refolded like curtains in the diapir.

the colored layers are deflected into a series of folds. On the other side of the step the flow accelerates again and the flow lines converge. The layers are thinned, the internal folds are stretched and sheets of salt begin to slide over one another. At the next step new folds are superposed on the old ones. Hence the namakier consists of a stack of overlapping tongues. At the tip of the namakier the stack is so flat that the folds are hardly visible; the tongues look almost like undeformed horizontal layers.

Similar effects probably take place below the surface, in the deep salt deposit that feeds the diapirs. Wherever salt migrating toward a diapir passes irregularities in the roof or floor of the deposit, flat-lying internal tongues may be generated. Tongues from a broad area of the deposit are carried into the narrow stem of the diapir, where each one is rotated to the vertical and constricted into curtainlike folds. (The effect is similar to the one created by pushing a napkin through a ring.) The folding and refolding of many generations of tongues accounts for the wide variety of intricate and beautiful patterns exposed in the roofs of salt mines.

A purely aesthetic appreciation of salt formations is hard to avoid when one has studied them for a while, and understanding the strange anatomy of these almost organic structures has its own intrinsic rewards. Yet as we have stressed above, such studies also have economic value. The possibility that some salt diapirs are shaped like mushrooms is highly significant in the search for oil and gas: it suggests that many diapirs may create considerably larger petroleum traps than had been thought. Just as the lobes of a mushroom diapir might entrain permeable rocks that would compromise the seal of a storage cavern, so might they also entrain or hide reservoir rocks laden with petroleum.

The fact that mushroom diapirs were discovered in the laboratory illustrates the importance of modeling. Mining and drilling can reveal only a limited amount of information on underground salt structures, and in general even those field data that are available have been underused, simply because investigators did not know how to interpret them. That is now changing, as physical and computer models are uncovering the fundamental processes that drive the upwelling of salt. In the future both petroleum geologists and designers of storage caverns will benefit from a much more solid understanding of the geologic structures they are exploiting.



ZAGROS COAST of the Persian Gulf is dotted with the crests of salt diapirs. The photograph is a composite of two images made with data from the Landsat thematic mapper; the region shown is about 100 kilometers west of the Strait of Hormuz and is about 150 kilometers wide. The folds parallel to the coastline are the Zagros Mountains. The salt diapirs resemble dark cauliflower heads. Their rise to the surface has been accelerated by the process that has raised the mountains: the ongoing collision of the Arabian and Eurasian plates. Three of the islands, which some day will be peaks on land, have salt cores.

Gallium Arsenide Transistors

Their speed holds promise for advanced computers and communication systems. It can be understood by plotting the energy and momentum of an electron propagating through a crystal of the semiconductor

by William R. Frensley

In semiconductor devices speed is of paramount importance. The fastest devices in existing computers can switch a current on or off in about a billionth of a second. Still faster devices are needed, however. They are essential to the construction of more powerful computers, and they would also make it possible to build new kinds of radars and communication satellites operating in the microwave band as well as at higher frequencies.

To meet these needs new semiconductor technologies are being developed in laboratories throughout the world. One way to speed the operation of semiconductor devices is to make them smaller, thereby shortening the distance electrons carrying a signal must travel. Another way to increase the speed of a semiconductor device is to increase the velocity of the electrons that flow through it. Technologies adopting this approach are generally based not on the traditional semiconducting element silicon but on a semiconducting compound, gallium arsenide.

Designers of gallium arsenide devices cannot rely solely on the extensive body of theory and practice that undergirds the established silicon technology. Atoms in a gallium arsenide crystal have nuclear charges and electron distributions that differ significantly from those of the atoms in a silicon crystal. Semiconductor physicists have therefore developed ways of representing the effects of those differences on the wavelike properties of an electron traveling through a gallium arsenide crystal. The depictions are based on numerical solutions of the quantum-mechanical equations that govern electron dynamics.

By means of such representations, as well as other techniques, it has become possible to understand and exploit the unique electronic properties of gallium arsenide. Chief among these properties is the fact that an electron traveling through a gallium arsenide crystal behaves as if it had a smaller mass than it appears to have in a silicon crystal. Hence an electric field of a given strength accelerates an electron faster in gallium arsenide than in silicon. This effect and other more exotic quantum-mechanical effects observed in gallium arsenide can be applied to make semiconductor devices whose high-speed performance cannot be duplicated by conventional silicon-based devices.

To appreciate the advantages of gallium arsenide, it is helpful to know how semiconducting materials are employed in a solid-state electronic device. The most familiar and generally useful device is the transistor, which actually includes several broad classes of devices. In its most basic form a transistor is an electronic switch: the flow of an electric current through the transistor is started or stopped as another, weaker current is applied to or removed from a particular section of the transistor. The current flow can also be controlled with greater precision, so that variations in the weak applied current produce corresponding variations in the current flowing through the transistor. In that case the transistor serves as an electronic amplifier: the stronger current flowing through the transistor mimics the pattern of the variations in the weaker applied current.

One of the simplest such devices, a field-effect transistor, or FET, illustrates the key factors determining a transistor's speed. A FET [*see illustration on page 82*] consists of two layers in a single semiconducting crystal: a nonconducting layer underlying a conducting layer, called the active layer, which is .1 or .2 micrometer thick. The active layer is usually made conducting by including relatively small numbers of impurity atoms that act as electron donors, which give up an electron to become positively charged.

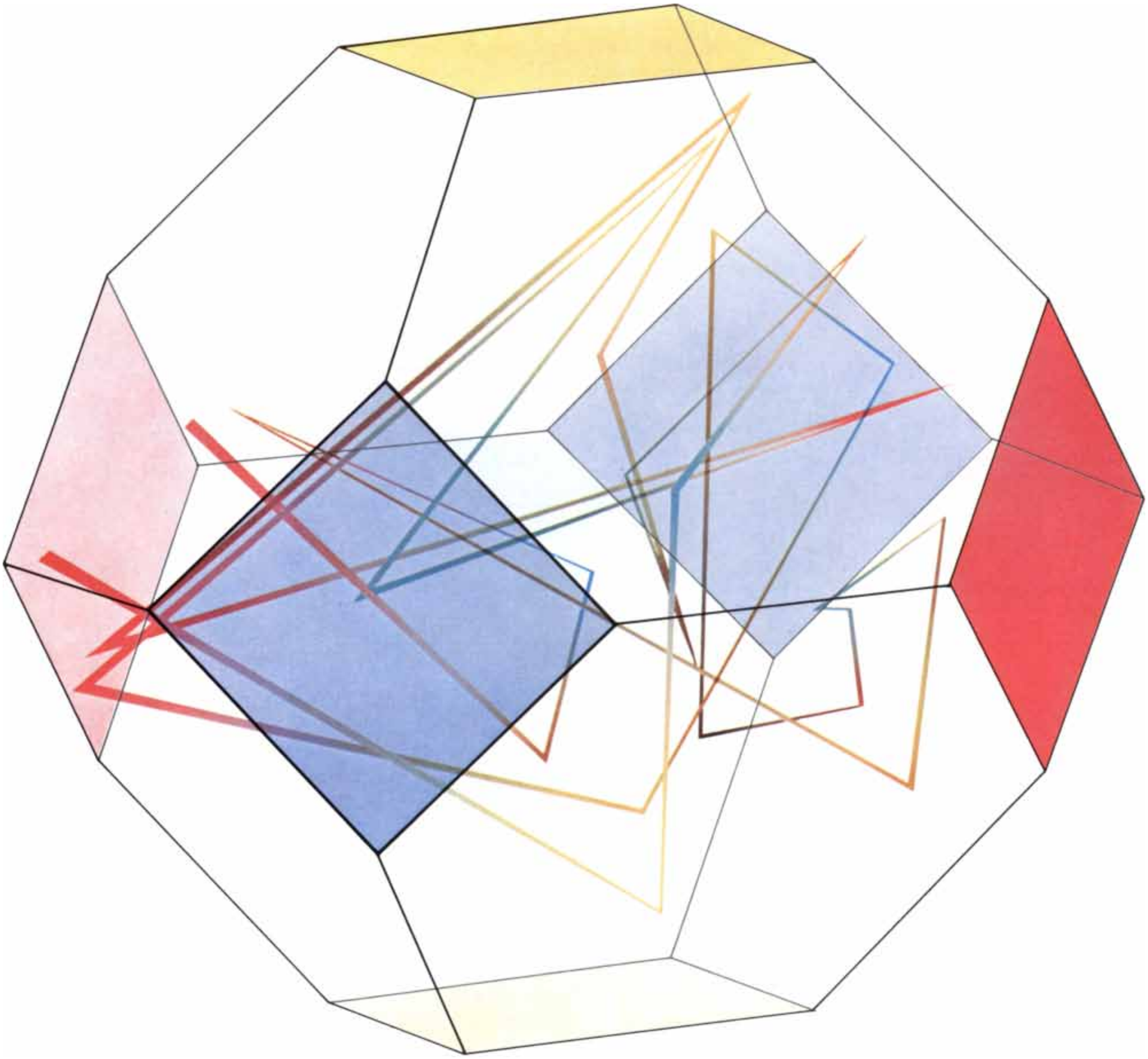
On the surface of the active layer are three electrodes called the source, the drain and the gate. A voltage applied between the source and the drain creates an electric field in the active layer, setting the layer's electrons in motion along the field lines. Normally the source is given a negative voltage with respect to the drain, so that electrons are injected into the active layer by the source, drift through the layer and are collected by the drain.

The gate is formed in such a way that the junction of the electrode metal and the underlying semiconductor results in what is known as a Schottky barrier. The key feature of a Schottky barrier is that the energy level of the electrons in the metal is much lower than the energy level of the electrons in the adjoining semiconductor. As a result any electron that drifts into the metal from the semiconductor will tend to be trapped in the metal. The electrons trapped in the gate are still attracted to the positive donors in the underlying semiconductor, however, and they remain close to the junction (within a few ten-thousandths of a micrometer), forming a negative surface charge on the electrode. As this negative charge builds up it repels electrons in the active layer, creating a depletion layer under the gate: a region in the active layer from which conduction electrons have been driven.

The depth of the depletion layer is controlled by the voltage applied to the gate. If the gate voltage is made negative, that is, if more electrons are supplied to the gate by the external circuit, the depletion layer will reach farther into the active layer. At a sufficiently negative voltage the depletion layer will extend all the way through the active layer. Then no current can flow between the source and the drain, and the transistor is said to be pinched off. If the gate voltage is made less negative, the depletion layer shrinks and more current can flow between the source and the drain.

How is such a transistor's speed of operation defined? That depends on the circuit to which it belongs. In a digital logic circuit, such as is found in computers and calculators, the speed of a FET is defined as the time required for the transistor to be pinched off. It can easily be shown that this reduces to the average time an electron in the active layer takes to travel the length of the gate.

The speed of an analog circuit, such as those found in radio and television receivers and transmitters, is defined somewhat differently. The function of a transistor in such a circuit consists not in turning the drain current completely on or off but in transforming small fluctuations in the gate current into corresponding large fluctuations in the drain current. The key parameter is the transistor's current gain, or amplification, which is defined as the ratio of drain current to gate current. The current gain drops as the fluctuations of the gate current increase in frequency. Speed in this case is defined as the ability to maintain high gain at high frequencies. Fundamentally, however, speed in both digital and analog circuits requires that the transistor respond quickly to a change in the gate current, and the primary way to



TRAJECTORY OF AN ELECTRON (*colored lines*) traveling through a crystal of gallium arsenide under the influence of an extremely strong electric field can be plotted in terms of the electron's momentum in three perpendicular directions. In such a "momentum space" the electron is confined within a polyhedron, known as the Brillouin zone, that contains all possible momentum values. The shape of a Brillouin zone depends on the structure of the crystal lattice; its geometric regularity reflects the regularity of the lattice. Here the color of a particular point on the electron's trajectory indicates the sum of its momentum components along each of the three differently colored axes. Because impurity atoms and phonons (quanta of thermal vibrations) scatter the electron, its path appears to be random. Yet if one were to sum the electron's momentum components over a long enough time, one would find that the electron in fact moves mainly in the direction of the electric field. The depiction is based on computer simulations done by Hisashi Shichijo of Texas Instruments, Inc., and Karl Hess of the University of Illinois at Urbana-Champaign.

increase the speed of a transistor in both applications is to decrease the length of the gate or to increase the electron speed or both.

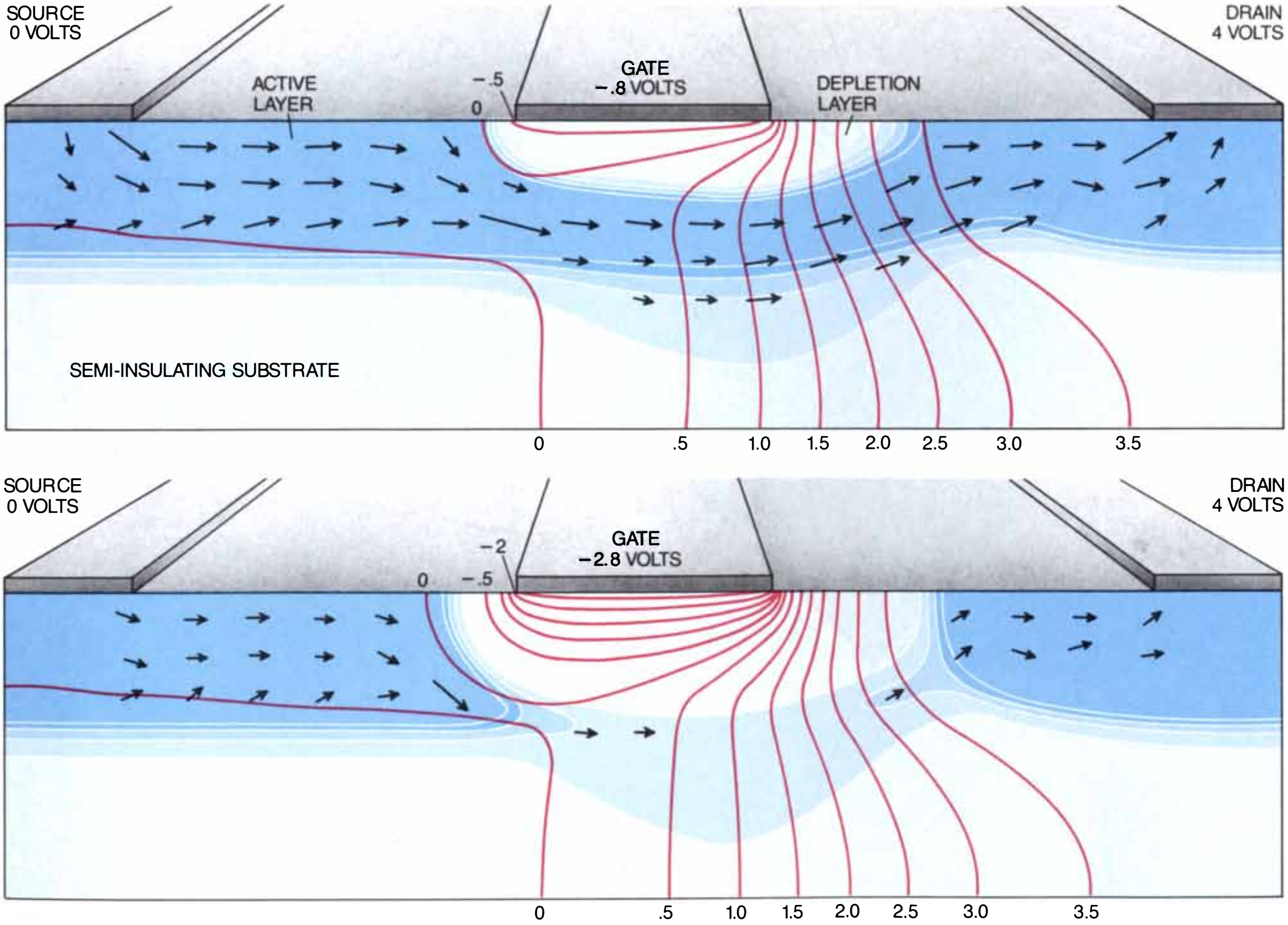
The gate electrode of a silicon-based FET in present-day integrated circuits is about a micrometer in length. Continued progress in semiconductor-fabrication technologies will undoubtedly lead to further reductions in gate size and therefore to increases in transistor speed. But in order to attain the highest possible operating speeds in transistors, the mobility of the electrons through the transistor needs to be increased as well. This can be achieved by making integrated circuits from semiconducting materials other than silicon.

To understand what determines the speed of electrons in a given semiconductor, one must consider how they move through the crystal structure. When a number of atoms are brought together to form a crystal of semiconductor, the discrete energy levels of the electrons surrounding the nucleus of each individual atom broaden into energy ranges, called bands. The energy levels that were occupied by the valence, or outermost, electrons in the atoms become what are called the valence bands of the crystal.

The electrons in the valence bands of a semiconductor, which are almost completely filled, form the latticework of chemical bonds holding the crystal together. A "forbidden energy" gap, in which there are no allowed electron states, separates the valence bands from unoccupied energy bands called the conduction bands. Because electrons in the conduction bands do not take part in interatomic bonding, any electrons in a conduction band are free to move through the crystal.

According to quantum mechanics, a particle, such as an electron, can also be thought of as a wave. In particular, an electron moving through a crystal can be treated as a propagating wave whose wavelength becomes shorter as its momentum increases. In a crystal an electron's total energy is a combination of its kinetic energy and the potential energy resulting from its interaction with all the other charged particles around it. Hence an electron's energy state depends on its wave function as well as on the periodic structure of the crystal lattice, made up of the constituent atoms' positive nuclei and their valence electrons.

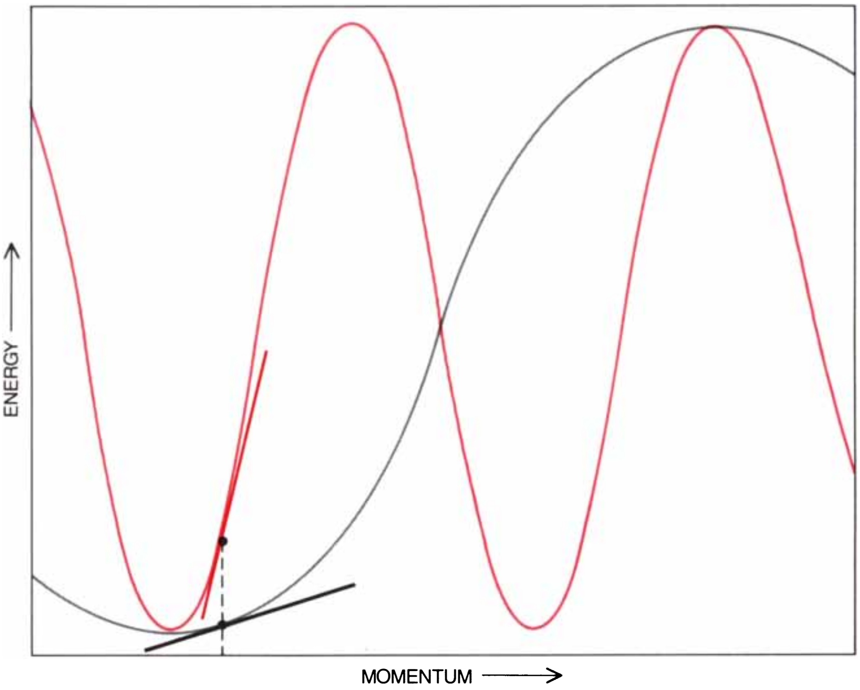
The interaction between an electron in a given state and all other electrons and positively charged nuclei in the surrounding material is too complex to be easily visualized, but with suitable approximations and the help of a computer the electron's energy as a function of its momentum can be



FIELD-EFFECT TRANSISTOR is fabricated from two layers of semiconducting material: the bottom layer is semi-insulating, whereas the top layer, called the active layer, is conducting. On the surface of the active layer three metal electrodes are formed. Electrons can readily flow between the active layer and the two outer electrodes, called the source and the drain. The middle electrode—the gate—is made so that it tends to repel electrons in the underlying semiconductor, forming a zone (called the depletion layer) that is devoid of conduction electrons. When a voltage is applied between the source and the drain (*top*), electrons flow from the source through the active layer to the drain. (Colored lines represent the voltage gradient in half-volt increments, and the arrows represent the current's strength and direction.) The current flow through the transistor can be regulated by adjusting the voltage applied to the gate, since a negative gate voltage will make the depletion layer extend deeper into the active layer. If the gate voltage is negative enough, the flow is stopped (*bottom*). The density of conduction electrons is indicated by the shading.

readily plotted [*see illustration at right*]. The curve's particular shape is a reflection of the interaction. Moreover, because the slope at a point on an energy-momentum curve is equal to the velocity of the electron in the state represented by the point, an energy-momentum graph allows one to see clearly how the electron's velocity depends on the band structure of a material: a steeper curve indicates a higher velocity. (I should mention here that the quantum-mechanical momentum of an electron, because it is a function of the electron's wavelength, is quite different from the familiar classical momentum. An electron at rest still has a wavelength, and therefore it is possible for an electron to have zero velocity and yet have positive momentum.)

In spite of their quantum-mechanical nature, electrons in a crystal are still governed by two fundamental principles of classical mechanics. The first one is that an external force acting on an electron (such as the force exerted by an electric field) changes the electron's momentum. An electron subjected to a constant force, produced by a constant electric field, will gain momentum at a constant rate. The second is that the change in the energy of an electron is equal to the applied force multiplied by the distance the electron moves.



VELOCITY OF AN ELECTRON as it moves through a particular material under the influence of an electric field can be determined from the slope of the curve showing the relation between the electron's energy and its momentum. The shape of the curve reflects the electronic properties of the material. The steeper the slope is, the faster the electron is moving. Hence two electrons (*dots*) traveling through two different materials may have the same momentum but different velocities, as is indicated by the slopes of the tangent lines. Because momentum is classically defined as the product of mass and velocity, physicists say that the electron having the higher velocity has a smaller "effective mass."

If a particle attains a higher velocity than another particle subjected to the same force for the same length of time, a classical physicist would say that the first particle has a smaller mass than the second. Such an interpretation, in fact, is applied in studying the motion of an electron in a semiconductor crystal. For a given applied force, the higher an electron's velocity is, the smaller its "effective mass" is said to be. In terms of the energy-momentum curves for electrons with the same momentum but moving in two different materials, one can say the electron has a smaller effective mass in the material that exhibits the steeper energy-momentum curve.

As long as the energy-momentum curve is bending upward, the slope is increasing and the electron is accelerating. When the curve begins to bend downward (as it does when the electron approaches the energy peak), the electron begins to decelerate. In fact, at the peak of the energy-momentum curve an electron's velocity is zero, because the slope of the curve is zero at that point.

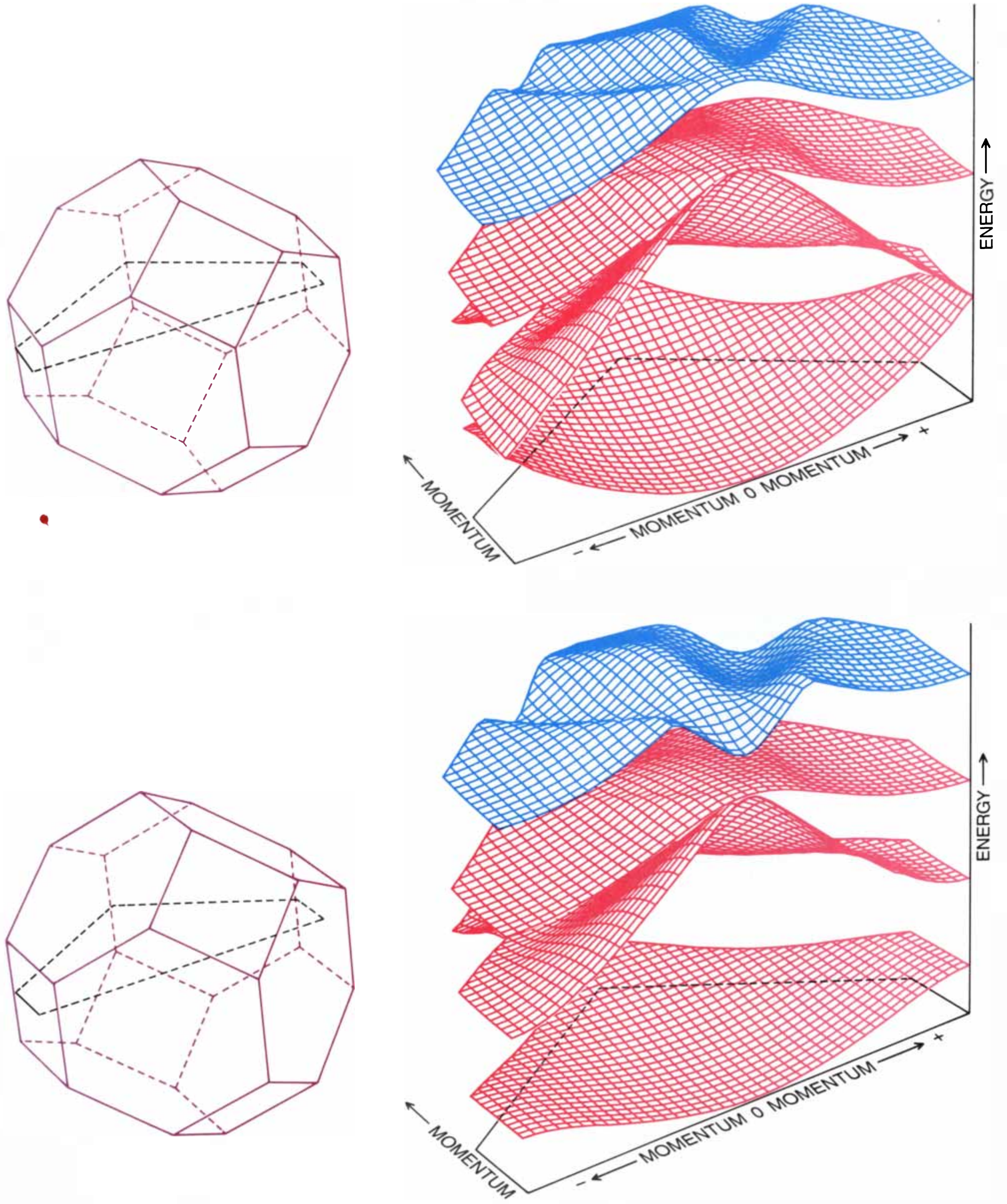
As the momentum increases further under the influence of an electric field thc velocity of the electron actually becomes negative, because the curve slopes downward. A paradox emerges on the downward slope: as an electron gains momentum in the direction of the field, it gains velocity in the opposite direction! In short, it moves in the direction opposite to that of the electric field. One way to interpret this behavior is to say the electron has acquired a negative effective mass.

This rather bizarre behavior reflects the finite range of energies covered by an energy band. If the electron continued to move in the direction of the electric field, it would continue to gain energy from the field and would soon gain more energy than is allowed by the band structure. When the electron begins to move in the opposite direction, it in effect returns the energy it has gained from the field.

It is mathematically convenient to combine an electron's wavelength and direction of propagation into a single quantity called the momentum vector. In this way all the possible momentum vectors of all electron states in a crystal lattice can be assumed to lie within a volume called the Brillouin zone, whose dimensions are not distance but momentum in three perpendicular directions. The shape of the Brillouin zone for a given material depends on its crystal structure rather than on its chemical makeup. In most semiconducting materials, including silicon and gallium arsenide, the Brillouin zone has the form of an octahedron whose corners have been lopped off, resulting in a figure with eight hexagonal faces and six square faces.

