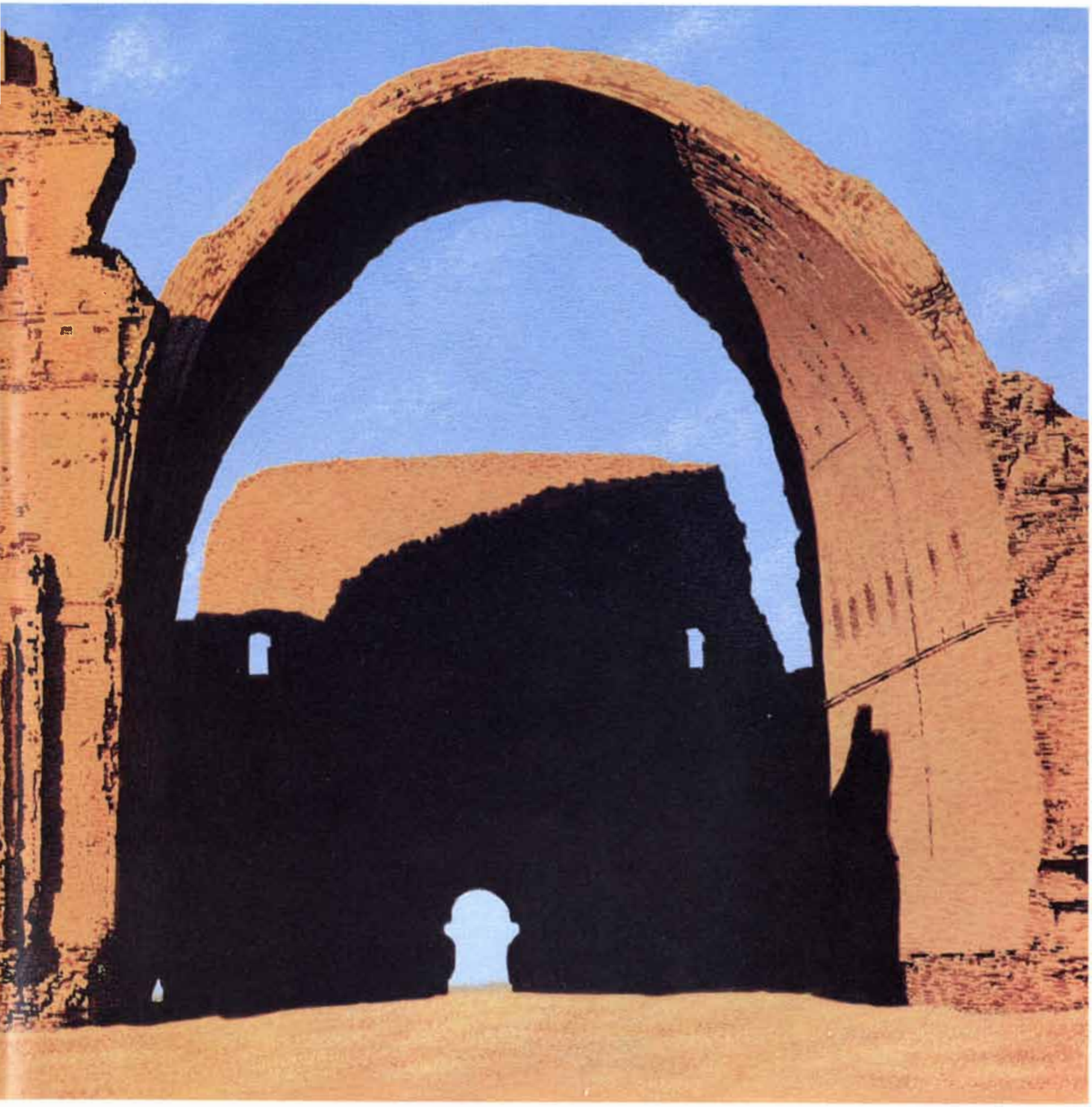
SCIENTIFIC AMERICAN



ANCIENT NEAR EASTERN VAULTS

$2.50

July 1987


BMW M6


BMW L6

Having a Grand Time...
Liqueur
Grand Marnier
Liqueur
PARIS FRANCE
For gift delivery anywhere call 1-800-CHEER-UP (except where prohibited by law).
Product of France. Made with fine cognac brandy 80 proof. ©1986 Carillon Importers, Ltd., Teaneck, N.J.

Established 1845 **SCIENTIFIC AMERICAN** July 1987 Volume 257 Number 1

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.

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 ______________________

THE COVER

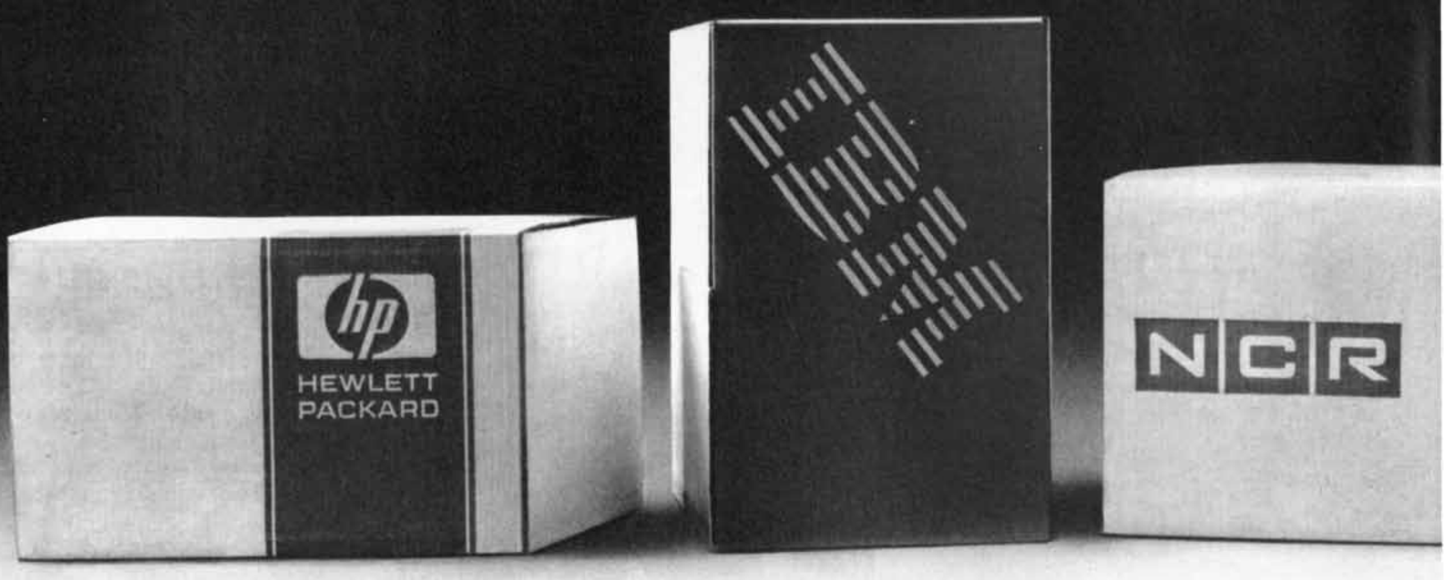
The cover shows the ruins of the Taq Kisra, a vast audience hall in the palace at Ctesiphon, south of Baghdad. Built by the Sassanians (a Persian people) sometime between the third and sixth centuries A.D., the Taq Kisra is roofed by the world's largest single-span vault of unreinforced brick. The vault spans 25.5 meters and is 28.4 meters high; it is the finest example of an ancient Near Eastern vaulting method known as the pitched-brick method (see "Arches and Vaults in the Ancient Near East," by Gus W. Van Beek, page 96). Unlike most Near Eastern arches and vaults, which were made of sun-dried mud brick, the Taq Kisra was constructed of lightly fired brick.

THE ILLUSTRATIONS

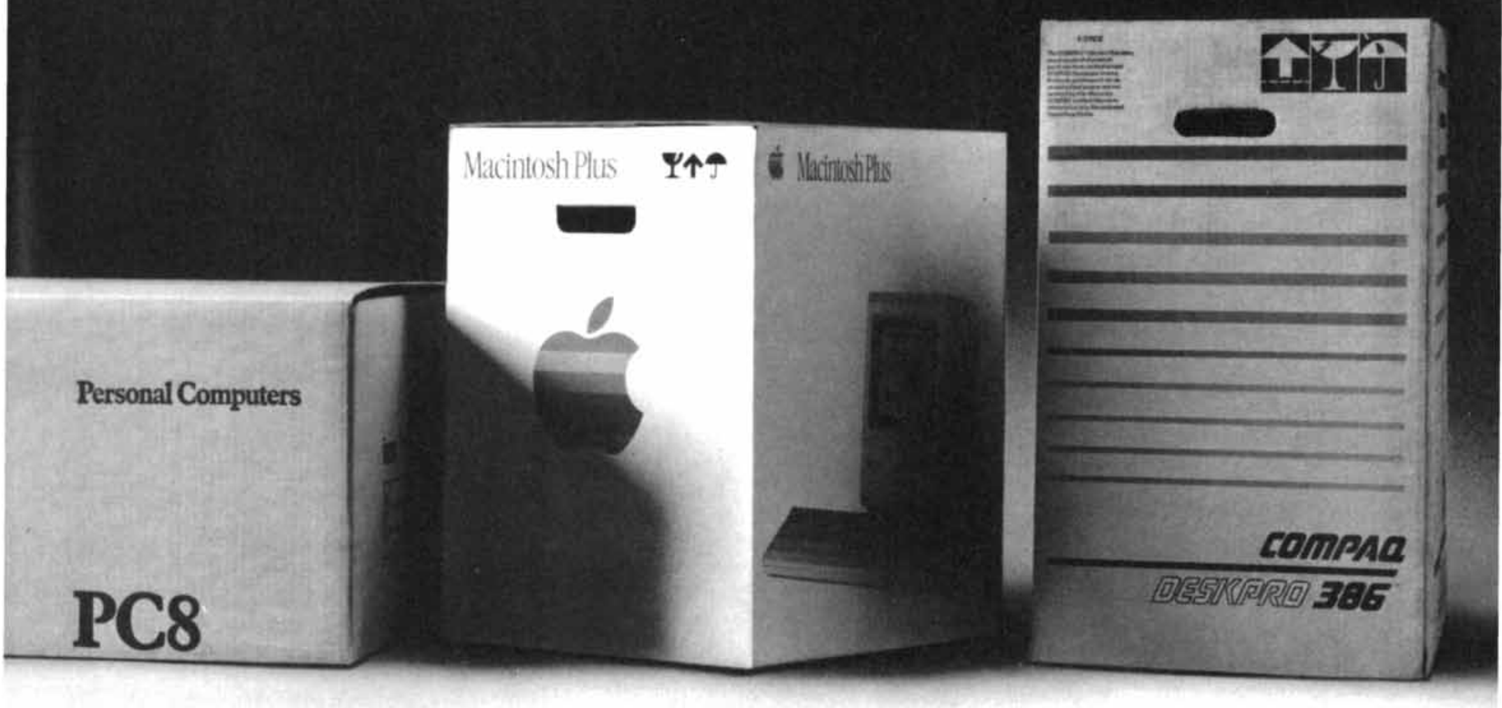
Cover painting by Marvin Mattelson

Page	Source
35	Dale E. Deemer, Allegheny Ludlum Steel Corp.
36–40	George V. Kelvin, Science Graphics
41	AB Sandvik Steel
43	Lubert Stryer
44	Tom Prentiss
45–50	George V. Kelvin, Science Graphics
53–54	Jet Propulsion Laboratory
63	George Retseck
64	Jet Propulsion Laboratory (*top*); Mark Showalter and Maren Cooke, Ames Research Center (*middle*); George Retseck (*bottom*)
65–66	George Retseck
69	U.S. Corps of Army Engineers
70	Andrew Tomko
71	United Press International
72	Michael Goodman
73	Stephen P. Leatherman
76	Michael Goodman
77	Robert Dolan
79	Samuel L. Howard
80–81	Tom Prentiss
82	Stanley F. Hayes, Willy Burgdorfer and M. D. Corwin, Rocky Mountain Laboratories (*top*); New York State Department of Health (*bottom*)
83	Sally Black
85	Idaho National Engineering Laboratory
86–89	George Retseck
91	Karl J. Niklas
92	Patricia J. Wynne
93	Karl J. Niklas (*top*), Patricia J. Wynne (*bottom*)
94	Patricia J. Wynne (*left*), Karl J. Niklas (*right*)
95	Karl J. Niklas
97	Gus W. Van Beek
98–100	Alan D. Iselin
101	Alan D. Iselin (*top left and bottom*), David B. Stronach (*top right*)
102	Gus W. Van Beek
103	Paul Logsdon
105–107	Michael Goodman
109	James P. Crutchfield
110	Edward Bell (*top*), Gordon Hughes (*bottom*)
111	Gordon Hughes

3M data cartridge available in five



tape backup is now popular models.



The makers of even the most popular PCs know it's insane to go to market without a reliable backup.

So the hunt began.

They each sat through hundreds of action-packed hours of blank, scrambled and frozen screens to see which backup system backed up best.

In the end, they each chose 3M data cartridge tape technology.

Why?

We've been covering computer and human errors almost longer than computers and humans have erred together.

Not only did we invent and patent data cartridge tape technology, we've had 16 years to make it better. Through every technological breakthrough, we've consistently proven to be the best way to back up data.

Just ask Apple.® And NCR.® And HP.® And IBM.® And COMPAQ.®

Trademarks/owner: Apple/Apple Computer Inc.; NCR/National Cash Register; HP/Hewlett-Packard; IBM/International Business Machines Corp.; COMPAQ/COMPAQ Computer Corp.

LETTERS

To the Editors:

In "Optical Neural Computers" [SCIENTIFIC AMERICAN, March] Yaser S. Abu-Mostafa and Demetri Psaltis convey some of the promise of this new and intriguing technology. We found the article unsatisfying in certain respects, however. Because of the lack of technical detail concerning how an optical neural computer performs its "closest match" operation during pattern recognition, the reader may misinterpret the true strengths and weaknesses of this technology.

One potential point of confusion arises when the authors state: "A simple algorithm will therefore never serve to solve a random problem," where a random problem is, for instance, recognizing an object such as a tree. We can point to the equations describing the physics of the hologram employed by the optical computer and say, "That is the algorithm." It appears that the authors are confounding the notions of algorithm and data. In an optical neural computer we view the algorithm as the mathematical description of the physics of the device. The data are the scenes that are recorded in the device. A conventional computer could in principle carry out the very same algorithm as an optical neural computer (albeit more slowly). The point we want to make is that this confusion between algorithm and data may leave the reader with the impression that an optical computer is doing something that a more conventional computer cannot do, and that this special (almost magical) performance allows it to solve hitherto unsolvable problems. This is not an accurate description of the optical computer's capabilities.

Another potential point of confusion arises when the authors state: "Even a young child can [distinguish a tree from a telephone pole]" and "A generalized definition of a tree based on the underlying regularity [such as possessing a trunk and branches] could lead to erroneous identifications." In fact, we would not find it particularly unusual if on first seeing a telephone pole a young child asked, "Why does that tree have wires hanging off it?" Nevertheless, the same child probably could tell apart a toy tree, a real tree in the backyard and a picture of a tree in a book. Those are distinctions that cannot necessarily be made on the basis of appearance alone, without an understanding of the context. The child is doing image understanding and not simple pattern recognition. The errors the child might make show that image understanding in human beings entails making speculative generalizations based on underlying regularities, and also remembering any exceptions to the rules (for example, the fact that telephone poles resemble trees but are not trees).

Furthermore, it is important to emphasize that the proposed optical neural computer is an analog computation device and *not* a general-purpose computer, even though it is faster at solving a particular class of problems.

In sum, it is certainly true that optical neural computers are an exciting technology, which holds great promise for performing pattern recognition and other artificial-intelligence tasks. Such computers can store and retrieve large volumes of information, perform a type of associative recall using noisy or incomplete patterns, and do all of this in parallel. In that sense the optical computer represents a potential advance in computer architecture and hardware. To imply, however, that such a device can "program itself" to solve pattern-recognition problems as young children do, in a manner that cannot be accomplished using a "simple algorithm," exaggerates the capacities of these devices and trivializes the capacities of human beings.

JAMES C. SPOHRER

W. ROBERT BERNECKY

New Haven, Conn.

To the Editors:

We are in general agreement with most of the writers' points. We fully agree, for instance, that a neural computer is well suited for a specific class of problems only (what we refer to as random problems) and in that respect is not a "general-purpose computer" in the traditional sense. Also, of course, the hologram is not doing anything magical—nothing that cannot be simulated by a digital computer.

The point is that an optical neural architecture can do it much more efficiently. That image understanding also plays an important role (another point brought up in the letter) is also true. On the differences between algorithms and data, however, we believe the letter may create confusion instead of clarifying valid points or reemphasizing them. We shall attempt to clarify the issue.

There is a difference between an algorithm that solves a problem and one that learns a problem. In the case of recognizing trees the solving algorithm is the final product, which can inspect a picture and decide whether or not it contains a tree. The learning algorithm, which makes use of the equations of the hologram, is given a large number of pictures and is told whether or not they contain trees. It uses those data in order to construct the solving algorithm.

To consider the equations describing the physics of a hologram as being identical with the algorithm by which an optical computer would recognize a tree is not appropriate. To do so is analogous to considering the equations of solid-state physics to be an algorithm for multiplication by an electronic computer.

YASER S. ABU-MOSTAFA

DEMETRI PSALTIS

To the Editors:

"Human-powered Watercraft," by Alec N. Brooks, Allan V. Abbott and David Gordon Wilson [SCIENTIFIC AMERICAN, December, 1986], was fascinating, but the drawing of a dugout canoe on page 124 hardly does justice to the possibilities of this type of craft. Such primitive dugouts do exist, and I have traveled in one at Mongu, rainy-season capital of the Litunga of Barotse, in Zambia.

A peak of human-powered watercraft technology, however, is represented by the waka of the New Zealand Maori. On February 7 I was privileged to be one of the paddlers on the *Waka Ngatokimatawhaorua,* which made the journey from Waitangi to Whangaroa—a distance of about 64 kilometers along the coast of Northland, New Zealand. The performance of this waka compares well with that of today's kayaks. Two modern seagoing kayaks accompanied the waka from the entrance of Waitangi harbor to the landing in Whangaroa harbor.

Ngatokimatawhaorua is roughly 38 meters long, and the 80 paddlers were able to complete the journey in 11½ hours. The hull of this waka is made of three dugout logs, cut from two kauri trees and lashed together. The joints are caulked with New Zealand flax. The vessel was built to traditional designs in 1940.

The standard European reference book on the waka is *The Maori Canoe,* by Elsdon Best. The main significance of the waka to the Maori is cultural. Maori song, ritual and tradition are steeped in the spirit of the waka.

TIM HASSALL

Whangaroa College
Northland, New Zealand

The Rolex Awards for Enterprise

For the Good of the Planet and its People

The Rolex Awards for Enterprise were established in 1976 to underwrite the projects of enterprising individuals who are committed to advancing the common good. Grants have been awarded in the fields of Applied Science and Invention; Exploration and Discovery; and the Environment. By helping to translate goodwill into action, The Awards not only serve mankind and nature as a whole, they also foster awareness of the moral as well as the intellectual dimensions of the spirit of enterprise. These are the 1987 laureates.

Expanding the frontiers of modern healthcare.

Jacques Luc Autran, founder of Seamen Without Frontiers, a philanthropic association, turned a fishing vessel into a floating hospital to bring healthcare and education to isolated societies, such as the islanders of the Maldives.

Preserving endangered species.

Stephen Kress's innovative techniques successfully repopulated near-extinct colonies of birds on Maine islands and will soon be employed in attempted recolonizations in the south of Japan.

New insights into evolution.

Entomologist Pierre Morvan will organize an expedition to Nepal to study the effects of extreme geographic isolation on the formation of particular species, and will complete a book on the subject.

Religious teaching and environmental education.

Nancy Lee Nash will glean teachings about nature from Buddhist writings and incorporate them into an educational syllabus to be used in Thailand colleges.

Demystifying an ancient culture.

Johan Gjefsen Reinhard will explore the peaks and lakes of the Peruvian Andes to determine the role of those ancient religious sites in the formation of pre-Colombian society.

The Spirit of Enterprise, The 1987 Rolex Awards, a book about the five laureates' projects, as well as 238 other outstanding proposals, is available in bookstores and from the publisher, Van Nostrand Reinhold.

Additional information about the awards is available from The Secretariat, The Rolex Awards for Enterprise, P.O. Box 178, 1211 Geneva 26, Switzerland.

ROLEX

$La_{2-x}B$

This is the formula for something that looks ordinary, but does something extraordinary.

It is a superconductor, a material that carries electricity with no loss of power. Until recently, superconductors worked only at extremely low temperatures, limiting their use.

In a major breakthrough, two IBM scientists discovered that this class of materials superconducts at a much higher temperature, raising the possibility of expanding this technology's use.

This discovery, by J. Georg Bednorz and K. Alex Müller, has sparked an explosion of research that could yield profound change. In fact, many think that if remaining obstacles can be overcome, superconductors could lead to major advances in many areas of human endeavor, including computers.

IBM is proud of its scientists' innovative achievement.