By means of a computer, I have been able to analyze slices through the center of the Brillouin zone: two-dimensional sections within which the momentum can change in only two perpendicular directions. By limiting the problem in that way I can depict each energy band in the material as a surface whose changes in height reflect changes in the electron's energy [*see illustration on next page*]. Such an energy-momentum surface is analogous to the earlier energy-momentum curves, which represent energy bands along only one dimension of momentum space. Hence an electron's velocity and acceleration are indicated respectively by the local slope and curvature of the surface.

The four valence bands of silicon and gallium arsenide are similar when they are depicted in this manner, the lowest band having an energy minimum at the center of the Brillouin zone (corresponding to zero momentum) and the other three bands having an energy maximum at the zone cen-



ELECTRON DYNAMICS in silicon (*top*) and gallium arsenide (*bottom*) crystals can be visualized if an electron's momentum is plotted along two dimensions in a region (*outlined in black*) within the material's Brillouin zone; the electron's energy is plotted along a third, perpendicular dimension. The red energy-momentum surfaces apply to the atoms' valence, or bonding, electrons, which cannot conduct current. The blue surface represents the mobile electrons of the conduction band, which are generally provided by impurity atoms in the crystal. The slope and curvature of the conduction-band surface reveal respectively the velocity and acceleration of current-carrying electrons. Such electrons tend to occupy the lowest energy states on the surface. In silicon the "valleys" in which the electron states collect (on either side of the front edge) are broadly curved, indicating that low-energy electrons have a large effective mass. The conduction-band minimum in gallium arsenide, in contrast, falls in the narrow central valley, indicating that low-energy electrons in the material have a small effective mass and can attain higher velocities than electrons in silicon under the influence of the same electric field. The figures were generated by the author, applying a technique developed by Marvin L. Cohen of the University of California at Berkeley and Thomas K. Bergstresser of the Sandia National Laboratories.

ter. The similarity between the valence bands of both semiconductors reflects the similarity of the bond latticework in gallium arsenide and silicon.

Many common transistors rely on localized electron voids in the valence bands (instead of on electrons in the conduction bands) to carry an electrical signal. The voids are produced by removing electrons from the upper valence band, often by introducing impurity atoms that have one electron fewer than the corresponding semiconductor atom. Because such electron "holes" behave like positively charged quasi particles, they can serve as charge carriers in transistors. Yet they are not suited for high-speed applications: the upper valence bands are relatively flat in silicon as well as in gallium arsenide. Hence the effective masses of the holes are rather large and their velocity is intrinsically slow.

Above the valence bands in a semiconductor is the forbidden-energy gap, and above the gap are the conduction bands. Because the valence bands are normally filled by the valence electrons, electrons added to the semiconductor crystal by implanted donor atoms must adopt states in the conduction bands, beginning with the lowest available energy states. It is in the lowest conduction band that high-speed charge transport takes place.

In silicon the lowest energy state in that band is found at nonzero values of the momentum: away from the center of the Brillouin zone. There are six possible values for the momentum vectors (of the same magnitude but heading in different directions) at the energy minimum. The valley around each of these states is broad: its sides slope gradually, indicating that electrons accelerate slowly and therefore have a large effective mass. The effective mass depends on the direction of motion and ranges from about the mass of a free electron outside a crystal down to a fifth of that value.

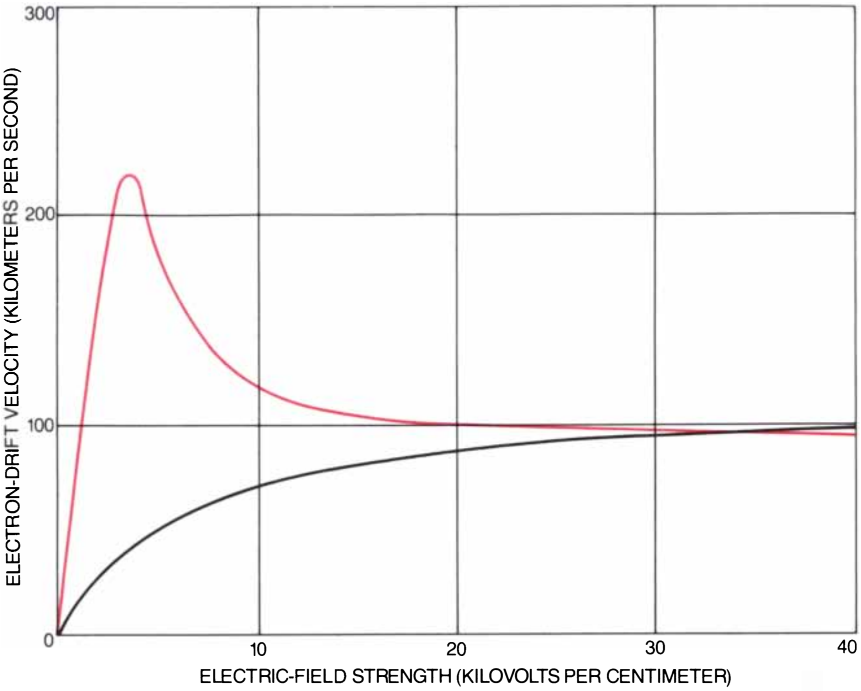
The shape of the lowest conduction band in gallium arsenide is different from that of silicon, owing to differences in the interaction of electrons with nuclear charges in the compound. The minimum energy falls at the center of the Brillouin zone, and so there is only one minimum. Moreover, the valley around the minimum is rather narrow, indicating a small effective mass: only .07 times the free electron mass. This small mass is one reason electrons move faster in gallium arsenide than they do in silicon.

Effective mass is not the only factor governing the movement of electrons through a semiconductor crystal, because they do not move unimpeded. Electrons collide with structural flaws in the crystalline lattice, charged impurity atoms and phonons, which are thermal vibrations of the crystal atoms. They therefore move not in free flight but in very short periods of free flight (lasting for perhaps a tenth of a trillionth of a second) between collisions. In fact, the motion of electrons in a semiconductor is a form of Brownian, or random, motion. Because an applied electric field accelerates electrons in a specific direction between collisions, on the average they drift in the direction of the field. Of course, the more rapidly an electron is accelerated between collisions, the higher its average drift velocity will be. Consequently the lower effective mass of electrons in gallium arsenide does increase the drift velocity.

Yet drift velocity also rises as the intervals between collisions increase. In this respect too the shape of the conduction band in gallium arsenide is advantageous. The rate at which collisions occur depends on the number of available energy states into which an electron can be scattered after a collision. In silicon the six broad conduction-band valleys provide many such states near the bottom of the conduction band. Gallium arsenide, on the other hand, has relatively few such sites near the narrow conduction-band minimum.

The consequence of the difference in conduction-band topography between silicon and gallium arsenide can be clearly seen by plotting the drift velocity of an electron as a function of electric-field strength in the two semiconductors [*see illustration above*]. Under weak electric fields the electron-drift velocity in gallium arsenide rises much faster than the velocity in silicon. (This is described by saying that gallium arsenide has a higher electron mobility than silicon.) The electron velocity in gallium arsenide reaches a peak value of about 200 kilometers per second and then decreases to about half that value as the field strength continues to increase. The velocity in silicon levels off at about the same value under much stronger fields.

The decrease in the electron-drift velocity observed in gallium arsenide as the electric field is increased (known as negative differential mobility) can be explained by the topography of the conduction band outside the central narrow valley: it resembles the conduction band of silicon in that there are several valleys where the electron has a large effective mass. Thus it is



ELECTRON-DRIFT VELOCITY, the average velocity of a conduction electron, is plotted as a function of electric-field strength for gallium arsenide (*color*) and silicon (*black*). At field strengths below about 30 kilovolts per centimeter the superiority of gallium arsenide for high-speed transistors is clearly evident. Under stronger fields, however, the conduction electrons in gallium arsenide scatter out of the central, narrow valley into broader, "satellite" valleys in the conduction-band surface (*see illustration on opposite page*). Because the satellite valleys resemble the conduction-band valleys of silicon, the drift velocity of electrons in gallium arsenide approaches that of electrons in silicon.

not surprising that the drift velocity approaches the velocity in silicon under stronger electric fields. As the electrons are accelerated in a sufficiently strong field, they gain enough energy to scatter into these broad "satellite" valleys, whose lowest energy states are somewhat higher than the minimum energy state of the central valley.

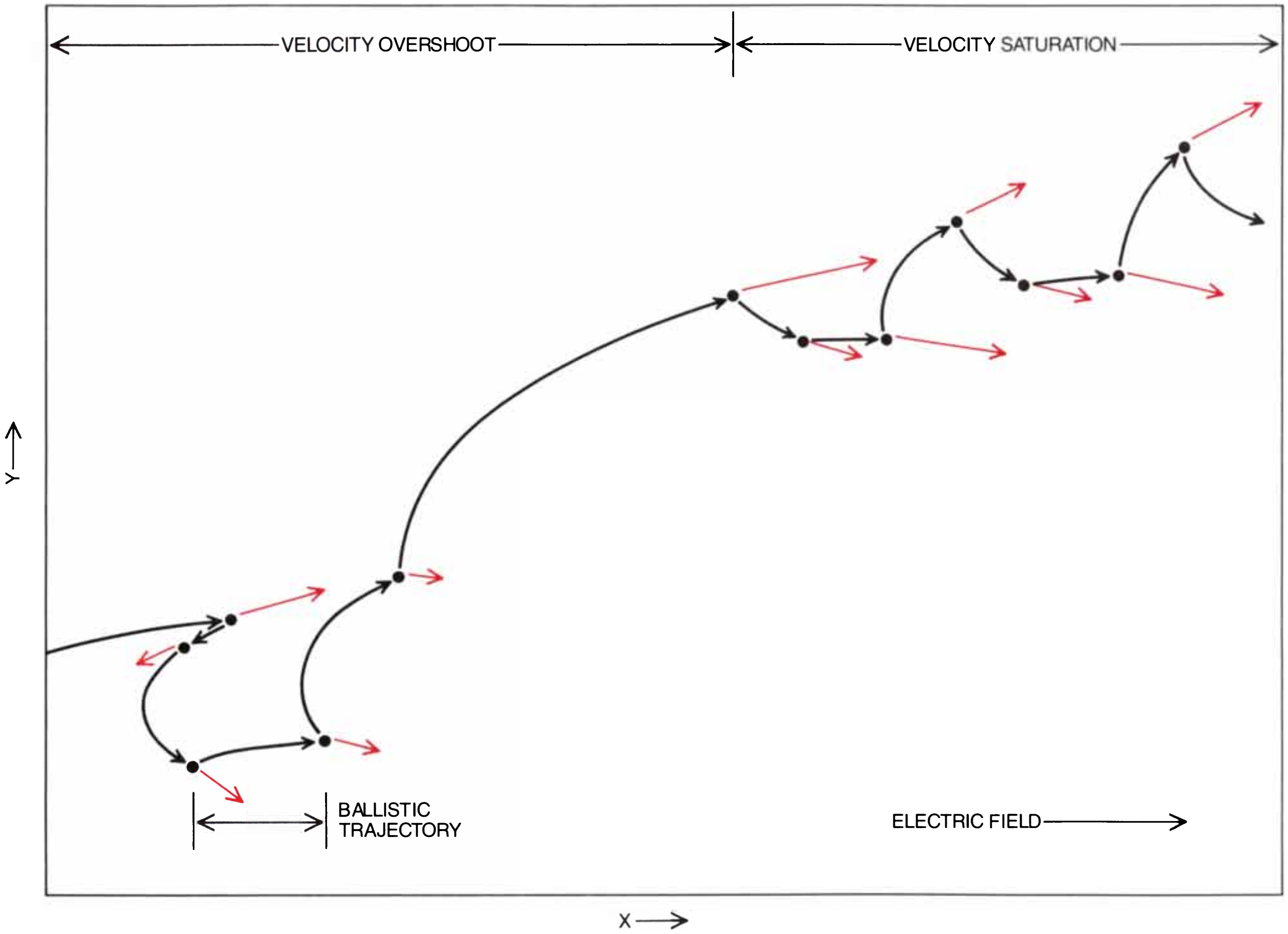
Under certain conditions the decrease in velocity with increasing electric-field strength leads to spontaneous current oscillations in gallium arsenide as well as in other related compounds. This phenomenon, which is the operational principle of certain solid-state microwave oscillators, is called the Gunn effect after John B. Gunn of the International Business Machines Corporation, who discovered these oscillations in 1963.

So far I have limited my discussion to phenomena that recommend gallium arsenide for use in high-speed conventional FET's. Yet as gallium arsenide integrated-circuit technologies move into volume production, more exotic phenomena that also occur in gallium arsenide are being exploited in the design of other solid-state devices. I shall discuss two of them here: velocity overshoot and ballistic transport.

In a FET that has a short gate length the electric field in the active layer changes drastically over a short distance, and an electron moving from source to drain experiences a sudden accelerative force. Indeed, the acceleration is so rapid that before a large number of energy-draining collisions can occur, the electron can reach velocities substantially higher than those measured over longer distances in uniform electric fields. This effect, called velocity overshoot, was discovered by Jacques G. Ruch of the Bell Telephone Laboratories in computer simulations of electron transport. The simulations also showed that the effect is much more pronounced in gallium arsenide than it is in silicon, because of the lower effective electron mass and the lower rate of scattering—as long as the electron state resides in the central conduction-band valley.

As a result of velocity overshoot, the drift velocity of electrons in gallium arsenide can actually approach 500 kilometers per second over distances of a few tenths of a micrometer. In fact, it is thought that velocity overshoot is largely responsible for the high-speed performance of most gallium arsenide transistors, although this is difficult to demonstrate directly.

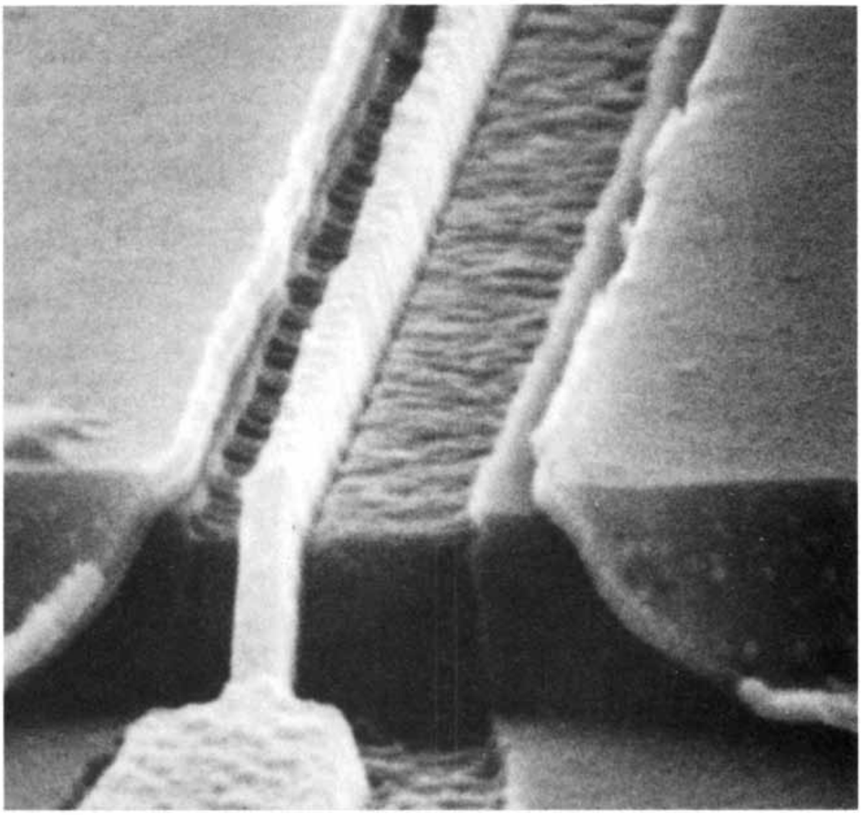
A related phenomenon that takes place over even shorter distances than velocity overshoot does is ballistic transport: the movement of electrons without a significant loss of energy. The effect, which was postulated by Michael S. Shur of the University of Minnesota and Lester F. Eastman of Cornell University, is again a result



SCATTERING by phonons (*red arrows*) causes an electron (*black arrows*) to follow an erratic path through a semiconducting material as it drifts in the direction of an applied electric field. Between scattering events the electron follows a ballistic trajectory, and so the motion of an electron that undergoes few scattering events is called ballistic transport. In gallium arsenide a conduction electron's initial state is in the central valley of the energy-momentum surface, where the effective mass is low. Such an electron can be accelerated so rapidly by an electric field that it attains a high drift velocity. Such "velocity overshoot" cannot be sustained, because at high velocities the electron is readily scattered out of the central valley into a satellite valley. There its effective mass is larger, and so it drifts at a much lower velocity and undergoes much more scąttering in a condition known as velocity saturation.

of the fact that collisions do not occur continuously; an electron can move some distance without suffering a collision. A semiconductor device can be made small enough so that the distance an electron must travel is less than the average distance traveled between collisions.

Ballistic transport in gallium arsenide has been observed only recently in semiconductor devices known as hot-electron transistors [see "Ballistic Electrons in Semiconductors," by Mordehai Heiblum and Lester F. Eastman; SCIENTIFIC AMERICAN, February]. Such transistors have an extremely short transit length because the electrons travel vertically through thin layers of semiconductor rather than horizontally along the active layer, as in the FET. The transit length of ballistic-transport devices has been as small as .035 micrometer. So far only experimental devices employing this effect have been made. A good deal of engineering remains to be done before ballistic-transport transistors can compete with other high-speed devices.



GALLIUM ARSENIDE TRANSISTOR made by Texas Instruments, Inc., has a gate electrode (*central ridge*) only .2 micrometer wide. The source electrode is just .2 micrometer to the left of the gate; the drain is farther away on the right. Such devices can generate electromagnetic radiation that has frequencies as high as 115 billion hertz.

Although it is clear that gallium arsenide has an inherent advantage over silicon in its higher electron mobility, gallium arsenide digital circuits will supplement rather than replace silicon circuits. Gallium arsenide is not likely to be employed in chips for low-cost, general-purpose computers, because silicon chips are significantly less expensive to produce. Nonetheless, many special-purpose digital systems (for communications and some military applications) have electronic components that have to operate at the highest possible speeds. Gallium arsenide integrated circuits will be incorporated in such systems, as well as in some of the supercomputers that are currently under development. Some military systems also require circuits that operate well at high temperatures and in the presence of high levels of radiation. Gallium arsenide circuits perform better than silicon ones under such conditions.

In analog circuits, however, the new gallium arsenide technology has already had a substantial impact. Because silicon transistors are too slow to work well at frequencies above about three gigahertz (billion cycles per second), gallium arsenide transistors have become the semiconductor devices of choice for applications over most of the microwave band, which extends to 30 gigahertz. If their gate lengths can be reliably reduced to a quarter of a micrometer or less, such transistors will also establish themselves as the dominant technology for devices operating in the millimeter-wave band, which extends from 30 to 300 gigahertz and is now largely unexploited.

Gallium arsenide FET's can be integrated with other necessary components (such as resistors, capacitors and inductors) onto a single chip called a monolithic microwave integrated circuit (MMIC). An MMIC, which is only a few millimeters on an edge, can perform functions that used to require bulky circuits consisting of vacuum tubes and waveguides. Many MMIC's could conceivably be fabricated on a single wafer, just as silicon logic circuits are now, lowering the cost of manufacturing them. Owing to their small size and weight, MMIC's will make it possible, for example, to build phased-array radar systems that can fit on an airplane. Such systems currently are the size of buildings.

Gallium arsenide MMIC's will also contribute to a reduction of the cost of current microwave systems, such as satellite receivers. Inexpensive home receivers will be an important part of direct-broadcast systems that are being developed by several countries. Such nationwide communication systems, which would broadcast television signals from satellites rather than from ground-based antennas, will operate at higher frequencies than current satellite communication systems do. The shorter wavelengths will allow the use of smaller dish antennas (about a meter in diameter), which are cheaper and easier to install.

The short wavelengths at which gallium arsenide transmitters and receivers could operate might lead to new kinds of applications. For example, "smart" munitions could be equipped with miniature radars that would enable them to home in on a target after being fired from a tank. The antenna for such a radar must be extremely small, because it has to fit into the bore of the tank's gun. If the radar is to have enough angular resolution to be useful for guidance, the wavelength of the radar signal must be several times smaller than the diameter of the antenna. This means that the radar must process millimeter waves, which gallium arsenide FET's can effectively generate and detect.

Gallium arsenide devices will certainly open a large part of the electromagnetic spectrum to new applications. At present these are mostly in the military domain, but as the cost of generating and detecting microwave signals comes down, new commercial and consumer applications will undoubtedly be found.

Gazelle Killing in Stone Age Syria

At Tell Abu Hureyra a band of hunter-gatherers started slaughtering entire gazelle herds 11,000 years ago. Hunting continued long after the emergence of agriculture, which sheds new light on that process

by Anthony J. Legge and Peter A. Rowley-Conwy

One of the most fundamental questions in archaeology is that of how agriculture came into being. How and why was a way of life based on hunting and gathering replaced by one based on plant and animal husbandry? The question is clearly complex, and no one site can provide a full answer. Yet some archaeological sites, by virtue of their location and the period when they were occupied, can serve as a window onto the emergence of agriculture.

One such site is Tell Abu Hureyra, which is on the banks of the Euphrates in northern Syria. Beginning about 11,000 years ago, in the Mesolithic period, the site was first settled by a group of hunter-gatherers. With only a single break, the site was occupied well into the succeeding Neolithic period. Because this is precisely the period when agriculture first emerged, Tell Abu Hureyra can offer much information about that process.

One of the most striking findings from Tell Abu Hureyra is that for perhaps as long as 1,000 years after the beginning of plant domestication there, hunting continued to play a crucial role in the community's subsistence. During that millennium the main source of animal protein was the mass killing of gazelles as the herds moved north in the early summer. Indeed, the mass-killing strategy may have been given up only when the gazelle herds were depleted.

Such unexpected findings throw a new light on the inception of agriculture. It is clear from Abu Hureyra that one part of the agricultural constellation (such as cultivation) may prevail long before the others (such as animal husbandry). In the interim hunting-gathering methods may coexist with agricultural ones. What is more, during that period hunting may be superior to animal domestication for obtaining animal foods. It takes time for such lessons to be assimilated, but the work at Tell Abu Hureyra has already begun to show how complex the origin of agriculture actually was.

The "tell" of Tell Abu Hureyra is a large mound containing the debris of human settlement during several thousand years. The mound's chief constituent is decayed mud walling from the occupants' houses. At Abu Hureyra this material was sufficient to form a substantial rise: the mound covers 11.5 hectares (about 28 acres) and is stratified to a depth of about eight meters. About one million cubic meters of deposits were included in the tell, which was excavated in 1972 and 1973 by Andrew M. T. Moore of the University of Oxford [see "A Pre-Neolithic Farmers' Village on the Euphrates," by Andrew M. T. Moore; SCIENTIFIC AMERICAN, August, 1979].

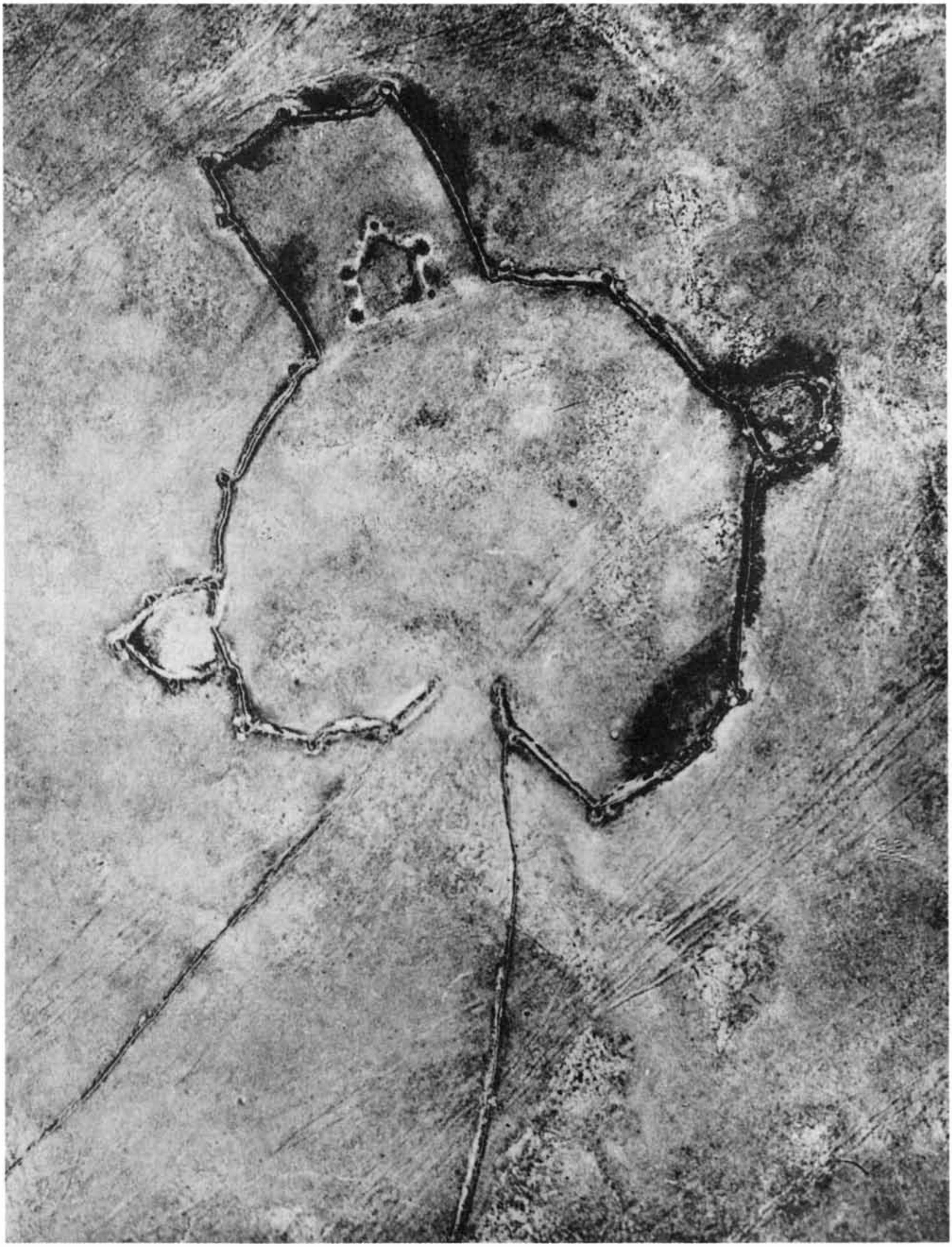
Moore's excavation was carried out on an emergency basis, because Tell Abu Hureyra was about to disappear from human history. As a result of the construction of the Tabqa Dam farther down the Euphrates, the tell was scheduled to be flooded in 1974 by the rise of Lake Assad. Since the mound looked promising archaeologically, a rapid but systematic excavation was planned. One of us (Legge) helped to plan the methods of recovery and provided machinery for retrieving substantial quantities of organic remains, both plant and animal.