Because innovation not only makes breakthroughs possible. It makes better products for our customers possible. IBM

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1937: "The family doctor of the horse-and-buggy era was largely dependent on his unaided five senses in making a medical examination. Solemnly he inspected your skin for eruptions, fingered your pulse and peered into your throat and ears. But the diagnosing of disease is no longer a matter of guess-work, based upon the observance of external symptoms. When you are ill today, your physician augments the testimony of his skilled senses by a multitude of laboratory tests that give him precise information about the nature and extent of your ailment. These tests enormously increase the accuracy of medical diagnosis, minimize the expense and duration of treatment and always enhance the possibility of making a cure."

"An interesting report recently received from the Northwestern University Traffic Survey Institute shows a definite and indisputable relationship between law enforcement and automobile accidents. By law enforcement is meant not mere arrest of motor vehicle law breakers but actual convictions. If only every motorist knew definitely that a violation would bring swift and sure retribution, each driver would be more careful."

"When, in December, 1903, Orville and Wilbur Wright made the flight that ushered in the Air Age, their flimsy biplane was powered by a homemade engine of 30 horsepower. In less than 34 years the genius of the engine builder has multiplied this power 50-fold and at the same time has so enormously improved the reliability, the smoothness and the all-around performance of the aircraft engine that it would indeed be difficult to recognize the modern engine as something akin to that which made history at Kitty Hawk."

"What are the possibilities, if any, of glass as a textile fiber? Glass for textile application is here. Broken into tiny filaments by steam under terrific pressure, glass is being assembled into strands, spooled into thread and yarn on modern textile machines and woven into cloth of pure glass. Such fabric finds its greatest use at this time as an insulating and filtering material. It is not unreasonable to assume that the day is not far distant when textile glass will be successfully combined to weave draperies, curtains and tapestries for theaters, ships and hotels to reduce fire hazards."

"A new all-time high for motor vehicle registration was made in 1936. Registrations in the United States totaled 28,221,291, of which 24,197,685 were passenger vehicles and 4,023,606 trucks and tractor-trucks."

SCIENTIFIC AMERICAN

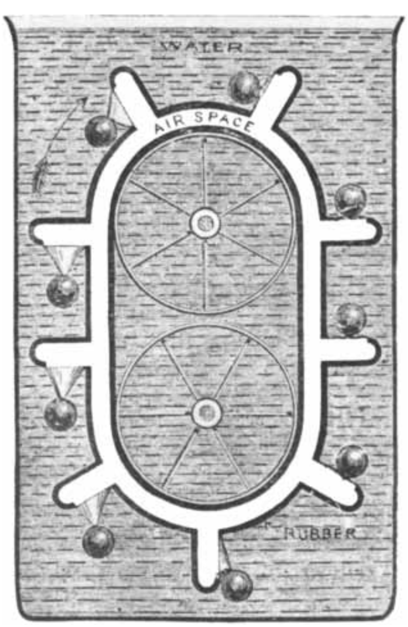
JULY, 1887: "Recent reports from Panama indicate that work will shortly cease on the line of the proposed canal. Unless the difficulties in the way have been greatly exaggerated, a further loan asked of the deluded investors, for the most part poor people, would be a cruel wrong, only serving to postpone impending disaster and raise hopes that cannot be realized. Now, when it is too late, the French investors will realize that the unfavorable reports from the United States that have followed the work from its very inception have not been animated by ungenerous motives. Reports say that both the French and United States governments will be asked in turn to take up and carry out the project of a canal at Panama, but it is not likely that either will do so."

"A remarkable paper by M. Cary Lea is now in course of publication in the *American Journal of Science,* in which the author gives the results of his laborious and extensive investigations concerning the salts of silver and their relations to the photographic image. The ease with which the author produces the most splendid colors, their richness and variety, is really marvelous. He reaches the gratifying conclusion that photography in natural colors is now among the possibilities of science."

"Shortly after the disastrous defeat of *Galatea* in American waters last fall, it was rumored that a final and desperate effort was to be made to wrest the famous America or Queen's cup from this side of the Atlantic, and to that end a new craft was being built. The rumor was finally confirmed, but the whole matter was kept shrouded in mystery until *Thistle* was launched. The ease with which she has defeated all the English crack yachts makes her a most serious competitor."

"The examinations as to tea, coffee and sugar conducted by Edward G. Love, Ph.D., for the New York *World* resulted as follows: of the samples of tea, 88 were not adulterated and 12 were adulterated, mostly with 'lie tea' and foreign leaves; of the samples of ground coffee, 72 were unadulterated and 28 were adulterated, mostly with chicory and peas; of the sugar samples, 98 were pure and only 2 adulterated with starch glucose. As to the weights of the samples, those of 270 were found to be correct whereas 30 were light."

"A correspondent describes the following 'perpetual motion machine': 'It is an endless rubber tube, with projections, on which are fastened thin rubber bags, and a small weight is attached to each bag. The bags are filled with air when the weight hangs down, and when it comes on top it presses the air out and through the hollow projection and tube into the next bag that comes in position. When placed over two wheels in water, the bags filled with air should be lighter and rise, while the other side with the air forced out should sink.' [Editor's note] Each bag, as it comes into position at the bottom of the left tube, will be filled with air expelled from a bag at the top. The weights will descend a certain amount, one in expanding and the other in contracting the bag. The lower bag has to be expanded against a heavy water pressure. Thus each weight will suffer 'lost motion' and a constant loss of power in this inoperative falling of the weights. Hence the machine will not move."



A "perpetual motion machine"

HOW GM IS TAKING THE LEAD IN QUALITY

NOT JUST WORLD CLASS—A NEW, SCIENTIFIC STANDARD OF EXCELLENCE.

People were calling the heartland of American industry "The Rust Belt." It was 1980. We said there would be an American industrial renaissance. And we meant it.

GM laid out a four-part strategy. And went to work.

At the heart of the strategy is the automobile, and at the heart of the automobile is the drivetrain. We were determined to set a new standard of excellence with the automobile in operation.

And we decided to do it the hard way. Instead of using engineering specifications or a survey method that fit well with our strengths and minimized our weaknesses, we asked you, our customer, what you want and need in an automobile in operation, what we call driveability.

Then we took the desires of the most demanding drivers, the 90th percentile, and called that standard our minimum. We said that every GM car, not just those that cost $25,000 or more, would have to meet the 90th percentile standard.

World class, which had been sufficient during the early stages of our strategy, was not a tough enough standard anymore. The GM Uniform Test Standard is a dynamic measure of excellence in engineering and manufacturing on a scale determined by the customer. It is the most rigorous test in the industry.

Across our entire 1987 production, from the Allante to the lowest priced car we sell, 96% of all GM vehicles tested meet the driveability expectation of the most demanding customers—the 90th percentile.

Here are some of the tests: After sitting out all night in low temperatures, the cars are checked for ease of starting and ability to back out of the garage and accelerate when cold. Then we check for idling at stop lights and smooth acceleration to 15, 25, 35, 45 and 55 mph maneuvers.

Every car is checked for performance in hard braking to a stop and then accelerating into traffic. And at the same time, the transmissions are evaluated for shift smoothness, noise, and overall operation.

When that's all done, we do it all over again under hot operating conditions.

The result: so far in 1987, 96% of all GM cars tested meet or exceed the demanding driver's standard.

On other aspects of quality—the fit and finish of the parts other than the engine and transmission—GM is also making enormous strides toward setting new standards of excellence.

And our goal is to keep raising the standard until GM cars stand above their competition in every category and every price range.

We've made this leap in quality here in America. In Fort Wayne, Ind., and Linden, N.J. In Michigan and Georgia and California. We are doing it here. With the best people and the best technology in the world.

We had a vision. We believed an American industrial renaissance was possible. And it is!

We are demonstrating to our fellow Americans in industry that timidity is not the answer. Retrenchment is no solution. At GM, we say, "Go for it!" And we have.

The vision is paying off.

Make Your Own Comparison

We invite you to visit any GM dealer to test drive any new GM car. Compare its driveability to your demanding standards. Then compare it to any of the cars built by our competitors.

For example: Take any expensive Japanese car, with air conditioning, automatic transmission, and so on, and compare it to a medium-priced GM car. You're the customer, you decide which car has the best driveability. See for yourself how GM's vision is paying off.

This advertisement is part of our continuing effort to give customers useful information about their cars and trucks and the company that builds them.

GM
MARK OF EXCELLENCE

Chevrolet • Pontiac
Oldsmobile • Buick
Cadillac • GMC Truck

THE AUTHORS

JULIAN SZEKELY ("Can Advanced Technology Save the U.S. Steel Industry?") is professor of materials engineering at the Massachusetts Institute of Technology. He became interested in innovative steelmaking techniques in 1977 while writing a Government report. Both his B.Sc. (1959) and his Ph.D. (1961) are from the Imperial College of Science and Technology in London. After working for a consulting firm in London and teaching at Imperial College he joined the faculty of the State University of New York at Buffalo in 1966. He moved to M.I.T. in 1976. Szekely joined the National Academy of Engineering in 1982.

LUBERT STRYER ("The Molecules of Visual Excitation") is Winzer Professor of Cell Biology at the Stanford University School of Medicine. He earned a B.S. at the University of Chicago in 1957 and an M.D. at the Harvard Medical School in 1961. Between 1961 and 1964 he held a Helen Hay Whitney research fellowship, spending part of his tenure as a visiting investigator at the Medical Research Council's Laboratory of Molecular Biology in Cambridge, England. He taught biochemistry at Stanford between 1963 and 1969, when he was appointed professor of molecular biophysics and biochemistry at Yale University. He returned to Stanford in 1976. Stryer is a member of the National Academy of Sciences.

JEFFREY N. CUZZI and LARRY W. ESPOSITO ("The Rings of Uranus") share a fascination with the structure and dynamics of planetary rings. Cuzzi is a member of the Voyager Imaging Team at the National Aeronautics and Space Administration's Ames Research Center. He has a B.S. from Cornell University, awarded in 1967, and a master's (1969) and a doctorate (1973) from the California Institute of Technology. He taught for one year at the University of Massachusetts at Amherst and then, between 1974 and 1976, was employed by the National Research Council as a National Academy of Sciences–NRC associate. In 1978, after two years in the Space Science Laboratory at the University of California at Berkeley, Cuzzi moved to Ames. Esposito got an S.B. at the Massachusetts Institute of Technology in 1973 and first worked with Cuzzi while writing his doctoral dissertation at the University of Massachusetts at Amherst; he earned his degree in 1977. Since then he has been in the department of astrophysical, planetary and atmospheric sciences at the University of Colorado at Boulder and has devoted most of his time to the Voyager mission. Esposito is chairman of the Voyager Rings Science Working Group.

ROBERT DOLAN and HARRY LINS ("Beaches and Barrier Islands") began collaborating about 15 years ago when they met through the U.S. Geological Survey. Dolan, professor of environmental sciences at the University of Virginia, has been investigating barrier islands for 25 years and has worked part time for the Geological Survey for almost as long. He has a B.S. (1958) from Southern Oregon College and an M.S. (1959) from Oregon State University. Dolan went to Virginia in 1965 after getting a Ph.D. from Louisiana State University. Lins is currently completing his doctoral dissertation at Virginia. He has been with the Geological Survey since 1969 and is currently a hydrologist in the Water Resources Division in Reston, Va. Lins received his bachelor's degree in 1971 at the University of Maryland and his master's in 1978 at the University of Delaware.

GAIL S. HABICHT, GREGORY BECK and JORGE L. BENACH ("Lyme Disease") are affiliated with the department of pathology at the State University of New York at Stony Brook. Habicht, an associate professor there, earned a Ph.D. from Stanford University in 1965. She held postdoctoral positions, first at Rockefeller University and then at the Scripps Clinic and Research Foundation, between 1965 and 1971, when she moved to Stony Brook. Beck got a B.S. from S.U.N.Y. at Albany in 1982 and then worked at the Veterans Administration Medical Center in Northport, N.Y. as a research biologist, before taking his present position as a research staff scientist at Stony Brook. Benach is a research scientist with the New York State Department of Health and also teaches at Stony Brook as an associate professor of pathology. His investigations focus on the diseases transmitted by ticks. Benach earned a B.S. (1966) and a Ph.D. (1971) from Rutgers University and then joined the Department of Health.

JOHANN RAFELSKI and STEVEN E. JONES ("Cold Nuclear Fusion") met in 1983 at the European laboratory for particle physics (CERN).

Rafelski, who got his Ph.D. from the University of Frankfurt in 1973 and did postdoctoral work at the University of Pennsylvania and the Argonne National Laboratory, has worked at CERN since 1977. Concurrently he taught theoretical physics, at Frankfurt between 1979 and 1983 and at the University of Cape Town until this summer. Rafelski has often collaborated with American investigators, in part through frequent visits to the U.S. National Bureau of Standards. Jones teaches physics and astronomy at Brigham Young University, where he received his B.S. in 1973. He was awarded a Ph.D. in physics from Vanderbilt University in 1978. After a year of postdoctoral work at Cornell University and the Los Alamos Meson Physics Facility (LAMPF), he did research on fusion at the Idaho National Engineering Laboratory. He went to Brigham Young in 1985. In addition to his work there Jones takes part in studies of muon-catalyzed fusion at several other laboratories including LAMPF, the Triuniversity Meson Physics Facility in Vancouver, the Rutherford-Appleton Laboratory in England and the Japanese National Laboratory for High-Energy Physics in Tsukuba.

KARL J. NIKLAS ("Aerodynamics of Wind Pollination") studies the biomechanical properties of plants as an associate professor of botany at Cornell University. He holds a B.S. in biology and mathematics from the City College of the City University of New York; his M.S. and Ph.D., awarded in 1974, are from the University of Illinois at Urbana-Champaign. He attended the University of London as a Fulbright-Hayes Fellow and in 1975 became a curator at the New York Botanical Garden. In 1978 he moved to Cornell, where he also holds joint appointments in plant biology and in ecology and systematics. In 1985 Niklas received a Guggenheim Fellowship to explore models of pollination.

GUS W. VAN BEEK ("Arches and Vaults in the Ancient Near East") has been curator of Old World Archaeology at the Smithsonian Institution since 1967 and has directed 11 field seasons at the Tell Jemmeh site in Israel. He has a B.A. from the University of Tulsa, a bachelor of divinity from the McCormick Theological Seminary and a Ph.D. from Johns Hopkins University, which he received in 1953. He remained at Johns Hopkins in the Arabian Publication Project for six years before joining the Smithsonian as an associate curator in 1959. Van Beek is currently writing a book on sun-dried mud-brick construction.

SCIENCE AND THE CITIZEN

Sweet Talk

Since its introduction in 1981 the artificial sweetener aspartame, sold as NutraSweet, has soared in popularity; it has taken the place of sugar in so many diet drinks and foods that by 1985 its annual U.S. consumption exceeded 3,500 tons. Yet recent experiments are re-igniting concerns expressed at the time of introduction that aspartame may affect brain chemistry, even when it is consumed in near-normal amounts. Several laboratories have evidence suggesting that large but not abnormal doses cause behavioral changes in human beings. Clinical investigators presented the new results to industry executives and Federal Government officials at an informal meeting in Washington in May.

One investigator, Paul Spiers of Beth Israel Hospital in Boston, gave preliminary results indicating that normal volunteers on a high-aspartame diet (equivalent to the amount of aspartame in about 12 cans of diet soda consumed in 24 hours by someone weighing 110 pounds) do worse than matched controls on standard tests of higher brain function; their performance (unlike that of the controls) got worse on successive tests. Subjects also complained more of headaches and dizziness when they took aspartame.

The effects are thought to be due to phenylalanine, an amino acid that is one of aspartame's three constituents. (It is also sold as a dietary supplement in "health food" stores.) Phenylalanine is a component of all protein, but brain levels are higher when it is taken in aspertame rather than in proteins, whose other amino acids impede its uptake. Phenylalanine has chronic neurotoxic effects: phenylketonuria, a disease of infancy that causes mental retardation and seizures if it is not treated, results from inherited defects in the enzyme phenylalanine hydroxylase. What is more, a number of healthy people—as many as one in 50 or 70—fail to metabolize phenylalanine efficiently because they have only a single adequate copy of the phenylalanine hydroxylase gene rather than the normal two copies. Blood phenylalanine levels of such individuals have been shown to reach peaks higher than is normal for a given dose.

Spiers's experiments used the dosage of aspartame (50 milligrams per kilogram of body weight per day) that is the "acceptable daily intake" established by the Food and Drug Administration. Results presented by Louis J. Elsas of the Emory University School of Medicine indicate that in individuals who have only one effective phenylalanine hydroxylase gene, phenylalanine slows brain waves at a dosage equivalent to only 34 milligrams per kilogram per day of aspartame. A can of diet soda typically contains about 200 milligrams of aspartame. According to Reuben K. Matalon of the University of Illinois College of Medicine in Chicago, doses of from 50 to 100 milligrams per kilogram per day of aspartame may be "not that uncommon." He found that when the higher dose is continued over a period of weeks, the phenylalanine level in 14 percent of normal subjects and in 35 percent of those who have only one functional gene exceeds the level considered to be excessive for pregnant women.

Commenting on the newly presented findings, a NutraSweet Company scientist stresses that the data are "preliminary." More specifically, he maintains that Matalon's study lacked a control group. Matalon counters that he has plenty of control data from earlier studies. The NutraSweet spokesman also points out that the Elsas study revealing changes in brain waves failed to detect behavioral effects of the kind reported by Spiers.

A possible explanation for any behavioral effects is suggested by experimental-animal results reported at the meeting by Matthew J. During of the Massachusetts Institute of Technology. In rats, he said, high doses of phenylalanine inhibit synthesis of the neurotransmitter dopamine. Several laboratories have found that high doses of aspartame can increase the frequency of seizures or lower the stimulation necessary to induce them in laboratory animals, although there is no evidence that it actually causes seizures.

The Food and Drug Administration's original approval of aspartame in 1981 was criticized because some scientists in the agency had reservations about the test data submitted by the then manufacturer, G. D. Searle & Co., but a subsequent independent review approved the FDA's decision. Searle says that no food additive has had as intense an evaluation as the one undergone by aspartame.

Richard J. Wurtman of M.I.T., the conference organizer, points out that because aspartame is classed as a food additive rather than a drug, adverse reactions do not have to be reported to any Government agency and continuing safety monitoring is not required by law. New legislation is expected to be proposed soon.

Family Ties

The evolution of the human body from the simian posture and proportions of our early ancestors to its current appearance is thought to have been gradual. A new fossil, unearthed in Olduvai Gorge, suggests that the transition may actually have been abrupt. Working in the sandy fossil trove east of the Serengeti Plains in Africa, investigators found bone fragments that reveal the proportions of a hominid who lived 1.8 million years ago. Although within 200,000 years her descendants would be recognizably human, the "healthy old lady" reconstructed from these fragments still looks very much like a long-armed early hominid.

"Olduvai Hominid 62" (OH 62) is an assemblage of more than 300 pieces of bone recovered last summer by a team led by Donald C. Johanson, who directs the Institute of Human Origins in Berkeley, Calif., and Tim D. White of the University of California at Berkeley. Recounting the discovery in *Nature,* the investigators report that they found weathered bits of skull, arm and leg bones in soil deposits known to be about 1.8 million years old. The bones were scattered on the surface and had probably been exposed for centuries.

On reconstructing the skull the workers recognized the profile of a hominid species called *Homo habilis,* which appeared in East Africa about two million years ago. Many such skulls had been collected, but the arm and leg bones of OH 62 provided the first evidence bearing on the height and body proportions of *H. habilis.*

An unexpected ancestral portrait emerged. Because *H. habilis* came into existence more than a million years after the celebrated biped Lucy, anthropologists had assumed that *H. habilis* hominids were built more like *Homo sapiens* than was *Australopithecus afarensis,* Lucy's species. Instead the remains of OH 62 are remarkably similar to Lucy's. A female who probably died at the ripe old age of 30, OH 62 was as short as Lucy, standing about three and a half feet tall, and her arms, like Lucy's, were quite long in relation to her height.

Johanson and White point out that height can vary dramatically within a single hominid species, but fundamen-

PATENT #4,198,494 ANSWERS HUNDREDS OF HOUSE CALLS EVERY DAY.

When the alarm goes out, chances are it will be answered by DuPont Patent #4,198,494—a remarkable fire-resistant material called Nomex® III aramid. The effectiveness of NOMEX is so well recognized that protective clothing made of it is worn by more than 70% of all career fire fighters across the country. It also protects U.S. military pilots and almost all professional race-car drivers.

However, NOMEX is more than just an incredible fire fighter. Through advanced technology, NOMEX can also be transformed into an incredibly strong yet lightweight honeycomb material. This technology has led to innovative uses that range from the interior panels on Boeing 747s to skis for downhill racing.

But this is just one of more than 80,000 patents that DuPont has earned over the last fifty years in health sciences, advanced electronics, transportation, agriculture, textile fibers, consumer and specialty products.

And even more important than the number is the contribution these ideas make. For thousands of fire fighters, and for the millions of people they serve, Patent #4,198,494 answers a very important call.

BETTER THINGS FOR BETTER LIVING.

DU PONT

REG U S PAT & TM OFF

tal body proportions rarely do. Hence it appears that, aside from the characteristic expansion in cranial capacity, hominid evolution stood still for more than a million years. Moreover, the ancestors of human beings must have experienced a tremendous spurt in growth over the next 200,000 years. By 1.6 million years ago *Homo erectus* had supplanted *H. habilis* in East Africa, and fossils from *H. erectus* females show that they attained heights of five feet or more; their knuckles no longer hung near their knees.

OH 62's remains raise two questions: Why did hominids stay the same for so long, and why did they then change so quickly? Johanson says the persistence of Lucy-like traits was probably not incidental. Long arms were made for climbing trees, and in a hostile environment trees held distinct advantages: access to food, means of escape from predators and assurance of a reasonably safe place to sleep.

In fact, Johanson is more intrigued by the critical period between 1.6 and 1.8 million years ago during which hominids left the trees to walk on the ground. Although increasing intelligence and reliance on stone tools could explain the anatomical and behavioral shift, he is not ruling out the possibility that changes in the physical environment may have necessitated a terrestrial existence. "Something big was happening in Africa" during that 200,000-year stretch, Johanson says, and he now plans to focus his work on the period that witnessed so many changes in early man.

Double Trouble

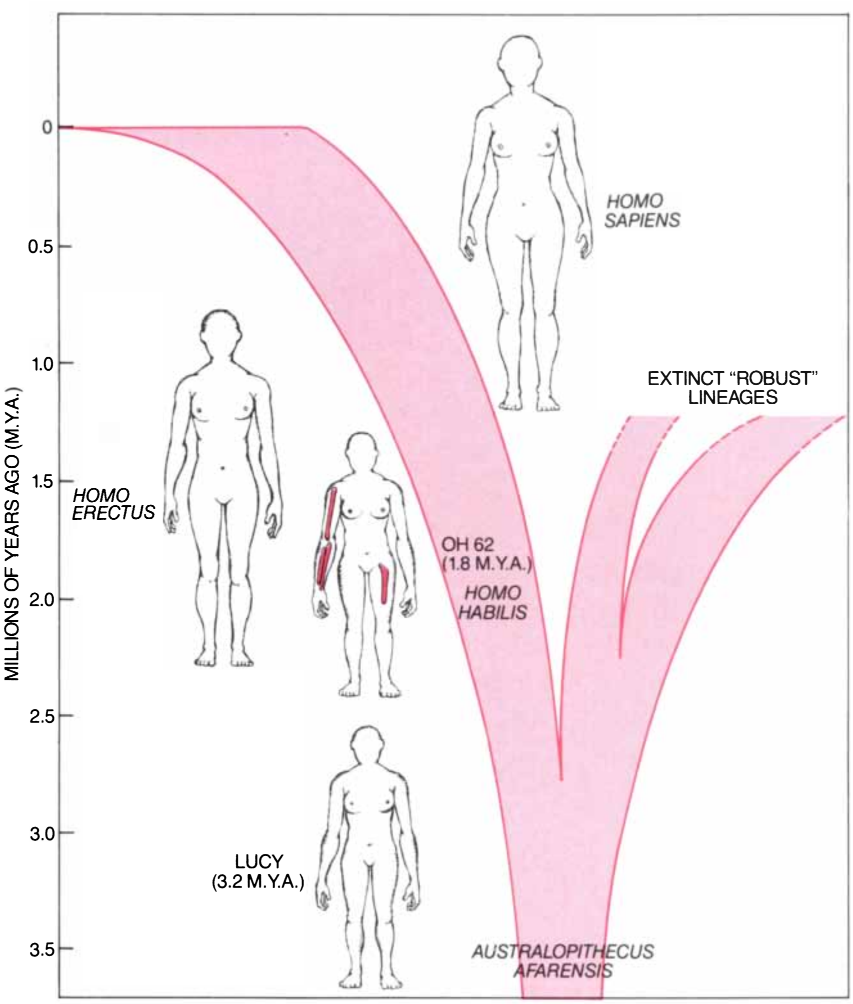
Just when Supernova 1987A seems to have exhausted its store of surprises, it serves up another. Observations made with sensitive light detectors suggest that the bright spark first spotted in the Southern Hemisphere on February 24 has a faint companion. Theorists say it is difficult to imagine how a stellar explosion could create two distinct sources of light. "I've invoked explanations, and so have a lot of other people," Jeremiah P. Ostriker of Princeton University says. "But I wouldn't call any of the explanations reasonable."

Two groups—one from the Harvard-Smithsonian Center for Astrophysics in Cambridge, Mass., and the other from the Imperial College of Science and Technology in London—have separately reported on the phenomenon. The evidence derives from observations they made with speckle interferometers, which log in digital form the arrival time and the position of individual photons striking a grid. These data, after extensive processing with a computer, enable investigators to pinpoint the origins of light sources with unparalleled accuracy.

The Harvard-Smithsonian workers, who first disclosed the finding, had installed their interferometer on the four-meter telescope at the Cerro Tololo Inter-American Observatory in Chile on two nights: March 25 and April 2. After analyzing the data for several weeks, the group announced it had distinguished a light source in the red end of the visible spectrum about one-twentieth of an arc-second from 1987A. Assuming that the Large Magellanic Cloud, the small galaxy in which Supernova 1987A exploded, is about 150,000 light-years from the earth, the gap corresponds to about 250 billion miles. To diverge this distance in one month (the approximate time that elapsed between the appearance of 1987A and the readings), two objects would have to fly apart from each other at approximately half the speed of light.

The companion, according to the Harvard-Smithsonian report, is fainter by about three magnitudes than 1987A, whose magnitude at the time of the sightings was just under 4 (the magnitude of the sun is 1). While considerably fainter than 1987A, the companion is much brighter than any object detected in the region before the supernova appeared.

Seeking confirmation of its results, the Harvard-Smithsonian team contacted the Imperial College workers, who were analyzing speckle data they gathered from the Anglo-Australian Telescope in Siding Spring, Australia. The British group, although it had not completed all its analyses, soon reported having found evidence of the companion source in observations made on April 14. The British workers, like the Americans, discerned the faint image only in the red end of the visible



RELATIVE PROPORTIONS of the hominid ancestors of humans are displayed on the family tree. The arm and leg bones of OH 62 suggest that body proportions changed little in the 1.4 million years separating Lucy (*A. afarensis*) and OH 62 (*H. habilis*), and then changed significantly between 1.8 and 1.6 million years ago with the advent of *H. erectus*.

spectrum; neither group discerned the companion in readings taken at blue wavelengths.

Astronomers have speculated that the dimmer image may be caused in any one of a number of ways: by a "light echo" bouncing off a cloud of gas emitted by the progenitor star before it exploded; by a faint companion to the progenitor that flared up when exposed to the supernova's intense radiation; by an asymmetry in the explosion that split the star into two separate pieces; by a pulsar, the remnant of the supernova, emitting a beam through a "window" in the expanding envelope, or by a massive object whose gravitational field acts as a lens, splitting 1987A's image. Investigators say only further observation will determine whether any of these theories is plausible.

Physical Sciences

Fractured Aphrodite

Geologically speaking, the earth's most distinctive characteristic is the moving tectonic plates that make up the lithosphere, its rigid outer shell. At a meeting of the American Geophysical Union James W. Head and Larry S. Crumpler of Brown University have reported what they consider strong evidence that our nearest neighbor, Venus, has plate tectonics too.

On radar images and topographic maps of Aphrodite Terra, a broad, elongated highland near the Venusian equator, the workers have spotted features that resemble the earth's midocean ridges, where plates spread apart and molten rock wells up to form new lithosphere. Aphrodite, they conclude, must be a boundary between two diverging plates.

The conclusion is not altogether surprising. In size and density Venus resembles the earth, and its interior, like the earth's, is thought to be heated by the decay of radioactive elements. Presumably the interior is ductile and is therefore susceptible to convection. Volcanoes have provided and may still provide an outlet at the surface for the convective upwelling of hot rock.

Nevertheless, many workers have believed convection in the Venusian interior is not organized in the large, strong currents needed to drive plate motion. Plate tectonics on the earth involves not only the upwelling of new lithosphere at spreading centers but also the subduction of old lithosphere at oceanic trenches. The topographic data gathered by the *Pioneer Venus* orbiter had shown no signs of a planet-girdling system of ridges and trenches, and some theorists had argued that the Venusian lithosphere would be too hot and buoyant to sink into the interior at a trench.

According to Crumpler, however, no one had looked closely enough at the data for Aphrodite Terra. He and Head examined both the *Pioneer Venus* data and high-resolution topographic profiles across Aphrodite made with the 300-meter radio telescope at Arecibo, Puerto Rico. In doing so they found some hitherto unobserved features: parallel, linear discontinuities—troughs or steps—that cut across Aphrodite and strongly resemble the fracture zones in the earth's midocean ridges.

Profiles parallel to the discontinuities show that the topography on opposite sides of Aphrodite is strikingly symmetrical: relatively small hills on one side are repeated on the other. The axis of symmetry lies along the crest of a linear rise that is perpendicular to the discontinuities and laterally offset by them, just as the axes of midocean ridges are offset at fracture zones. All of this indicates to Head and Crumpler that new lithosphere has been created at the crest of Aphrodite and has spread to both sides, in directions parallel to the discontinuities.

A single isolated spreading center is hardly conceivable; indeed, the two workers say they have recently found evidence that the spreading center extends beyond Aphrodite, reaching at least two-thirds of the way and possibly all the way around the planet in its equatorial region. If new lithosphere is created near the Venusian equator, then old lithosphere must be destroyed by subduction near the poles, but so far no evidence of a subduction zone has been found.

Not everyone is convinced by the findings of Head and Crumpler. Raymond E. Arvidson of Washington University, for instance, thinks "the data are equivocal." He and the workers at Brown agree, however, that the spacecraft *Magellan* will test the hypothesis when it begins mapping Venus in 1990. The high-resolution maps will enable investigators to estimate the age of the Venusian crust from the number of large meteorites that have bombarded it. If Head and Crumpler are right, the age of the surface, and hence the number of craters, should increase progressively from the equator to the poles.

Occult Occulters

Variations in the radiation from quasars about 1,000 megaparsecs away (a distance more than 30,000 times the diameter of the Milky Way) have given evidence of a new kind of interstellar object within our galaxy: formations of ionized matter larger than most stars but much smaller than most other interstellar structures (such as the enormous cloudlike structures found throughout the galaxy).

Ralph L. Fiedler, Brian K. Dennison and Kenneth J. Johnston of the Naval Research Laboratory and Antony Hewish of the Cavendish Laboratory in Cambridge, England, examined seven years' worth of daily measurements of the intensity of the radiation at 2.7 and 8.1 gigahertz coming from some 36 distant radiowave sources, most of which were quasars. They report in *Nature* that three of the sources showed episodes of an unusual kind of variation: the radiation at 2.7 gigahertz rose briefly, dropped sharply to a level much lower than normal (where it stayed for a few months) and then rose steeply above normal for a short time (about a week) before dropping back again to its normal level, where it remained.

During the most significant of these episodes the radiation at 8.1 gigahertz also behaved strangely. Instead of remaining relatively constant it jumped up and down between normal levels and very high levels, returning to its usual constancy at the same time that the 2.7-gigahertz radiation became bright again.

It is unlikely that the variations were due to any change in the quasars themselves. Instead, Fiedler and his fellow workers suggest that during each episode some object passed between the source and the earth, briefly interfering with the path of the radiation. The object could be a formation of ionized material, which would have refracted the quasar's radiation. The refraction would have caused the source to appear dim in the 2.7-gigahertz band; radiation scattered by refraction would have been observed on the earth before and after the dimming was observed, accounting for the brief intervals of high intensity recorded immediately before and after the dimming.

The sharp spikes at 8.1 gigahertz could be the result of refractive focusing: irregularities in the structure of the object could focus the quasar's light sharply on the earth. Such focusing should be rare—it can occur only if the earth is in the "focal plane" of the refractive object—which would explain why sharp 8.1-gigahertz peaks were seen during only one of the three episodes.

The investigators were able to calculate the angular velocity of the object putatively responsible for the most significant episode by noting how long it took for the intensity of the 2.7-giga-

hertz radiation to drop. They then assumed that the object's velocity is less than about 250 kilometers per second (the speed of the fastest-moving clouds in the galaxy); on that assumption they calculated that the object should be less than about 1,300 parsecs away. The duration of the dimming at 2.7 gigahertz suggests the object is less than seven astronomical units across (1 A.U. is the distance from the earth to the sun).

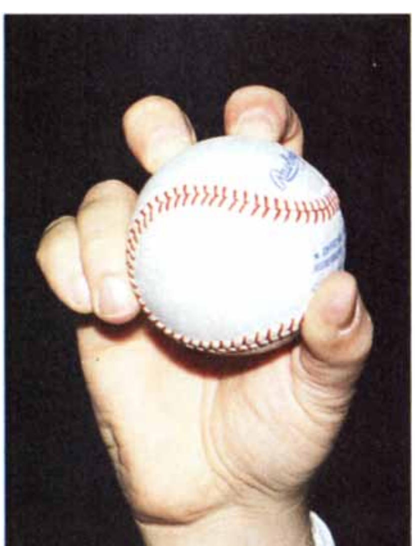
Because the workers found three events in observations of so few sources made over a relatively short time, they reason that the occulting objects could be quite numerous. If the objects fill the galactic halo uniformly, there might be as many as 120 of them per cubic parsec—which would make them 1,000 times as plentiful as stars.

Knuckleballs

Experienced baseball fans know when a knuckleball pitcher is at work: he does not fling the ball so much as he lobs it, and as the ball floats across home plate the batter and even the catcher often lunge wildly in their attempts to hit or catch it.

What makes the knuckleball so elusive? Six years of research by Joel W. Hollenberg, professor of mechanical engineering at the Cooper Union for the Advancement of Science and Art in New York, and a succession of his students have turned up some answers. The erratic behavior of the ball, they report, results from its very slow spin and from the aerodynamic properties of the baseball's raised stitches.

Hollenberg's project began in 1982 when two of his students interviewed Phil Niekro, a master knuckleballer now pitching for the Cleveland Indians. The students learned that most pitches—such as fastballs, curveballs, sliders and screwballs—spin because pitchers wrap their fingers around the ball and, in releasing it, whip them across its surface. To throw a knuckleball a pitcher grips the ball with his fingertips (the protrusion of the knuckles gives the pitch its name) and releases it with a shotputlike motion that minimizes spin.


KNUCKLEBALL is gripped with the fingertips and thrown with a pushing motion.

The investigators tested baseballs in a wind tunnel at Cooper Union. They found that the drag exerted on a baseball by the air varies according to the ball's velocity and the orientation of its stitches. By gluing strips of silk to the ball the workers learned that drag tends to be higher when the flow of air around the ball is laminar, or smooth, and lower when the flow of air is turbulent, or rough.

The airflow around a slowly spinning knuckleball may shift from a largely laminar state to a turbulent state or vice versa one or more times during its 60-foot, six-inch flight from the pitcher's mound to home plate. Hollenberg thinks the knuckleball is most likely to change direction during a "transition" between the two states, when the flow is laminar on one side of the ball and turbulent on the other. If the ball changes direction in the second two-thirds of its flight, the batter is hard pressed to adjust.

With data from their wind-tunnel tests and from experiments done at other institutions, Hollenberg and his students have devised a computer program that predicts how a knuckleball will act given its position, direction, spin and velocity at the moment of release. The program also calculates the effect on the pitch of wind, temperature, barometric pressure and humidity. The model predicts, for example, that knuckleballs should become less lively as temperatures rise.

The next step, Hollenberg says, is to test the computer program's predictions. All he needs to do that is a machine that can throw knuckleballs, and equipment for filming and then analyzing the pitches. "We're hoping the good knuckleball fairy will descend and help us," Hollenberg says. Who might the good fairy be? Hollenberg replies: "George Steinbrenner."

Shock of Impact

Did a gigantic meteorite strike the earth 66 million years ago at the end of the Cretaceous period, when the dinosaurs and many other creatures became extinct? The possibility was first raised in 1979, when Luis W. Alvarez, Walter Alvarez and their colleagues at the University of California at Berkeley reported that a thin layer of clay dividing Cretaceous strata from those of the succeeding Tertiary period contains anomalous amounts of iridium, an element that is rare in the earth's crust but abundant in meteorites. The Alvarez group proposed that the iridium was deposited by the impact of an asteroid some 10 kilometers in diameter, which raised a globe-girdling dust cloud that blotted out the sun and caused the mass extinction.

Eight years later the hypothesis is still being debated. Since iridium is also abundant in the earth's mantle, an excess of the element can in principle be explained without invoking extraterrestrial mechanisms. Several workers, among them Charles B. Officer and Charles L. Drake of Dartmouth College, have proposed that intense volcanic activity took place at the end of the Cretaceous, causing major extinctions and erupting large amounts of iridium. Writing in *Science,* Bruce F. Bohor, Peter J. Modreski and Eugene E. Foord of the U.S. Geological Survey now report the global distribution of a tracer less equivocal than iridium: quartz grains that were apparently scarred by the shock of an impact.

The Geological Survey workers had previously identified such shock-metamorphosed quartz, in which the crystal structure is visibly disrupted in layers paralleling the crystallographic planes of the mineral, at a single site in Montana. The quartz grains were abundant precisely at the iridium anomaly, and they showed multiple, intersecting sets of disrupted layers—a feature previously seen only in grains from meteorite impact sites. Disrupted layers also occur in quartz from regions of crustal deformation and explosive volcanism, but such grains bear only one set of parallel layers.

Possibly the shocked quartz found in Montana could have been eroded from some preexisting impact crater and deposited along with iridium from a volcanic source. Now, however, the Geological Survey workers have identified grains of shocked quartz in the iridium layer at seven additional sites around the world: five in Europe, one in New Zealand and one on the floor of the Pacific Ocean. The global coincidence of shocked quartz and iridium points to a common causative event.

The size of the quartz grains as well as their structure argues that the event was a massive impact. The grains Bohor and his colleagues studied are too large (in some cases more than .1 millimeter in length) to be carried very far by atmospheric circulation. Even if explosive volcanism could have produced the shock features, Bohor points out, the material would have settled out of the volcanic cloud relatively near the eruption. To account for the global distribution one would have to conceive of a large number of volcanoes erupting at once around the world. A single impact, in contrast, could have released enough energy to loft debris out of the atmosphere on ballistic paths that might well have carried it around the globe.

The findings not only bolster the impact hypothesis but also add detail to the scenario. Quartz is abundant only in continental crust, and so its presence in the layer of fallout might indicate the asteroid struck one of the continents. The largest grains of shocked quartz were found in the western U.S.; reasoning that the largest grains would have fallen to earth close to the impact site, Bohor suggests that the asteroid struck somewhere in North America. Why has the crater not been found? Some 66 million years of mountain building, glaciation, erosion and sedimentation may have obliterated the scar, Bohor thinks.

Officer believes such confidence is unwarranted. He refers to an unpublished finding by Neville L. Carter of Texas A&M University that minerals bearing the kind of shock features associated with known volcanic eruptions occur throughout a 400,000-year sequence of strata spanning the Cretaceous-Tertiary boundary, and to a study by James H. Crocket of McMaster University in Ontario suggesting that high levels of iridium were also deposited over a similar extended period. The geologic record, Officer maintains, reflects prolonged, intense volcanism and not a sudden impact.

BIOLOGICAL SCIENCES

Growth Switch

Understanding the genetic control of cell division has been something of a holy grail for molecular biologists because such control is fundamental not only to normal growth and replacement of tissue but also to the various pathological growth processes known collectively as cancer. Yet the human genes controlling the successive events of the cell cycle have been remarkably elusive. Now a human gene that appears to play an important role in the control of cell division has been isolated at the Imperial Cancer Research Fund in London.

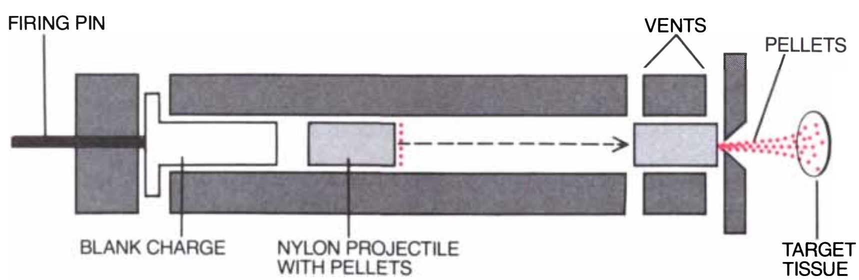
Melanie G. Lee and Paul Nurse identified the cell-division gene not on the basis of its function in human cells but by finding that it switches on cell division in yeast. They knew that yeast cells having a mutation in a particular gene (called *cdc2* in the yeast *Schizosaccharomyces pombe*) stop dividing at two characteristic points in the course of cell division; the protein encoded by the gene must perform some vital function at those points. The London investigators decided to look for the human gene encoding an analogous protein by means of a complementation test: by looking for a human gene that would do for the mutated yeast cell what the yeast cell could not do for itself. They gambled that as critical a process as cell division would be similarly controlled in animal and vegetable cells.

Lee and Nurse report in *Nature* that fortunately *S. pombe* is able to express the genes in a widely used human c-DNA "library," that is, the yeast's protein-synthesizing machinery can read the human DNA and translate its message into protein. The workers persuaded *cdc2*-defective *S. pombe* cells to take up and express one or another of the many thousands of human genes in the library and picked out the yeast cells that were able to divide and multiply. From those cells they isolated the human DNA that had been taken up: the human *cdc2* equivalent.

The DNA sequence of yeast *cdc2* and its human equivalent proved to be very similar. The fact that the sequences have changed little over the billion years since the distant ancestors of yeast and human beings went their separate evolutionary ways indicates that the genes are indeed of fundamental importance, Robert A. Weinberg of the Massachusetts Institute of Technology's Whitehead Institute for Biomedical Research points out. Both genes encode moderately heavy protein kinases, which are enzymes that attach phosphate groups to other proteins. Such enzymes were already suspected of playing an important role in cell division, and a defect in their activity might underlie some cancers.