The organic remains found in the excavation reflect how the tell dwellers subsisted. Their modes of subsistence are in turn closely connected to the location of Abu Hureyra. The mound lies at the edge of the valley of the Euphrates. On one side is the floodplain of the river; on the other is dry, level steppe covered by grasses and small shrubs. The area currently receives only about 200 millimeters of rainfall each year, which puts it near the limit of the zone in which farming can be carried out without irrigation.

The current climate of the area is not much different from the one prevailing in 9000 B.C., when the site was settled. (The dates in this article are uncalibrated radiocarbon dates.) In Mesolithic times the village may have had between 200 and 300 inhabitants. Beginning in about 8000 B.C. there seems to have been a hiatus in occupation that ended with a resettlement in about 7500 B.C., during the Neolithic period. The population of the settlement in the Neolithic (the period traditionally associated with the rise of agriculture) was larger than it had been in the preceding epoch. The Neolithic village may have had between 2,000 and 3,000 inhabitants.

The modern methods of excavation employed at Tell Abu Hureyra made it possible to trace very precisely the changes in subsistence that occurred through the Mesolithic and Neolithic. Large samples of sediment were put in flotation devices; Gordon Hillman of the Institute of Archaeology at the University of London is analyzing the plant remains thus collected. All excavated earth was put through a sifting device with a one-centimeter grid that trapped larger objects such as bones. Of 60,000 identifiable bone fragments, 40,000 have so far been analyzed in our work, which is supported by the Science and Engineering Research Council of Britain.

What has all this bone told us? The first conclusion is that as a source of animal protein the steppe was a far more important environment than the river valley. Throughout the Mesolithic and into the early part of the Neolithic, gazelles are by far the most important animal-food species. Other species are represented, including onagers (a species much like the wild ass), sheep, goats, pigs, deer and wild cattle. Gazelle bones, however, make up about 80 percent of the total. The primary species is the Persian gazelle, or goitered gazelle, *Gazella subgutturosa*.



DESERT KITE is a structure for the mass killing of gazelles that is found in Syria, Jordan, Saudi Arabia and the Sinai desert. The photograph shows a kite at Dumayr, about 30 kilometers northeast of Damascus. It was probably built of stone. The gazelles were driven between "training walls" (*bottom*) into the large enclosure, where they were killed. Three small enclosures weie attached to the main enclosure. One has been replaced by a rectangular structure (*top*). The small circles dotting the walls may have been hiding places for hunters. The photograph was made in 1930 by Père A. Poidebard, a pioneer aerial photographer in Syria.

Gazelles generally do not need large quantities of either water or green food such as fresh leaves and shoots. Since they are well adapted to dry regions, they can readily survive on the arid steppe. Pigs and deer, on the other hand, need to drink often; cattle, sheep and goats must drink at least every few days. As a result, all these species prefer the river valley to the steppe. Their relative scarcity among the remains at Abu Hureyra suggests that the valley species were hunted to a low level early in the occupation and thereafter the steppe provided the major part of the animal economy.

Our analysis of the thousands of bones from the tell not only showed that the gazelle was the predominant species but also revealed a great deal about how these small, fleet creatures were killed. In our work we noted that the gazelle remains included the bones and teeth of many very young animals. Not all stages of growth, however, were represented. We began to examine the remains in an attempt to find out which stages of growth were present and which were missing.

One excellent source of information was the teeth, which develop in a predictable pattern. Persian gazelles are born with (or soon develop) three temporary "milk" molars on each side of the jaw. The milk molars last somewhat longer than a year, and during this time two permanent molars erupt behind them. At a little more than one year the milk molars are replaced by permanent premolars. The third permanent molar comes up at about the same time, yielding the full complement of six permanent teeth on each side of the jaw. Because gazelles eat tough herbaceous food, often mixed with grit and sand, the enameled crown of the milk molars is steadily worn down through the first year. The wear on the milk teeth offers a good measure of a young gazelle's age.

To find the age distribution of the young gazelles at Tell Abu Hureyra we measured the third milk molar. This is the largest of the milk teeth, and it has an easily recognizable form characterized by three cusps, or biting surfaces. When the height of the middle cusp was measured, we found that the teeth could be divided into two distinct groups. The first group showed no wear and the roots were not fully formed. In the second group the enameled crowns were heavily worn.

Unfortunately *Gazella subgutturosa* is now almost extinct in the wild in the Near East, and so an understanding of this pattern cannot be obtained there. The same species survives, however, in the Turkmenian region of the U.S.S.R. From 1942 through 1947 V. G. Geptner of the Moscow State University Museum of Zoology collected a large series of gazelle skulls there. Geptner's collection is housed at the museum, and on a recent visit to Moscow one of us (Legge) examined the skulls. Measurement of the crown heights of the third milk molars confirmed what we had already begun to suspect: the unworn milk molars from Abu Hureyra were those of newborn animals, whereas the teeth that were heavily worn came from animals killed at one year of age.

That finding was reaffirmed by examining the growth and fusion of limb bones. The calcaneum, or heel bone, was well suited to our purpose. As its common name suggests, the calcaneum forms part of the heel. (Since the gazelle essentially runs on its toes, however, the calcaneum is well off the ground.) Like the third milk molar, the heel bone is an easily recognized specimen showing clear signs of growth. One sign is that the body of the bone lengthens and thickens considerably during the first year of life. Another is that at about 14 months the process (a small extension to which a tendon is attached), which begins as a sepa-

TELL ABU HUREYRA was a large mound on the banks of the Euphrates in northern Syria. The people in the foreground are settled (nonnomadic) Bedouins; the camels are carrying cotton. The mound consisted largely of the collapsed mud walls of a prehistoric settlement lasting from 9000 to 6000 B.C. The authors conclude that the prehistoric community killed gazelle herds by means of some type of desert kite. The structure on the mound is part of a nearby village. The hole and the piles of earth are from an excavation carried out in the early 1970's. After the mound was excavated it was submerged by the damming of the Euphrates.

rate piece of bone, fuses to the main structure.

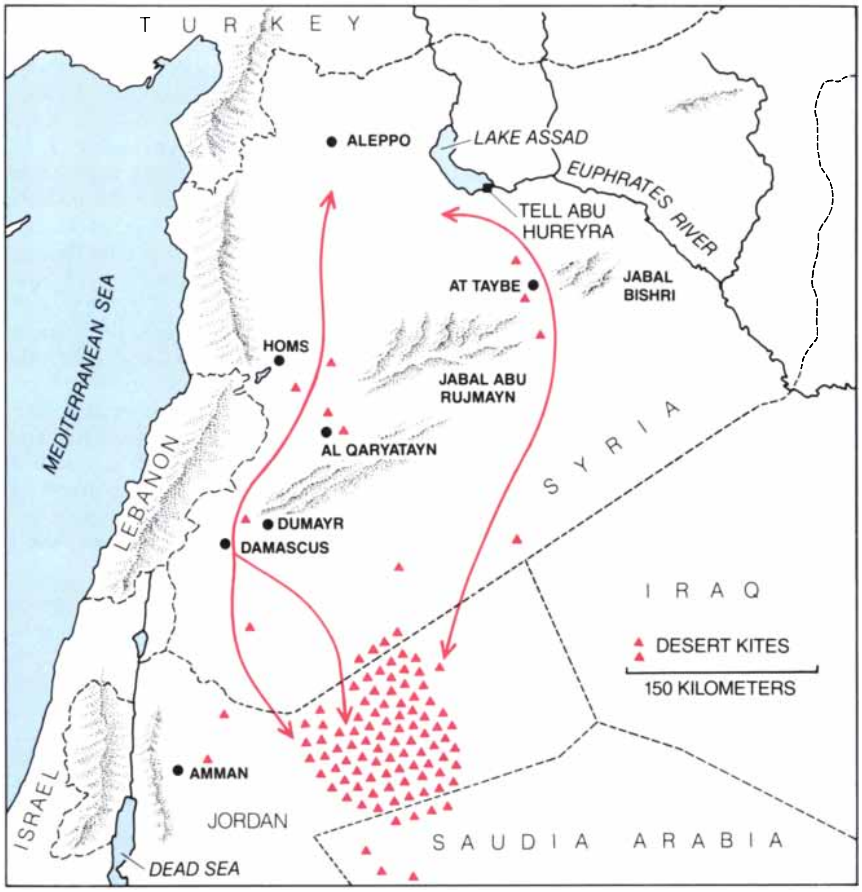
The bones of the young gazelles corresponded nicely to the pattern observed among the milk molars. One group of heel bones included small specimens with the process unfused; these came from newborn animals. A second cluster also had the process unfused but were almost as large as the adult bones; these represented yearlings. The most numerous group was made up of large bones with the process fully fused; these came from adult gazelles. (Later analysis of the permanent teeth confirmed the presence of adults of all ages as well as newborns and yearlings.)

How is it possible for there to be a division of bones and teeth into those from newborns, yearlings and adults? Part of the answer lies in the fact that the life cycle of a gazelle herd is highly synchronized. All the births in a herd take place in a short period. If the herd is hunted at the same time each year, the dead animals will have a characteristic age distribution that includes groups of young animals a year apart in age. This is what must have happened at Tell Abu Hureyra: the killing was a seasonal affair. The presence of newborns among the prey indicates the time of hunting was just after the females gave birth. Records left by early European travelers show that gazelles near Abu Hureyra gave birth in late April and early May.

Not only was the killing seasonal but also the object of the hunt was the herd rather than individual gazelles. The pattern of bones at the tell is most unlikely to have stemmed from the hunting of individual animals. When hunters stalk a herd and select individual gazelles, they take young animals in prime condition—often mainly males. At Abu Hureyra, however, we found bones from every age group, including the very young and the very old. This pattern undoubtedly resulted from a killing technique in which an entire herd was taken at once.

Now, mass-killing methods are well known in prehistory and also among hunter-gatherers in historic times, such as the Indians of the North American plains. The Plains Indians exploited various methods for taking herds of buffalo and pronghorn antelope. These methods all shared one simple plan: the herd was driven toward an enclosure or over a concealed pitfall, where it could be efficiently slaughtered. In some instances the animals were driven between "training walls" that converged on a destination chosen by the hunters.

Whether it is worthwhile for a group of hunter-gatherers to employ such methods depends on several factors. Mass killing is primarily a strategy for open land, where large herds offer suitable prey. Many hunters are needed to drive the herds across the land and to operate the traps where the killing is done. These three criteria (open country, large herds and enough hunters) were satisfied at Tell Abu Hureyra in the Mesolithic and early Neolithic periods.



GAZELLE MIGRATIONS in Jordan and Syria took place on a north-south axis. From the distribution of desert kites and other evidence the authors reconstructed two routes (*color*). In late spring the gazelles moved to the north, where the young were born; at summer's end they returned south. The western route is speculative. The eastern route ends near Tell Abu Hureyra, where the herds were hunted on arrival in April and May.

In spite of the fact that mass-killing strategies were well documented in North America, until recently they were not much considered in the Near East. In recent years, however, archaeologists have begun to recognize that structures for killing gazelle herds are widespread in Jordan, Syria, Saudi Arabia and the Sinai desert. Such structures were in use quite late—European travelers saw them in operation in the 20th century—but initially neither their extent nor their function was clear. The wide distribution of such structures was not recognized until airmail routes were established and pilots began photographing archaeological sites from the air. In 1929 Group Captain L. W. B. Rees, a pilot on the Cairo-Baghdad run, published the first account in the journal *Antiquity.* Rees dubbed the structures "desert kites" because of a supposed resemblance to a child's kite as seen from the air.

Desert kites vary in form and in size, but they do share some common elements. There is generally a central enclosure, which can range from a few meters to 150 meters across. The enclosure has a narrow entrance from which stone walls diverge; the walls may extend for several kilometers across the desert. In some kites the enclosure and the walls are made of large stone slabs set on edge; in others they are made of piled boulders.

Rees and the observers who followed him in the 1930's and 1940's were military men. They gave the kites a military interpretation as enclosures for the defense of people and livestock. Since then it has become clear that the kites actually served for mass killing. The open end of the V formed by the walls often lies near small valleys through which the gazelle herds moved. The hunters turned the herd into the opening and stampeded the terrified animals into the walled killing enclosure, which was often concealed over the brow of a low hill.

According to recent archaeological work and travelers' accounts, the kite was operated in one of two ways, depending on its location and material. In Jordan the predominant material was basalt boulders from the lava that carpets much of the northern and eastern part of the country. There the killing enclosure has built into its stone walls small round hiding places shaped like the visible part of a water well. Svend W. Helms of the Institute of Archaeology of the University of London has suggested that in these niches hunters waited concealed to shoot the trapped gazelles with bows and arrows.

In Syria, where the ground is not covered by ancient lava flows, a different tack was taken: pitfalls were dug outside the main enclosure. J. L. Burckhardt, a 19th-century traveler in the Near East, saw kites with pitfalls being operated near the Syrian village of Al Qaryatayn. He recounts that in the wall of the enclosure were low places through which the frightened gazelles leaped, falling into the adjoining pits. Gazelles injured in the pitfalls were quickly killed and processed by the removal of the limb bones. The meat was then salted and dried for storage, and the surplus would undoubtedly have been consumed throughout much of the year.

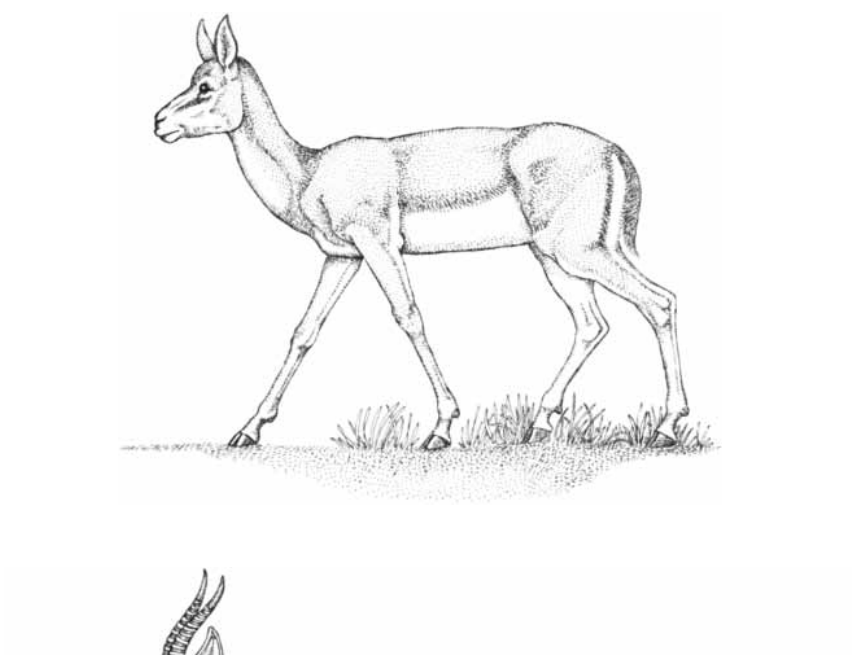
It seems evident that some such structure must have been employed at Tell Abu Hureyra during the millenniums when the great quantity of bone accumulated. Moreover, the kite must have been quite close to the village. Our collection includes a good sampling of all the bones of the body. The limb bones are not underrepresented as they would be if the kite had been far from the village. Historical records show that when the gazelles were partially processed at the site before transport, the limb bones were abandoned at the killing place.

As yet we have no direct evidence of the mass-killing structure. At the time the tell was excavated no archaeologist expected to find evidence of hunting on such a large scale in a community where the beginnings of agriculture were already present. Therefore the area was not surveyed for the remnants of a kite. All is not lost, however. Although the tell is flooded, much of the nearby land is not. On the basis of the terrain one can make a good guess about where the kite must have been. In the near future we intend to return to Abu Hureyra and try to find the kite's remains.

Even without the remains of the kite, it is possible to deduce much about how the mass killing fitted into larger ecological and social patterns. One such pattern is the seasonal movement of the gazelle herds. As the archaeological remains show, the gazelles were killed at Tell Abu Hureyra during a brief period in the early summer. That period is likely to mark the northernmost point in the herds' annual migration.

The Persian gazelle is remarkably well adapted to arid conditions. Indeed, for the most part the animal can get the moisture it needs by eating dry plants when the dew is on them. There is, however, at least one exception to the rule: to produce milk, females must have access to water or to green leaves, stems and shoots. To satisfy this need the gazelle herds moved north out of the desert during the spring. On reaching the somewhat moister environment near the Euphrates valley, the pregnant females gave birth. The herds spent a few months in their northern territory before heading south again in July.

The topography of the region, accounts provided by travelers and the distribution of desert kites suggest that the gazelle herds followed two main routes in their annual north-south migration. A (somewhat speculative) western route followed the valley extending northeast from Damascus to the district of Homs, which lies west of a low mountain range called Jabal



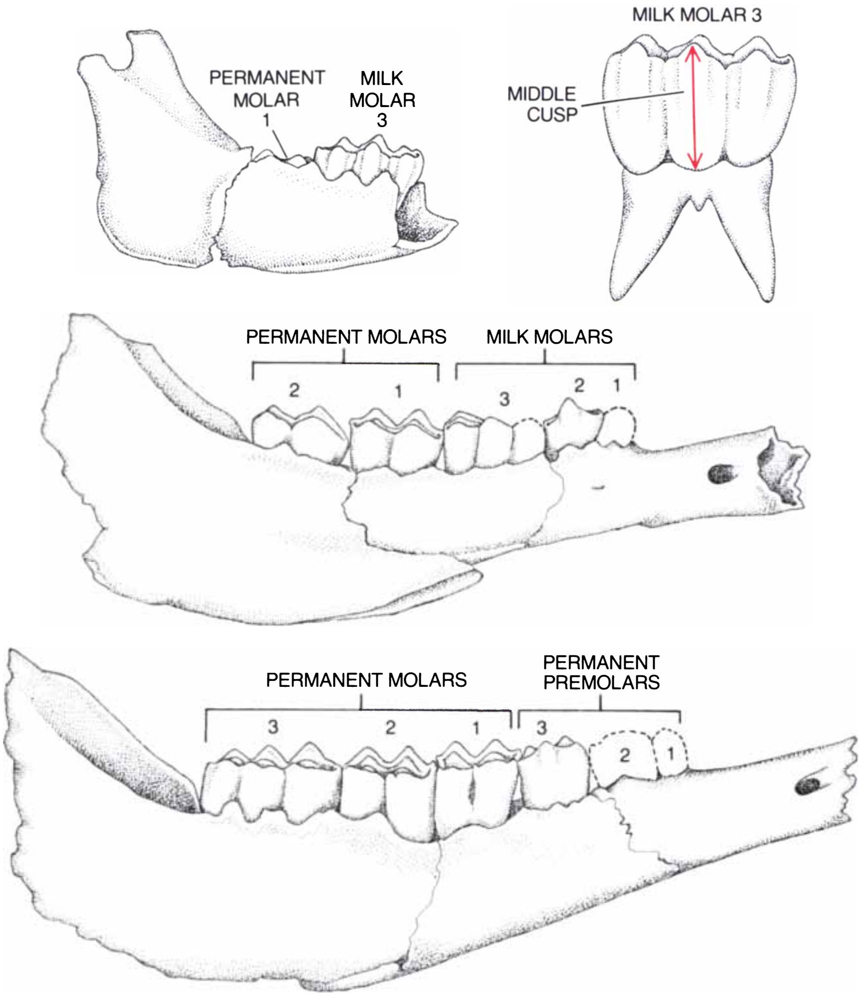
PERSIAN GAZELLE (*Gazella subgutturosa*) was the main prey of the hunters at Tell Abu Hureyra. The male (*bottom*) stands 60 centimeters at the shoulder. The Persian gazelle is now extinct in the wild in Jordan and Syria as the result of modern overhunting.

Abu Rujmayn. The kites seen by Burkhardt lie along this route.

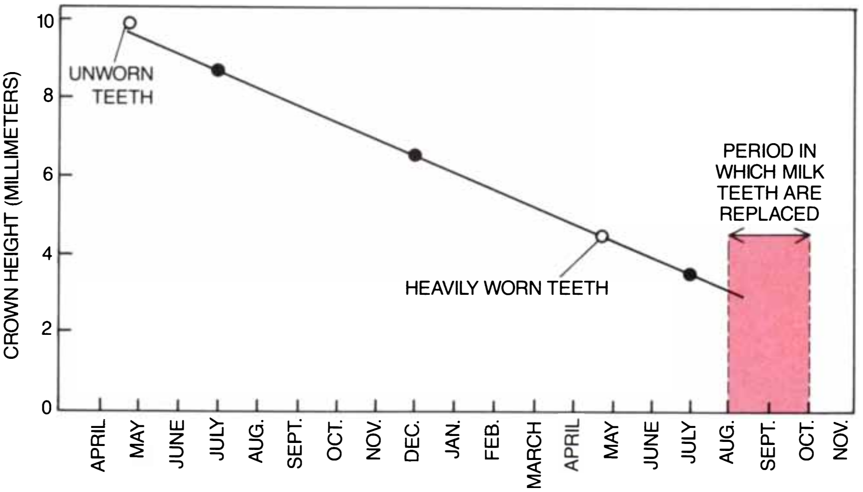
The other and more definitive route lies to the east of Jabal Abu Rujmayn. Between that range and the adjoining one called Jabal Bishri lies a lowland opening that was once exploited by caravans on their way from Aleppo to the Persian Gulf. As historic accounts make clear, the gazelle herds also used this passage. Just to the north of the pass lies the village of At Taybe. There in 1604 the Portuguese traveler Pedro Teixeira saw kites being used for hunting gazelles. The kites were different from the ones seen elsewhere in Syria or in Jordan. Instead of stone training walls they had rows of rag pennants hung on slender poles. The gazelles, frightened by the motion of the pennants, were driven between the poles toward the enclosure.

In spite of the difference in the form of the training walls, the kites seen by Teixeira were operated in much the same manner as the stone-walled variety were, and the two types coexisted in the same region. A survey in the 1930's by Père A. Poidebard, a pioneer of aerial photography in Syria, revealed stone-walled kites to the south of At Taybe.

This concentration was intriguing to us because it is the group of desert kites closest to Abu Hureyra. Its presence confirms that mass killing was a subsistence strategy practiced in the region of Abu Hureyra, at least in historic times. Although the gazelle herds were hunted at many points along their migratory route, the northernmost point where they could have been systematically killed was Tell Abu Hureyra. After reaching that point the herds broke up into small groups for the summer, and mass-killing methods would no longer have served until the herds reassembled in July to move south.



GAZELLE TEETH yield clues to the age of animals killed at Tell Abu Hureyra. The three jawbones are from the mound. At birth the animal has three "milk" (temporary) molars. The upper jawbone, from a newborn, retains the third milk molar; behind it the first permanent molar is visible. During the first year permanent molars come in. The middle jawbone, from a yearling, has three milk molars and two permanent ones. At about 14 months the milk molars are replaced by permanent premolars. The bottom specimen is from an adult. It has three permanent premolars and three permanent molars.



WEAR ON THIRD MILK MOLAR shows when the gazelles were hunted. During the animal's first year the crown of the third milk molar is steadily worn down. The rate of wear was computed by measurements of modern gazelle skulls (*solid circles*). The third milk molars from Tell Abu Hureyra (*open circles*) come from newborns or from yearlings. Hence the gazelles must have been killed in April and May, when the young were born.

The picture pieced together so far is quite informative about the preagricultural subsistence strategy at Tell Abu Hureyra and also about the rise of agriculture there. If only the mass killing were considered, one might assume Tell Abu Hureyra was a seasonal camp for gazelle hunting. Other evidence, however, makes it clear that the site was permanently occupied from the beginning. Not the least of such evidence is the huge mound, whose scale strongly implies settled village life. In addition, plant remains from the Mesolithic levels of the tell indicate that a wide variety of potential plant foods were being collected, and the collecting seasons of these plants cover most, if not all, of the year.

Were elements of agriculture already present at Abu Hureyra during Mesolithic times? It is not impossible that they were. Among the plant remains Hillman is analyzing are charred seeds of wheat. Yet the wheat is a wild form of the type called einkorn, and there is no direct evidence for its cultivation. Stronger evidence comes from the finding of goat and sheep bones. Radiocarbon accelerator dating carried out at the University of Oxford has confirmed that some of the goat and sheep bones can be dated to the Mesolithic. But were these wild animals, hunted for food as the gazelles were, or domesticates, maintained for clothing and milk as well as for meat?

That is a complex question. Today no wild sheep or goats are found on the lowland steppe in the vicinity of Abu Hureyra. Wild sheep and goats shun such sites in favor of higher altitudes and, particularly in the case of goats, steeper terrain. Yet such highland zones may be the only places where these animals can survive the ecological pressures humanity exerts. In earlier epochs wild sheep and goats may have lived near Abu Hureyra. In any event, it is not easy to distinguish the bones of wild sheep and goats from those of their domesticated descendants merely by size and shape.

Yet if one argues that wild sheep and goats were present in the lowlands near Abu Hureyra, it must also be noted that these species are rare at other sites of the Near Eastern Mesolithic. There are no other tell sites excavated by recent methods where the earliest layers are as old as those at Abu Hureyra. Earlier sites in the region (dating from 50,000 to 10,000 B.C.) are small and lack the tell structure characterizing later sites such as Tell Abu Hureyra. Sheep and goats are rarely found at the early sites, which have yielded only a few hundred sheep bones among them. It seems likely that something unusual was going on at Abu Hureyra, perhaps entailing early domestication techniques.

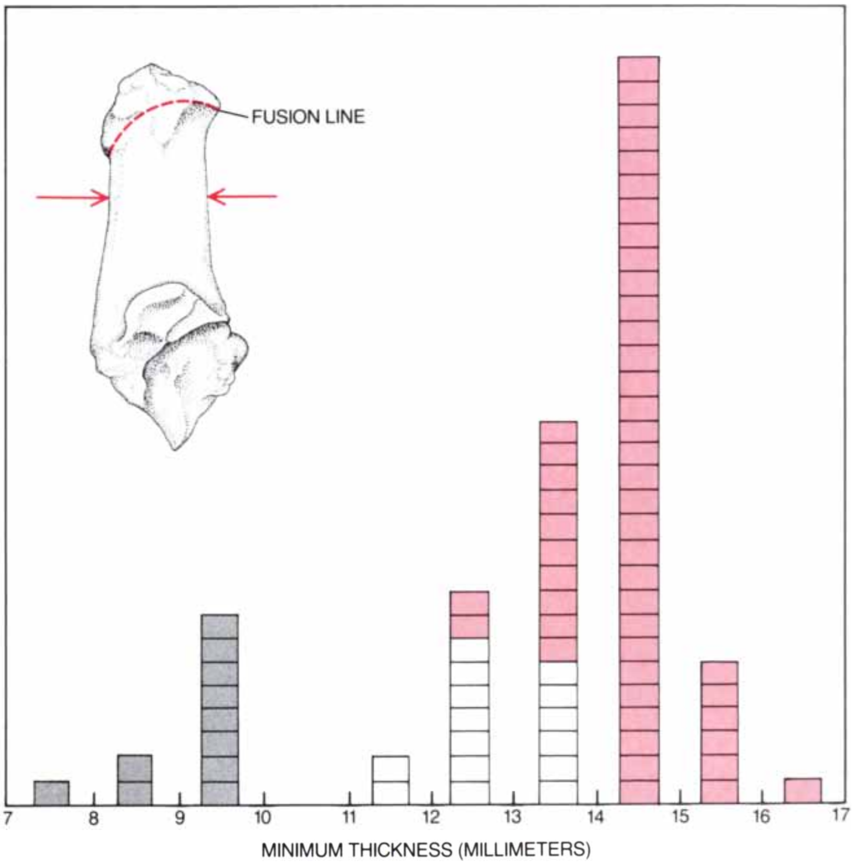
Even if the sheep and goats were domesticated, however, their significance in the overall animal economy of the site during the Mesolithic was relatively small. Together sheep and goats provide only 10 percent of the bones found in Mesolithic levels; as we noted above, gazelles provide 80 percent. Remarkably, that pattern continues into the early Neolithic. Seasonal killing of gazelles continues, and their bones still make up between 70 and 80 percent of the bone at the site. Yet this was a time when plant cultivation was already widespread in the Near East and, as Hillman has shown, large-scale cultivation of plants was under way at Tell Abu Hureyra.