Shotgun Marriage

The delicate task of introducing new genetic material into living plant cells has been accomplished by a seemingly blunt instrument: a .22-caliber shotgun. Microscopic tungsten pellets coated with RNA or DNA, fired by the gun at about 1,000 miles per hour, penetrate the cell wall without killing the cell, and the genetic material is expressed, that is, translated into protein. With further develop-



MICROSCOPIC PELLETS coated with genetic material are fired into plant cells by a miniature shotgun (*top*). The photomicrograph shows a single pellet inside an onion cell.

ment the method promises to insert genes into plant cells faster and more efficiently than other methods devised to date, facilitating research into the expression of foreign genes in plant cells and eventually providing a new technique for genetic engineering to improve crop plants.

The shotgun method is described in *Nature* by Theodore M. Klein, Edward D. Wolf, Ray Wu and John C. Sanford of Cornell University and its New York State Agricultural Research Station at Geneva, N.Y. The coated pellets, which average four micrometers (thousandths of a millimeter) in diameter, are suspended in a drop of water that is applied to the front end of a cylindrical nylon plug. The explosion of a gunpowder-filled blank cartridge drives the plug down the gun barrel. The plug is stopped dead by a steel plate at the end of the barrel, but a stream of thousands of pellets passes through a one-millimeter hole in the plate to impact on the target, which in the first trials was a one-centimeter-square patch of tissue from the surface of an onion bulb. The pellets punched neat holes in the walls of some epidermal cells and lodged in the cells.

The Cornell team tested the gun's effectiveness by firing RNA of the tobacco-mosaic virus (TMV) into the cells. The RNA was translated by the cells to form live virus particles: crystalline viral inclusion bodies could be seen in some 35 percent of the cells containing pellets, and extracts of the cells, applied to tobacco leaves, gave rise to typical TMV lesions.

The investigators went on to attempt actual gene transfer by shooting plasmids incorporating the gene for the enzyme chloramphenicol acetyltransferase (CAT) into onion cells. They were able to show that the foreign gene was expressed in some of the onion cells—that the foreign enzyme was synthesized by the cells—by demonstrating that a cell extract had significant CAT activity in an enzyme assay: it converted most of the chloramphenicol present into its acetylated derivatives.

The epidermal cells targeted in the early tests are relatively large and cannot be regenerated to form complete plants. Klein reports that the Cornell team has recently achieved similar results with much smaller, regenerable cells, opening the way to genetic-engineering experiments.

Climate Control

Marine algae would not at first sight appear to have much influence over the formation of clouds thousands of feet above the ocean. Yet a group of atmospheric scientists has recently put forward the surprising notion that there is indeed a strong relation: algae, they suggest, may be part of a thermostat that regulates global temperature by controlling the brightness of clouds and hence the amount of sunlight bounced back into space. The authors think their theory may explain why changes in the earth's climate over the past 4.5 billion years have been comparatively modest, even though the sun is widely believed to have increased in brightness by about 30 percent over that period.

The theory is proposed in *Nature* by Robert J. Charlson of the University of Washington and James E. Lovelock of the Coombe Mill Experimental Station in Launceston, England, and their co-workers. Marine algae are known to excrete a chemical called dimethyl sulfide (DMS), which evaporates and is then oxidized in the atmosphere. Among the reaction products are submicroscopic particles (about .01 micrometer in diameter) of various sulfate salts. The investigators believe that over the remote ocean this "sulfate aerosol" provides most of the seed particles necessary to trigger the condensation of cloud droplets.

There are indications that DMS is produced fastest by the kind of algae that inhabit warm, salty and brightly illuminated water. If this is in fact the case, the algae may complete a negative-feedback loop. As the ocean becomes warmer they would produce a growing number of seed particles, thereby encouraging the formation of clouds consisting of more numerous but smaller droplets. Such clouds reflect back into space a relatively large percentage of the sun's radiation and so would counteract the initial warming. Theory suggests that ocean-surface temperature may be surprisingly sensitive to the number of particles produced by algae, because (in contrast to the situation over land) suitable particles are in short supply. Charlson believes "we have identified an important missing aspect of the climate equation."

Do more seed particles mean more reflective clouds in real life? Research by Sean A. Twomey of the University of Arizona and independently by James A. Coakley, Jr., of the National Center for Atmospheric Research in Boulder, Colo., tends to support the idea: particle-rich smoke from ocean-going ships causes a measurable increase, visible as bright streaks on satellite photographs, in the reflectivity of overlying clouds.

Why should plankton bother to make DMS? Charlson and Lovelock invoke Lovelock's "Gaia" hypothesis, according to which life evolved so as to create conditions favorable for its existence: they suggest that DMS synthesis "grew to become an unconscious benefit to the system." This kind of explanation is rejected by most biologists, who argue that natural selection penalizes "altruistic" traits that benefit other species to the detriment of genes controlling the trait. But there are other and more plausible explanations for DMS synthesis, and the idea of a feedback loop involving plankton is being taken seriously.

MEDICINE

Female Trouble

It is widely known that women live significantly longer than men (78.2 versus 71.2 years on the average). Nevertheless, being female has its risks. According to recent reports, women who consume even moderate amounts of alcohol may double their risk of breast cancer; women who undergo surgical removal of the uterus and both ovaries, without receiving estrogen therapy, may double their risk of coronary heart disease.

Data linking alcohol consumption to breast cancer are presented in two independent reports in the *New England Journal of Medicine*. In one study a team led by Arthur Schatzkin of the National Cancer Institute analyzed data collected for the Epidemiologic Follow-up Study of the first National Health and Nutrition Examination Survey. A total of 7,188 women from 25 through 74 years of age were interviewed and followed over a 10-year period, by the end of which 121 members of the group had developed breast cancer. The data suggest that women who take an average of three drinks per week show a 40 to 50 percent increase in the risk of developing breast cancer and that women who have one drink or more a day increase their risk by 100 percent.

In a second study, carried out by Walter C. Willett of the Harvard Medical School and his colleagues, 89,538 female nurses were asked to quantify their drinking habits on written questionnaires. Then they were followed over a four-year period, during which 601 of them were diagnosed as having breast cancer. After adjusting for factors known to influence the onset of the disease, the Harvard team found that women who took one drink or more per day had a 50 percent greater risk of contracting the disease than women who drank only occasionally or not at all.

To be sure, women who drink alcoholic beverages may be at higher risk

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.

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.

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.

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 #BL-1450. Or, for more information, see your authorized Lotus Dealer, or write Lotus Development Corp., 90 Annex, Atlanta, GA 30390-03070.

Lotus Manuscript™

for breast cancer for reasons other than the intake of alcohol, and both groups concede that more research on the relation between breast cancer and alcohol is needed. The fact remains that an association between drinking and the disease has been shown in almost every study conducted to date, and adjustment for known risk factors has not materially altered the association.

Saxon Graham of the State University of New York at Buffalo suggests in an editorial in the *New England Journal* that "women at especially high risk for breast cancer, such as those who are obese, who have had few children, who were first pregnant when they were older than 25, or whose mothers had breast cancer, should curtail their alcohol ingestion."

Statistics linking menopause to coronary heart disease are also intriguing. For many years it has been thought that women who experience either natural menopause or artificial menopause induced by hysterectomy (surgical removal of the uterus) are at higher risk for coronary disease than premenopausal women. Because hysterectomy is the most frequently performed major operation in the U.S., this has been cause for alarm.

New findings published in the *New England Journal* should put women at ease. According to a team led by Graham A. Colditz of the Harvard Medical School, the increased risk stems from a sudden drop in the level of the female sex hormone estrogen, caused by the removal of both ovaries, and not from menopause itself.

Results from questionnaires sent to 121,700 women between the ages of 30 and 55 suggest that for women who undergo natural menopause (and therefore experience the gradual decline in estrogen levels associated with reduced ovarian function) there is no appreciable increase in the risk of coronary disease. Women who undergo simple hysterectomy or who have the uterus and only one ovary removed also incur no increased risk.

On the other hand, women who undergo both hysterectomy and bilateral oophorectomy (removal of both ovaries) more than double their risk of coronary disease if they do not receive replacement estrogen. If they receive estrogen, they are at no greater risk of coronary disease than premenopausal women of the same age.

The news for women who are concerned about coronary heart disease, then, is largely good. The report clearly states that the more gradual decline in estrogen characteristic of natural menopause is "not associated with any meaningful increase in risk" and that women who have both ovaries removed can avoid any increase in risk by means of estrogen therapy.

Modeling Calamity

As the toll of AIDS mounts, hospital systems in the worst-affected areas are showing signs of strain. Proposals have recently been made to devote some hospitals in New York City and San Francisco to the exclusive care of AIDS patients, and the cost of the epidemic in New York City could reach $1 billion per year by 1991. San Francisco General Hospital is struggling under the burden of a probable 18,000 AIDS outpatient visits this year.

How far and fast will the epidemic spread? No one knows. Even conservative estimates put the cumulative total of U.S. cases at more than 300,000 by the end of 1991. The total cost of AIDS in the U.S. in 1991—including lost productivity—will exceed $66 billion, according to a study that has been endorsed by Congress's Office of Technology Assessment.

Uncertainty tends to breed apprehension—an effect the nature of AIDS readily amplifies. Last February, Stirling Colgate, a senior fellow at the Los Alamos National Laboratory, wrote to Solomon J. Buchsbaum, executive vice-president of the AT&T Bell Laboratories and chairman of the White House Science Council, presenting a mathematical model of the spread of the disease. Colgate's model projected a 10 percent level of infection in the U.S. within about nine years, even assuming that transmission will fall twofold (because of changing sexual practices) every four years. He called for the declaration of a national AIDS emergency and a massive mobilization of resources, together with "saturation" public education and compulsory and universal periodic testing. Colgate's views were soon attracting the attention of the leadership of the National Academy of Sciences and of the Institute of Medicine as well.

Buchsbaum replied that a heterosexual population containing a large fraction of monogamous couples would not be alarmed enough to support such measures. Colgate has now written again to each member of the Science Council directly. President Reagan's decision to call for wider AIDS testing and to appoint a presidential commission presumably reflects growing concern within the Administration.

In the meantime Colgate, together with his co-workers James M. Hyman, E. Ann Stanley and Thomas G. Marr, are building a detailed supercomputer model that will simulate transmission, chiefly by sex and intravenous drug abuse, taking into account the frequency of such exposures. The model will eventually incorporate such details as the rate of pair formation and breakup, the frequency of sexual contacts and the prevalence of homosexuality and bisexuality.

The model represents the change in infectivity of an AIDS carrier over time: after a brief peak at the time of first infection, infectivity is thought to fall to a low level before rising again as overt disease develops, often from five to seven years later. Most AIDS cases so far, the Los Alamos group maintains, have been transmitted during the initial infectious peak; future transmission at later stages might lead to large increases in infection.

The group maintains that its preliminary investigations suggest "the safety of the heterosexual population is largely an illusion" based on the different time scales of spread in different groups. When the infection level reaches 10 to 15 percent, the Los Alamos researchers say matter-of-factly, "society starts to break down."

Yet there is still little evidence of a dramatic spread of infection beyond contacts of high-risk groups in the U.S. as a whole. Critics also point out that Colgate's projections took no account of social barriers between separate subpopulations. Colgate concedes that his initial time scales may need to be doubled, but the Los Alamos group thinks social barriers will provide only a temporary respite.

Meade W. Morgan, head of statistics for the AIDS program of the Centers for Disease Control in Atlanta, cautions that although models may be useful for evaluating strategies, any predictions "should be taken with a grain of salt" because they are sensitive to specific assumptions. The CDC epidemiologists dispute the Colgate group's expectation that all infected people will eventually die from AIDS; the CDC's official estimate of the fraction that will contract AIDS within five years is still from 20 to 30 percent, according to Morgan. But he expects the figure to rise with time and says, "I would not want to say [the Los Alamos group] is wrong."

TECHNOLOGY

Fourier Transformation

An algorithm has been devised that may be able to supplant the Fourier transform, one of the most widely used mathematical tools of modern science and engineering. Fourier transforms, which convert raw data into pliable mathematical equations, help in-

vestigators to analyze a host of phenomena—from chemical reactions to the spectra of stars. Ronald N. Bracewell of Stanford University, who devised the algorithm, says it can do the same work as Fourier transforms in about half the time.

The transforms are named for the French mathematician Jean-Baptiste-Joseph Fourier. In 1815 he published a paper that showed how any curve, no matter how irregular or aperiodic, can be represented by the sum of a set of sine functions, which when plotted create uniform, periodic waves. Peaks in the curve can be attained by stacking waves whose crests coincide; flat areas can be represented by combining waves whose crests and troughs cancel one another.

Fourier transforms remained largely a tool for specialized scientific applications until the 1960's. Then James W. Cooley of the Institute for Advanced Study in Princeton and John W. Tukey of the Bell Telephone Laboratories showed how the calculations that constitute a Fourier transform can be reduced by applying a technique that divides data into readily manageable segments. Computers made this technique—which came to be known as the fast Fourier transform—even easier to perform, and its use spread.

Bracewell's algorithm builds on the work of another Bell Laboratories scientist, Ralph V. L. Hartley, who in 1942 proposed an alternative to the Fourier transform. Whereas Fourier transforms involve complex numbers, which combine real numbers with imaginary ones (such as $\sqrt{-1}$), Hartley's equations involve only real numbers. Computing with complex numbers takes much more time than computing with real numbers, Bracewell explains.

In 1982 Bracewell wrote a paper on how Hartley's ideas might lead to an improvement on the Fourier transform. Urged by a Stanford colleague to go further, Bracewell created an algorithm that carries out a "fast" version of the Hartley technique. Bracewell uses the algorithm for his own research into the periodicity of sunspots. When run on his personal computer, he says, the algorithm can process 1,024 data points—records of the fluctuations of sunspots over time—in six minutes; a fast Fourier transform used to take 10 minutes to do the same job.

Bracewell calls the algorithm, which he has recently patented, the Hartley transform. The U.S. Patent Office, however, refers to it as the Bracewell algorithm, and others call it the Hartley-Bracewell algorithm. The history of mathematics, Bracewell says, has made him aware of the tenuous connection between names and original invention. He notes that the Fourier transform was probably discovered by Leonhard Euler, a Swiss mathematician, in the 1770's. And, astonishingly, versions of the fast Fourier transform have been found in notes written by the German mathematician Carl Friedrich Gauss in 1805.

Keeping It Together

Diffraction, the bending of waves around edges, is a ubiquitous phenomenon. In the case of light waves it accounts for the blurriness of a shadow's outline and limits the resolving power of optical instruments. It also causes a normal laser beam, however tightly it is collimated, to diverge. Although the effect can be minimized (primarily by resorting to smaller wavelengths of light and beams of larger diameter), it was thought such spreading could not be entirely eliminated. Three investigators at the University of Rochester have now shown, both in theory and in practice, that diffractive spreading of a light beam is in fact easily avoided.

The workers began with the fact that ordinarily a laser beam's intensity is not radially uniform but has what is known as a Gaussian profile: the intensity is highest at the center of the beam and falls off continuously toward the fringes. Owing to diffraction, a beam having such a profile quickly broadens as it travels through space, so that the central intensity peak becomes progressively less pronounced. Yet James Durnin, Joseph J. Miceli, Jr., and Joseph H. Eberly point out in *Physical Review Letters* that the equations governing diffraction in fact allow light beams with profiles conforming to an entirely different class of functions to maintain an extremely narrow central intensity peak that is not subject to diffractive spreading.

The simplest of these functions is known as the zeroth-order Bessel function of the first kind and is designated J_0. Bessel functions are one of the commoner functions in mathematical physics; they are often applied to describe the behavior of systems having axial symmetry, such as organ pipes and drumheads. It is well known that these functions can describe the pattern characteristic of electromagnetic radiation traveling through a cylindrical waveguide. Still, investigators had not considered shaping a light beam's intensity profile in the form of the J_0 function in order to attain diffraction-free propagation in space.

To generate such a beam Durnin and his colleagues shone laser light onto a thin annular slit placed at the focal plane of a lens. The light passing through the slit to impinge on the lens can be viewed essentially as coming from an infinite number of point sources arranged in a circle and radiating in phase. Such a radiation pattern is collimated by the lens into a beam that has the characteristic J_0 intensity profile: a needlelike central peak surrounded by a series of much smaller concentric ridges.

Because so far only a small fraction of the energy of a light source can be converted into J_0 form, such beam-forming optics are not likely to find a role in laser weapons or other laser devices that rely on powerful pulses of light. "If your goal is to transfer sheer power from one place to another, you would probably do as well with a Gaussian beam," Durnin says. Moreover, although the J_0 beam does not undergo diffractive spreading, in practice it does have an effective maximum range, determined by the diameter of the lens. (In most situations, however, the maximum propagation range for a J_0 beam will be much larger than the range of a conventional laser beam that has a comparable intensity-peak radius.) In view of these limitations, the investigators believe the applications of J_0 beams are more likely to be found in high-precision operations such as tracking, pinpoint welding and the alignment of optical components.

Superscope

A system of microscopy has been developed that takes "hard," or highly energetic, X rays from a synchrotron beam and produces three-dimensional images with a resolution approaching one micrometer—1,000 times finer than the medical CAT (for computer-assisted tomography) scanner. The new device was developed at the Exxon Research and Engineering Company by Brian P. Flannery, Harry W. Deckman, Wayne G. Roberge and Kevin A. D'Amico.

The system makes use of a beam at the Brookhaven National Laboratory's synchrotron light source, together with a high-resolution X-ray detector and advanced computational procedures for reconstructing images, to extend tomography to the microscopic domain. For this reason the developers call it three-dimensional X-ray microtomography. Because the system exploits the capability of a synchrotron source to provide virtually monochromatic X rays at precisely adjustable energies just above and below characteristic atomic absorption features, the developers have obtained images that highlight the distribution of particular elements in a sample. Hence they believe microtomography will become a unique probe for studying the microstructure of a wide range of materials and biological tissues.

Tomography in medicine employs X rays to scan a planar section of a patient. The measurements are mathematically analyzed to create a cross-sectional map: a reconstructed image of the bone and tissue structure. Tomography overcomes the ambiguity of radiographic images, such as standard medical X rays, in which objects along a line of sight are projected as a common shadow image.

Microtomography depends on the same physical and mathematical principles. Obtaining accurate images at microscopic resolution, however, requires an enormous increase in the intensity of the X-ray beam. To decrease the size of the pixels (basic picture elements) by a factor of 10^3, which is necessary to get from the one-millimeter resolution of conventional tomography to the one-micrometer resolution of X-ray microtomography, the X-ray brightness must be increased by 10^{12} for a three-dimensional image. Lee Grodzins of the Massachusetts Institute of Technology pointed out some years ago that a synchrotron X-ray source would provide an adequate flux for creating high-resolution images. This potential could be realized, however, only by developing new techniques for X-ray imaging and by advances in data processing to handle the vast amounts of computation necessary to reconstruct three-dimensional images. Those are the goals now achieved by the Exxon group.

X-RAY MICROTOMOGRAPHY system made this radiograph of a tiny insect of the thrips order. False colors were chosen to emphasize the insect's internal structure.

The developers believe X-ray microtomography will become an important probe and diagnostic tool for investigations in materials science, biology and medicine. In particular, they see it as a means of studying contained systems under conditions of temperature, pressure and environment resembling those found in such diverse systems as reactors and living organisms.

OVERVIEW

Growing Up

Del and Rey, Johnny and Jimmy, and nature versus nurture.

Most parents probably assume that their baby will sit up, grasp objects, walk and develop other skills sooner if actively encouraged than if left alone. Yet many child psychologists maintain that infants develop more in response to an unalterable sequence of genetic cues than to parental coaxing or other forms of environmental stimulation. In journals and textbooks, proponents of the theory frequently cite two "classic" experiments carried out on twins in the early 1930's. The experiments are offered as evidence that depriving an infant of human interaction will not permanently retard his or her physical development and that, conversely, training will not advance it.

This view of the experiments has been challenged by Micha Razel, a child psychologist associated with the Weizmann Institute of Science in Rehovot, Israel. Razel contends that the experiments actually showed that environmental stimulation is crucial to healthy development: restrictions on such stimulation that were imposed by the experiments, he says, left two subjects physically disadvantaged, perhaps permanently.

The experiments, both of which began in 1932, were performed separately by two young psychologists who later became prominent members of their profession: Wayne Dennis, then an assistant professor at the University of Virginia, and Myrtle B. McGraw, then a researcher at the Columbia-Presbyterian Medical Center in New York. With the help of his wife, Marsena, Dennis studied female fraternal twins named Del and Rey, the daughters of an indigent Baltimore woman. For 13 months, beginning when the twins were 36 days old, they were kept in a room in the Dennises' home.