Not only was the vegetable aspect of agriculture in operation but also its animal side seems to have been present. In the early Neolithic the samples of sheep and goats are large enough for the crown heights of their third milk molars to be measured in the same way as those of the gazelles. Such measurements show that, unlike the gazelles, sheep and goats were not killed seasonally. On the contrary, they were slaughtered throughout the year. This is what one would expect of a farming settlement where sheep and goats are domesticated on a permanent and systematic basis. Nevertheless, the importance of these species remains low, and the gazelle continues to be predominant.

This is interesting and perhaps puzzling as well. Why should sheep and goats—which are strongly associated with early domestication—retain a minor role long after agriculture has taken hold? One reason must have been the presence of abundant gazelle herds that were readily available on an annual basis. If a year's supply of meat can be obtained within a few weeks each spring, there would seem to be little reason to take up the arduous practice of large-scale herding.

This favorable situation did not last forever. Sometime in the seventh millennium B.C. the animal economy at Tell Abu Hureyra underwent an abrupt and profound change: gazelles exchanged positions of relative importance with sheep and goats. In a brief period gazelles declined until they contributed only 20 percent of the bones at the site. In the same period sheep and goats increased until they provided the same proportion the gazelles once had: 80 percent.

It appears this dramatic shift was linked to events far to the south of the tell. It is probable that when the taking of gazelle herds began at Abu Hureyra, desert kites were not widespread in Syria and Jordan. Indeed, few archaeological sites are contemporary with the settlement of Tell Abu Hureyra in 9000 B.C. during Mesolithic times. Midway through the Neolithic period,



GROWTH OF CALCANEUM, or heel bone, confirms seasonal killing. During the gazelle's first year the calcaneum (*inset*) thickens considerably. At about 14 months the process (an extension where tendons attach) fuses completely. Calcaneums from Tell Abu Hureyra are of three types. Those from newborns (*gray*) are thin, with unfused processes. Bones from yearlings are thicker but unfused (*white*). Adult calcaneums are thick; the process is fused (*color*). Such grouping implies seasonal rather than ongoing hunting.

however, kites had become far more widespread.

Early aerial surveys, combined with recent fieldwork by Alison Betts and her colleague Helms, reveal an extensive concentration of kites in northern Jordan. In addition to being very densely concentrated the kites are linked: the training walls of one kite overlap those of the next, forming great chains that can extend for tens of kilometers over the plain. In this arrangement the kites must have devastated the gazelle herds.

Such sites are difficult to date, since they stand on rocky ground where sediments—a prime tool of dating—accumulate slowly. There is, however, evidence that some of these kites were built at about the time the residents of Abu Hureyra switched from gazelles to sheep and goats. Betts has described flint tools of roughly that date from kites in eastern Jordan. We have seen similar material from kites near Azrak, an oasis 60 kilometers east of Amman, at sites surveyed by A. N. Garrard of the British Institute at Amman for Archaeology and History. Neolithic stone tools have also been reported at kites in northern Saudi Arabia. The tools frequently lie only in the kites; sometimes they are even found in the hiding places where hunters may have stood.

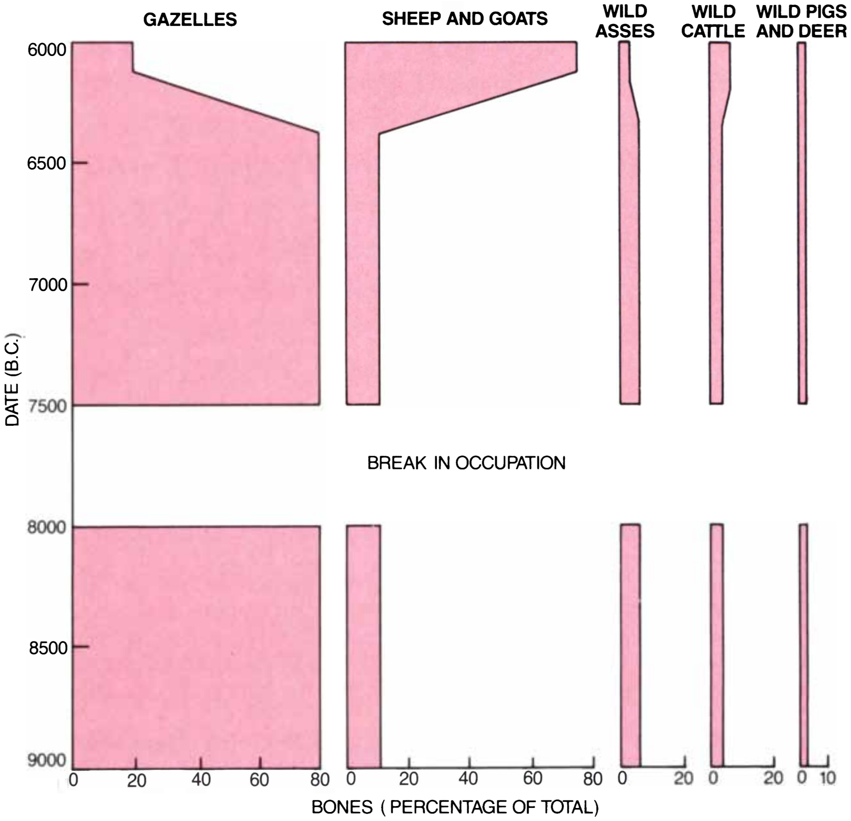
This array of evidence suggests that in the seventh millennium B.C. desert kites spread rapidly across the steppe in northern Jordan. Of course, the taking of any herd through mass kills may lead to overkilling, and it seems that this is what happened. In the Mesolithic and the earliest part of the Neolithic the people of Tell Abu Hureyra were one of the few groups hunting gazelles by mass-killing methods. Somewhat later in the Neolithic the extension of kites into northern Jordan may have put too much pressure on the herds. The result may have been a decrease in the number of gazelles and perhaps even a disruption of the overall migratory pattern. The reduced availability of gazelles, in turn, could have forced the community at Abu Hureyra to fall back on husbandry of sheep and goats.

Thus after a millennium-long delay large-scale animal husbandry emerged to complement plant cultivation, and the village of Tell Abu Hureyra entered the phase of full-blown agriculture. One of the most notable things about the site is the level of detail at which this process can be followed. Such detail is attributable largely to two factors. One factor is the enormous wealth of organic remains, both plant and animal. The other is the continuity offered by the presence of an agricultural community on the same site as its predecessor hunting-gathering village.

It is no accident that both types of village were on that spot. The location was clearly favored by circumstances that encouraged large-scale permanent settlement very early. The plants of the river valley combined with the animals of the steppe to provide a stable year-round food supply. In the long run the stability of this hunting-gathering economy was punctuated twice by rapid change. The first change was the introduction of plant cultivation. Then, 1,000 or more years later, herded animals equally abruptly replaced the ones that had been pursued across the steppe.

Several lessons can be drawn from this history. One lesson is that the inception of an agricultural way of life need not be a unitary process. A concept much discussed in archaeology in recent decades is that of the Neolithic Revolution, which entails the establishment of permanent settlements and the origin of agriculture. The idea of a revolution suggests a sweeping change coming all at once. The prehistory of Tell Abu Hureyra, however, suggests not a single, sweeping change but a stepwise process in which key elements (a sedentary way of life, cultivation and animal husbandry) come into being one at a time over an extended period.

There may also be another way in which the revolutionary picture is deficient. One of its underlying assumptions is that agriculture is inherently superior in productive power to hunting-gathering modes. In the long run that is undoubtedly true. Yet in a particular locale during the period when agriculture emerges, the old methods may be superior. At Abu Hureyra the ease of mass killing clearly delayed the introduction of herding on a large scale until the rewards of hunting were reduced by distant events. Domestication does not simply take over and replace old systems immediately. Instead it runs parallel with them until it is drawn on to meet new demands. A great contribution of Tell Abu Hureyra is that it has helped archaeologists to understand this complex and fundamental process.



BALANCE OF ANIMAL SPECIES at Tell Abu Hureyra reveals a dramatic shift in the seventh millennium B.C. The width of the bars indicates the fraction of bones accounted for by the main species at the site. From 9000 B.C. until shortly after 6500 B.C., 80 percent of the bones come from gazelles and 10 percent from sheep and goats. Then there is a rapid reversal: sheep and goats increase to more than 60 percent, whereas gazelles fall to 20 percent. The shift may have been caused by a depletion of the gazelle herds.

Air Pollution by Particles

Acidic particles in the atmosphere are known to reduce visibility and damage materials. Ingenious methods have now demonstrated that the main source of the particles is the combustion of fossil fuels

by Robert W. Shaw

In many parts of the world the façades of architectural treasures are eroding; soil and lakes are becoming abnormally acidic, causing vegetation and fish populations to dwindle or disappear. Much of the damage is now attributed to acid rain: rain or snow carrying dissolved acids.

Yet wet precipitation is not the only way pollutants reach the earth from the atmosphere; diffusion and other processes enable acidic gases and particles to find their way to the ground even under dry conditions. Many environmental scientists suspect that dry deposits are as destructive to materials and the environment as tainted rain or snow, and they suggest that an improved definition of "acid rain" would include both wet and dry pollution.

Dry deposits are less widely discussed than acid rain as such because for a long time less was known about them. In the past 10 years, however, many investigators have uncovered striking new information about the sources and possible effects of atmospheric particles. Indeed studies tracing particle samples to their sources have helped to quash the notion that natural emissions from swamps, volcanoes or trees might be responsible for much of the acid fallout around the globe. It is now beyond question that even in rural areas acid deposition (both wet and dry) almost always stems from the activity of human beings: primarily the combustion of fuel for power, industry and transportation.

When coal and certain other fuels burn, they emit many substances, including carbon particles (if combustion is inefficient) and sulfur dioxide (SO_2) gas. In addition the high temperatures of combustion cause nitrogen in the air to combine with oxygen, yielding nitrogen oxide (NO_X) gases. When the gases encounter water or related molecules in the atmosphere, they form sulfuric acid (H_2SO_4) droplets and nitric acid (HNO_3) gas, both of which are readily dissolved in earthbound rain. If the atmosphere is relatively dry, nitric acid tends to remain in the gaseous state but sulfuric acid tends to form minute particles. These bits sometimes reach the ground in rain, but they and other particles often settle out of the air on their own.

Early studies of atmospheric particles provided little insight into the nature or causes of acidic deposition because the bits were difficult to measure and analyze. Rain captured in a collection bucket generally reflects the amount and composition of precipitation in an area, but unless airborne particles are gathered with specially designed equipment the resulting samples may not yield similarly reliable information. Moreover, the particles often consist of a mixture of substances, making it quite a challenge to identify their components and hence their sources.

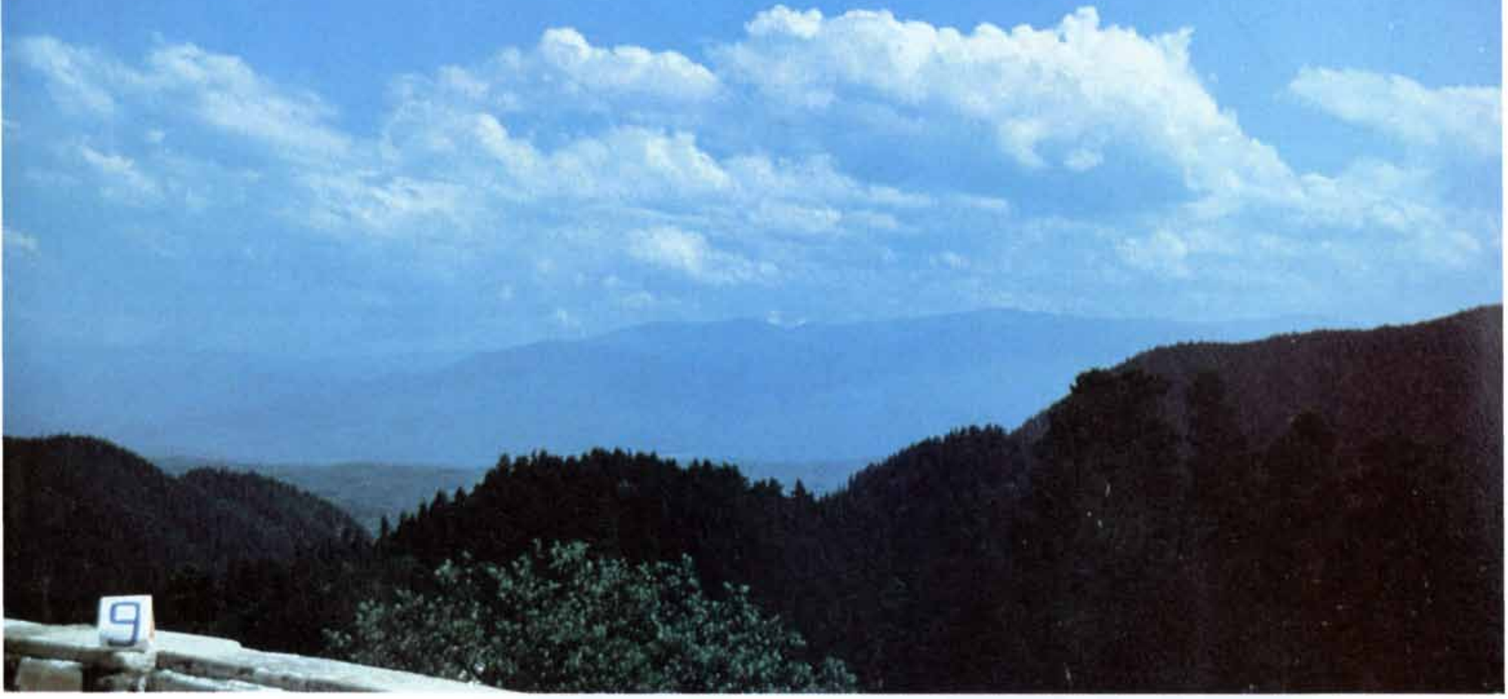
In spite of such obstacles the U.S. Environmental Protection Agency determined during the late 1970's that atmospheric particles were important enough to warrant a renewed effort at collection and analysis. In addition to soiling buildings and otherwise damaging materials, they were thought to contribute to the haze that reduces visibility over large areas of the eastern U.S. in the summer. With the help of investigators involved in instrument design, analytic chemistry and statistics, a group of us at the EPA therefore set about developing an arsenal of compatible devices and techniques that would enable us to gather airborne particles efficiently, determine their composition and make informed inferences about the materials giving rise to them.

In order to collect multiple samples reproducibly we needed an automated collector. The machine would have to capture samples in a way that reflected the total concentration of atmospheric particles in an area. It would also have to divide its "catch" into two groups—fine particles (smaller than one micrometer in diameter) and coarser ones—because earlier studies had demonstrated that the two sizes were qualitatively different. (A micrometer is a thousandth of a millimeter.)

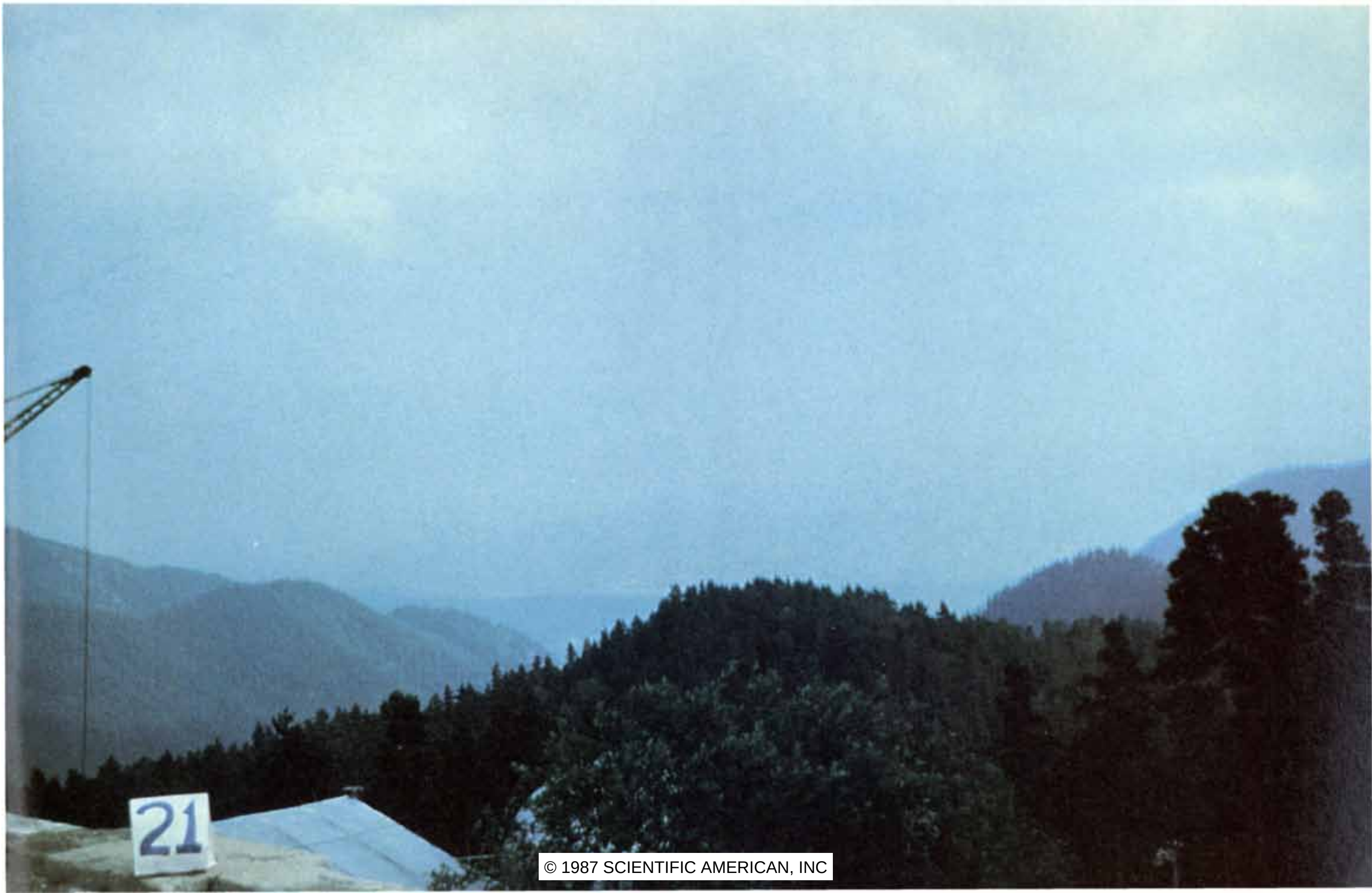
The realization that particles of different sizes might have significantly different natures emerged from statistical manipulations of data. When workers first attempted to study dry deposition, they often counted the particles in various size intervals. They found that the number increased as diameter decreased. The pattern was interesting, but it revealed little about the nature of the samples.

Then about 20 years ago investigators studying the smog in Los Angeles decided to look at the total mass of all the particles in different size ranges. In every one of hundreds of samples they found an apparently natural division: particles with diameters in the range of from .1 to one micrometer (fine particles) accounted for a large fraction of the mass, and those with diameters in the range of from one to 100 micrometers (coarse particles) accounted for the rest of the mass [*see top illustration on page 98*]. This consistent finding strongly suggested that fine and coarse particles not only are different in size

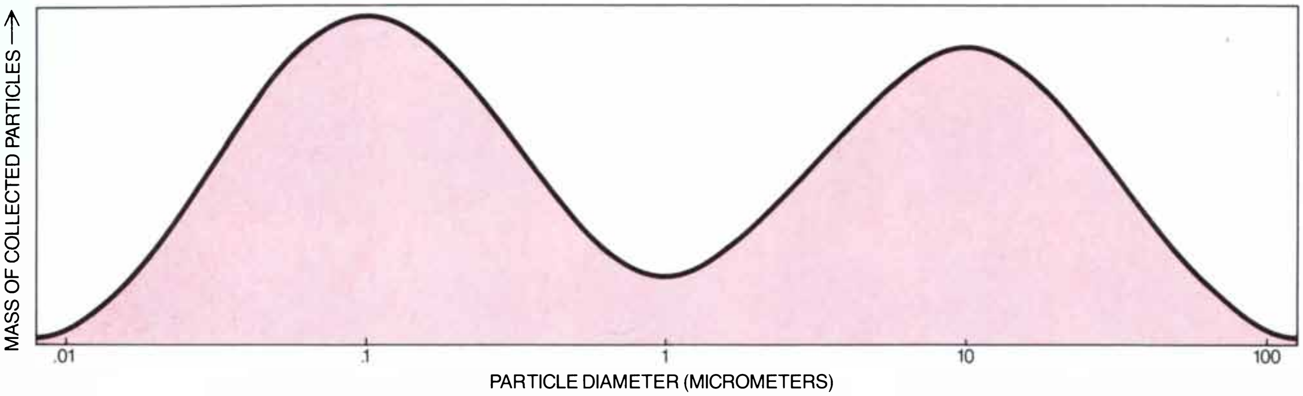
MOUNTAIN AIR of Soviet Georgia was significantly clearer on July 9, 1979 (*top*), than on July 21 (*bottom*). Increased haze and reduced visibility late in the month stemmed in part from a marked rise in the concentration of atmospheric particles containing sulfate (SO_4^{2-}). The size of these often acidic particles makes them effective scatterers of light. The presence of sulfate, which is rarely produced by natural processes, was a surprise. It meant the air in the region was more polluted by distant industry than had been thought.



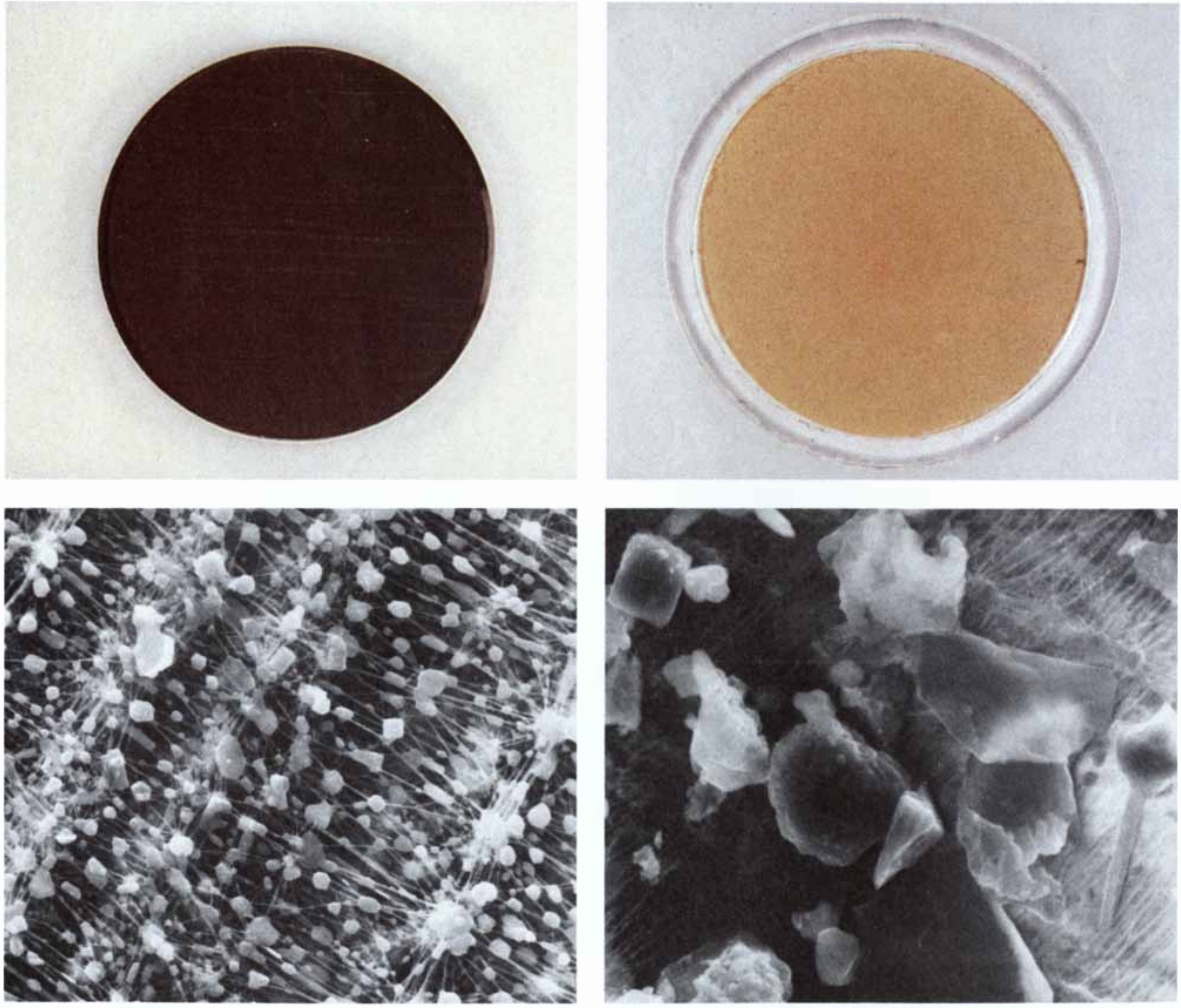
9



21



TWO-HUMPED CURVE is characteristic of atmospheric particles worldwide. It emerges when the diameter of sampled particles is plotted along a horizontal axis (on a logarithmic scale) and a curve is constructed above the axis so that the area under the curve is proportional to the mass of the collected bits within any size interval. The two humps reflect the fact that particles smaller than one micrometer in diameter differ qualitatively from larger ones. The fine bits typically are acidic products of combustion and are formed by accretion. The coarse chunks generally are nonacidic and form by the mechanical breakdown of natural substances.



FINE AND COARSE SAMPLES collected from the air often have a strikingly different appearance. In a specimen gathered at a construction site in North Carolina (*top*) the fine fraction (*left*), which is largely a product of diesel fuel burned in earthmovers, is black. The coarse fraction (*right*), which consists mostly of clay churned up by the wheels of the vehicles, is orange. Similarly, when samples collected from urban air are viewed with an electron microscope (*bottom*), the small particles (*left*) appear smooth and even but the coarse ones (*right*) are jagged. At larger magnifications the shape of a particle may suggest its origin. David L. Johnson of the State University of New York College of Environmental Science and Forestry in Syracuse made the micrographs.

but also constitute different classes of materials.

Many studies have since confirmed that this is so. The work has shown that the fine particles in the atmosphere evolve primarily from chemical processes, notably combustion reactions, and are essentially acidic. They therefore require the most intensive scrutiny in many analyses. In contrast, the coarse particles derive mainly from the mechanical breakup of naturally occurring materials, such as soil, and they are dominated by nonacidic substances. The processes governing the growth of a particle in the atmosphere are not thoroughly understood, but they appear to be self-limiting: growth ceases when the diameter approaches one micrometer. Mechanical processes, on the other hand, generally cannot break substances into bits smaller than one micrometer in diameter.

The challenge of both collecting a representative sample and dividing it into two groups was solved by inventing a machine known as a virtual impactor, which exploits the different abilities of coarse and fine particles to change direction when they are in motion. An airstream draws particles from the environment into the machine. Then the stream forks. Part of it continues to flow straight; the rest essentially turns a corner, moving 90 degrees to the side before flowing in the original direction again.

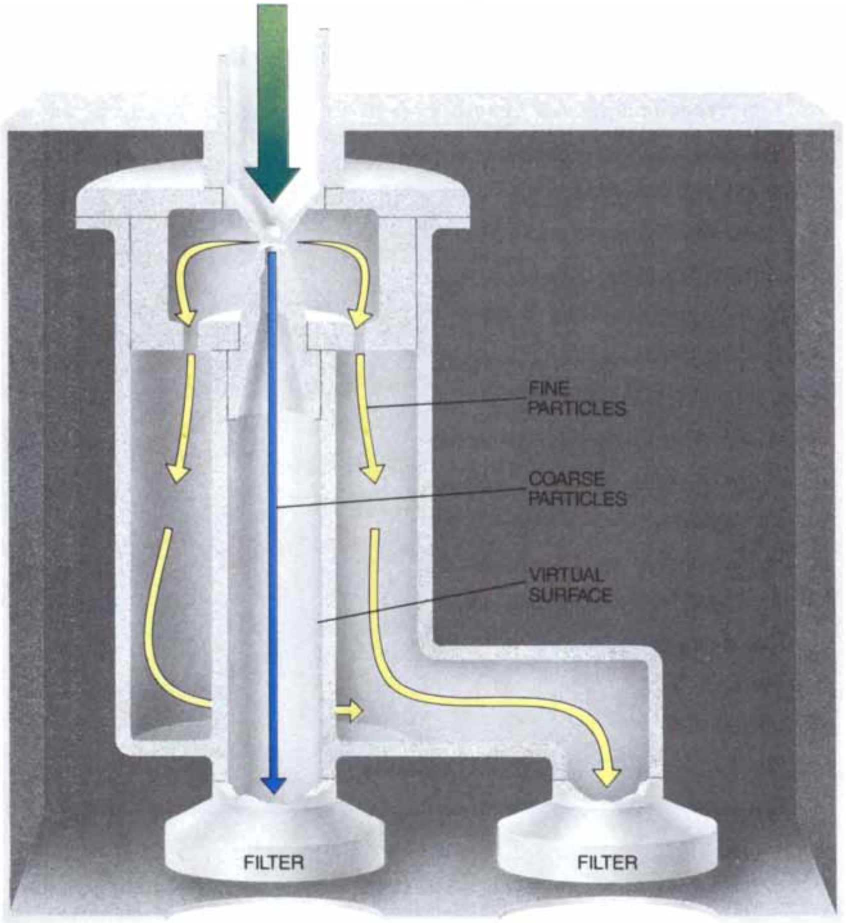
The larger chunks trapped in the airstream are too heavy to make the sideways detour, and so they flow with the straight part of the stream into a chamber (called a virtual surface). In the chamber their movement is slowed so that they can drift gently onto a Teflon collection filter and distribute themselves evenly across the surface. (In earlier devices, fast-moving particles often bounced off filters, distorting the final measurements.)

The finer particles, in contrast, can make abrupt changes in direction. They follow the curving current and gather uniformly on a filter in another part of the impactor. Automated and loaded with magazines of collection filters, virtual impactors can operate unattended for a month or more, accumulating samples on a schedule chosen by the operator. We enlisted Teflon for the filters because glass fiber, the earlier material of choice, reacts with gases in a way that creates new particles—and misleading results. Teflon does not react with atmospheric gases.

After a sample is gathered its mass must be determined so that the concentrations of individual substances in the atmosphere can be calculated. The mass can be measured directly by weighing the filter before and after a sample is collected, but this process is too cumbersome to perform repeatedly. We therefore decided to take advantage of the beta gauge, a tool borrowed from nuclear physics. The filter and sample are placed between an emitter and a detector of beta particles, which are electrons that are released by certain radioactive materials. The gauge determines the mass by counting the number of electrons that pass through the filter. The larger the mass of a sample is, the smaller the number of electrons that reaches the detector will be.

The next routinely applied test, X-ray fluorescence spectroscopy, is also borrowed from nuclear physics. In spectroscopy we pass X rays rather than beta particles through the sample and thereby identify its atomic elements. The radiation stimulates the specimen to emit X rays of varying energy levels. Each energy level is characteristic of a particular element in the sample, and the intensity of the radiation (the number of rays emitted) is a measure of the element's concentration. The method, one of several we could have chosen, is highly sensitive and can measure many elements at once. Like the beta gauge, it leaves the particles essentially unchanged for future tests.

X-ray fluorescence spectroscopy is quick and easy to do and so is employed routinely, but it provides only a rough guide to the composition of the captured sample. Although it indicates which elements are present, it cannot specify, say, that the "molecular species" of the sulfur is sulfate (SO_4^{2-}). Nor can it indicate that specific elements are combined in the same particles. Such information is important because the chemical state of an element affects both its behavior and its impact on the environment; it also helps to



COLLECTION DEVICE known as a virtual impactor captures atmospheric particles in an airstream and then separates them into two groups by size. The bigger particles, which are relatively heavy, cannot change course when part of the stream detours to the side; instead they flow straight down into a chamber (a "virtual surface"), where their movement is slowed so that they will not bounce when they reach a waiting collection filter. Most of the smaller particles follow the detoured stream, coming to rest on a filter of their own.

suggest the element's material source.

In order to obtain added specificity, other analyses—some of them destructive of the sample and some not—must be done. Which tests a team chooses depends on the objectives of the study and the funds available. Nondestructive approaches we found to be valuable are optical and electron microscopy, which are time-consuming but can reveal the species of particles that have known morphologies.

In recent years workers have sometimes also done X-ray-diffraction analyses of samples. The scattering of X rays by a crystalline material reveals its three-dimensional structure, which in turn may indicate the material's source. For instance, the technique has revealed the presence in certain specimens of mullite: a variety of aluminum silicate that forms only at high temperature, such as during the combustion of coal. This material has a distinct crystal structure that differentiates it from other forms of aluminum silicate found in soils and clays.

These nondestructive methods are valuable, but investigators generally have to analyze a sample chemically in order to determine the state of its constituent elements. For example, dissolving a specimen in water causes ions to dissociate, making it possible to measure the concentration of hydrogen ions and so determine the overall acidity of the sample. (By definition an acid is a substance that releases hydrogen ions in solution.) We also depend on this approach to obtain such information as whether nitrogen is present as nitrate (NO_3^-) or as ammonium (NH_4^+), or if sulfur is present in the form of sulfate.

These direct manipulations sometimes reveal the sources of certain particles, but workers generally have to rely on complex statistical analysis and a certain amount of subjective judgment to pin down such sources. Particularly informative methods include factor analysis, which was developed by psychologists in the early decades of this century, and "chemical-element balance," which was developed for the study of atmospheric particles in the early 1970's. Factor analysis helps to identify the likely material sources of the elements in a sample, and chemical-element balance indicates the relative contribution of each source to the sample.

In factor analysis a computer scans all the compositional data relating to samples collected at different times. It then identifies those elements whose concentrations appear to increase or decrease in synchrony with one another. Elements that vary together may have the same origin (and probably are present together in particles). To suggest the sources of the related elements, the computer then compares the elements that are clustered statistically with a list of possible sources and their "footprints": the element combinations they typically emit.

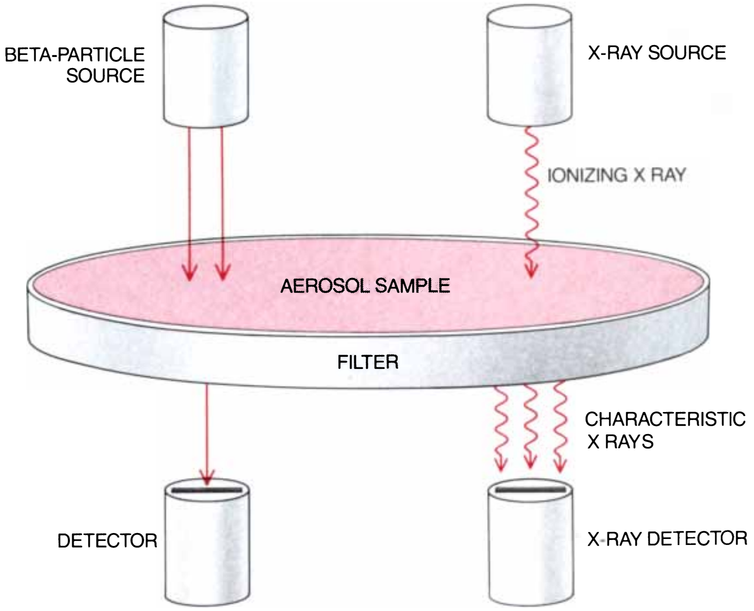
Imagine that the concentrations of the elements selenium and sulfur are measured on four consecutive days and that both substances double on the second day, triple on the third and fall to half of their original levels on the fourth. The finding that the elements vary together would imply they might derive from the same material. A scan of a source list would then reveal that selenium and sulfur are components of coal, indicating that coal may account for the presence in the sample of both elements.

In doing an analysis of chemical-element balance workers compile a list showing the typical ratios of elements produced by various putative sources. Such a list might indicate, for instance, that the ratio of selenium to sulfur in the coal of a region is one to 1,000. Having measured the concentrations of selected elements in a specific sample, the workers might then predict, on the basis of the standard ratios in the source list, what the concentrations of other elements in the sample should be if they are from the same source.

If a particle sample is found to include .01 microgram of selenium and the source list indicates that selenium is a trace element in coal, the investigators might hypothesize that all the selenium comes from coal and that the amount of coal-derived sulfur in the sample should be 10 micrograms. If this is in fact the amount of sulfur in the sample, the workers might conclude that all the selenium and all the sulfur probably came from coal.

In practice, of course, the elements could be produced by various sources. Chemical-element balance copes with this complication by positing (on the basis of the known ratios) several different elemental concentrations that could be generated by the suspected sources. It then finds the "mix" that best fits the actual concentrations in the collected samples. Such manipulations might reveal, for instance, that 80 percent of the fine fraction of a sample is a by-product of coal combustion and 15 percent comes from motor vehicles. Or it might show that 94 percent of the lead in a specimen is from motor vehicles, 4 percent from the burning of refuse and 1 percent from the burning of coal.

With our new set of tools and techniques in hand, our group at the Environmental Protection Agency established a temporary laboratory sev-



TWO TECHNIQUES make it possible to analyze particles without removing them from collection filters. To determine the mass of a sample (*left*) workers insert the particle-laden filter between a source that emits beta particles and a detector that counts them. As the mass increases, the number of particles that can penetrate the sample decreases. To determine the atomic elements in a specimen workers may also separately carry out X-ray fluorescence spectroscopy (*right*). X rays passed through the sample cause each element to emit characteristic X rays. The energy levels of the rays reveal the identity of the elements; the intensity of the rays (the number emitted) reflects the concentrations.

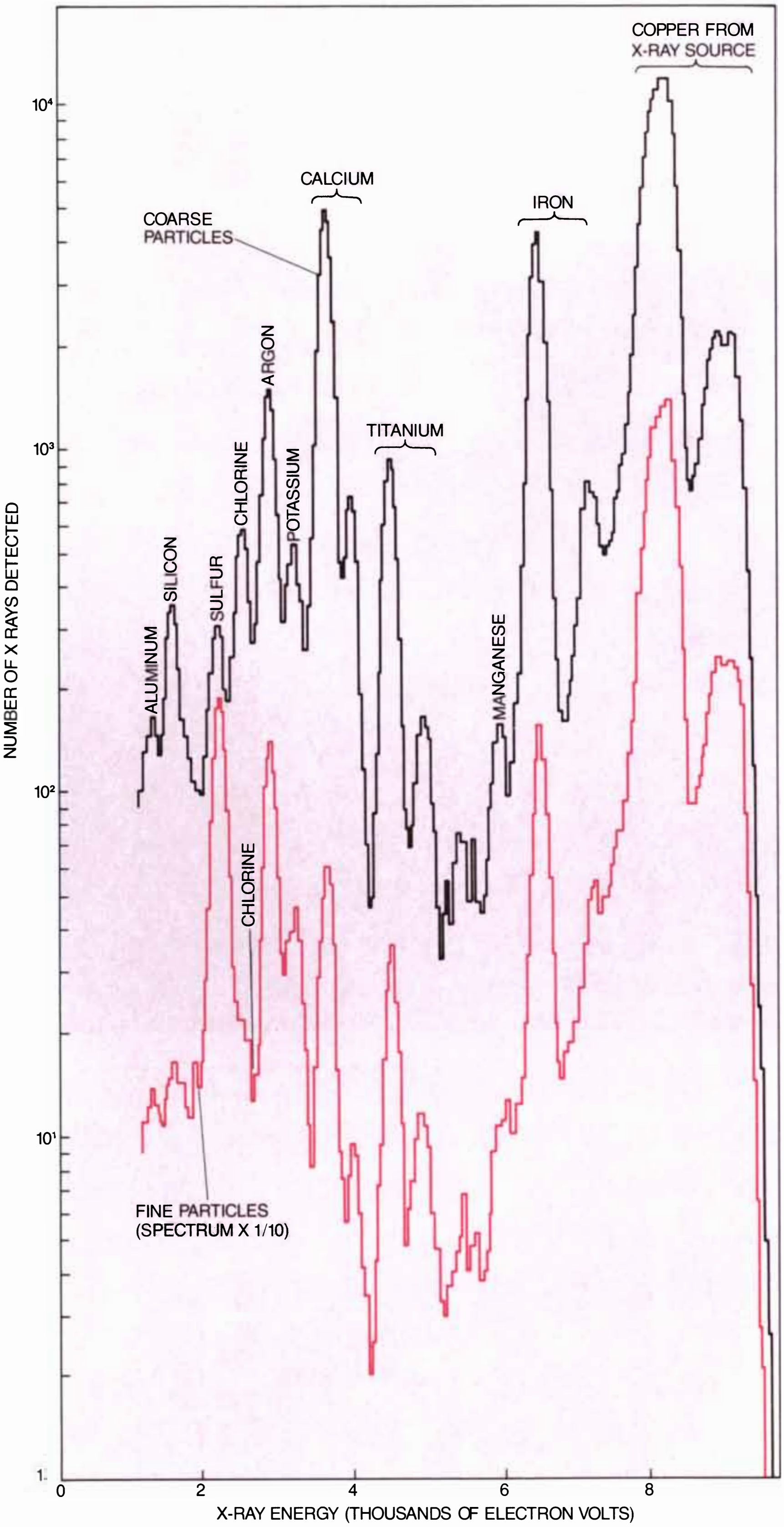
eral years ago in the Great Smoky Mountains National Park in Tennessee. Our mission, which was part of a much larger effort to study air quality in the park, was to determine the relative contributions of natural emissions (from vegetation) and emissions from motor vehicles and industry.

At the time many people believed the haze in the area was caused by fine particles produced when hydrocarbon gases emitted by evergreens reacted with other substances in the environment. X-ray fluorescence and chemical studies showed, however, that hydrocarbons contributed little to the mass of our samples and that acidic sulfate dominated. Indeed, the levels of sulfate were much higher than those typically found in pristine environments, forcing us to conclude that fuel combustion of some sort had affected the atmosphere in the park.

The finding in the Smoky Mountains caused much surprise, as did a similar study conducted under the auspices of a joint U.S.-Soviet environmental research program. Our goal in this instance was to study the formation of particles by chemicals emitted from evergreens. To minimize interference from pollution we traveled to a remote astronomical observatory in the mountains of Soviet Georgia, a region known for its clear air as well as for its impressive stretches of evergreen forests.

For a month our team and environmental scientists from the Soviet Union measured both gases and particles. The gas studies showed that only low levels of sulfur dioxide, nitrogen oxides and halocarbons (the class of gases that includes the ozone-depleting refrigerants) were present. Such a finding usually indicates clean air. The mixture of particles in the atmosphere was, however, remarkably like the one in the Smokies—mostly sulfate, with only a small fraction of hydrocarbons. In fact, we were not able to obtain enough hydrocarbons to do the study that had originally been planned.

Significant amounts of sulfate-containing particles in what were regarded as two pure areas raised the obvious question: What was their main source? We cast our vote for coal burned in distant regions because the by-products of combusted coal are usually released from high smokestacks. This practice minimizes air pollution in the immediate vicinity, but it also allows sulfur dioxide gas to remain in the atmosphere long enough to both travel and combine with other substances to form particles. Consistent with this theory, we have found that the sulfate particles in Soviet Georgia, an area extremely distant from industry, are

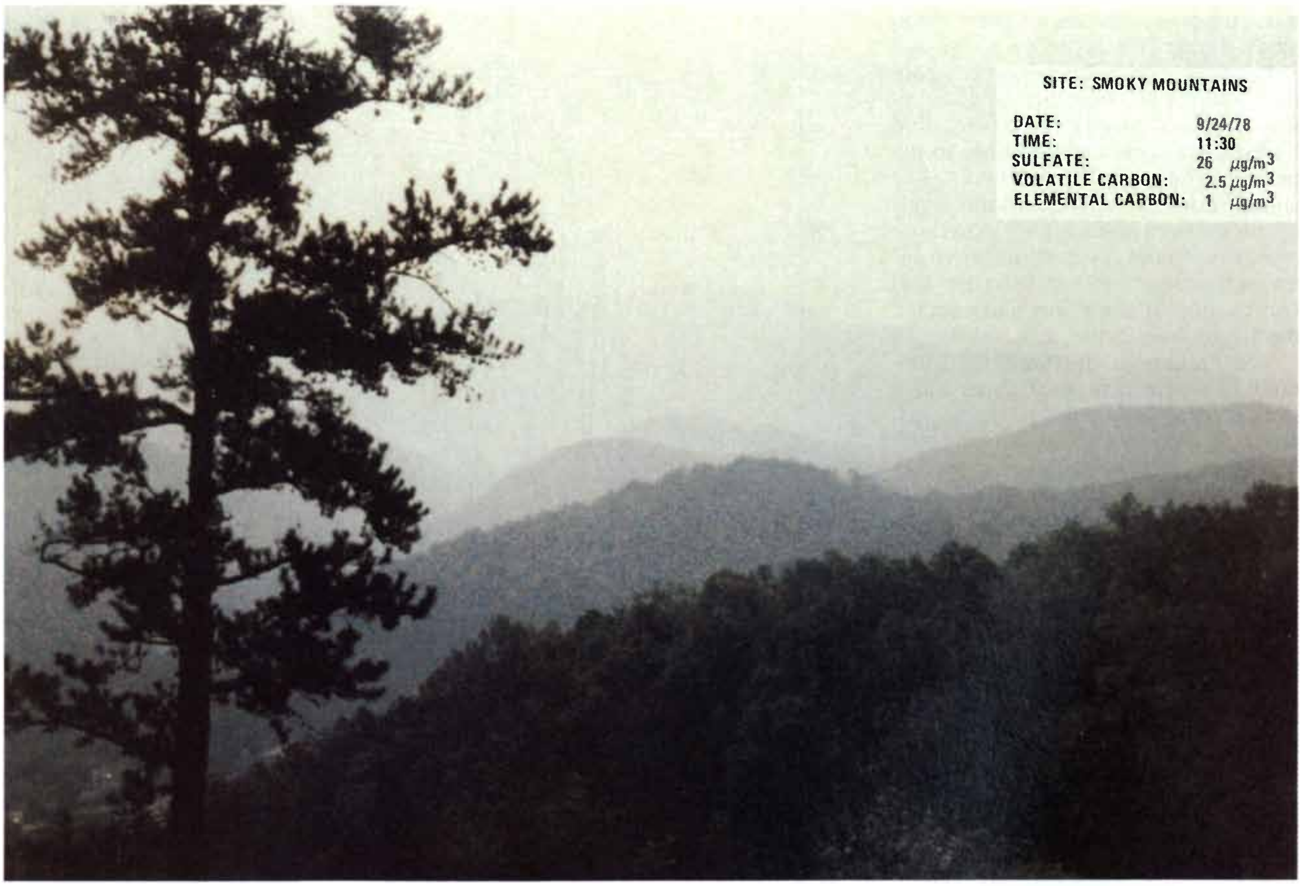


X-RAY SPECTRUM of particles collected in St. Louis by Thomas G. Dzubay of the U.S. Environmental Protection Agency shows that sulfur is the most abundant element in the fine fraction. Sulfur in fine particles typically derives from coal. (The number of X rays produced by each element in the fine fraction is determined by multiplying the value on the vertical axis by 10.) Silicon, calcium and iron, all typical components of soil dust, are major constituents of the coarse fraction. The argon peaks reflect air in the spectrometer.

older than those in the Smokies. In the eastern U.S. roughly a fifth of the sulfur in the atmosphere takes the form of particles; in Soviet Georgia there is 30 times more sulfur in particles than in sulfur dioxide gas.

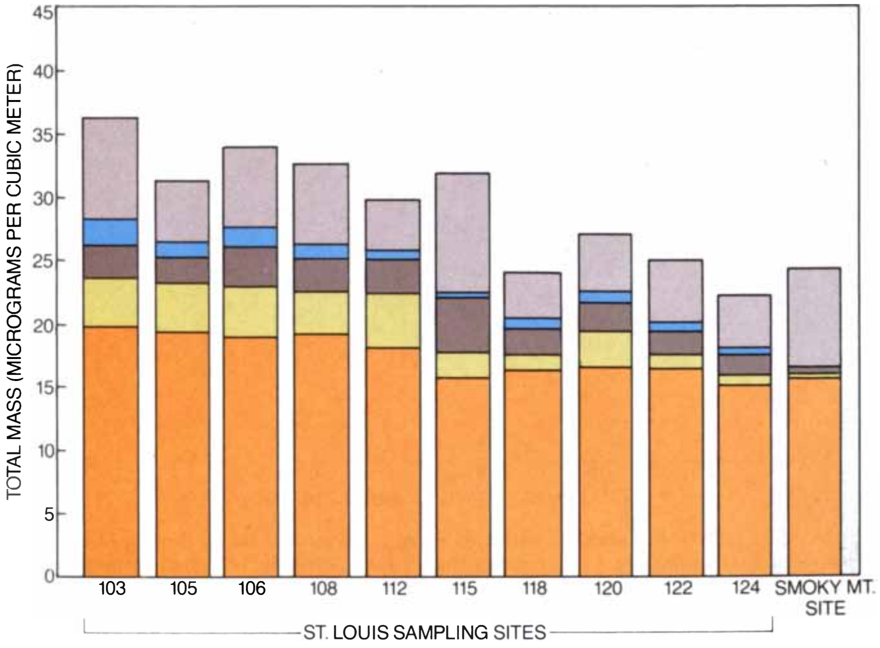
Although we suspected that coal was the major source of the sulfate in the Smoky Mountains and at our Soviet camp, we did not find a conclusive link between power plants and atmospheric particles until we did a 16-month study in the Ohio River valley. This region has the greatest concentration of coal-burning power plants in the U.S.

For that study we wanted to do a best-case analysis. We therefore established our sampling stations in rural areas of Kentucky, Indiana and Ohio—away from cities, roads and the smoke plumes of power plants. We found that on the average about 50 percent of the fine-particle mass was sulfate; the to-



SITE: SMOKY MOUNTAINS

DATE:	9/24/78
TIME:	11:30
SULFATE:	26 $\mu g/m^3$
VOLATILE CARBON:	2.5 $\mu g/m^3$
ELEMENTAL CARBON:	1 $\mu g/m^3$



HAZE over the Great Smoky Mountains (*top*), captured photographically by Robert K. Stevens of the EPA, is mostly due to sulfate particles in the air. A statistical test known as chemical-element balance, which suggests the sources of particle samples on the basis of their known components, indicates that much of the sulfate in the Smoky Mountains probably is a by-product of coal burned at distant power plants. The test also reveals (*bottom*) that the concentration of coal-derived particles in the Smokies (*bar at right*) is not much lower than the concentrations found at 10 sampling sites in St. Louis, a distinctly industrialized city. Dzubay collected the data.

tal concentration was nearly equal to what was found in industrialized cities in the region. In this instance we were able to link the sulfate to coal combustion definitively because we found a consistent association between the concentration of sulfate and the trace element selenium. No other probable sources of selenium exist in or near the Ohio River valley.

Sulfate is now known to contribute most of the fine-particle mass over much of the eastern U.S. and in many other regions. Because the sulfate-containing bits are often highly acidic, it is entirely possible that they also damage materials and alter the acid-base balance of ecosystems, much as acid rain does.

There are also some indications that such particles, because of their size and consequent light-scattering characteristics, may affect the temperature balance between the atmosphere and the earth's surface, causing perturbations in the weather. By coincidence, fine particles in the atmosphere are of just the right size to be highly effective at scattering light. Model calculations indicate that this scattering can both cool the surface of the earth (by preventing sunlight from reaching it) and heat the upper atmosphere. These alterations might in turn change the natural motions of air masses and encourage the formation of temperature inversions: patterns in which the temperature increases, rather than decreases, with increasing altitude. Inversions cause air (and the pollutants it contains) to stagnate.

Such effects have yet to be established with certainty, but another—and very tangible—effect has been proved: again by scattering light, high concentrations of sulfate-rich particles in the atmosphere diminish visibility. Proof that sulfate contributes to poor visibility comes from many sources, including our studies in Soviet Georgia and the Ohio River valley. An analysis of visibility levels and coal consumption in the U.S. shows that visibility has decreased significantly as coal consumption has increased. (Visibility has also decreased in many areas around the world, including the Arctic.)

In Soviet Georgia we found that visibility decreased markedly whenever the concentration of sulfate particles increased. In the Ohio River valley the fraction of sulfur that appears in particulate form rises dramatically from winter to summer, in parallel with an increase in the incidence of haze. Sulfate concentrations probably increase in the summer because there are longer periods of daylight and higher temperatures and humidity. These conditions increase the rate at which sulfur dioxide gas converts into sulfate particles. At the same time more frequent episodes of air stagnation in summer allow pollutants in the atmosphere to accumulate.

Sulfate is the dominant component of fine atmospheric particles in much of the U.S., but a recent year-long study in the Los Angeles area shows that carbon can be a major component of the fine fraction as well. Indeed, in Los Angeles it accounts for most of the fine-particle mass. This particulate carbon appears to be a direct by-product of fuel burned by motor vehicles and industry and does not come from hydrocarbon gases released by trees or other sources. Modern power plants that burn fuel efficiently do not emit carbon bits, but inefficient furnaces—and wood stoves that heat homes in many parts of the country—do.

The carbon particles are a combination of both soot (pure carbon) and complex organic (carbon-based) molecules. Like sulfate particles, they scatter light and decrease visibility; they also absorb light, making carbon-containing hazes appear dark. Soot itself does not appear to pose a threat to health, but certain of the organic molecules that adhere to it may be. On the other hand, the organic molecules may have at least one potentially positive role: they are probably characteristic of the particular fuels that produce them and so may serve as clues to the carbon's source.