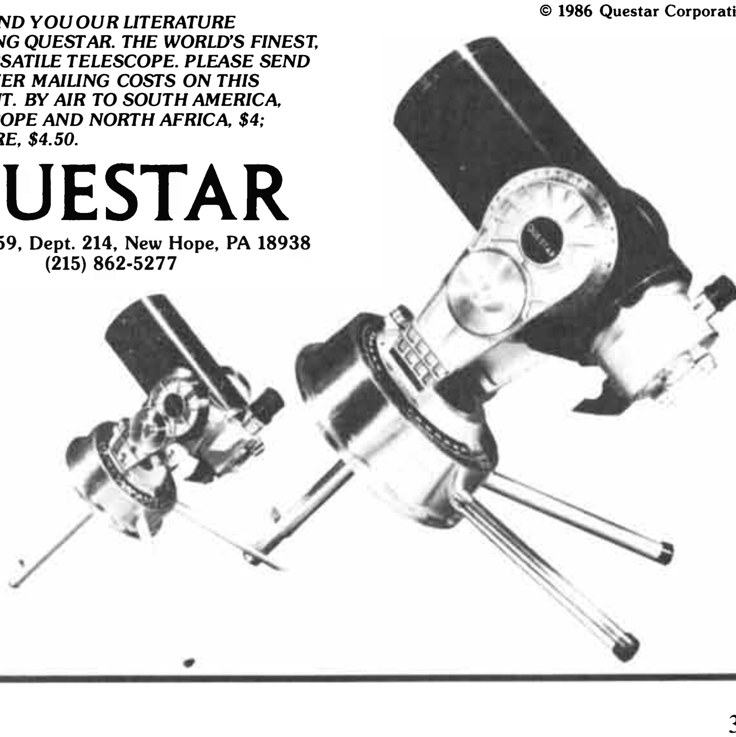
In *The Journal of Genetic Psychology* in 1935 Dennis wrote about the "semi-isolation" of Del and Rey: "No toys were introduced until the subjects were eleven months of age," he explained. "To restrict practice which might influence sitting, the infants were kept almost continually on their backs in the cribs.... During the first six months we kept a straight face in the babies' presence, neither smiling nor frowning, and never played with them, petted them, tickled them, etc." Throughout the experiment the twins' view of each other was blocked by a screen between their cribs.

In 1938, in the same journal, Dennis declared that his experiment showed "the infant within the first year will 'grow up' of his own accord." In support of this conclusion he noted that both Del and Rey developed many "responses"—such as laughter, bringing the foot to the mouth and crying in response to sounds—at roughly the same ages as a group of 40 "control" children reared at home.

Yet a chart in the same paper revealed that Del and Rey were slower than the slowest of the control group in developing such key skills as crawling, sitting and standing. Although Dennis stated that a period of training at the end of the 14 months brought the twins up to normal levels in these skills "rather promptly," his data indicated that Del could not stand or walk unassisted until she was more than two years old. When the twins were four years old, Rey's I.Q. was 107 but Del's was only 70; moreover, a physical examination of Del at the age of six showed she suffered from a partial paralysis of her left side. Dennis dismissed Del's "general retardation," calling it "in all likelihood due not to the experiment but to some organic deficiency."

Dennis, who died in 1976, apparently never reported on the twins' later development. He noted in 1951 that after the experiment ended Del and Rey did not return to their mother, who "has shown little feeling of responsibility for them," but lived "at times under the care of relatives and at times in an institution for children."

McGraw's twins, Johnny and Jimmy, were from a poor family in Brooklyn. (McGraw chose the twins in the belief they were identical, but as they grew she concluded they were fraternal.) Only Jimmy, who was considered the "control," was restricted, whereas Johnny was given extensive training. Describing Jimmy's situation, McGraw stated: "He was allowed not more than two toys at a time and was left in his crib undisturbed. He was,



however, by no means isolated, since his crib was merely behind a screen in a busy nursery." The experiment was conducted in McGraw's laboratory not around the clock but for eight hours a day, five days a week over a period of two years. When they were not with McGraw, the twins were at home with their family.

In 1933 McGraw noted in *Psychological Bulletin* that Johnny had achieved no significant superiority in such basic skills as grasping, sitting alone and walking; she concluded that "certain traits are subject to little if any modification through practice." But McGraw went on to state that "other types of behavior can be greatly expanded through training." Indeed, when Johnny was less than a year old, he could climb a 61-degree incline, swim underwater and roller skate, McGraw wrote, whereas Jimmy's "extreme caution often interfered with his motor performance."

After two years Jimmy was given intensive training to see if he could catch up to Johnny, but (except for one skill, tricycle riding) Jimmy's performance "was never as good as Johnny's." When the twins were 10 years old, Johnny's physical superiority persisted: "It is not that he runs faster, or climbs higher," McGraw wrote in an article in the *New York Times Magazine* in 1942, "but that he shows more ease and grace."

McGraw, who is now 88 and living in Hastings-on-Hudson, N.Y., maintains that she has always been forthright about the advantages Johnny gained from his early training. Johnny, she recalls, was still more agile than Jimmy the last time she tested them, when they were 22 years old. Asked why other investigators have reported that Johnny's advantage was ephemeral, she observes: "I think people read what they want to."

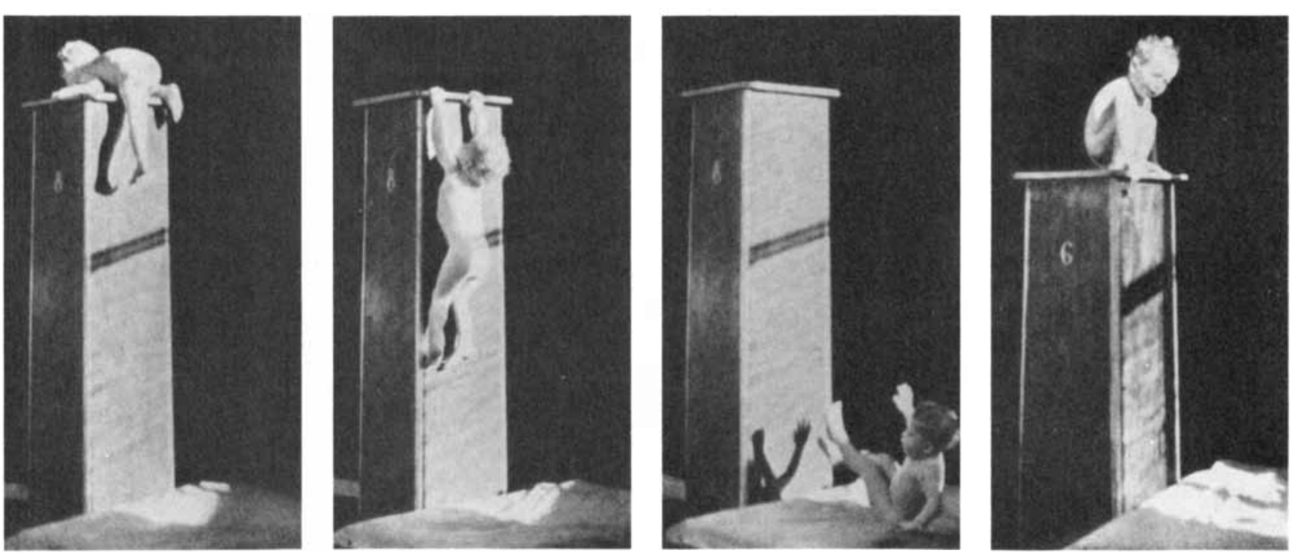
The following citations typify the way in which the Dennis and McGraw experiments have been interpreted. In the *Handbook of Infant Development,* published in 1979, Jerome Kagan of Harvard University wrote: "After [Del and Rey] returned to a normal environment they eventually displayed a growth function that was normative." Alfred L. Baldwin of the University of Rochester stated in his textbook *Theories of Child Development,* published in 1980, that Johnny's training "gave him only a small, temporary advantage over Jimmy, his untrained twin."

Razel, an advocate of training for infants, learned of the Dennis and McGraw experiments in 1973 while working for his doctorate in psychology at New York University, soon after his wife bore their first child. His thesis adviser, suggesting that the attention Razel was giving his child was excessive, referred him to the experiments. "I couldn't believe what he said the results were," Razel recollects, "and so I decided to see for myself."

Razel's review of the experiments—which emphasizes Del's retardation and Johnny's long-term physical advantage over Jimmy—was published in 1985 as a chapter in *Advances in Applied Developmental Psychology,* edited by Irving E. Sigel of the Educational Testing Service. Now Razel is hoping to publish an article that calls for a follow-up study of Del, Rey, Johnny and Jimmy, to determine what the long-term results of the experiments have been. Sigel says he is considering publishing the article in *Journal of Applied Developmental Psychology,* which he edits. "The training of infants is a hot issue," he says, "and Razel is offering a point of view that tends to be minimized."

Was it wrong to conduct the experiments? J. McVicker Hunt, himself a pioneering experimental psychologist who knew both Dennis and McGraw, says: "It would be utterly wrong to repeat this kind of thing today. But ethics depend on what you know, and they didn't know that what they were doing might be harmful." Edward F. Zigler of Yale University warns that Razel's belief that training is crucial to a child's development can also be taken too far. "It would be awful to tell parents that what they do makes no difference," Zigler notes. "But if parents are working their heads off trying to make a brilliant child out of an ordinary one, that can hurt the child too."

McGraw says she was always more concerned with simply observing her subjects than she was with resolving the nature versus nurture debate: "We have to learn to get away from these dichotomies." She insists that she has no regrets about the experiments—that Jimmy was not deprived any more than many children would be in a normal home setting and suffered no real ill effects. Jimmy, who is now a businessman living with his wife and children in upstate New York, occasionally visits her, McGraw says, whereas Johnny, who lives with his mother in Florida, has long kept himself incommunicado. "Jimmy knows Johnny's address," McGraw says, "but he won't tell me what it is."



ADVENTUROUS JOHNNY, 21 months old, climbed down from a stand more than five feet high; Jimmy, his untrained twin, refused to budge from a shorter stand (*right*). The photographs appeared in the 1935 book *Growth: A Study of Johnny and Jimmy.*

Established 1845 SCIENTIFIC AMERICAN July 1987 Volume 257 Number 1

Can Advanced Technology Save the U.S. Steel Industry?

Emerging technologies that reduce the costs of making iron and steel may help smaller American mills to counter stiff foreign competition, but entirely new processing techniques are needed for the long term

by Julian Szekely

The widely publicized problems of the U.S. steel industry—persistent operating losses, bankruptcies, forced mergers and massive layoffs—are borne out by grim statistics: since 1982 the industry's operating losses have amounted to some $6 billion and its employment roster has shrunk from 500,000 in 1975 to fewer than 200,000 today. In explaining the industry's dismal state observers have suggested many causes: poor management, self-serving labor unions, outdated technology, competition from overseas and the replacement of steel by materials such as aluminum and fiber-reinforced plastics.

Yet such hardships are not unique to American steel producers. Similar—if not quite so acute—problems also beset the steel industries of Japan and western Europe. Indeed, projections of world steel production indicate stagnation and decline in the industrialized countries posed against rapid growth in the steel industries of developing countries. According to one estimate, the steel-production capacity of developing countries is expected to increase by 20 percent between 1985 and 1990 while the capacity of the U.S. erodes by about 3 percent. Can anything be done to reverse or at least slow the downward trend in American steel production?

Unfortunately the problems confronting the U.S. steel industry are the result of a combination of worldwide trends, organizational weaknesses and technological issues that allow no single ready solution. The introduction of new steelmaking technologies might nonetheless play a critical role in revitalizing this important industry. Incremental technological improvements, in conjunction with other measures to increase output or reduce the work force, can increase the efficiency and product quality of existing plants, and such improvements should be undertaken to ensure the continued viability of the industry, at least in the next several years.

To ensure its long-term economic health, however, I believe the industry will have to switch to the manufacture of novel, high-value-added products. This requires the development and implementation of an entirely new range of innovative steelmaking technologies—specifically technologies for creating and producing a wide range of high-priced "customized" steels.

In considering the impact such technologies could have on the U.S. steel industry it is imperative to recognize that the industry is not monolithic. There are actually three types of steel producers: integrated producers, minimills and specialty-steel mills. They differ markedly in breadth of operations, machinery and susceptibility to economic hardship. The integrated producers, which start with iron ore and coal and end up with a wide assortment of shaped steels, account for some 70 percent of all ordinary grades of steel manufactured in the U.S. The minimills, which reprocess scrap steel generally into a limited range of low-quality products, produce most of the rest. The specialty-steel mills, whose output accounts for only 5 percent of U.S. production (but a much higher percentage of total revenues), are similar to minimills in that they tend to be smaller than the integrated producers and are based on scrap, but they manufacture much more expensive products than minimills do and commonly have an active in-house research-and-development effort.

The worst economic troubles afflict the integrated producers. By and large their equipment is old, is less automated and does not incorporate many of the latest evolutionary refinements in steelmaking technology. The minimills as well as the specialty-steel mills have succeeded in avoiding the worst of the economic difficulties. Indeed, some of them are quite profitable. Both take advantage of new technology for refining and casting steel as soon as the technology becomes available. The minimills deliberately concentrate their production in rather narrow and localized market niches; the specialty-steel mills, on the other hand, preserve flexibility in their operations in order to fulfill a customer's particular specifications.

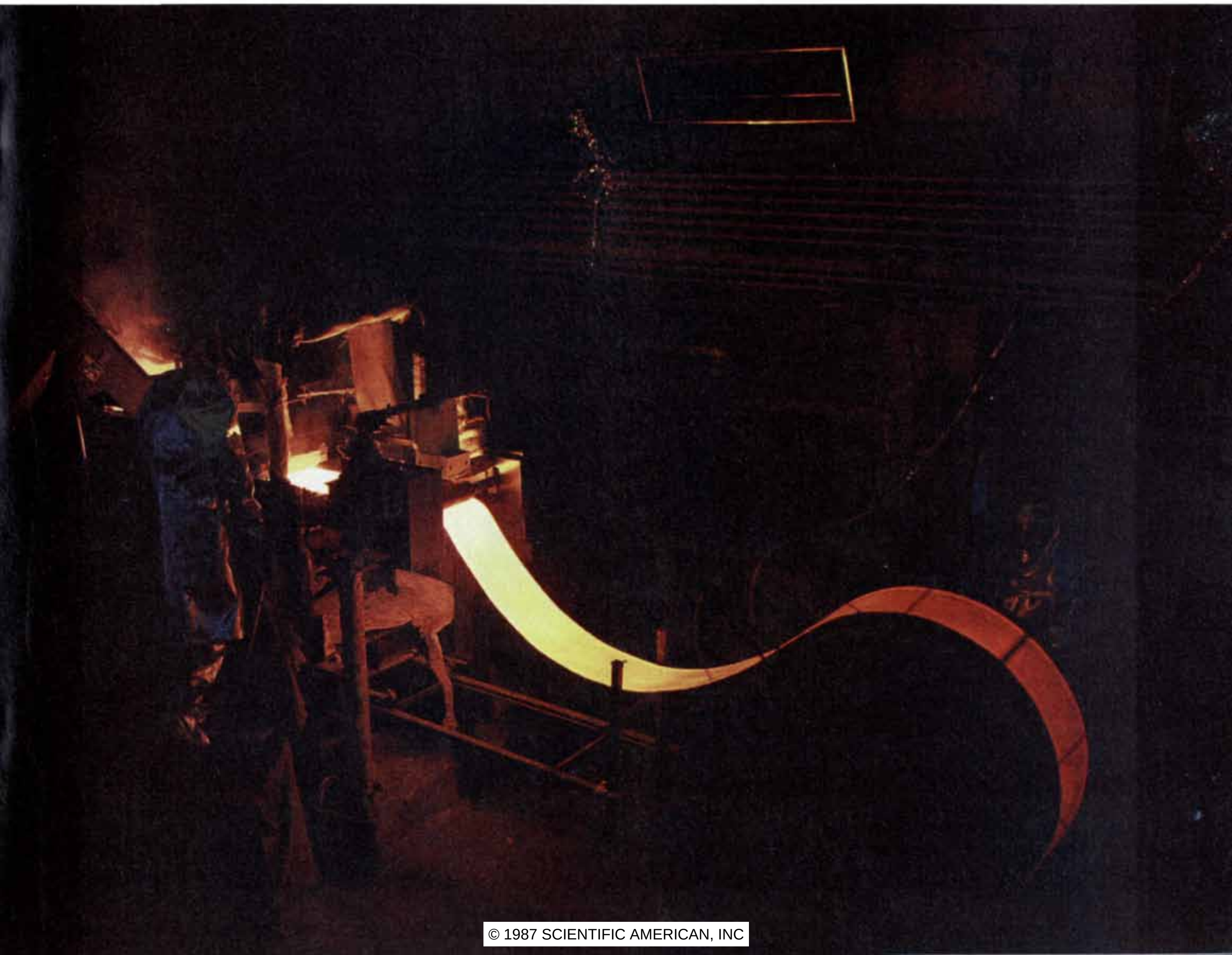
DIRECT CASTING produces stainless-steel strip from molten metal at the Allegheny Ludlum Steel Corporation Research Center. Such casting saves energy, time and labor compared with older processing.

Integrated producers are in a sense victims of the inflexibility of the enormous enterprise to which they have been historically committed. An integrated steel plant's operations begin with the mining of coal and iron ore. The coal is heated in the absence of oxygen to transform it into coke (an operation that has many adverse environmental effects). The ore is finely ground and agglomerated into small lumps. The treated raw materials are fed into a blast furnace, where the iron oxide in the ore is reduced (stripped of oxygen) and smelted to yield pig iron. Crushed limestone is also loaded into the blast furnace to form a liquid, called slag, that captures most of the waste products of the process.

The pig iron is turned into steel by a controlled oxidation reaction in which undesirable impurities (such as carbon, silicon and manganese) are made to combine with oxygen. At this stage the steel may be alloyed: mixed with other materials to invest it with certain properties. The molten steel is then cast either in batches into oblong ingots weighing between 50 and 100 tons or continuously into slabs (which are about 10 inches thick and between 30 and 40 inches wide), blooms (which have about a 20-inch-square cross section) or billets (which have a square cross section between three and 10 inches on a side). The ingots, slabs, blooms and billets must be repeatedly squeezed between sets of rollers to deform them into finished shapes such as plates for ships, sheets for automobiles, structural members for construction, and rods for the manufacture of wire or tubes.

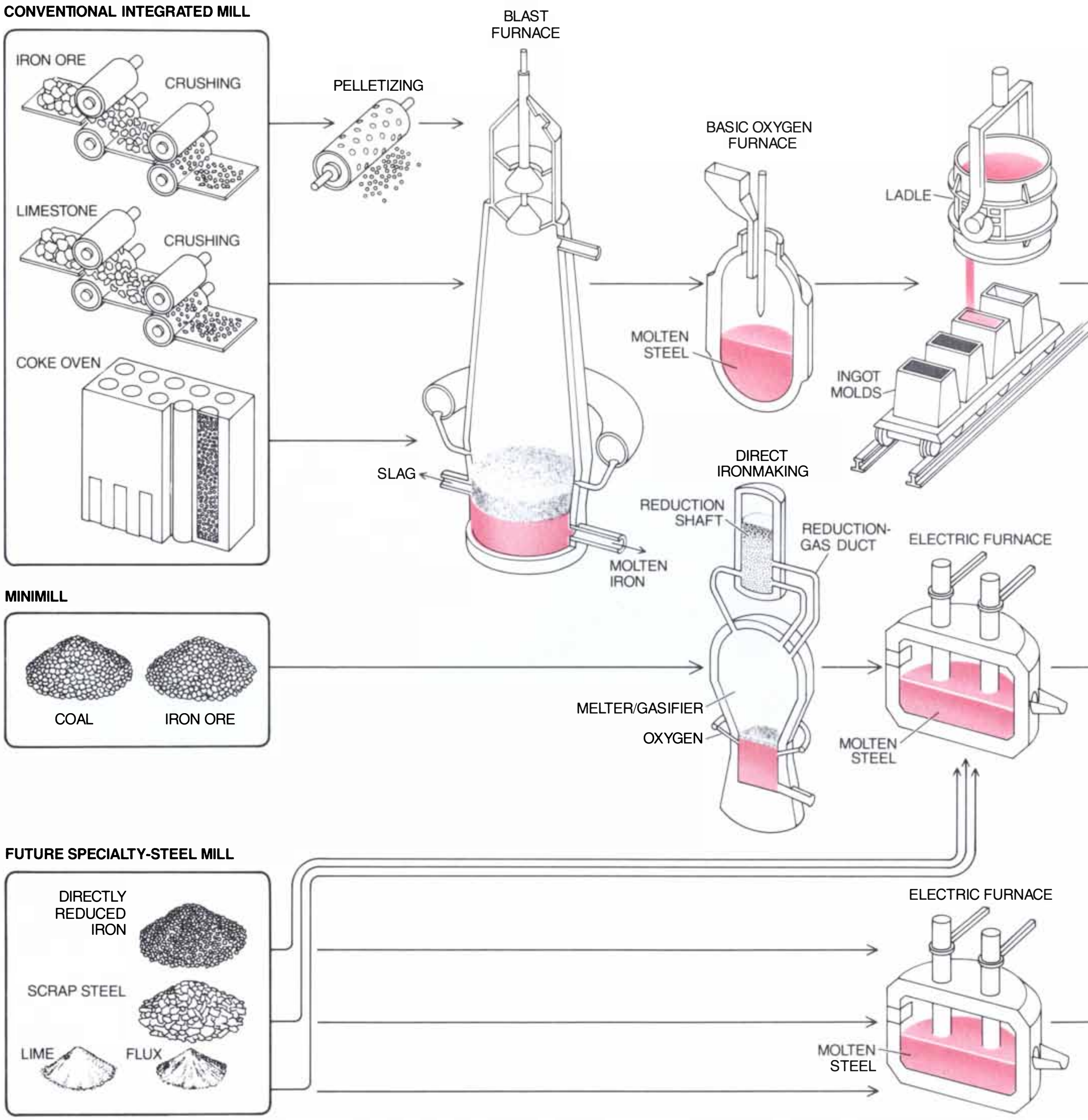
A typical U.S. integrated steel plant may manufacture between two and five million tons of finished steel products per year. To do this it requires about 25 gigajoules of energy and about six man-hours per ton. (One gigajoule is roughly equivalent to the energy obtained through the complete combustion of 100 pounds of coal.) Usual production costs (excluding capital costs) are between $400 and $500 per ton. Labor is the major cost associated with the operation of the rolling mills; energy is the major cost in the earlier coking and iron- and steelmaking operations. The capital cost of building an integrated plant is estimated to be as high as $1,600 for each ton of annual capacity. The capital outlay, then, for a typical integrated plant can easily amount to several billion dollars, approximately equally divided between the "front end" operations (coke-, iron- and steelmaking) and the finishing operations (rolling).