Our studies and those of others have shown that sulfate and carbon compounds produced by combustion are usually the major components of fine atmospheric particles in any large geographic region. Occasionally, however, unexpected substances appear. The methods described above are well suited to detecting and explaining such aberrations.

One December a few years ago we were astonished to find that chlorine accounted for some 20 percent of the mass of fine-particle samples collected in scattered Middle Western locations for several days. Chlorine can be found in marine air as sea salt, but it normally contributes no more than .1 percent to the particles in the air over the continent. Nevertheless, samplers in Wisconsin, Kansas, Missouri, Kentucky, Indiana and Ohio all showed dramatic and virtually simultaneous rises in chlorine.

Puzzled, we considered various potential sources of chlorine. The salt (sodium chloride) that is spread on roads after a snowfall was one possibility, and our samples did have sodium levels that might be expected if the chlorine had come from salt. Yet no snow had fallen and no roads had been salted near the study sites. Burning of vegetation would certainly elevate chlorine levels in an area, but it would also boost the levels of potassium and soot—and these substances were not abundant. Measures of other substances revealed that in most respects the air at the collection sites appeared to be unusually clean: the concentrations of sulfur, nitrogen oxides and ozone were much lower than normal, as was the concentration of sulfate in the fine-particle mass.

These data suggested a surprising conclusion: rather than representing an unusual episode of pollution, the chlorine signaled an even more unusual intrusion of clean air. Calculations done by workers at the National Center for Atmospheric Research supported our finding and suggested that sea air had traveled from the northern Pacific Ocean, down across Canada and into the Middle West.

A similar event tested the sleuthing abilities of Thomas G. Dzubay of the EPA. After collecting many samples of atmospheric particles in St. Louis, he noted that aluminum, potassium, strontium, antimony and barium were significantly more abundant in one sample than in the others. It occurred to Dzubay that these elements are all present in fireworks, and he looked at his record of collection dates. Sure enough, the unusual sample was gathered between noon and midnight on the Fourth of July.

Such a discovery shows that the complexity of particles, once a stumbling block to their analysis, is actually a benefit: by noting the elements or compounds that occur together, investigators can often deduce which materials spawned them. Perhaps someday the analyst's tools will be sharpened enough to establish not only that the burning of coal or other fuel accounts for the atmospheric particles in a given locale but also that the fuel was burned at a particular site.

In the meantime an enormous body of data leaves no doubt that fuel combustion is the main source of acid deposition and particulate fallout. This, and the fact that pollution travels across national boundaries, can no longer be questioned. Now policymakers in government, industrial managers and citizens throughout the world confront the challenge of determining what, if anything, is to be done to augment existing pollution controls, how much money should be spent and who must pay the bills.

THE AMATEUR SCIENTIST

Music and ammonia vapor excite the color pattern of a soap film

by Jearl Walker

The glistening bouquet of colors in a soap film arises from optical interference. The light waves being reflected from the front surface of the film interfere with the ones being reflected from the rear surface. This phenomenon provides scope for a number of demonstrations.

Suppose the film is illuminated by light of a single wavelength, which is to say a single color. If the interference is constructive (the two sets of waves from a particular part of the film are in step), that part of the film looks brightly colored. If the interference is destructive (the waves are out of step), that part of the film is dark.

A variety of splotches and bands of color are created when the film is bathed with a beam of white light, which consists of many wavelengths and therefore of many colors. Wherever you see a certain color, light waves with the corresponding wavelength interfere constructively. Waves at other wavelengths interfere destructively, so that the associated colors are pale or absent. The thickness of a region of the film determines what wavelengths interfere constructively there. Since the thickness usually varies, the film has an array of colors.

If the film is vertical, the liquid in it moves slowly downward, mainly because of gravity. The result is a smooth variation in thickness from top to bottom. At any given height the thickness is approximately constant horizontally. Therefore you see horizontal bands of color, each band related to a certain thickness of the film. If the film is disturbed so that the thickness changes, the colors fluctuate, perhaps wildly.

To demonstrate such fluctuations Andreas Kay of Cologne in West Germany has recently devised a way to project the colors of a soap film by making the film function as a mirror. To prepare a mount for the film, Kay cut away the interior of a lid from a can of Sir Winston tea. (Although any lid might serve, the shape of the one from Sir Winston tea is advantageous.) When the film is mounted vertically, the fluid that slowly flows to the bottom collects in a gutter around the border of the lid and then drains off the mount. If the mount has no gutter, the draining fluid collects at the bottom of the film and distorts it.

In Kay's design the film serves as a concave mirror reflecting light onto a screen where the interference colors can be examined. Kay dips the lid into a soap solution, mounts it on the can and then positions the can in a beam of light with the film vertical. Wetting a straw in the soap solution, he pierces the film with the straw and sucks some of the air out of the can. Because of the reduced air pressure in the can, the external air pressure pushes the film inward, making it concave. The curvature determines the film's focal length as a concave mirror. Removing more air makes the film curve more and shortens the focal point.

The film reflects only a small part of the light illuminating it—approximately 3 percent of what reaches the front and back surface. The rest of the light enters the can. Unless it is prevented from reflecting back out, it will mask the faint interference colors reaching the screen. To reduce the glare Kay paints the interior of the can flat black. He says the glare will be virtually eliminated if the back of the can is also fitted with an inclined black wall that serves as a light trap. When light is reflected by the wall, it travels farther into the can and is reflected several more times, being partially absorbed at each point of reflection.

Kay illuminates the film with a brilliant slide projector adjusted for a long focus and aligned so that the light reflected from the film passes through a lens system. He positions the equipment and the screen to maximize the clarity of the image on the screen. In addition he fine-tunes the curvature of the film by means of a syringe that fits snugly through a hole in the can. As the absorption of light by the paint warms the air inside the can, the air pressure increases, thereby decreasing the film's curvature. Kay removes some of the air by pulling the plunger outward, restoring the curvature.

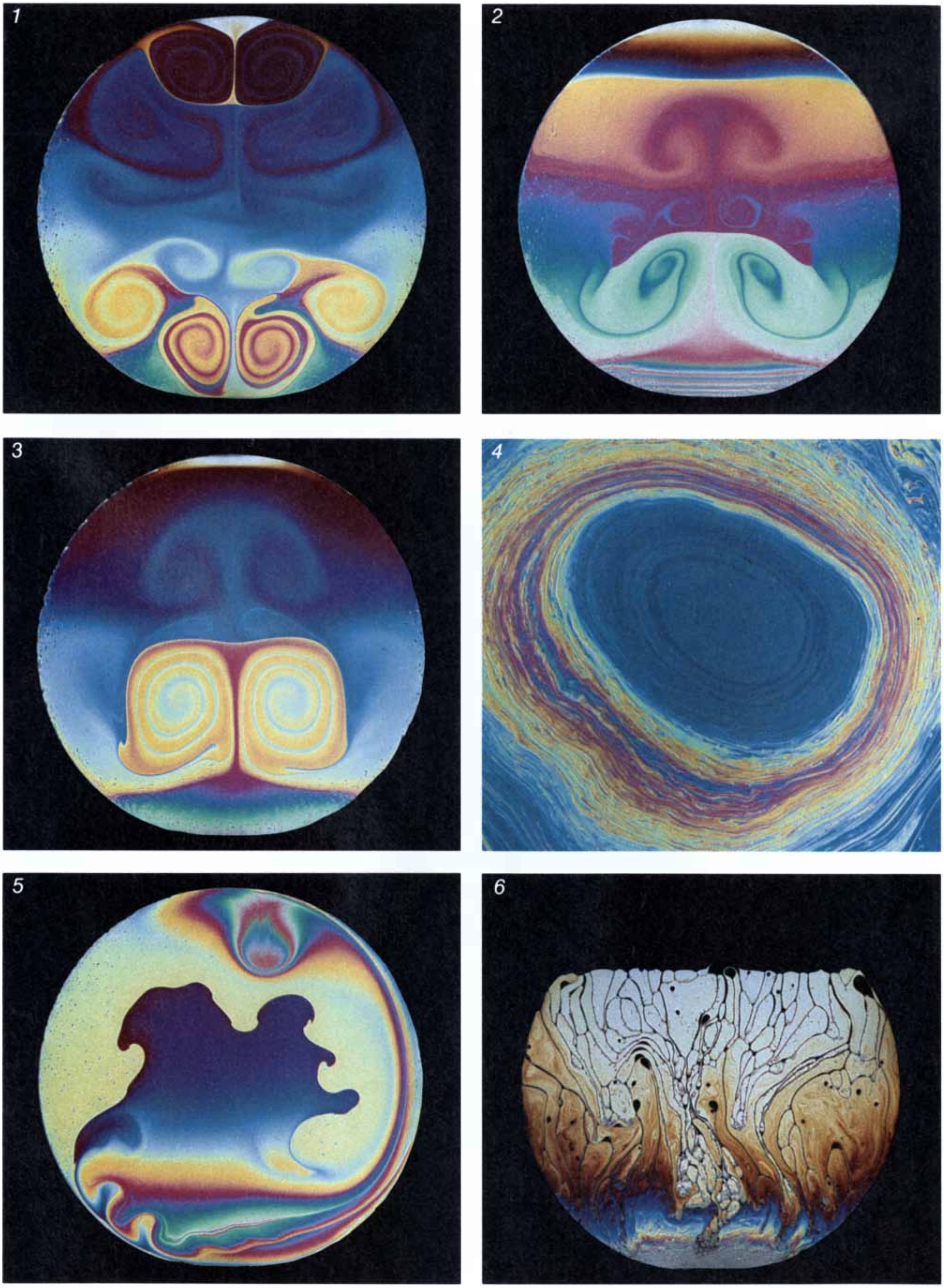
Kay finds that to produce a clear image he must pass the reflected light through the lens system, which he took from another slide projector. The system, which has a focal length of 150 millimeters, is placed at the focal point of the film so that nearly all the light reflected by the film passes through the lenses. Normally the plastic barrel that holds the lenses extends well beyond them so that it can slide into the projector. Unless the extension is removed it will block some of the light from the film. Kay cuts off the extension near the lenses.

To animate the colors on the screen Kay plays loud music on a bass speaker placed near the film. The fluctuations in air pressure distort the film, varying the direction in which light is reflected. The fluctuations also change the thickness of the film and thus shift the colors that are created by constructive interference of the reflected waves. The colors dance over the screen somewhat in time with the music. Kay has found that the strong beat of rock music provides a better show than the fuller and less percussive sound of other music.

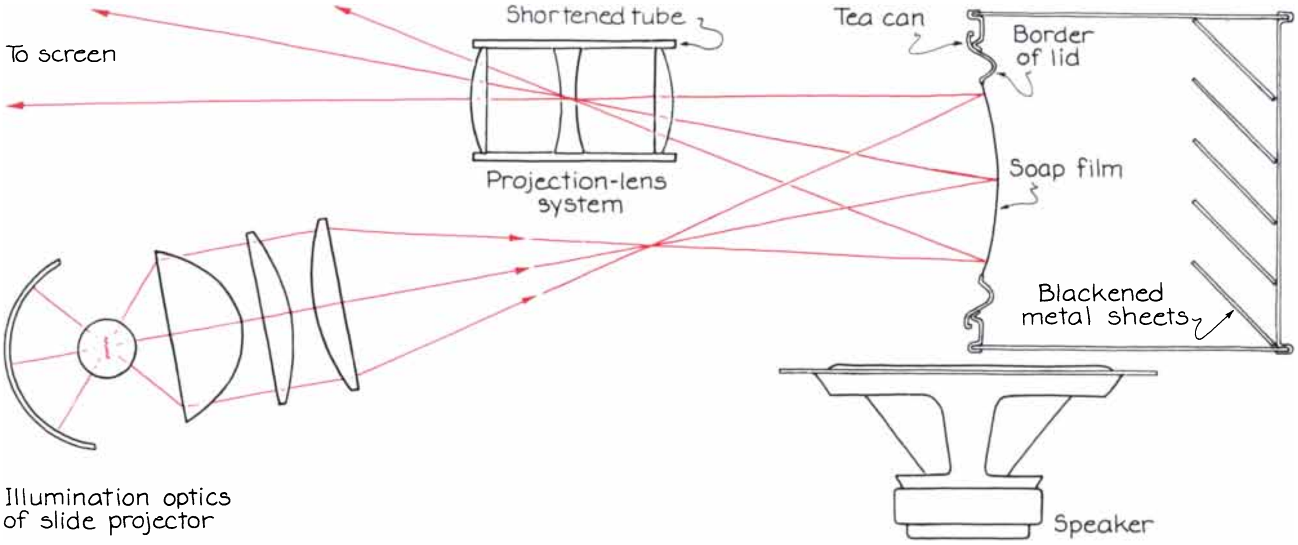
In some frequency ranges the film begins to resonate with the sound. Parts of the film then vibrate vigorously. The movement creates beautiful patterns of vortexes and symmetrical streams in the film and in the image on the screen. When the film is exposed to loud ultrasound, the patterns on the screen become crazed. Small sprays leap from the film. Some of the activity in the film is created by streams of air forced across the film by the pulsating speaker.

Kay's soap solution consists of 1.4 grams of triethanolamine, 100 grams of glycerin (which is supplied at 85 percent dilution) and two grams of oleic acid. After mixing the chemicals he stores the solution in an airtight bottle kept in the dark so that air does not oxidize the oleic acid. The solution should sit for 24 hours before films are prepared. By then it should be clear. If it is not, it probably contains too much oleic acid. Such a solution causes small drops to form on the film. Adding more triethanolamine to the solution fixes the problem. Kay's films become thin within minutes and last for about an hour.

Kay has ways other than rock music to arouse a dance of colors on



Andreas Kay's soap-film patterns from sound (1–4) *and ammonia* (5); *"critical fall"* (6)



Kay's arrangement for projecting an image of a soap film

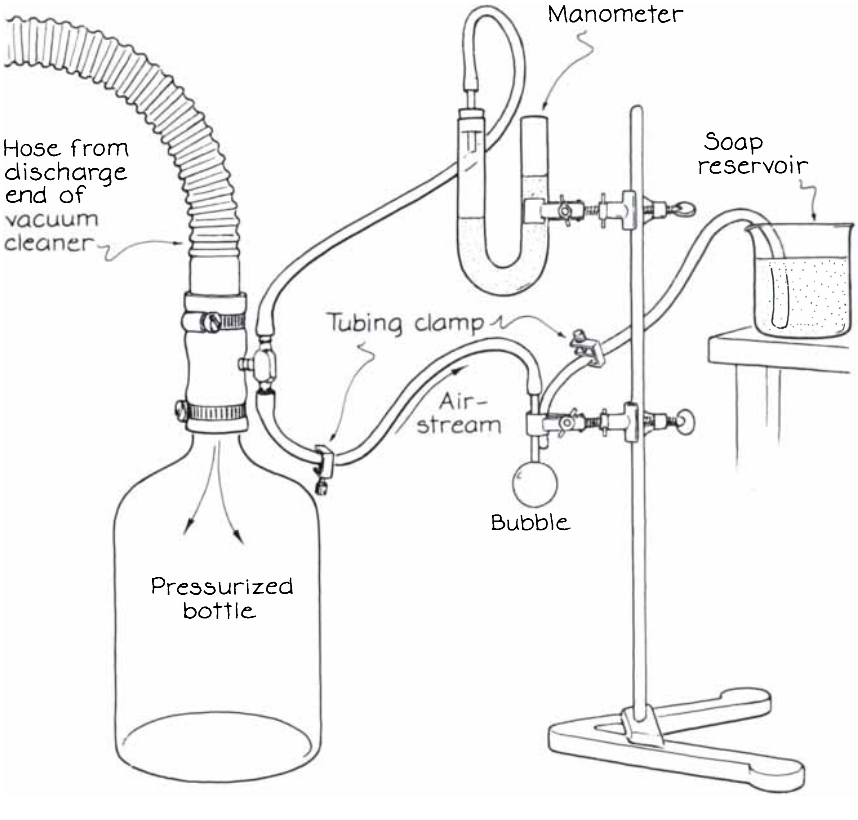
the screen. A whiff of ammonia drives the colors into a frenzy. Ammonia increases the surface tension of the film wherever it is absorbed. As the affected region pulls inward on the surrounding film, its weight increases until it is finally so bloated that it falls along the film. On the screen the fall seems to drag along entrained colors.

By changing the liquid solution to 15 grams of triethanolamine, 60 grams of glycerin and 25 grams of oleic acid Kay demonstrates a thinning behavior called critical fall. Several minutes after the film is mounted it begins to thin dramatically at the top. Soon it is so thin that the light reflected from the front surface is always out of step with the light reflected from the rear surface. The film is then black. The black region marches to the bottom of the film, preceded by a kinetic display of convection in the regions that are still colored.

Jurjen K. van Deen of The Hague also studies soap films. In particular he has wondered how one ever manages to blow a soap bubble from a ring—something children commonly do. As the bubble escapes from the ring, how does it avoid breaking because of the hole left in its side by the ring? The hole must close just as the bubble escapes, but how?

To answer the riddle van Deen tried to photograph a bubble blown from a ring by an assistant. He found that unless he used a high-speed camera and was careful to steady the ring, the photographs were always blurry and ill-timed. He then worked out an arrangement that is quite steady. A soap solution siphoned from an elevated reservoir flows through a tube mounted on a laboratory stand. The rate of flow is regulated by a clamp. Attached to the tube is a second tube through which air flows. The air tube extends downward past the end of the solution tube. When the soap solution leaves the solution tube, it drains to the opening of the air tube. In that position it forms a film.

The air is supplied by a 25-liter plastic bottle kept under pressure by the discharge from a vacuum cleaner. Another tube clamp regulates the airflow. A water-filled U-tube monitors the air pressure. All connections with the bottle are made through a specially fitted piece van Deen attached to the top of the bottle. The solution is synthetic dishwashing soap, diluted with water to a ratio of 1 : 5 or 1 : 10, with



Jurjen K. van Deen's soap-bubble generator a dash of glycerin and a teaspoon of sugar added. Photographs are made in strobe light to keep the time of illumination brief (from 30 to 50 milliseconds). Longer periods of illumination yield blurry photographs.

After the soap solution forms a film over the end of the air tube, the airstream inflates the film into a bubble. As the bubble steadily increases in size its weight eventually overpowers the surface tension that makes it adhere to the tube. In due course it breaks free. The airstream then begins to inflate another bubble. Bubbles form in a steady procession. As a result the camera and the flash can be triggered accurately. Van Deen controls the growth rate of the bubbles by adjusting the flow of air; he is able to establish the size of each bubble by adjusting the flow of the soap solution.

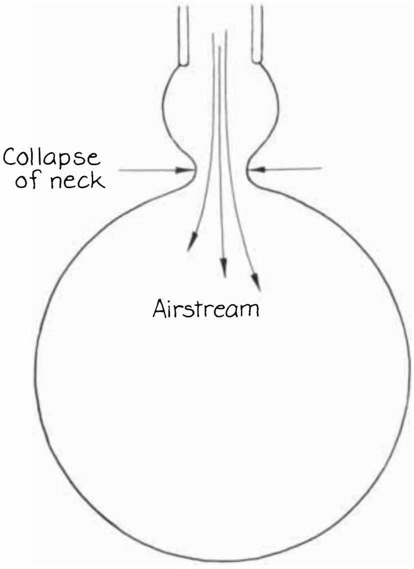
With this arrangement van Deen discovered how a bubble survives its launching. As the bubble grows it remains attached to the tube by an umbilical section that is initially cylindrical but narrows as the weight of the bubble increases. Just as the neck collapses onto itself, the bubble breaks free; thus the hole in the side of the bubble closes. Surface tension jerks the rest of the umbilical section back onto the tube to form a new film for the next bubble. Similar events no doubt take place when a child blows a bubble from a ring.

If the flow rate of the air in van Deen's experiment is high, the bubble may be blown away before it has had time to form properly. He therefore rests it on a ring. In this arrangement his photographs reveal that the umbilical section sometimes develops several necks along its length before fully collapsing. He suggests that a study of the dynamics of the umbilical section might be rewarding.

Great fun can be had with a new toy called Bubble Thing, manufactured by David Stein, Inc., of New York. The mechanism produces enchanting bubbles, several meters long, that look like amoeba. The unofficial length record is about 15 meters.

The device consists of a plastic tube with a loop of narrow cloth ribbon at one end. One side of the loop is attached to the end of the tube and the other side is attached to a runner that slides along the tube; the function of the runner is to alter the width of the loop. A weight hangs from the bottom of the loop.

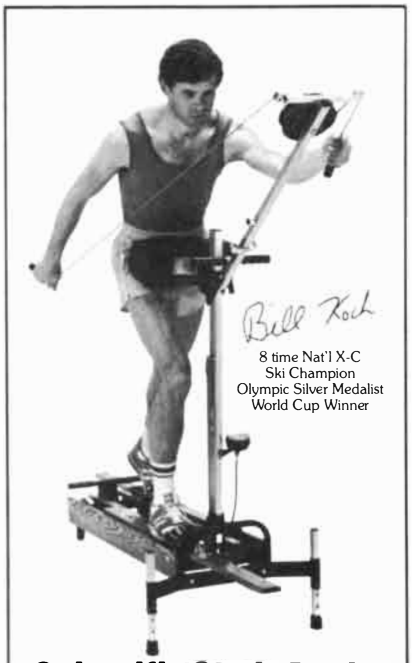
Following the instructions, I mix one unit of Joy dishwashing soap with 10 units of water and a quarter unit of glycerin (which I obtained from a pharmacy). Closing the loop by moving the runner to the far end of the tube, I dip the loop into the solution, thoroughly soaking it. Lifting the device from the solution, I hold the tube horizontally while slowly moving the runner to open the loop and expose the soap film that stretches over it. On a breezeless day one walks backward to inflate the film. Otherwise the breeze will do it. The bubble grows, stretching away from me as more fluid is drawn out of the ribbon by surface tension. If the breeze is too strong or I walk too fast, the film pops, leaving gossamer threads floating in the air or gently gliding to the ground.



The neck of a bubble

When I succeed, a tubular bubble stretches for many meters, wriggling like some wild animal. To free it I close the loop, allowing an umbilical section to form, narrow and finally collapse on itself. If I wait too long, the bubble breaks away without closing the hole at the loop. In this case it disintegrates slowly. When I release the bubble properly, the hole closes.

As the bubble floats, surface tension acts to wrestle the tube into a sphere so that the surface area is minimized. This action fights against the force of gravity and the nonuniform distribution of liquid in the bubble. In the ensuing dance of shapes the bubble may burst. If it survives the battle of forces and the buffeting by the breeze and if it avoids certain disintegration as a result of touching something solid, it floats down the street like a zeppelin. Through all these movements it glistens with delicate colors that change constantly.



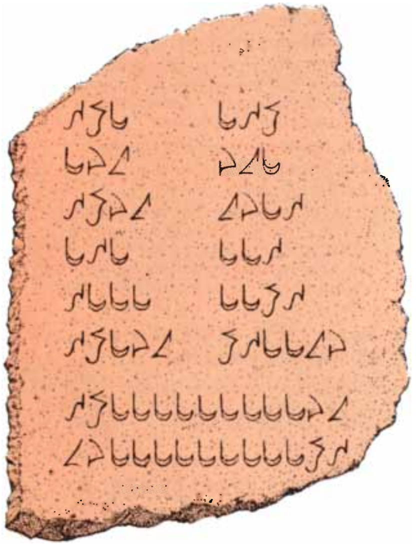
COMPUTER RECREATIONS

Word ladders and a tower of Babel lead to computational heights defying assault

by A. K. Dewdney

No one knows how or why Martian civilization vanished, but a cave at the foot of Olympus Mons holds a sad remnant of this ancient and wonderful culture. At the end of the cave a tablet reminiscent of the famed Rosetta stone is shrouded by a dusty fabric of unknown composition. The fabric is pulled aside, revealing what seems to be a kind of dictionary: a list of words on the left half of the tablet is matched by a list of words on the right. Two sentences are engraved at the bottom of the stone. The sentences are related in a peculiar way. The first one can be transformed into the second one by repeated substitution of words from the dictionary: each word in the sentence on the left-hand side of the dictionary can be replaced by the corresponding word on the right.

Martiansentencescontainnospaces. Indeed, none are wanted owing to the fluid and expressive character of Old Martian; often there are many ways to divide a Martian sentence into words. As a general rule, however, only a few divisions make sense in a given context. The substitution process can be illustrated employing the first sentence of this paragraph. If *tainnospaces* appeared in the left-hand column of the dictionary and *cealwisdom* appeared as the corresponding word on the right, the latter word could be substituted for the former to obtain a new and valid sentence: Martiansentencesconcealwisdom.



Martian dictionary of wisdom

The wise ones of ancient Mars held that all information worth learning could be achieved by starting with one basic sentence and substituting words according to the dictionary of wisdom, the only surviving fragment of which lies in the cave. The basic sentence is the first of the two at the bottom of the Olympus Mons stone shown on this page. In general a computer will not be able to verify whether a given sentence can be derived from the basic one. In other words, there is no possibility of writing a computer program (no matter how large or how fast the computer is) that will decide correctly, for each dictionary and two words (or sentences) of input, whether a translation is possible from the first word into the second. I shall explain why below. In the meantime, can readers transform the basic sentence into the one displayed underneath it? (This particular transformation can be achieved.)

The Martian translation problem is one of an entire family of puzzles that call for the transformation of words, sentences and even entire paragraphs into other words, sentences and paragraphs. The early development of the subject is due in part to the Reverend Charles Dodgson, otherwise known as Lewis Carroll. Among the many mathematical and symbolic pastimes Carroll invented during his Oxford walks is a transformation called a word ladder. In a word ladder one starts with two specific words in the language of one's choice. The first word is called the source and the second the target. Can one transform the source word into the target by changing one letter at a time?

If the two words have the same number of letters, the task is trivial. But can one ensure that all the intermediate character strings are also words? That is a horse of a different color, but a mount that people in the wordplay academy can easily ride. One can in fact start with *horse* to illustrate the process. In one step the word can be changed into *house.* Another step yields *mouse.* By this alchemy it would seem a horse could be changed into almost anything whose name has five letters. Is that possible?

To answer such a question in general, it is necessary to construct a transformational network. As an enemy of overly impressive terminology, I have renamed it a word web. The complete web for the English language would be a monstrous affair: every English word would be present in the web as a unique dot. To satisfy the requirements of the word ladder two dots would be joined by a line if their corresponding words differed by exactly one character. To find the words to which *horse* could be changed, for example, one merely traces through the web from *horse,* adding to a list all words encountered on the journey. A complete exploration of such paths would be tantamount to plucking the web by the point *horse* and pulling up with it all the points connected to it. The whole would probably look like a poorly mended fishnet. Of course, the great majority of words would not get pulled up with *horse;* any word not having five letters would stay below, for instance.

Would all the five-letter words come up? That is to say, is it possible to build a word ladder from *horse* to any other five-letter English word? Although I must say I do not know the answer to this question, judging from past reader response I would be willing to bet that someone will find it. The requisites are a dictionary that has been put in machine-readable form and a program that searches the dictionary for connected words. Every time a word is added to the list of all words connected to *horse,* the program deletes it from the dictionary in memory and runs through the entire list again, comparing each word on it with each word encountered in the steadily diminishing dictionary. Every time a dictionary word differs by one letter from a word on the list, it is added to the list and the process begins again. Different dictionaries will undoubtedly yield different results, will they not? Perhaps. On the other hand, two reasonably complete dictionaries might give the same result. In other words, the webs might be so thick that their topology would be practically the same.