Japanese integrated producers, in contrast, have pioneered ultramodern but conventional steelmaking plants that can manufacture a ton of steel with about 80 percent of the energy and less than half of the man-hours required by American producers. Another significant difference between the steels manufactured by integrated plants in the two countries is the markedly better quality of the Japanese steels in terms of chemical purity, cor-



rosion resistance, mechanical properties and dimensional tolerances. The difference in quality is attributable in large part to the application by the Japanese of "top of the line" computer-controlled processing equipment and to the constructive attitude of both labor and management. Such facts make it tempting to conclude that the older, labor-intensive machinery still operating in U.S. integrated plants is at fault for the poor performance of the American industry. To a certain extent this is true, but it cannot explain why Japanese integrated producers are also experiencing economic trouble.

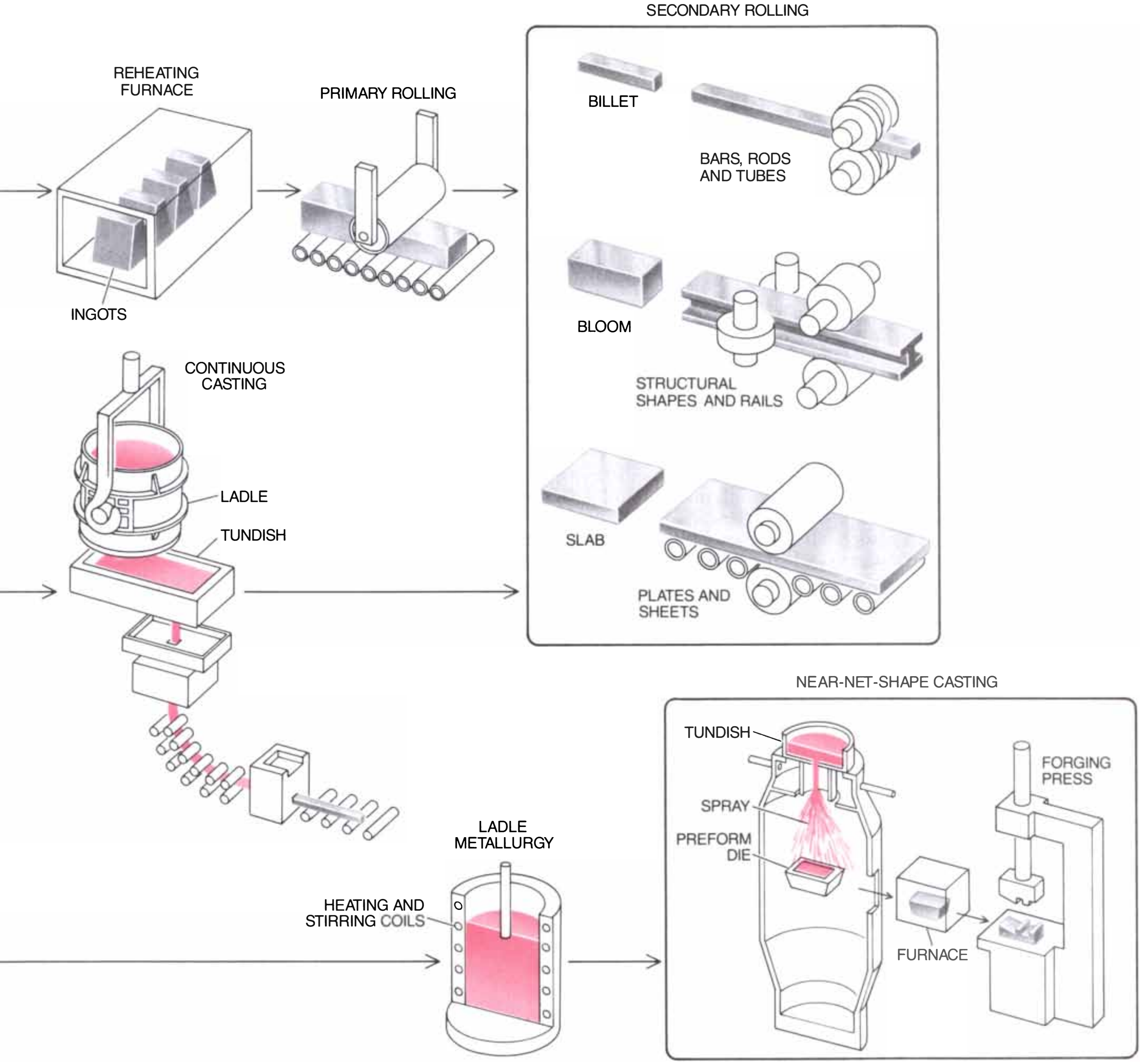
The fact is that steelmaking as practiced by integrated producers—in any country—is an inherently inefficient process that is still rooted in the 19th century. Because integrated producers



MANUFACTURE OF STEEL PRODUCTS varies markedly according to plant type: integrated mill, minimill or specialty-steel mill. Operations at integrated mills encompass both "front end" and finishing processes. The front end includes the preparation of coke, iron ore and limestone for charging a blast furnace, where the ore is reduced (stripped of oxygen) and heated to yield liquid iron and slag (a waste product), as well as the refining operations that convert the iron into steel. Finishing refers to the casting, reheating and rolling operations that transform the molten steel into shaped products such as rods, tubes, structural beams, plate and sheet. Minimills generally dispense with most of the front end of steelmaking and begin with steel scrap, but they could soon supplement the scrap with molten iron from direct-ironmaking plants. Direct ironmaking represents a substantially smaller in-

are ultimately reduced to a common technological denominator, other factors dominate the economics of steelmaking, namely labor costs. It is in this regard that developing countries have a significant comparative advantage: they generally have substantially lower wages. Hence it is developing countries (such as Korea and Taiwan) rather than developed countries (such as Japan) that undercut American steel in both the domestic and the international markets.

Minimills have been able to compete successfully with both domestic and foreign integrated producers, because they have dispensed almost entirely with the energy- and capital-intensive front end of steelmaking: the iron-smelting process, including the mining and preparation of the raw ma-



vestment than a blast furnace does and relies on iron ore and coal (rather than coke) as raw materials. The alternative is to buy directly reduced iron: iron from modern plants that reduce iron ore without melting it. Minimills were among the first steel mills to exploit continuous casting: the casting of billets, blooms or slabs directly from the molten steel. The process bypasses the casting of ingots and many of the reheating and rolling steps still carried out at some integrated mills. Specialty-steel mills are also primarily scrap-based, but unlike minimills, which tend to produce low-grade steels for a local market, they produce high-quality alloyed steels to customers' specifications. A future specialty-steel mill might impart still higher quality to its products by further refining and purifying the steel-alloy melt in special ladle-metallurgy stations and then casting the product in virtually its final shape.

terials and the blast-furnace operation. Their capacity lies in the range between 300,000 and 1.2 million tons per year. More important, their products require only about two or three man-hours per ton of steel, and the capital cost of a minimill installation is in the range of only between $200 and $300 for each ton of annual capacity. Minimills also tend to be well run and benefit from excellent relations between labor and management.

In addition, minimills have found a profitable way to market steel products: they sell their finished products locally, thereby reducing transportation costs, and concentrate on a limited range of shapes and sizes within a narrow group of products that can be manufactured economically [see "Steel Minimills," by Jack Robert Miller; SCIENTIFIC AMERICAN, May, 1984]. Such products include reinforcing rods for concrete (which are made from the lowest-grade steel), wire rods and bar products of assorted cross-sectional dimensions. More recently, however, minimills have been making inroads into lines of higher-grade products.

Minimills generally start with scrap steel (although they may supplement it with reduced iron) and carry out only the refining, casting and rolling operations. One key process that has made the economical operation of minimills possible is continuous casting, on which all minimills rely. (In contrast, only about half of the U.S. integrated producers have continuous casters.)

In assessing what technology can realistically do—indeed, must do—to improve the economic state of the steel industry one should bear in mind the factors that constrain the competitiveness of American steel producers. In the case of the integrated producers these factors amount to excessive labor, energy and capital costs, as well as manufacturing inflexibility. There are also some definite limits to the further expansion of minimills, set by the availability of scrap, by its impurity content (which prevents certain high-quality grades of steel from being produced) and by the cost of extensive finishing facilities to make flat-rolled products.

These technological and economic considerations should ultimately determine which technological avenues are pursued. One technology that appears attractive in this respect is direct ironmaking: the one-step production of molten iron from iron ore and coal. Direct ironmaking promises major savings in capital outlays and operating costs over the conventional blast furnace and its concomitant cokemaking and ore-agglomeration facilities, which currently burden the integrated producers.

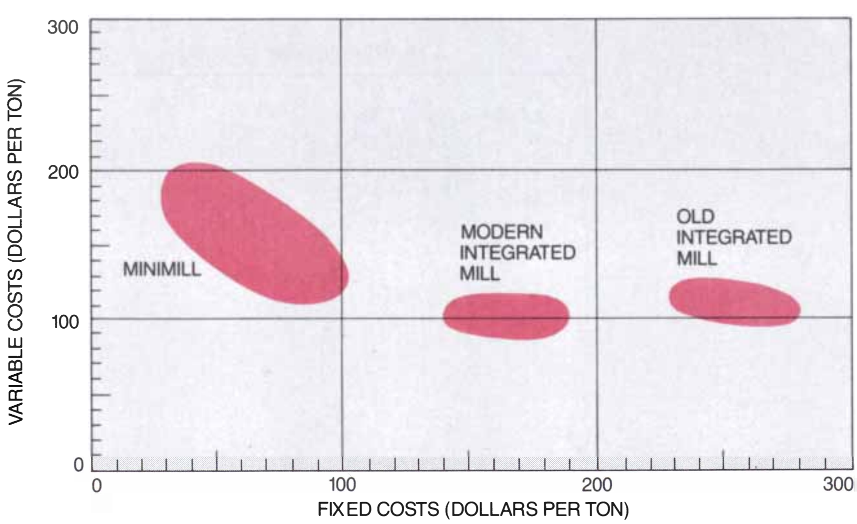
A similar reduction in the number (and cost) of separate processing stages is the driving force behind another family of new technologies: those that seek to transform molten steel in one step to a finished product. These technologies would free the capital, energy and labor costs tied up in the sequence of finishing operations. Normally, for example, thin sheets and plates must be rolled from ingots or slabs in a rather laborious and time-consuming process. If steel sheets a few millimeters thick could be made directly in a continuous caster, minimills would have a cost-effective way to enter the lucrative flat-rolled market, which they have until now largely shunned.

Direct ironmaking has progressed to or beyond the pilot-plant scale; installations have already been built that are capable of producing hot, molten iron economically at production levels of from 30,000 to one million tons per year. Various direct-ironmaking processes are being perfected, but in essence all of them consist of a vessel in which iron ore (or partially reduced iron ore) and coal are charged along with an injection of oxygen. The end product is molten iron, similar to that produced in the blast furnace. Coal performs a dual function in that it acts both as a reducing agent and as a fuel to generate the heat necessary to melt the iron.

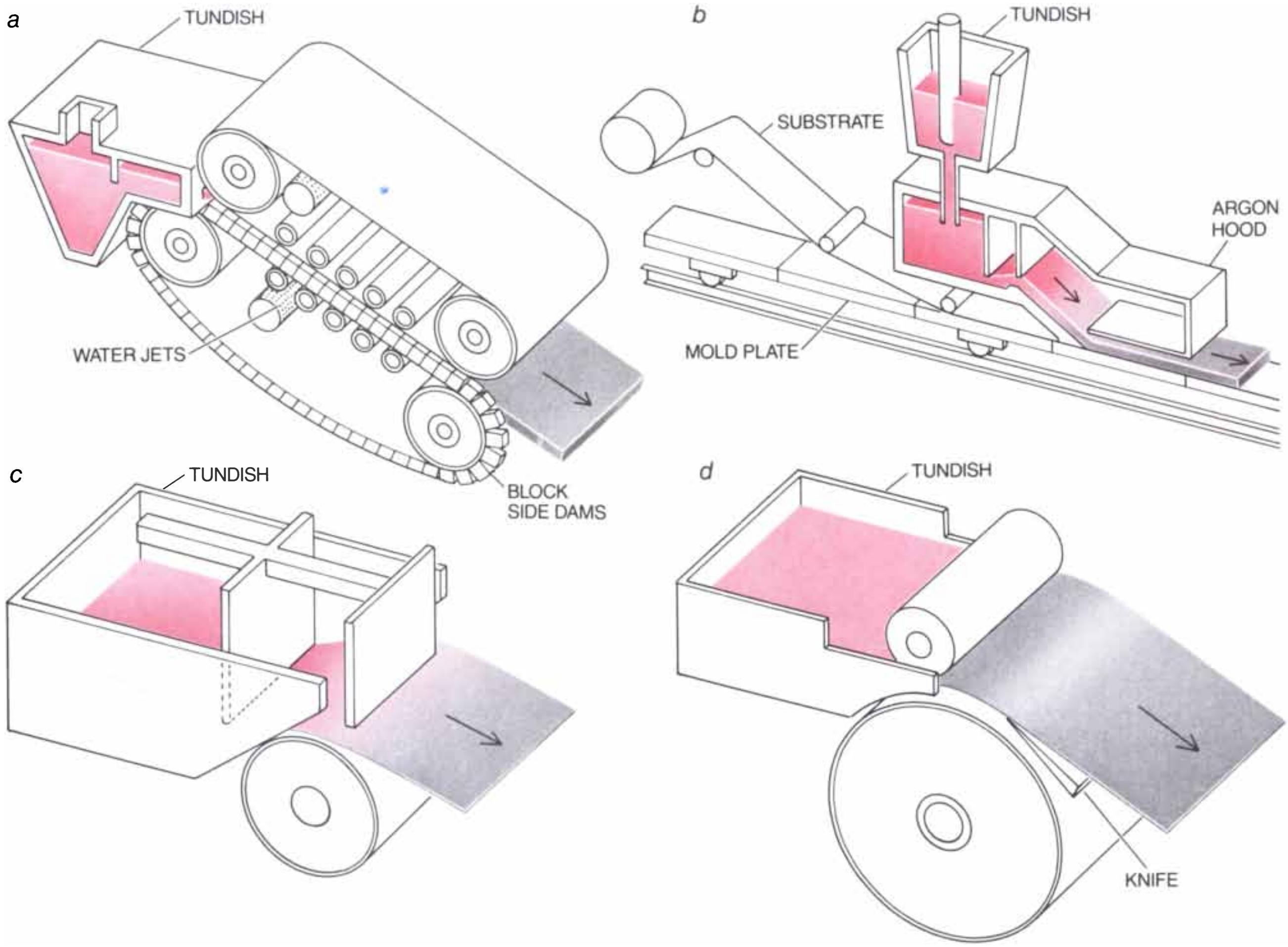
Direct ironmaking might appeal to minimills, since it is a potential source of iron that is not contaminated with "tramp" metallic elements, such as copper, lead and zinc, that are commonly found in scrap. It may also appeal to the integrated producers as a supplementary ironmaking facility or possibly as a replacement for old coke- and ironmaking facilities, which may be forced to close in response to environmental concerns.

There are technical problems in the realization of direct-ironmaking processes, however. They include the development of materials capable of withstanding the extremely corrosive conditions and high temperatures in the vessel and the maintenance of a delicate balance between several simultaneous chemical reactions in the system. There are also some economic obstacles. Currently there is excess worldwide ironmaking capacity; consequently integrated producers, which can ill afford capital outlays, are not likely to modernize their installations with direct-ironmaking facilities—even if it would result in more efficient and cheaper production of iron.

A key drawback of direct ironmaking for minimills is that it produces iron rather than steel and a steelmaking furnace would be a necessary adjunct. Since minimills already have steelmaking facilities, it might make more economic sense to use imported directly reduced iron (reduced without smelting) rather than invest in the construction of a direct-ironmaking facility, if pure iron is needed for special applications. In the light of these considerations, direct ironmaking may fill certain niches in the U.S. steel industry



TYPICAL RANGE OF COSTS per ton of annual capacity for an old integrated mill, a modern integrated mill and a minimill can be expressed in terms of fixed and variable costs. Fixed costs are expenditures for such items as capital, manpower and maintenance, which do not vary significantly once a plant is operating. Variable costs include expenditures for raw materials, other supplies and energy, which may change from year to year. Since minimills do not have to wait as long as integrated mills to recoup the investment in setting up operations, they can consider replacing old equipment with new sooner.



DIRECT-CASTING TECHNOLOGIES are being developed that would make it possible to continuously cast steel plate, which is between 12 and 50 millimeters thick (*a, b*), or sheet, which is one millimeter or less thick (*c, d*). This type of casting would reduce production costs and enable small, nonintegrated mills to compete directly with large, integrated producers in the flat-steel market.

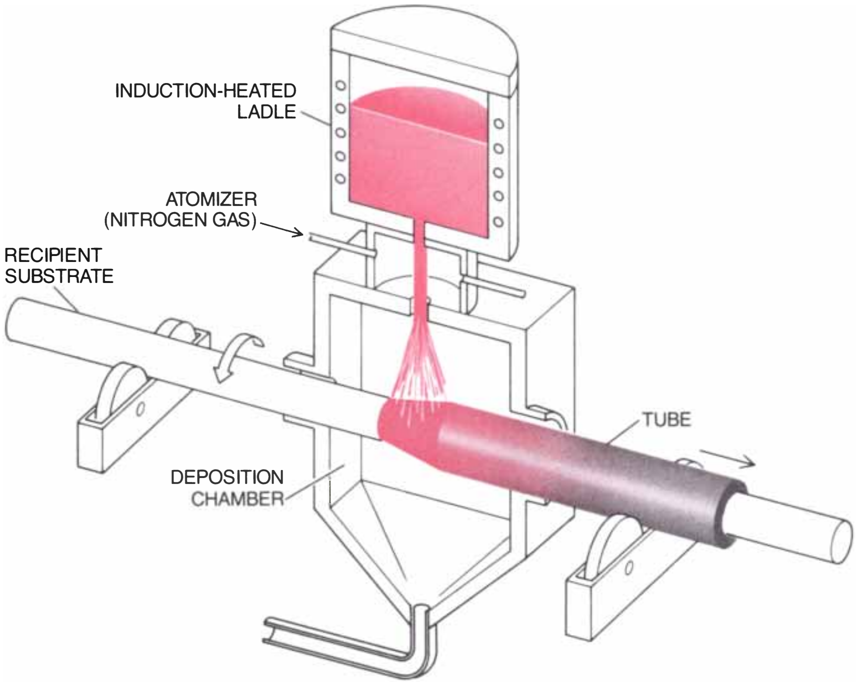
but is unlikely to have a broad impact throughout the industry.

The direct casting of plates and sheet [*see illustration above*] is sure to have a more important role than direct ironmaking in the U.S. steel industry, particularly in the minimill sector. On the face of it, direct-casting concepts are tremendously attractive because they make it possible for minimills to eliminate all or most of the rolling operations and the high associated labor cost. These technologies may appeal to outmoded integrated plants too, because in going directly from ingot casting to continuous thin-slab or sheet casting, they could circumvent a large fraction of their rolling operations. The more modern integrated plants, which already have continuous slab casters, would not benefit as much, since in their case only a relatively small part of the rolling-mill operations would be eliminated by switching to thinner shapes. Nevertheless, it is a safe bet that a substantial fraction of the steel produced in the world in about 10 to 15 years will be processed by direct casting.