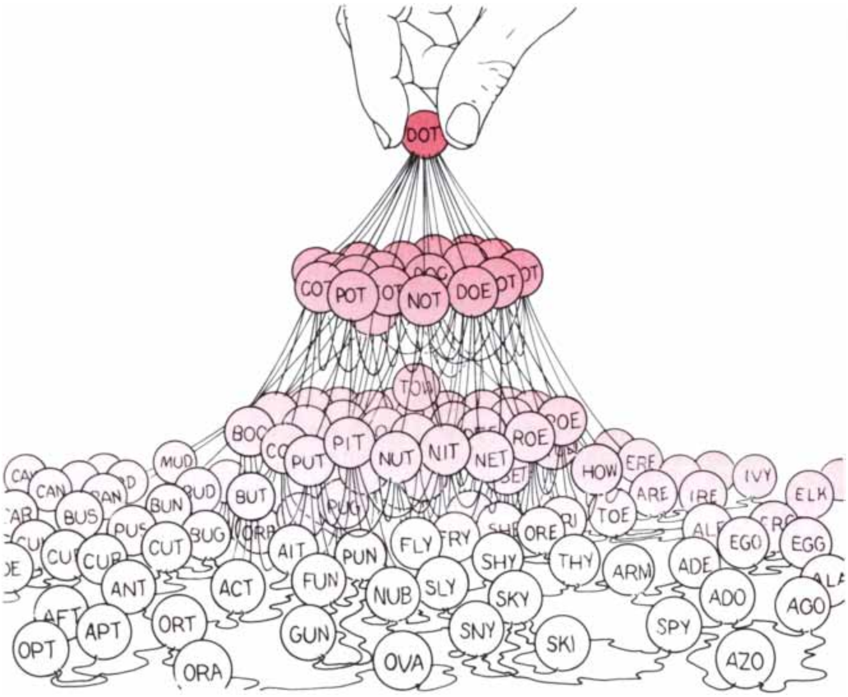
The question of whether *n*-letter words are connected can be answered almost by hand for small values of *n*. Certainly all one-letter words are connected within one web. The case for two-letter words can also be decided here and now. According to my Scrabble dictionary, the two-letter words are *aa, ad, ae, ah, ai, am, an, ar, as, at, aw, ax, ay, ba, be, bi, bo, by, da, de, do, ef, eh, el, em, en, er, es, et, ex, fa, go, ha, he, hi, ho, id, if, in, is, it, jo, ka, la, li, lo, ma, me, mi, mu, my, na, no, nu, od, oe, of, oh, om, on, op, or, os, ow, ox, oy, pa, pe, pi, re, si, so, ta, ti, to, um, up, us, ut, we, wo, xi, xu, ya* and *ye*.

A little work reveals that all the words in the web are connected. First, all two-letter words that begin with the same letter are mutually connected. This enables one to lump all words beginning with the same letter together. It remains only to note that the *a* lump is connected to the other four vowel lumps, that the *b* lump is connected to each of the other 16 consonant lumps and that the *a* lump and the *b* lump are themselves connected.

Armed with only two low-order test cases, there are people who would boldly conjecture that all *n*-letter English words are connected. The proposition typifies the excitement of abstract wordplay. Readers with computers might attempt to test the proposition with other low values of *n*, such as $n = 3$ [*see illustration on this page*]. Those without computers must be content with turning *hate* into *love*. Better still, they could attempt to convert *evil* into *good*, achieving valuable moral and logo-logical insights in the process.

The game of word ladders provides a simple example of a so-called textual transformation system. The two major ingredients of such a system are a set of texts to be transformed and a rule of transformation. The texts might be words, sentences or paragraphs, which are really just strings of characters, no matter how the set of texts is chosen. The rule of transformation consists of a procedure for replacing part of a given string by another string. In following the procedure a dictionary is usually consulted.

In word ladders, for example, the set of strings consists of all English words in a given dictionary. The procedure takes any word submitted to it and generates another word that differs from the first word by one letter and is also found in the dictionary. Here the dictionary is a simple list of words.



The three-letter word web

Most published dictionaries have entries that consist of words "defined" in terms of other words. Another kind of transformational chain—one that is more sophisticated than the word ladder—is implicit here. Imagine, for example, that one starts with a single English word, looks up its meaning and makes a list of all the words found in its definition. Then one looks at the meanings of those words, makes a combined list of the words and continues to repeat the process. Since English has a finite number of words (I cannot vouch for other languages), the procedure soon develops a cyclic character, so to speak: certain words begin to recur. They are defined, in part, in terms of themselves. One could call them primal words. They are not to be heralded with primal screams but should be repeated silently with something like awe: perhaps they constitute the indefinable conceptual foundation of English.

The project, or something much like it, was carried out by Robert Amsler of Bell Communications Research. Amsler used a dictionary that had been modified so that the definition of every word was reduced to a set of genus words serving to identify the broad classes of things to which the word belonged. The genus words were supplemented by differentia: words that serve to distinguish a given word from other members of the genus. Utilizing a program similar to the one described above, Amsler traced through his special dictionary and found basically what the philosophers of language predicted he would find: primitive words, all of which were used in the ultimate definitions of other primitive words. Examples included *food, person, thing, instrument* and *group*. The primitive words occupied the ultimate branches of what has been called a tangled hierarchy, something like a forest in which the branches of different trees have grown together. The primitive words lie "upward" in the direction of increased generality.

Amsler's fundamental idea is that dictionaries have more structure and reveal more about language than people commonly suppose. It would be surprising if the relatively simple genus-differentia scheme did not produce a few anomalies. A barbecue, for example, turned out to be "an animal roasted over an open fire." The barbecue thus found itself on a list somewhere between *aardvark* and *zebra*.

Amsler's research has applications for data-based retrieval schemes in which a user might want to retrieve instances of a general class. He is currently rebuilding his tangled hierarchy on the *McGraw-Hill Dictionary of Scientific and Technical Terms*.

A related transformational chain has been explored by Ron Hardin, a research scientist at the AT&T Bell Laboratories in Murray Hill, N.J. Hardin has devised a number of interesting and amusing textual transformation games over the years. His lat-

afraid – apprehensive – expectant – hopeful – confident – **unafraid**
classified – secret – mysterious – unidentified – **unclassified**
dress – groom – tidy – clear – wipe – take off – divest – **undress**
even – level – uniform – regular – periodic – spasmodic – **uneven**
perfect – pure – unvarnished – unfinished – rough – **imperfect**
potent – dynamic – animated – excited – nervous – weak – **impotent**
rational – clever – witty – funny – silly – brainless – **irrational**
safe – undamaged – mint – untarnished – polished – slippery – **unsafe**
social – friendly – peaceable – inoffensive – retiring – **antisocial**
valid – convincing – plausible – specious – unsound – **invalid**

Ron Hardin's tower of Babel

est preoccupation involves *The New Collins Thesaurus.* The user of a thesaurus commonly looks for synonyms of a word or for a conceptual class of words that he or she has in mind. One might imagine that each word inhabits a small cloud of words that mean practically the same thing. Yet the clouds are not all widely separated from one another. Some clouds overlap others. As a consequence, although certain pairs of words are not in themselves mutual synonyms, they can share common synonyms. Therein lies the source of much mischief. Consider the following chains unearthed by Hardin. Each word is found in its predecessor's synonymic cloud:

acceptable → so-so → ordinary → inferior → rotten → unacceptable

reliable → steadfast → obstinate → wayward → unpredictable → unreliable

Hardin has generated several thousand such chains by computer. In each chain a word is gradually converted into its antonym. The great majority of the chains generated by Hardin are no longer than these examples. The exercise reminds us that human languages are fluid, that the meaning of words depends heavily on context and that every word has a certain ambiguity that is both a blessing and a curse. Consider the case of a lawyer cross-examining a witness about the character of a defendant:

"You have stated in testimony that Watson is a reliable man. Would you not also describe him as 'steadfast'?"

"Yes."

"Indeed, would you not say that from time to time the accused was obstinate in his steadfastness?"

"Uh—I suppose so."

"At times, certainly, his obstinacy amounted to a form of waywardness, did it not?"

"I'm not sure."

"Not sure? Come, come, Dr. Finch, to be obstinate and wayward are practically the same thing."

"I suppose."

"A wayward man is an unpredictable man, in short, an unreliable man. You have just contradicted your earlier testimony."

An additional sampling of Hardin's word-reversal sequences is included [*above*]. His use of *The New Collins Thesaurus* is not quite as straightforward as I have made it sound. The thesaurus is organized under headwords. When a user of the thesaurus wants a new word for a given context, he or she looks up the headword having a meaning closest to the one intended. Listed below the headword are the related words organized by sense of meaning. For example, under the headword *express* one finds senses of meaning that relate to speech and speed. Hardin's main problem was to connect senses in a logical and consistent way through headwords. As Hardin puts it, "If one supposes that the distinctions being made when senses are distinguished ought not to disappear when propagated...there is no consistent way to connect the thesaurus."

Hardin therefore devised an algorithm that systematically examined all pairs of related headwords and computed the strength of the connection between the various senses in which each could be taken. The senses that had the highest degree of connection were then incorporated in a kind of word web spun by the algorithm. In this way the distinctions between senses tended to remain well defined, although sometimes they became blurred. When the web was complete, a second algorithm that finds the shortest path between any two senses was deployed. It was then an easy matter for Hardin to specify various pairs of words to create his own tower of Babel, after that biblical edifice whose completion was thwarted by a confusion of tongues. Hardin's tower is a mini-tome with special attractions for linguists and lawyers.

An earlier textual transformation of Hardin's was mentioned in this department in September, 1985. There I quoted a poem called "Topeka Beagle Buffers," a textual transformation of the familiar nursery rhyme "Peter Piper." The secret of its construction is conceptually no more difficult than the one underlying the tower of Babel.

Here is how Hardin transformed the verse beginning "'Twas the night before Christmas" into another verse beginning "Tweeze denied beef worker isthmus," for example. First he decomposed all the words of the original into a string of phonemic tokens: basic sounds grouped together according to rough similarities. Having done the same for some standard dictionary, he had a computer regroup the phonemes into new words by ignoring boundaries between words in the original text when necessary. Before the madness was complete Hardin and some colleagues at Bell Laboratories had to polish by hand certain rough spots in the computer version. The resulting poem leaves the Christmas celebrant with a strangely confused notion of just what the original was about [*see illustration on opposite page*].

It remains now to return to the Martian translation problem. Given a dictionary of permissible substitutions and two words (however long), why is it impossible for a computer to decide in general whether the second word can be obtained from the first one by a sequence of substitutions from the dictionary? The problem is called Thue's word problem after the Norwegian mathematician Axel Thue.

Many people outside the field of computing (and even a few within it) do not realize there are some things computers simply cannot do. One of the first people to suggest that computers have inherent limitations was the Englishman Alan M. Turing, a founding father of computer science. Turing conceived of a simple kind of abstract computer called (not by him) a Turing machine. The device is capable in principle of carrying out any computation a currently existing or conceivable computer is capable of doing. Turing showed it is impossible for one Turing machine to decide, given another Turing machine and its input, whether the second machine will ever halt or not. In more concrete terms, it is impossible to write a program that, given a second program and data as input, will decide whether the second program will ever stop manipulating that data.

Turing's theorem had at least one practical application. There once was a systems expert at a certain educational institution who decided that a mainframe computer was spending too much time executing freshman computer-science programs. A major reason for the delays, it seems, was that many of the programs contained unintentional infinite loops. If the programs could be scanned by an infinity-detecting program in advance of execution, much time might be saved. Alas, the infinity-detecting program was impossible, as a student of Turing's was quick to inform the expert.

Given the unsolvability of the so-called halting problem, it is relatively easy to demonstrate to the man and woman on the street that the Martian translation problem is also unsolvable. The demonstration proceeds by showing that the halting problem can be converted into the Martian translation problem.

The conversion is conceptually simple to make. The computer is encoded with a long string of symbols. The string essentially describes the initial contents of the computer's memory and working registers. A second string is then constructed. It indicates that the computer is in a halted state; whatever computation was to have been done is complete. The program to be tested for halting is translated into a set of rules for converting the first string into the second by means of substitutions. Each intermediate string then symbolizes one step of the computer's operation as it proceeds, one hopes, toward completing the processing of the data presented to it. The rules use substitution to move this virtual computer from state to state by following the program.

If, for example, the current content of a register is to be stored in a certain memory location, the rules substitute the new content of that memory location in the part of the current string that represents the memory location in question. The transformation procedure will succeed in obtaining the second string from the first if, and only if, the computer in question ultimately halts its processing of the data provided. Since according to Turing's theorem the computer will never halt, the transformation cannot be made.

Even without such a fancy argument, the unsolvability of the Martian translation problem is strongly suggested by the following observation: if the source word can be transformed into the target word, a program that systematically tries every conceivable substitution will eventually come on the right ones. But if the transformation is impossible, how will one ever find out? How long should one wait? Readers unable to solve our special example of the Martian translation problem will have to wait one month.

In May I discussed the ways in which computers are employed in the stock market. The discussion moved in three stages from the periphery to the heart of the market. First, individual investors who have access to computers may buy any of hundreds of programs devoted to the technical or fundamental analysis of stocks. Closer to the exchange, large trading institutions and brokerages use computers to calculate the sale or purchase of stock futures options (the option to buy or sell a stock on some future date at a price set in the present). Closer yet, there is the possibility of replacing the stock specialist, either in whole or in part, by a computer program. A specialist is the human being who manages all the trading in a particular stock at a particular exchange. Can the specialist's role be filled by a computer program that handles all buy and sell transactions, changing the price in some appropriate way as it goes? I appealed to the "experts" among our readers and I was not disappointed.

Aaron C. Brown is a financial consultant from New York who wrote a Ph.D. thesis on technical analysis in the finance department at the University of Chicago some years ago. He warns against engaging in technical analysis. My caveat about the practice was a little weak for Brown's taste: "There is overwhelming evidence that [such analyses] do not work in the sense of generating excess return. More important, they are dangerous." Particularly dangerous, says Brown, is the strategy of buying a stock when its price drops in anticipation of a near-term rise. The strategy is cousin to the gambling system that doubles the bet after every loss, he cautions. An investor might well buy all the way down a stock's long decline, each time expecting an upturn in price.

Charles J. Higgins of the Department of Finance and Computer Information Systems at Loyola Marymount University in Los Angeles has examined a particular technical method called the trend-line charting technique, which is similar to the trading-channel technique I discussed in May. As far as Higgins is concerned, transaction costs make all the difference. In the price data studied by Higgins, a majority of securities did not produce abnormal returns significantly in excess of the average market. "However, many securities did produce significant abnormal returns...when transaction costs were low or nil." This points to an opportunity, says Higgins, for specialists, floor traders and other members of an exchange (paying little or no transaction fees) to achieve significant returns.

In the 1960's and 1970's the physicist M. F. M. Osborne made a study of trading methods. He taught some courses on stock-market finance at the Graduate School of Business Administration at the University of California at Berkeley. According to Osborne, who is now living in Hillcrest Heights, Md., trading algorithms have been in use for some time, at least as a set of rules applied by the human trader. Osborne, who has studied a sequence of progressively more sophisticated trading algorithms, has sent two volumes of notes on the subject. He remarks that things just get tricky for the specialist who tries to make undue profits from a position at the center of the market.

There is an old two-part jingle that begins, "What is life?" This September 21 through 25, participants in the first Artificial Life conference at Los Alamos, N.M., will find many answers—given in terms of abstract systems, various computer simulations, hardware, software and, possibly, some wetware (chemical systems) to boot. There is still time for observers with a special interest in the subject of artificial life to contact Christopher Langton, the organizer of the conference, for details. Langton can be reached at the Center for Nonlinear Studies, MS B258, Los Alamos National Laboratory, Los Alamos, N.M. 87545.

TWEEZE DENIED BEEF WORKER ISTHMUS

Tweeze denied beef worker isthmus, winnow Trudy how's,
Knot agreed juries during, gnaw Tiffany moss.
This talking swear unbided Gemini wit cairn
Hint opus scenic (alas!) sinewy dare.
Unjelled runner nozzle tools smuggling deer butts
Well fissions unshoe kerplunks thence endear huts.
Anemometer cur chiffon dyeing mayhap,
Adjust subtle warp reins fairy loin winger snap.
Winnow taunted launderer roast sachet glitter
Ice brine bromide bet deucey woodwinds schemata.

The first few lines of "Tweeze denied beef worker isthmus"

BOOKS

London's miasmas, deep-water navigation, mother's milk and strong atomic bonds

by Philip Morrison

THE BIG SMOKE: A HISTORY OF AIR POLLUTION IN LONDON SINCE MEDIEVAL TIMES, by Peter Brimblecombe. Methuen & Co. ($39.95).

Smoke too was a gift from Prometheus—an unintended side effect. This brief, broadly informed, anecdotal and yet scrupulously documented account centers on the problem in one archetypal place. The tale opens far afield by presenting the oldest evidence we have of the effect smoke can have on human beings: soot-blackened mummified lung tissues from Egypt and Peru, from the Aleutians and the Canaries. The tissues bear witness to long years of inhaling smoke from the indoor fires that warmed and cooked for people in huts and lodges.

We have no lung samples from early Britons, but with a clever optical device a medical researcher has photographed the maxillary sinus cavities in 4,000 British skulls from all periods since the Bronze Age. Sinusitis is diagnosed from the roughened texture of the bony floor of the sinus. The Anglo-Saxons suffered it most, when they passed through a few cold, damp centuries. In Roman days most cooking had been done outdoors, and Bronze and Iron Age huts were draftier. A partial technical fix was found: the chimney. Before the 16th century only noble English houses boasted such structures; charms to cure smoky fires were cheaper, and they were popular palliatives.

In London and other crowded medieval cities the citizens were oppressed by their neighbors' smoke as well as their own. The fuel-consuming trades were still small in scale, and such cottage industrialists as bakers and dyers burned not much more fuel than the ton of wood estimated to have been consumed for domestic purposes per house per year. Only the limekilns went on to a new scale of smoke generation: mortar for buildings is a heavy product, and production took place at the urban market, the fuel being brought upriver. The iron forges took plenty of blame: a medieval satire aimed at the "sooty smoked smiths" ends with an angry "Christ punish these horse-shoe benders, / Who cake our clothes and ruin our night's sleep." Yet the limekilns burned thousands of tons of fuel per year, a single forge only a few.

The importation of cheap coal hauled as ballast from the mines of Tyneside—sea coal, it was called—led to the first action by a London government against air pollution. In 1306 smoke abatement was mandated by a proclamation directly banning the use of sea coal by limekilns, under threat of "grievous ransoms."

By then the pattern of pollution was already determined. The population grows; people live more densely, constrained for a long time within the area enclosed by city walls. Wood fuel becomes dearer as nearby forests are cleared for farming. For a time lighter, smokeless charcoal is shipped to the city from forests farther off. Then sea coal attracts heavy users, its unfamiliar sulfurous fumes particularly resented by visitors and the gentry.

Restoration London was home to the breezes of environmental awareness as it was to the smoke of the Industrial Revolution. John Evelyn's classic *Fumifugium* put the case to Charles II. Evelyn's proposed bill to remove the smoky trades from the city was properly drawn up, but "we must assume that it was dropped." (Our author reflects that "environmental idealism often runs up against economic considerations.") The Royal Society debated but was of little practical help. One Henri Justel, F.R.S., published in *Philosophical Transactions* of 1687 the design of a stove that consumed its own smoke. He reported its performance under one stringent if homely test: "Coal steept in Cats-piss makes not the least ill scent."

The gathering and analysis of data on urban public health started with John Graunt, the gifted draper whose 1662 study of the Bills of Mortality (along with royal support) brought him his F.R.S. He argued that changes in plague. mortality must signal poor air. He was interested too in the increasing incidence of rickets in London, shown here in a simple graph. The doubling in the number of deaths follows the curve of coal use, but the causes were surely manifold: the gray climate of the time, the soot-laden skies, the food shortages. A tight link between air quality and health is hard to document at any time, except during acute crises.

Fog darkens and swirls through the literary and artistic accounts of Victorian London. The era's prose, cartoons and drawings are eloquent in evidence. The fog frequency peaked about when Holmes and Watson lived in Baker Street. During World War I, T. S. Eliot's Prufrock still could see the "yellow fog that rubs its back upon the windowpanes."

New fuels, new motors, new legislation and the slow rise of comprehensive monitoring mark the recent past. Effective action against the big smoke followed the disaster of early December, 1952, when 4,000 Londoners died in a week from a few stagnant days of a foggy inversion that stubbornly trapped the smoke of a million chimneys. Both the Conservative and the Labour parties then supported air-pollution reform, and the new Clean Air Act was passed in the summer of 1956. A table lists a dozen major London smogs from 1873 to 1982; peaks of fatalities remained statistically significant up to 1975.

Now at last both domestic and industrial sources are under control. Smokeless zones have been set up within which dark smoke is banned; they now cover 90 percent of London. "There has been remarkably little resistance." Air pollution has decreased by about 80 percent. London is sunnier now—and ironically the effluents of its gasoline engines are precursors to a number of irritant photochemical products, once the pungent particulars of Los Angeles County.

THE LAST NAVIGATOR, by Stephen D. Thomas. Henry Holt and Company ($22.95).

On the loneliest reaches of the blue Pacific or well off the shoulder of Brazil, small seaworthy craft move under sail on every bearing. At any time they number some 1,000 vessels, scattered worldwide. They are not armed, they carry no cargo, they trail as a rule neither lines nor nets, and from time to time some of them race. These are the oceangoing yachts of the world.

Prehistory offers a surprising parallel to that intrepid if haphazard fleet. The Pacific tracks they now sail in sport were pioneered long ago by the Neolithic peoples of the sea, who broadcast the seeds of settlement among the isolated archipelagoes of coral. Affinity is inevitable, for however a fiber-glass hull differs from a breadfruit log, life for a small crew alone on blue water enforces such a sharp departure from the grounds of everyday existence that the exacting craft unites its devotees.

In Micronesia, south of Guam, among the eastern atolls of the Carolines that dot 500 miles of ocean, people of the sea continue today their traditional venturing over the Pacific swell; for them the outrigger voyaging canoes and the skills of safe guidance are part of daily life. In an unusually self-revealing, honest and moving book, a Boston navigator tells of his stays among the clans of Satawal, citizens of a new nation of many islands. He brought no ethnographer's credentials at all, but instead the hands and mind of a man at home topside and in the boatyard, handy with chain saw, Dacron and sextant. He linked his modern professional experience, and his inner search for a worthwhile life, to the discipline of a few well-initiated senior navigators.

Stephen Thomas studied within the navigation school of the legendary Wareyang, who once on Pulap (not far east of Satawal) was one of the first to practice navigation, which some say was taught to humans by the spirit of the rainbow. The teachers of the art are no shadowy informants but men whom we come to know well as their young and deeply engaged guest seeks to share their daily tasks and joys through the year. They are Mau Piailug, the most celebrated of Satawal navigators, and his younger brother Uurupa, a master canoe builder (and a medical technician, fluent in English).

A landlubber can gain some view of the navigation of the Carolines from this narrative of their "talk of sailing." Fifty pages of drawings, diagrams and lists of data outline what every navigator knows: the stars that rise and set all around the horizon, the seaways between islands aligned by guide stars, the changing bearings along a number of the main routes, a list of weather clues, a reinforcing set of mnemonic aids and a batch of visual headings and ranges on landfall, each matched to an opening into the lagoon of an atoll.

A Carolines skipper has no written texts, no maps, no tables; he has what his head holds. He must be able to access the data under fair wind and foul, when the canoe is awash with breaking surf, when he is cold and hungry and afraid. Yet all that meticulous geometry is not the task itself, only its skeleton. The small boat moves in the great ocean; as these seamen model their voyage, the sea bears its islands past the stationary canoe as the voyage continues. "You see," said Uurupa, "some old men know much more navigation than I do, they know all the...talk. But they cannot navigate because they cannot keep the movement...clear in their minds."

There we glimpse a little of the mind of the navigator. To see something of his heart, we need even more. Young Stephen Thomas, a philosophy graduate who had been drifting his way into a career among yachts, crossed some five years ago from the Galápagos to the Marquesas Islands, a couple of lonely weeks at sea. "The habitable globe was the length and breadth of my small boat; the sphere of human action was everything I or my crewman did. All else was chance. Each act took on huge proportions: tying a knot in a line, plotting a sextant shot, resting for several hours before a night watch. Any act could set off a chain of events ending in death." Piailug spoke his heart in more epic strain: "To be a *palu* [navigator] you must have three qualities: [fierceness, strength, and wisdom].... If you are not fierce...you will be afraid of the sea...of losing your way.... [Strength] is almost the same. It means 'strength directed by thought'.... The knowledge of navigation brings all three....a *palu* is a *man*."

The reader perceives by and by that Thomas did not go back to the atolls just to record the terms of art that are the talk of sail. His voyage was above all an interior journey. He sought from the famous captain a fathership he felt he had missed: sailing directions for the sea of life. That sense of engagement informs the entire book, opening it to many readers who perhaps would not be fascinated by the craft of the navigator.

Sturdy Piailug is a man in his fifties. He has opened new sea routes long avoided by the navigators of his atoll; he has shown his skills on test voyages over half of the Pacific, and even on television. He and his wife have 16 children; today responsible parents are often anxious. The house needs to be extended; the young men in high school, bound for college, are not inclined to take the long, risky apprenticeship of the *palu* (even though the fishing on the empty atoll of West Fayu, a short sea leg away, is still a mainstay of Satawal). The church and the rock-music tapes, the uncertainty of the old taboos and the abundant liquor are signs of the swiftness of a voyage, where any man, even a navigator, might for a time lose his way. "Don't they see," Piailug says of his chiefs, "that soon, very soon, change will crash on this island like a wave?" Navigation will change beyond what we foresee, with new gifts we give ourselves. But as long as our species sails, there will be no last navigator, and those who have made good use of the gifts of Wareyang will have fully worthy successors.

One part of this lore is quietly—almost covertly—held. To our minds it is strange. With the usual detail, by course and range, it lists certain quite unlikely fixed marks, called *epar,* that are said to be present always at the right place. They include circling frigate birds, a jumping yellowfin tuna

Piailug teaching a class in celestial navigation

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 $1.25 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 0036-8733/87. $1.25 + 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 953, Farmingdale, N.Y. 11737-0000. Telephone inquiries: 1-800-227-5782. The date of the last issue on subscriptions appears at the right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on 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 ______________________

(Piailug himself saw that one in about 1960 or so), a ray with a red spot behind the eyes. All have particular places and names. Thomas tried to argue it out with Uurupa, a coolly logical thinker. "'Always there?' he asked himself tentatively. 'Yes, I really believe that.... But remember what I told you? I said our master taught us to voyage for islands and not to see *epar*.... The spirit...gave us the *epar* to use *only when we are lost*.'"

It is the same here on the other ocean. A guitarist has a new song, the radio announces. Its title is explicit, an old comfort from the artists, patent fiction, yet dear in all the world: "Death is a robber, but he can't catch me." Use only when you are lost.

BREASTS, BOTTLES AND BABIES: A HISTORY OF INFANT FEEDING, by Valerie A. Fildes. Edinburgh University Press, distributed by Columbia University Press ($30).

The nail test was set down by Pliny the Elder. You put a drop of the milk on your thumbnail; the right stuff spreads gently and retains its form when rocked a little. "Milk which runs off immediately is watery, whereas milk that stays together like honey...is thick." Such objective perceptual clues represent the model of conscious intervention by human beings into the subtly adaptive biochemical system through which mothers feed infants, a system human beings largely share with the entire mammalian class. It is daunting how often reasonable inferences from our partial knowledge have proved damaging in practice.

This chunky volume combs the past for every kind of clue to how infants were fed and why during the preindustrial centuries of Europe, from the first printed medical books in about 1500 to about 1800; its story ends before the era of 19th-century science, nutrition from Liebig and infection control from Pasteur. A general introduction touches on the universal issues from Sumer to Galen, Avicenna and the Ayurvedic writings, partly because the theories of Renaissance medicine in Europe drew heavily on the ancients. The feeding of infants is followed from maternal breast-feeding through wet-nursing to mixed feeding, hand-feeding and on to weaning; the text stays close to its sources, but it never accepts their quirks uncritically.

The thoughtful comparative lists and small-sample statistics that effectively support the lively text (mostly from the safe enclosure of a long appendix) begin with the old writers. In a model of historical exposition the force of argument grows as information increases.

In many parts of the world it is held that the mother's first milk, called colostrum, which persists for three or four days after delivery, is a bad substance. It does not pass Pliny's test. The artists show us what was done: a print portrays a nurse feeding the newborn with a spoon (honey and oil were popular and are mentioned in the Old Testament), and a Tintoretto shows a woman offering the newborn St. John her breast while the mother rests in bed. Before 1673 all writers condemn colostrum as harmful: it was a tradition. The change in ideas began in Paris with an influential work that did not recommend colostrum but did report that some people thought it was useful. Medical author after author modifies the old stance: the stuff is not harmful, and it may do some good. By 1748 William Cadogan boldly recommends that no infant be given anything by mouth before it is put to the breast; harm will come to any child who is denied colostrum.

A dual feedback loop had been noted. The purging of the dark greenish meconium from the newborn intestine is aided by early suckling. And milk fever, with a high mortality, arises a few days after delivery in many mothers who do not breast-feed. The residual milk serves as a focus for infection, which is often made worse by clumsy efforts to remove the milk. Early breast-feeding reduced milk fever to unimportance.

Today it is recognized that maternal antibodies supplied in colostrom protect the infant against infection for about six weeks. A table of infant mortality (up to one year) shows a steep fall among English ducal families after 1750, and parish records confirm the idea that much of this fall was realized during the first month of life. Neither improvements in midwifery, recourse to cow's milk, better nutrition nor any other hypothesis yet proposed fits the evidence (that the gain was registered in early survival) as well as a turn to breast-feeding does.

Subtler still, mothers who breast-feed immediately after giving birth have a stronger emotional bond to an infant than those who may not even see the child for days, and who must then learn how to give suck with distended breasts. It is argued that the concept of the special nature of infants and children, with explicit concern for their welfare, began in Britain in this period. "The change in neonatal feeding practices cannot be disregarded as one factor in the change of attitude of British society towards its children."

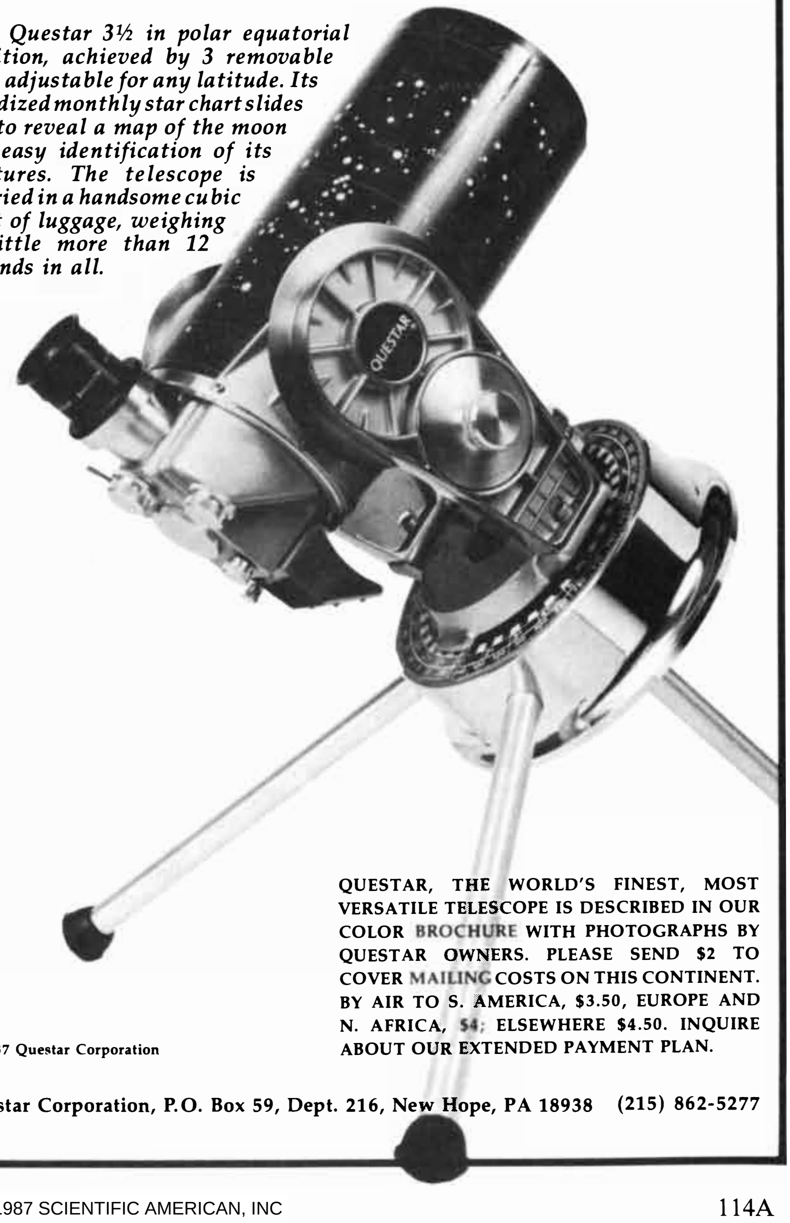
Jump to the other end of infancy: weaning. What is the age of weaning? Here Dr. Fildes offers much original material. First of all, three kinds of weaning age are found in books: the age recommended by a sagacious author, the age said to be common among the people and the age at which children were actually being weaned. The ancient sources—Galen or the Bible or Maimonides—tend to agree on two or three years. (One individual wet-nursing contract in Roman Egypt stipulated 16 months.) The medical books of the 16th and 17th centuries follow the old authors, but factual ages can be found at least for a sample of known, named children. For instance, the six little ones of John Dee, scryer and scientist at court to Elizabeth I, were put to a wet nurse. Their median age of weaning was 14 months. Almost the same median age is given for the five breast-fed children of a clergyman in the mid-17th century. The median remains the same for more than 40 named children of the author's sample—the largest yet published—in confirmation.

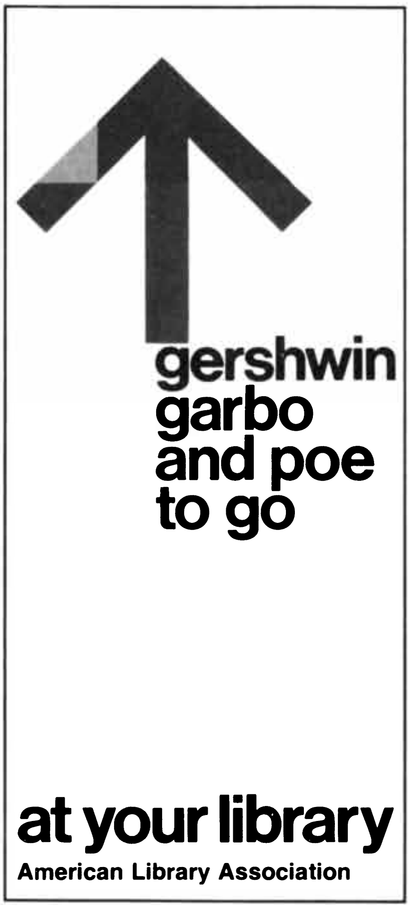
The picture is coherent. The upper- and middle-class women who had depended on a wet nurse during the earlier period were largely breast-feeding by the 18th century. Hand-feeding had become practical and socially acceptable. Whereas earlier the children had been weaned at an age when they could share the diet of the family, in the late 18th century they were weaned while still without teeth, to take milk mixtures from feeding vessels. Some 50 formulas are listed here, the paps and panadas of three centuries: flour, bread, milk, broth.

There is a clear risk in today's world as before. Breast milk, from the mother or from a wet nurse, is a good and pure food. Anything less direct—as the preparer of sterile foodstuffs knows—carries the danger of acute gastrointestinal infection. "The errors in infant nutrition made during industrialisation of western societies are being repeated today. Even though we have the knowledge,...the infant mortality in the Third-World societies today is identical to that of 18th-century London: up to 70 percent of infants did not survive to their second birthday."

STRONG SOLIDS, THIRD EDITION, by A. Kelly and N. H. Macmillan. Oxford University Press ($89).

"This book is about the highest attainable static strengths of solids." It is a tough and technical monograph (first reviewed in this column in 1967), so buoyant with the successes its authors report that it is full of pleasures—many in images—for the determined general reader. Three distinct classes of strong solids are now at hand: metals, glasses, ceramics and composites.





Each receives a long chapter or more of quantitative review, treating the theoretical concepts, data on tests and achievements and a sampling of the calculations.

At bottom the subject is atoms. The graceful bend of the strong fly rod and the failure of the bridge girder alike have atomic roots, long though the chain from atom to macroscopic artifact may be. The recognition that the work required to tear apart two adjoining layers of atoms in a perfect crystal can give an estimate of maximum tensile strength is an old and powerful idealization, and with it the book opens. On that basis we learn that a strong solid should stiffly resist even very small extensions and should have small, closely packed atoms. Ultimate failure in real life is a matter of stress seeking out every weak point, of flow, dislocations, cracks, notches and statistical fiber failure; theoretical cleavage stress is a design criterion for the wonderful one-hoss shay: the whole crystal parts at once.

The argument is decades old; with it atoms entered the domain of the structural engineer. But it is almost in awe that we read of the confirming work in the 1980's. The computer simulates an array of a few thousand atoms. The forces of atomic interaction are explicitly modeled and the array is set free to evolve, each atom moving in response to external forces from its neighbors, fully obedient to the laws of Newton. The shape and defects of the initial array and the boundary stresses applied can be specified. "Typically, each atom...is moved in turn a few thousands or tens of thousands of times," its position computed anew 10 times within each period of atomic vibration. The "experiment" simulates atomic motions over some 10 picoseconds. One modeled glass fiber displayed good agreement with the observed stiffness of the actual glass under study as it extended linearly—to fail (not by neat cleavage) at a stress about half what had been estimated.

Real crystals as well as computer images have been tested against the ideal one-hoss-shay limits. Some time ago tiny single crystals of cadmium were selected under the electron microscope to be flaw-free, with smooth surfaces and without cracks or inclusions. These little whiskers were then tensioned as they grew "in a series of beautiful and amazingly reproducible experiments." They did their stuff just as expected. From this platform of simplicity the authors reason that very strong solids must have the densest network of strong, three-dimensional, directional atom-to-atom bonds. If the bonds do not knit into an interlocking and rigid frame, some atoms can slip away. Only seven elements pass the test: silicon is the heaviest among them, beryllium the lightest.

In the larger world we cannot hope for the flawlessness that was built into the deacon's masterpiece. The tip of a real microcrack can be seen in a dazzling electron-microscope image, made possible by heroic thinning of the film of silica glass that was its host. It looks just like the parabola that preatomic elastic theory assumed to be the shape of the end of a small crack. The failure of brittle solids by the propagation of such tiny internal cracks was rationalized in energy terms in a brilliant paper by A. A. Griffith in 1920, and now we can see it at work. Many details remain to be grasped. Surface notches and the dislocations that arise from missing or extra parts of lattice planes are more complicated but still commonplace flaws, at least qualitatively manageable in theory and practice.

The first class of strong materials treated is the newest: strong glasses and ceramics. The subtle statistics of a population of independent flaws have been worked out and beautifully tested. The same ideas apply to reinforcing fibers. The rule of thumb is that the bigger the sample, the more numerous the flaws.

Understanding is almost all. Contact flaws—tiny scratches and nicks—are fatal to glass fibers. Beware of fine quartzy dust! Once you make sure that your glass rods are not contact-damaged by dust or by their fellows, they will be reliably strong. High-performance ceramics are prepared by mixing pure micron-size particles with the binder and then shaping and firing to make a dense polycrystalline body. Pores, surface flaws and grain size after growth must all be controlled. The results are imposing, if not yet cheap. Surface compression is one clear way to go: thoughtful quenching and glazing have worked, increasing strength up to fivefold.

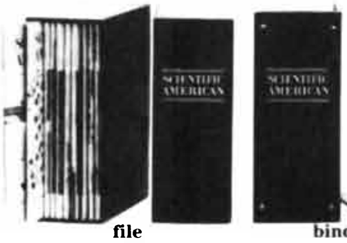
Another path is to block internal cracks. The most successful version is called transformation toughening. Inclusions, calculated to undergo a volume-expanding phase transformation within the temperature range encountered during processing, are incorporated into the microstructure. Zirconia (zirconium oxide) is the best example. That substance undergoes a density change at moderate temperatures that is unsuitably large, so that pure zirconia is not a success as a structural material. When it is alloyed with oxides of calcium or magnesium, however, its higher-temperature phases are stabilized to lower temperatures. Then sufficiently small particles remain metastable and can be transformed into a less dense phase by the push of an approaching crack tip. The particle volume increases, generates compression in the small region around itself and thereby stops crack growth.

The Age of Metals is not yet past. The same level of ingenuity and understanding achieved for ceramics has been attained for metals, which are still the chief strong materials. They are the most versatile ones too, mainly because so many alloys—solid solutions—can be created from the many metallic elements. The theoretical limiting strength of iron in tension is about 3,000 pounds for a wire one millimeter in diameter. Cold-drawn wire of steel can be bought that has a third of that ideal strength. No other engineering materials are now used at as large a fraction of their theoretical strength as the metals are.

The control of internal cracks by the ceramist is matched by the metallurgist in controlling the motion of dislocations, which is important for the toughness of the ductile metals. The main trick is to create particulate obstacles to dislocation movement by manipulating the materials present in the grains. Usually such manipulations make tiny hard inclusions, so that the toughened metal is difficult to machine or weld. One carbon-free steel, an alloy mainly of nickel and iron, can be formed while it is cold, however, before being reheated to induce the strengthening precipitation. Such maraging steels, as they are called, are tough, workable and strong, but they are not corrosion-free.

The last class of strong materials consists of the fiber-reinforced composites. Here the statistical theories of failure have been well tested, and they are applied to bundles of all forms: random lengths, parallel long fibers and the like. The matrix is important, of course. It is now possible to design a composite whose tensile breaking strength along the fiber direction is close to that of the fibers themselves, even if they are small and brittle.

The technology of making and using strong fibers is lively; a review of the art closes this book, but one is left certain of changes to come. The last pages tell of ongoing efforts to incorporate minute, flawless single-crystal whiskers, the strongest solids we know, as fibers for the composite. Silicon carbide whiskers, too short to be spun, can be aligned in thick fluids and serve, for example, to reinforce aluminum alloys. An open and graceful architecture based on strong solids will appear someday, a long time later than strong tools, skis and aircraft.



BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

STARS & STRIPES

Optimum Windward Performance of Sailing Craft. John S. Letcher, Jr., in *Journal of Hydronautics,* Vol. 10, No. 4, pages 140–144; October, 1976.

Comeback: My Race for the America's Cup. Dennis Conner with Bruce Stannard. St. Martin's Press, 1987.

Performance Predictions for *Stars & Stripes.* James C. Oliver, John S. Letcher and Nils Salvesen in *Transactions of the Society of Naval Architects and Marine Engineers,* Vol. 95, in press.

COLLISIONS BETWEEN SPINNING PROTONS

Energy Dependence of Spin-Spin Effects in p-p Elastic Scattering at $90°_{\text{c.m.}}$. E. A. Crosbie, L. G. Ratner, P. F. Schultz, J. R. O'Fallon, D. G. Crabb, R. C. Fernow, P. H. Hansen, A. D. Krisch, A. J. Salthouse, B. Sandler, T. Shima, K. M. Terwilliger, N. L. Karmakar, S. L. Linn, A. Perlmutter and P. Kyberd in *Physical Review D,* Vol. 23, No. 3, pages 600–603; February 1, 1981.

Polarized Scattering Data Challenge Quantum Chromodynamics. Bertram M. Schwarzschild in *Physics Today,* Vol. 38, No. 8, pages 17–20; August, 1985.

Energy Dependence of Spin Effects in $p\uparrow + p\uparrow \rightarrow p + p$. G. R. Court, D. G. Crabb, I. Gialas, F. Z. Khiari, A. D. Krisch, A. M. T. Lin, R. S. Raymond, R. R. Raylman, T. Roser, K. M. Terwilliger, K. A. Brown, L. G. Ratner, D. C. Peaslee, P. R. Cameron, J. R. O'Fallon, T. S. Bhatia, L. C. Northcliffe and M. Simonius in *Physical Review Letters,* Vol. 57, No. 5, pages 507–510; August 4, 1986.

Colour Theory in a Spin. Harry J. Lipkin in *Nature,* Vol. 324, No. 6092, pages 14–16; November 6, 1986.

THE CAUSES OF DOWN SYNDROME

Observations on an Ethnic Classification of Idiots. J. Langdon H. Down in *London Hospital Clinical Lectures and Reports,* Vol. 3, pages 259–262; 1866.

Down's Syndrome as a Model Disease. Charles H. Scoggin and David Patterson in *Archives of Internal Medicine,* Vol. 142, No. 3, pages 462–464; March, 1982.

Down's Syndrome and Alzheimer's Disease: A Review. C. Oliver and A. J. Holland in *Psychological Medicine,* Vol. 16, No. 2, pages 307–322; May, 1986.

THE CLONAL-SELECTION THEORY

Croonian Lecture: On Immunity with Special Reference to Cell Life. Paul Ehrlich in *Proceedings of the Royal Society of London,* Vol. 66, No. 432, pages 424–448; July 24, 1900.

The Natural-Selection Theory of Antibody Formation. Niels K. Jerne in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 41, No. 11, pages 849–857; November, 1955.

A Modification of Jerne's Theory of Antibody Production Using the Concept of Clonal Selection. F. M. Burnet in *The Australian Journal of Science,* Vol. 20, No. 3, pages 67–69; October 21, 1957.

The Acceptance and Rejection of Immunological Concepts. David W. Talmage in *Annual Review of Immunology,* Vol. 4, pages 1–11; 1986.

SALT TECTONICS

Gravity, Deformation and the Earth's Crust. Hans Ramberg. Academic Press, 1981.

Age, Budget and Dynamics of an Active Salt Extrusion in Iran. C. J. Talbot and R. J. Jarvis in *Journal of Structural Geology,* Vol. 6, No. 5, pages 521–533; 1984.

External Shapes, Strain Rates, and Dynamics of Salt Structures. M. P. A. Jackson and C. J. Talbot in *Geological Society of America Bulletin,* Vol. 97, No. 3, pages 305–323; March, 1986.

Weakening of Rock Salt by Water during Long-Term Creep. Janos L. Urai, Christopher J. Spiers, Hendrik J. Zwart and Gordon S. Lister in *Nature,* Vol. 324, No. 6097, pages 554–557; December 11, 1986.

GALLIUM ARSENIDE TRANSISTORS

Electron Dynamics in Short Channel Field-Effect Transistors. Jacques G. Ruch in *IEEE Transactions on Electron Devices,* Vol. ED-19, No. 5, pages 652–654; May, 1972.

Physics of Semiconductor Devices. S. M. Sze. John Wiley & Sons, Inc., 1981.

Millimeter-Wave GaAs FET's Prepared by MBE. B. Kim, H. Q. Tserng and H. D. Shih in *IEEE Electron Device Letters,* Vol. EDL-6, No. 1, pages 1–2; January, 1985.

Ultrafast Chips at the Gate. Herb Brody in *High Technology,* Vol. 6, No. 3, pages 28–35; March, 1986.

GAZELLE KILLING IN STONE AGE SYRIA

Prehistoric Hunters of the High Plains. George C. Frison. Academic Press, 1978.

Jawa: Lost City of the Black Desert. S. W. Helms. Cornell University Press, 1981.

Black Desert Survey, Jordan. Alison Betts in *Levant: Journal of the British School of Archaeology in Jerusalem and the British Institute at Amman for Archaeology and History,* Vol. 15, pages 1–10, 1983; Vol. 16, pages 25–34, 1984; Vol. 17, pages 29–52, 1985.

AIR POLLUTION BY PARTICLES

Smoke, Dust and Haze. S. K. Friedlander. John Wiley & Sons, Inc., 1977.

Aerosols: Anthropogenic and Natural, Sources and Transport. In *Annals of the New York Academy of Sciences,* Vol. 338; 1980.

Receptor Models. Glen E. Gordon in *Environmental Science & Technology,* Vol. 14, No. 7, pages 792–800; July, 1980.

Characterization of the Aerosol in the Great Smoky Mountains. Robert K. Stevens, Thomas G. Dzubay, Robert W. Shaw, Jr., William A. McClenny, Charles W. Lewis and William E. Wilson in *Environmental Science & Technology,* Vol. 14, No. 12, pages 1491–1498; December, 1980.

THE AMATEUR SCIENTIST

Dynamic Processes in Soap Films. Karol J. Mysels in *Journal of General Physiology,* Vol. 52, No. 1, Part 2, pages 113s–124s; July, 1968.

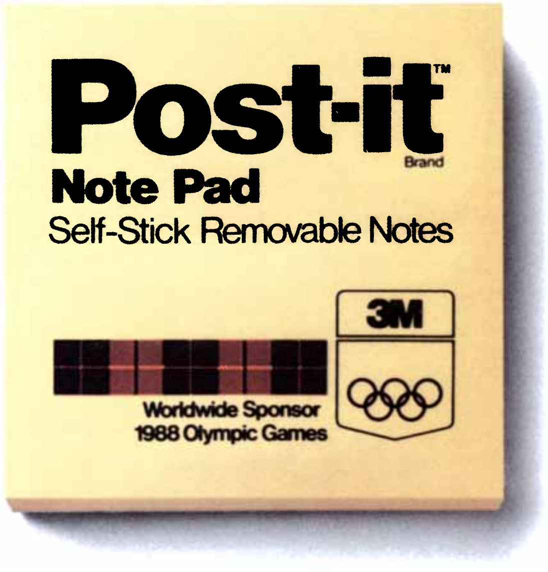
COMPUTER RECREATIONS

The Theory of Computer Science: A Programming Approach. J. M. Brady. Chapman and Hall, Ltd., 1977.

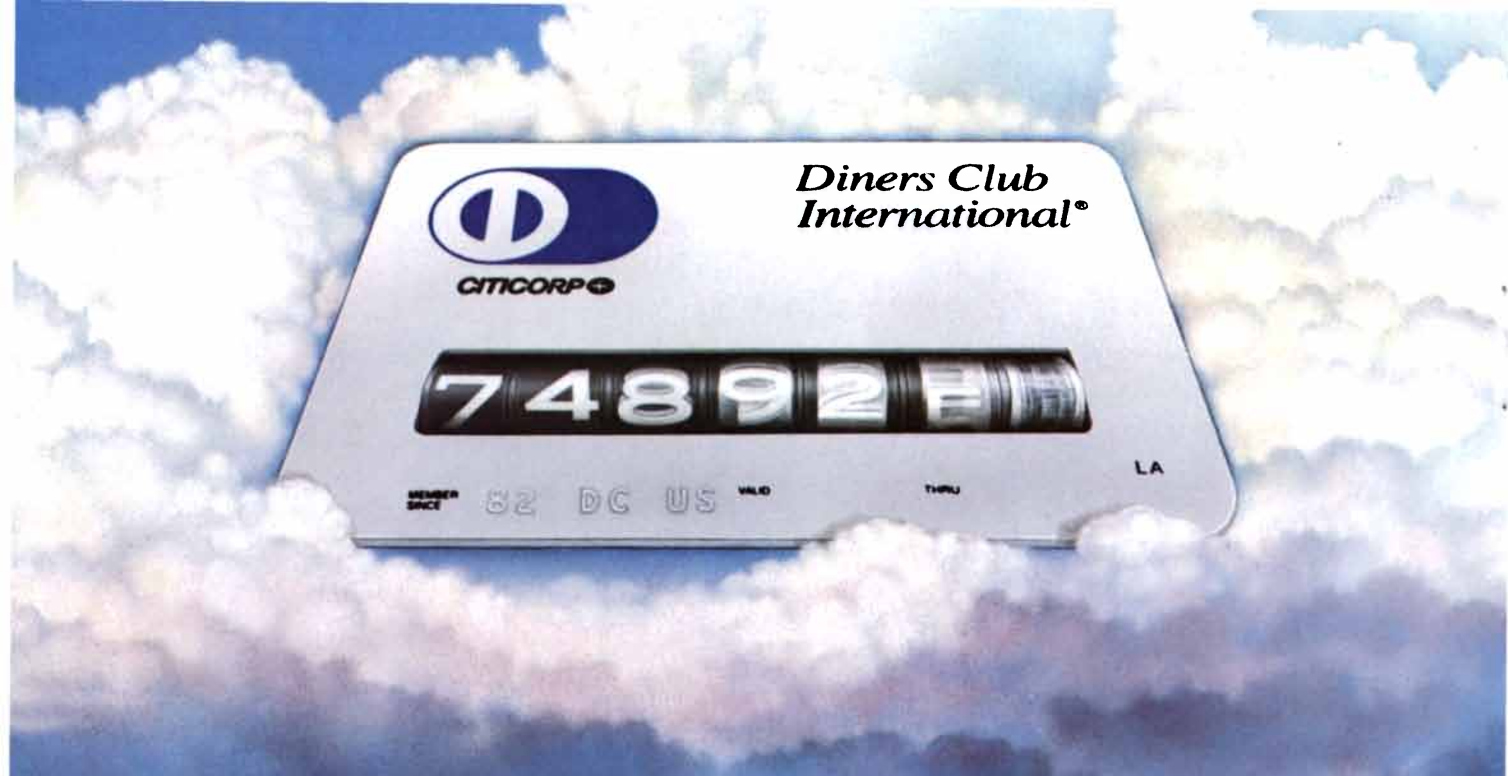
The Official Scrabble^{R} Players Dictionary. Pocket Books, 1978.

The Oxford Guide to Word Games. Tony Augarde. Oxford University Press, 1984.

ONE MINUTE MANAGERS NEED TEN SECOND MEMOS.



When People Count On You,
Count On Post-it™ Notes.
BRAND


Diners Club International®
CITICORP
748921
MEMBER SINCE 82 DC US VALID THRU
LA

Diners Club International®
CITICORP
3812 345678 0009
R C ROBBINS
82 DC US
LA