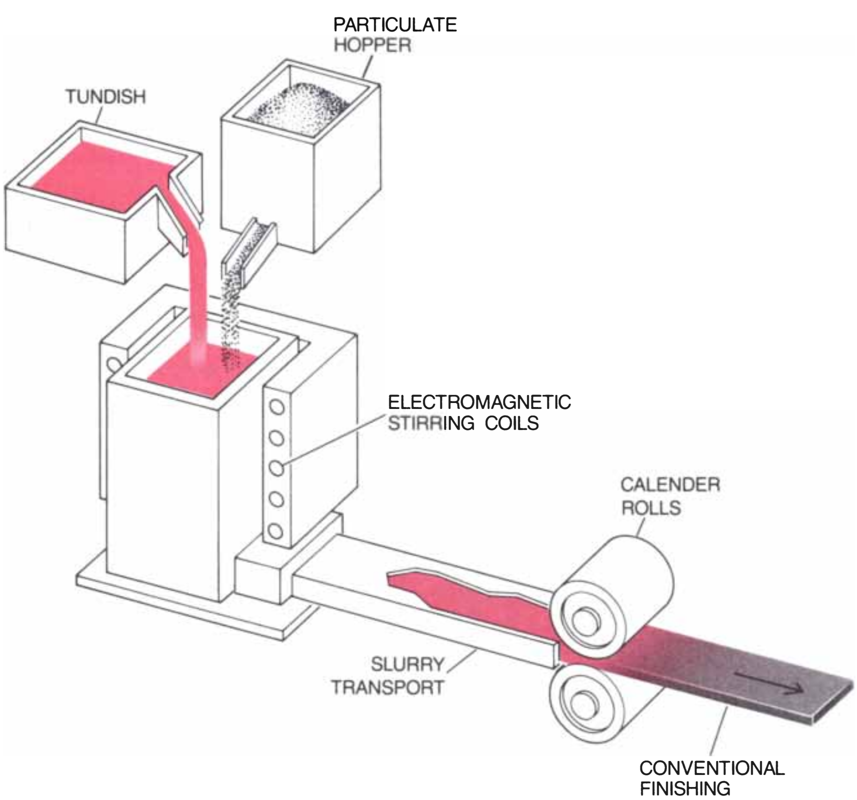
Yet this technology has a number of problems that still have not been completely solved. The reduction in mechanical working made possible by direct casting puts much greater demands on the purity of the molten steel. Traditional steelmaking, in which ingots are extensively processed by deformation, was wasteful, but it was much more tolerant of variations in the composition of the steel in the ladle. In modern casting as practiced by minimills one cannot afford any waste and consequently must be careful to ensure that the composition and dimensions of the steel meet precise specifications.

For example, in the process of continuous casting the surface of ordinary steels is contaminated by the atmosphere. In the case of slabs, which range in thickness between 200 and 300 millimeters, the problem is easily solved by scraping off a layer between half a millimeter and one millimeter thick from the solidified slab. In the continuous casting of sheet, which is typically no more than one millimeter thick, such a treatment is obviously impossible. In this respect specialty-steel mills, which cast alloys that resist such contamination, have an advantage.

The principal economic impact of direct-ironmaking and -casting technologies would be a reduction in the cost of making products that already are available. The cost reduction, estimated at between 15 and 40 percent, could have a marked effect on the competition between minimills and integrated producers in this country. It is unlikely, however, that these technologies (even if they were coupled with a restructuring of the steel industry in favor of the minimills and specialty-steel mills) would restore a long-term, competitive edge to the U.S. steel industry in the domestic market—let alone in the international market. Rival mills in developing countries have access to the same steelmaking technology, can take advantage of much lower intrinsic wage rates and may benefit from government policies that



SPRAY CASTING is a radically new technology that would enable steel mills to manufacture high-quality steel products having superior mechanical properties and resistance to corrosion. In the so-called Osprey process shown here molten steel is atomized by jets of nitrogen gas and deposited on a rotating cylindrical substrate to form a seamless tube.



RHEOCASTING, a process whereby a partially solidified melt is kept castable by vigorous agitation, may someday rely on electromagnetic forces to stir the melt. Such mixing would also make it possible to distribute reinforcing particles or fibers evenly in the semisolid melt. The resulting material—a metal-matrix composite—would exhibit a high strength-to-weight ratio. It is the kind of product future steel mills will have to make.

enable them to ignore market forces.

To ensure itself a share of the future world steel market the U.S. steel industry will have to develop radically new methods for making products that are not yet widely available. In the light of current and likely future economic realities such products would have to acquire a substantial value from unusual processing. The products would include steels with coatings or surface modifications providing superior resistance to corrosion and wear, fiber-reinforced steels and new steel alloys with unique microstructural properties imparted by rapid-solidification or powder-metallurgy techniques. Naturally, sophisticated processing technologies are necessary to make these steels, but the products would command a price high enough to justify the capital (and development) expenditures. A minimum threshold value of about $1,000 or more per ton might serve as a rough guideline.

One technology that stands out in this regard is that of "near net shape" casting. The idea is to carry direct casting even further. In the so-called Osprey process, for example, a stream of molten steel is atomized by means of high-speed gas jets, and the small droplets of liquid metal are sprayed onto a solid substrate. By adjusting the movement of either the nozzle or the substrate it is possible in principle to build up solid bodies of any given shape.

This has been shown to work rather well in the production of tubes: the molten metal is simply sprayed onto a rotating cylindrical substrate. Similarly, a wider stationary nozzle can spray layers of steel onto a moving belt to produce strip or sheet at a rate of between five and 10 tons per hour. Because the steel droplets solidify quickly as they are deposited, bulk materials can be produced that have desirable mechanical properties and corrosion resistance. Moreover, the rapid solidification that makes these processes work has been shown to invest steel with a fine-grained, homogeneous microstructure that may eliminate the need for any subsequent mechanical working.

Another area where research might yield fruitful results is rheocasting: the mixing and casting of partially solidified melts. The key to rheocasting is the constant agitation of the melt. The agitation causes the melt to remain fluid and therefore castable even if the melt has a high (say 30 to 50 percent) solid content. (The same principle is exploited in the making of ice cream.)

When steel is cast in this way, its

crystalline grains can be made to be small and uniform, endowing the steel with desirable structural and mechanical properties. Moreover, the forceful mixing may be particularly effective for dispersing fibers or other particulate matter in the steel. A modified rheocasting technique could therefore produce metal-matrix composites that have excellent mechanical properties and strength-to-weight ratios. Since the amount of heat that needs to be extracted from the agitated melt in order to solidify it completely is considerably less than the heat of solidification of molten steel in conventional ladles, the rheocast steel can also be cast more readily in near-net shapes.

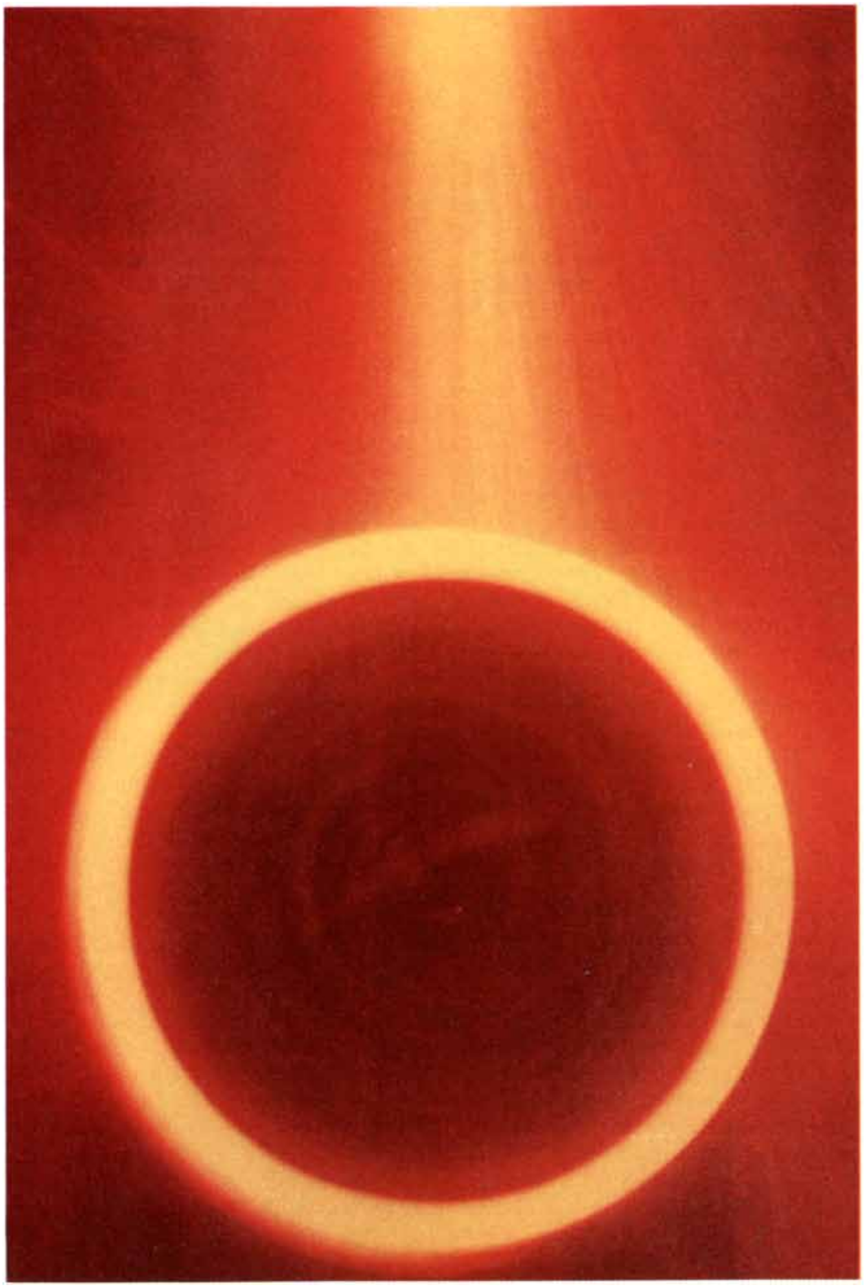
Although the internal microstructure of a piece of steel determines its mechanical properties, it is the piece's surface layers that directly affect its corrosion resistance and certain other characteristics. Hence future steel mills should include some type of facility that would apply one of several ways of treating the surface of steel to give it special properties. One method is laser glazing, in which a high-power laser beam is passed rapidly over a material's surface to melt surface layers. The molten layers are immediately cooled by the underlying bulk material, and so the end result is a surface microstructure that can be modified without affecting the underlying steel.

Other ways to achieve special surface properties are plasma spraying and magnetron sputtering. In plasma spraying a high-temperature, high-velocity ionized gas that is produced by subjecting the gas to an electric arc carries a coating material to the surface to be treated. Plasma spraying makes it possible to apply heat-, wear- or corrosion-resistant coatings to metals. In magnetron sputtering a charged-particle beam is directed at a melt placed next to the workpiece. As the particles impinge on the melt they sputter the melt onto the piece.

To be sure, both plasma spraying and magnetron sputtering have existed for several years. Recently, however, the phenomena underlying them have come to be better understood (largely through computer modeling) and superior heat-resistant materials have become available; one can now consider applying the techniques in a precisely controlled way to the processing of tonnage materials, such as steel sheet for automobiles.

In examining the condition of the U.S. steel industry one may conclude that no simple "technology fix" will save the entire industry. Although there are many novel and promising technological ideas on the horizon, the benefits to be derived from their implementation are not likely to be uniformly distributed throughout the industry. Emerging technologies such as direct casting and spray forming would be particularly well suited for the needs and the scale of operation of minimills and specialty-steel mills. If these concepts are successfully implemented, the minimills will provide an even more formidable competition to the integrated producers—domestic and foreign alike. Loosening the reins of one's imagination, one can plausibly envision ladle metallurgy, near-net-shape casting and surface modification ultimately as integral parts of a relatively small, highly automated "utopian" steelmill, where molten steel is produced directly from iron ore and other raw materials.

Indeed, it seems unlikely that the integrated producers with large, inflexible operations will survive in their present form beyond the 20th century. Yet if nonintegrated producers end up exploiting these new technological developments in order to grow at the expense of the integrated producers, the national economy may be the ultimate beneficiary.



OSPREY PROCESS (*see top illustration on opposite page*) **produces an incandescent tube, which is seen here head on. The photograph is from AB Sandvik Steel of Sweden.**

The Molecules of Visual Excitation

When a rod cell in the retina absorbs light, a cascade of reactions results in a nerve signal. That cascade has now been worked out in molecular detail. A key intermediate is a protein called transducin

by Lubert Stryer

This is an exciting time in the investigation of vision. Many years ago William A. H. Rushton of the University of Cambridge wrote: "Molecules respond to light as do people to music. Some absorb nothing. Others respond by the degraded vibration of foot or finger. But some there are who rise and dance and change partners." At the time Rushton wrote, his description was largely poetry: it was not known precisely which molecules are involved in the response of the retina's photoreceptor cells to light; nor was it known how those molecules interact.

In the past decade, however, experiments in many laboratories (including my own) have revealed the molecular basis of visual excitation. The molecules that participate in the response are known, and the basic scheme of their interactions has been worked out. That detailed biochemical work has shown Rushton's description to be a prescient one. The molecules that form the basis of the response to light do indeed "rise and dance and change partners" in a remarkable cascade that lies at the root of vision.

The molecular cascade that has been worked out so carefully in the past decade has its seat in the photoreceptor cells of the retina. The photoreceptor cells are of two types, which are called rods and cones because of their characteristic shapes. Rod cells make it possible to form black-and-white images in dim light; cones mediate color vision in bright light. The human retina contains three million cones and 100 million rods. The electrical signals generated by the rods and cones are processed by other retinal cells before being transmitted to the brain by way of the optic nerve.

My interest in the molecular basis of vision was originally stimulated by certain striking properties of the rod cells. As receptors, rods have attained the ultimate in sensitivity. A rod cell can be excited by a single photon, which is the smallest possible quantity of light. The cascade of molecular reactions amplifies this minute piece of information into a signal that is useful to the nervous system. What is more, the degree of amplification varies with the background illumination: rod cells are much less sensitive in bright light than in dim light. As a result they function efficiently over a wide range of background illumination.

I was also attracted to rod cells because their exquisitely sensitive sensory system is packaged in a distinct cellular subunit that can readily be detached and studied. The rod cell is a long, thin structure divided into two parts. The outer segment contains most of the molecular apparatus for detecting light and initiating a nerve impulse. The inner segment is specialized for generating energy and renewing the molecules needed in the outer segment. In addition the inner segment includes a synaptic terminal that provides the basis for communication with other cells. If an isolated retina is gently shaken, the outer segments fall off and the machinery of excitation can be studied in a highly purified form. This feature has made the rod cell a great gift to biochemists.

The outer segment of the rod is a narrow tube filled with a stack of some 2,000 tiny disks. Both the tube and the disks are made up of the same type of bilayer membrane. The outer (or plasma) membrane and the disk membrane, however, have different functions in the reception of light and the generation of a nerve impulse. The stacked disks contain most of the protein molecules that absorb light and initiate the excitation response. The outer membrane serves to convert a chemical signal into an electrical one. Much of the dramatic new work on visual excitation has been devoted to tracing the process that links the molecules of the disk membrane with those of the outer membrane.

Among the most important of the molecules associated with the disk membrane is the one called rhodopsin. Rhodopsin is the photoreceptor protein of rod cells, the molecule that absorbs a photon and makes the initial response in the chain of events that underlies vision. Rhodopsin has two components, which are called 11-*cis* retinal and opsin. 11-*cis* retinal is an organic molecule derived from vitamin A. Opsin is a protein that has the capacity to act as an enzyme. The absorption of a photon by 11-*cis* retinal triggers the enzymatic activity of opsin and sets the biochemical cascade in motion.

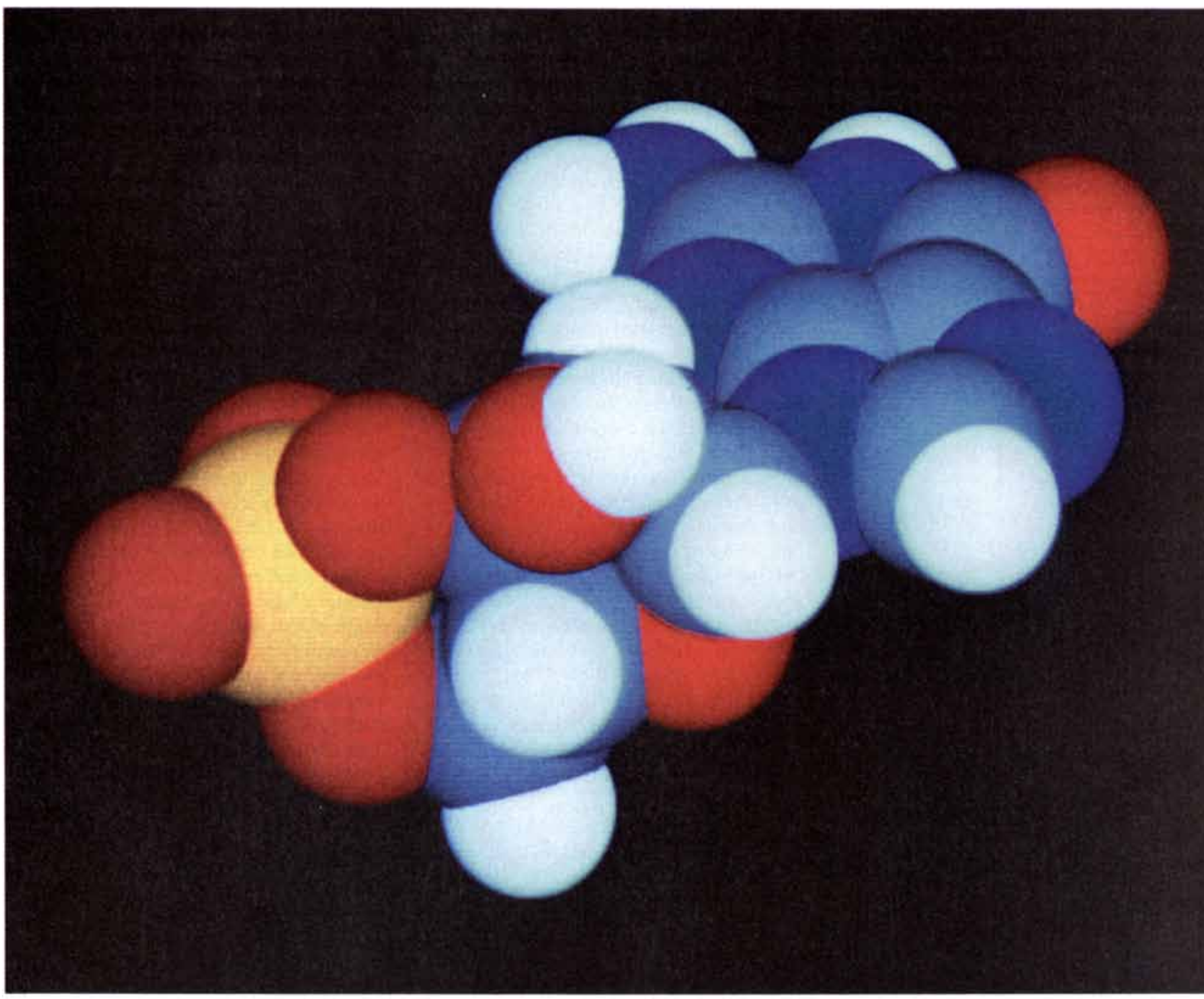
Opsin is a single polypeptide chain of 348 linked amino acids. Recently the amino acid sequence of opsin was worked out in the laboratories of Yuri A. Ovchinnikov of the M. M. Shemyakin Institute of Bioorganic Chemistry in Moscow and Paul A. Hargrave, then at Southern Illinois University. That work has provided considerable information about the three-dimensional structure of the protein, which is threaded through the disk membrane. It appears that opsin has the form of seven helixes (of the type known as alpha-helixes) arranged vertically in the membrane and connected by short nonhelical segments [*see illustration on page 45*]. Attached to one alpha helix is a single molecule of 11-*cis* retinal, which lies near the center of the membrane, its long axis aligned with the plane of the membrane.

This arrangement leaves retinal nested at the center of a complex and highly structured protein environment. That environment (among other factors) is responsible for "tuning" retinal by influencing the spectrum of radiation it can absorb. Whereas retinal by itself in solution absorbs most intensely at a wavelength of 380 nanometers (in the ultraviolet part of the spectrum), rhodopsin does so at 500 nanometers (in the green). This shift is an excellent one from a functional point of view, because it matches the absorption spectrum of rhodopsin with the light reaching the eye.

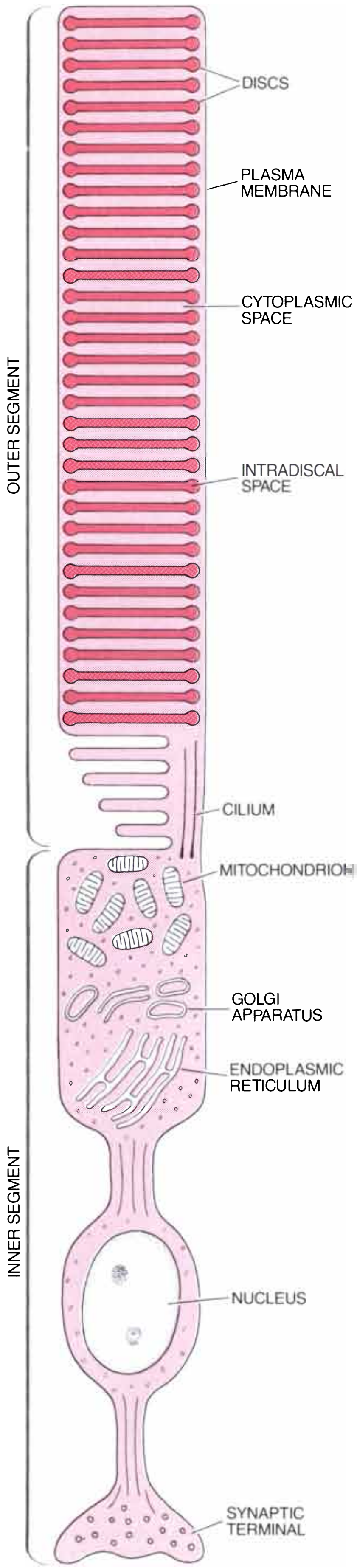
What happens when 11-*cis* retinal absorbs a photon? The general answer is that the molecule is isomerized. Isomers are molecules having the same atoms but different shapes. Indeed, the label 11-*cis* designates a particular isomer of retinal. The backbone of retinal is a string of carbon atoms; 11-*cis* indicates that the hydrogen atoms attached to carbon atoms 11 and 12 lie on the same side of the chain. This configuration forces the chain to bend between carbons 11 and 12. In another isomer, called all-*trans* retinal, the hydrogens attached to carbons 11 and 12 lie opposite each other and the carbon backbone is straight.

Now, as long ago as 1957 George Wald and Ruth Hubbard of Harvard University identified the initial molecular event in vision by showing that when 11-*cis* retinal absorbs a photon, it is converted into the all-*trans* form. The energy of light straightens out the bend in the chain of carbon atoms. In this motion rhodopsin is quite responsive: absorption of a photon leads to isomerization about half of the time. In contrast, spontaneous isomerization in the dark takes place roughly once in 1,000 years. The contrast has valuable consequences for vision. When a photon strikes the retina, the rhodopsin that is struck reports the event with high efficiency, while the millions of other rhodopsin molecules in the cell remain silent.

More than a decade after the work of Wald and Hubbard several advances revealed something about what happens at the termination of the excitatory cascade in the outer membrane. The plasma membrane is selectively permeable to ions, which carry a net electric charge. As a result there is a difference in electric potential between the inside of the rod cell and the outside. In the resting state the inside of the cell is about 40 millivolts (mV) negative with respect to the outside. In 1970 some elegant electrophysiology by Tsuneo Tomita of Keio University as well as by William A. Hagins and Shuko Yoshikami of the National Institutes of Health showed that following illumination the potential difference increases. The increase varies with the strength of the stimulus and the background illumination; the maximum potential difference is −80 mV.



CYCLIC GMP (3′,5′ cyclic guanosine monophosphate) is the transmitter in the rod cells of the retina that is directly responsible for generating a nerve impulse. Cyclic GMP includes atoms of nitrogen (*dark blue*), carbon (*light blue*), hydrogen (*white*), oxygen (*red*) and phosphorus (*yellow*). The phosphorus atom forms part of the ring structure for which the molecule is called cyclic. When the ring is intact, cyclic GMP holds open channels for sodium ions in the outer membrane of the rod cell. When the ring is enzymatically cleaved, the sodium channels close and the electrical properties of the membrane change, giving rise to a nerve impulse.



ROD CELL is divided into two parts that have specialized functions. The apparatus for detecting light is in the outer segment, which holds a stack of some 2,000 disks derived from the plasma membrane. The inner segment contains organelles for making specialized molecules required in photoreception. When light strikes the disk, molecules there are modified. The signal is sent by a chain of reactions to the plasma membrane. It travels through the plasma membrane to the synaptic terminal, from which it is sent to other retinal cells.

The increase in potential difference, which is known as a hyperpolarization, is due to a decrease in the permeability of the membrane to sodium ions (which carry a positive charge). After the general nature of the hyperpolarization had been worked out, my colleague Denis A. Baylor showed that the absorption of a single photon blocks the influx of millions of sodium ions by closing hundreds of channels for sodium in the plasma membrane [see "How Photoreceptor Cells Respond to Light," by Julie L. Schnapf and Denis A. Baylor; SCIENTIFIC AMERICAN, April]. Once the sodium channels close, the light-induced hyperpolarization is passed along the outer membrane to the synaptic terminal at the other end of the cell, where the nerve impulse arises.

These fundamental results at each end of the biochemical cascade posed a stark question: What happens in between? How does the isomerization of retinal in the disk membrane lead to the closure of sodium channels in the outer membrane? The plasma membrane of the rod cell is physically distinct from the disk membranes. Hence the signal must be carried from the disks to the outer membrane by a transmitter. Since the absorption of a single photon can lead to the closing of hundreds of sodium channels, many transmitters must be formed per photon absorbed.

What is the transmitter that carries the excitation signal? In 1973 Hagins and Yoshikami proposed that calcium ions, sequestered in the disk in the dark, are released on illumination and diffuse to the plasma membrane to close sodium channels. This attractive hypothesis generated much interest and many experiments. Recent work has shown, however, that although calcium ion has a significant role in vision, it is not the excitatory transmitter. Instead the transmitter is a molecule called 3′,5′ cyclic guanosine monophosphate, or cylic GMP.

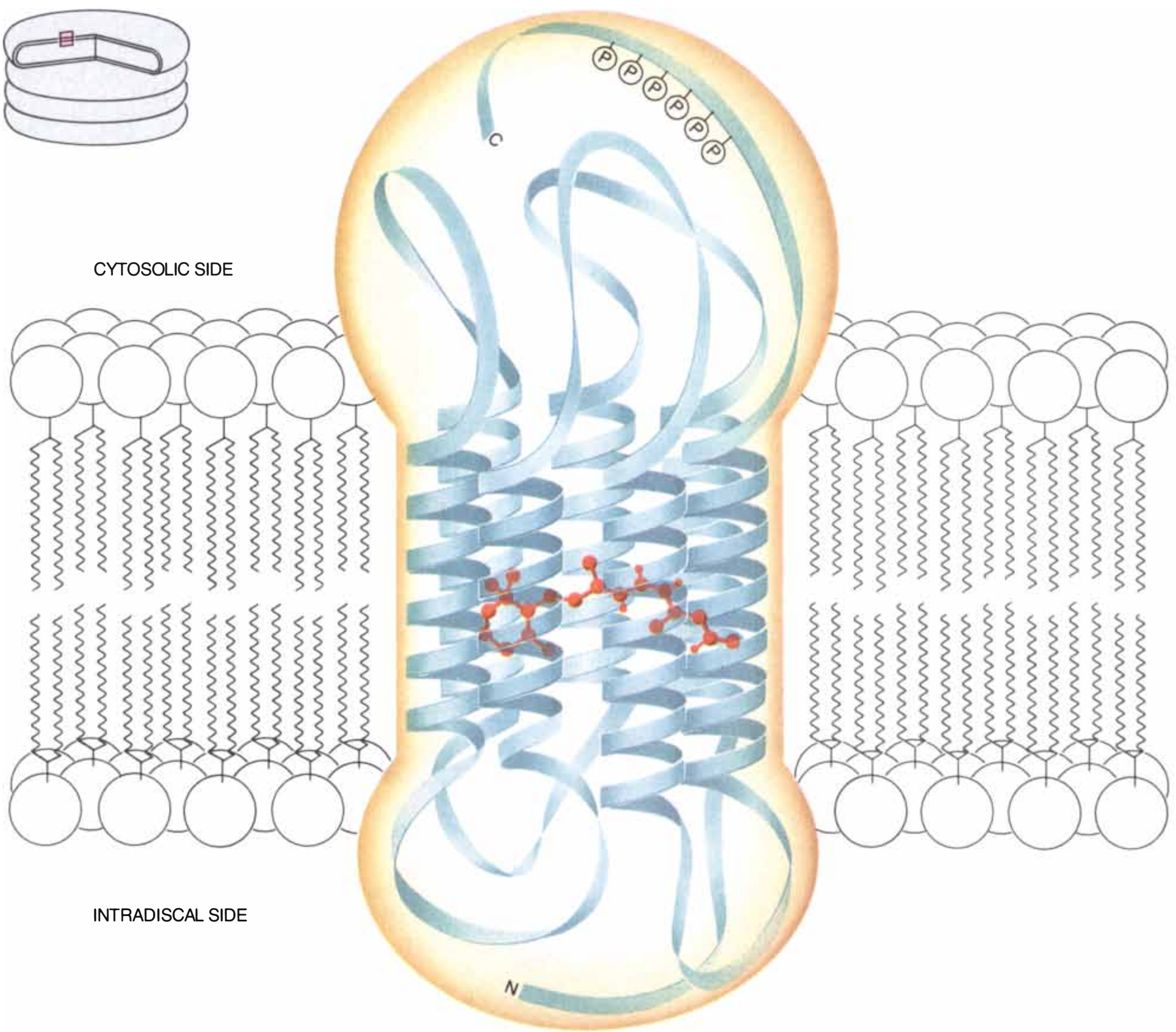
The capacity of cylic GMP to act as a transmitter is closely related to its chemical structure. GMP is a nucleotide of the type that forms the subunits of RNA. Like other nucleotides, it has two components: a base and a five-carbon sugar unit. In the case of GMP the base is guanine; nucleotides containing guanine are known as guanyl nucleotides. The word cyclic indicates that the carbons designated 3′ and 5′ in the sugar molecule are joined by a phosphate group. The link that joins the two carbons—known as a phosphodiester bond—forms a ring. When the ring is intact, cyclic GMP is capable of keeping the membrane sodium channels open. When it is cleaved by an enzyme called a phosphodiesterase, the sodium channels close spontaneously.

Several steps intervene between the excitation of rhodopsin and the enzymatic cleavage of cyclic GMP. When retinal absorbs a photon and opsin is activated, rhodopsin in turn activates an enzyme called transducin. Transducin, whose action was first elucidated in my laboratory, is a key intermediate in the excitatory cascade. Transducin in turn activates a specific phosphodiesterase. The phosphodiesterase then opens the ring of cyclic GMP by inserting a water molecule into it (a process known as hydrolysis). Although this pathway is not difficult to describe in outline, unraveling it and understanding its physiological significance required a wide variety of experiments in many laboratories.

In 1971 Mark W. Bitensky and William H. Miller of the Yale University School of Medicine found that light markedly reduces the level of a cyclic nucleotide in rod outer segments. Subsequent studies showed that the reduction was due to the hydrolysis of cyclic GMP by a phosphodiesterase specific to that nucleotide. At that time the calcium hypothesis was still quite strong, however, and it was by no means clear that cyclic GMP had much direct influence on the excitatory response. Then in the late 1970's two seminal findings were made; those were the ones that initially aroused my interest in cyclic GMP as a transmitter candidate.

At a research conference in the summer of 1978 Paul A. Liebman of the University of Pennsylvania reported his finding that a single photon could trigger the activation of hundreds of phosphodiesterase molecules per second in preparations of rod outer segments. Earlier work, carried out in the presence of the nucleotide adenosine triphosphate (ATP), had shown much less amplification. Liebman observed that considerably more amplification could be obtained with guanosine triphosphate (GTP).

Guanosine triphosphate is a nucleo-



RHODOPSIN MOLECULE, embedded in the membrane of the disk, receives light and initiates the excitatory cascade. Like the plasma membrane from which it is derived, the disk membrane is a bilayer of lipid (fatty) molecules. Rhodopsin has two components: 11-*cis* retinal and opsin. Opsin is a protein that has the form of seven helical structures threaded through the membrane, connected by short linear segments. 11-*cis* retinal (*red*) lies near the center of the membrane, attached to one helix. The absorption of a photon (a quantum of light) by retinal alters retinal's shape and activates rhodopsin. This model for the structure of rhodopsin was proposed by Edward Dratz of the University of California at Santa Cruz and Paul Hargrave of Southern Illinois University.

tide closely related to the noncyclic form of GMP. Instead of having a single phosphate group attached to its 5′ carbon, however, it has a chain of three phosphates bound to each other by phosphodiester linkages. The energy stored in those bonds provides the basis for many cellular functions. For example, the splitting off of one phosphate group converts GTP into guanosine diphosphate (GDP) and liberates considerable energy. In this way the cell obtains energy needed to drive chemical reactions that are otherwise energetically unfavorable. Liebman's key finding was that this process seemed to be at work in the activation of the phosphodiesterase, where GTP is an essential cofactor.

On the way home from the research conference where Liebman had presented his report, I visited Yale and had the good fortune to stay overnight at Miller's home. After an excellent dinner he showed me some intriguing experimental records. Miller and his co-worker Grant Nicol had succeeded in injecting cyclic GMP into the outer segment of intact rod cells. The results were striking. Cyclic GMP quickly reduced the potential difference across the plasma membrane. Not only that, the injected nucleotide sharply increased the delay between the arrival of a light pulse and the hyperpolarization of the membrane. The simplest interpretation was that cyclic GMP opened sodium channels, which then remained open until cyclic GMP was degraded by the light-activated phosphodiesterase.

That hypothesis was an alluring one, but direct proof was lacking. On returning to my laboratory I discussed the new findings with Bernard K.-K. Fung, a postdoctoral fellow. Fung shared my enthusiasm for cyclic GMP as a transmitter. Together we decided to explore the molecular mechanism by which the phosphodiesterase that cleaves cylic GMP is activated. Liebman's discovery that GTP was necessary for activation was significant in our thinking, because it suggested that a protein capable of binding GTP might be a significant intermediate in activation. We began to look carefully at what happens to GTP in rod cells.

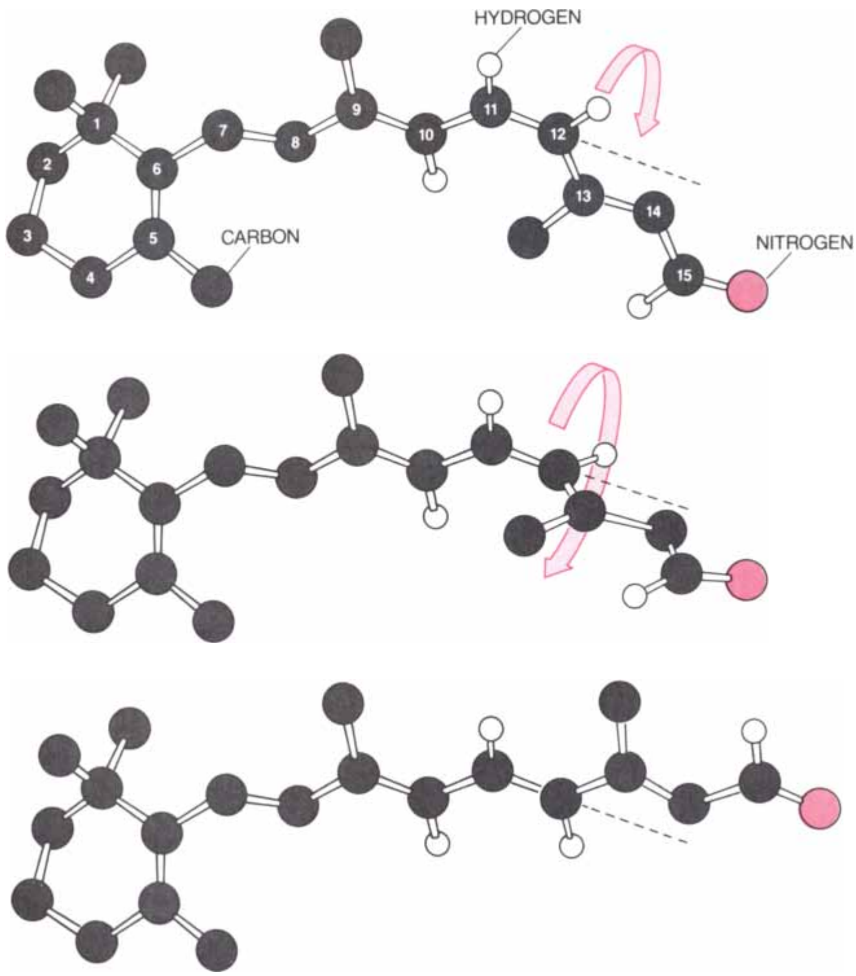
Our first experiment was designed to detect the binding of GTP and its chemical derivatives to rod outer segments. Radioactively labeled GTP was incubated with rod outer-segment fragments. After several hours the preparation was washed over a filter that retained membrane fragments or molecules as large as proteins but allowed small molecules such as GTP (and its relatives) to pass. We found that a substantial amount of the radioactivity was bound to the membrane. Further work showed, however, that the molecule bound to the membrane was not GTP but GDP.

Those results strongly suggested that there was a protein in the rod-cell membranes that is capable of binding GTP and then converting it into GDP by cleaving off one of its phosphate groups. It seemed increasingly likely that such a protein was a key intermediate and also that the conversion of GTP into GDP might be driving the activation process. While this work was in progress Walter Godchaux III and William F. Zimmerman of Amherst College reported finding just such a protein in rod-cell membranes. It was not clear, however, how their finding fitted with the other pieces of the puzzle that were then emerging.

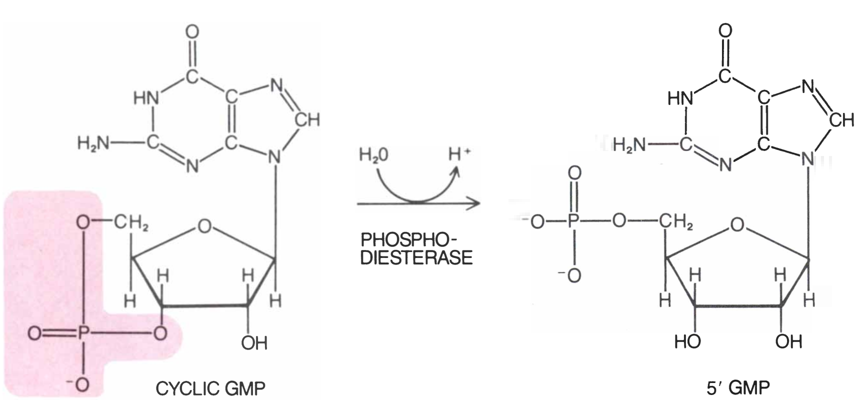
One of the striking aspects of the activity Fung and I observed was that not only was there something in the membrane capable of binding guanyl nucleotides but also when the membranes were illuminated, GDP was released. Moreover, the release of GDP from the membrane was strongly enhanced by the presence of GTP in the surrounding solution. Now a hypothesis began to emerge to simplify this welter of data. It seemed one part of the activation process involved an exchange of GTP for GDP in the membrane. That was why the release of GDP was so markedly enhanced by GTP: the release of a GDP molecule depended on the substitution of a GTP for it. Later perhaps GTP might be converted into GDP.

More and more, the exchange of GTP for GDP seemed to be near the heart of the activation process. For one thing, that exchange is highly amplified. We measured the effect of light on the uptake of GTP by the membrane and found that the photoexcitation of one rhodopsin molecule led to the binding of about 500 molecules of a GTP analogue (chosen because it resists being reduced to GDP and therefore enabled us to isolate the exchange step). The discovery of this amplification was exciting, because it pointed toward an explanation of the overall amplification characterizing the excitatory cascade.

This central result led us to propose that there is a protein intermediate in the excitation cascade that can exist in two states. In one state the protein binds GDP; in the other it binds GTP. The substitution of GTP for GDP is the signal for the protein's activation. That exchange is triggered by rhodopsin; in turn it serves to activate a specific phosphodiesterase. The phosphodiesterase then closes the sodium



RETINAL CHANGES ITS SHAPE after absorbing a photon. In 11-*cis* retinal, the form taken in the dark (*top*), the hydrogen atoms attached to carbons 11 and 12 lie on the same side of the carbon backbone, which forces the backbone to bend. Absorption of a photon causes the chain to rotate between carbons 11 and 12. That rotation straightens out the end of the chain, yielding a different chemical form: all-*trans* retinal (*bottom*).



RING OF CYLIC GMP is opened by an enzyme called a phosphodiesterase. The ring (*color*) includes a phosphorus atom in a link called a phosphodiester bond. The enzyme inserts a water molecule into the bond, cleaving it to yield a molecule called 5′ GMP.

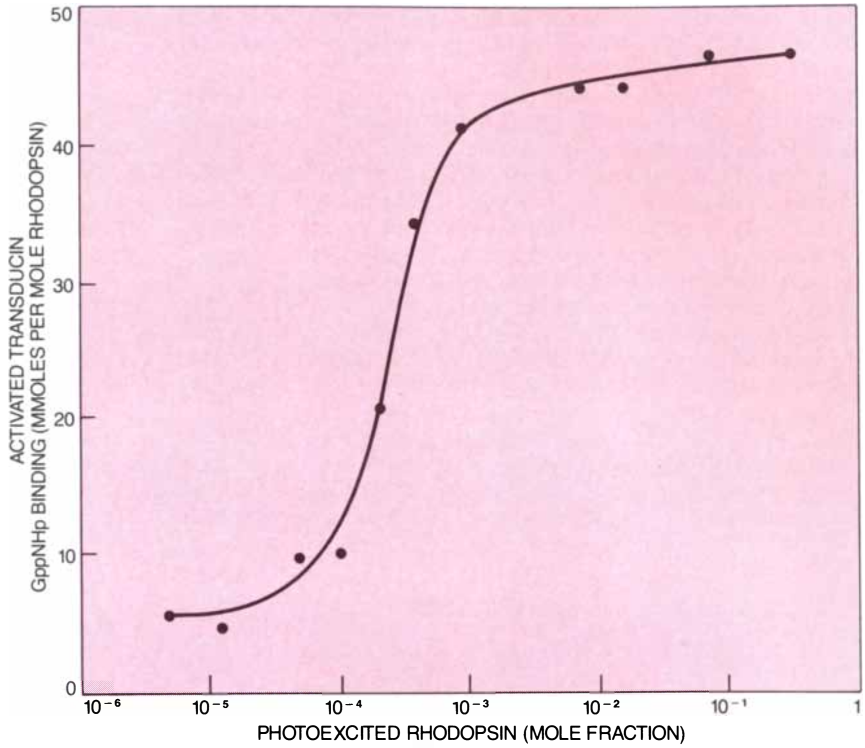
channels in the plasma membrane by cleaving cyclic GMP. Building on the work of Hermann Kühn of the Institute for Neurobiology at the University of Jülich, we soon isolated the postulated protein. It was given the name transducin in recognition of the fact that it mediates the conversion of light into an electrical impulse, a process known as a transduction. Subsequent work showed that transducin consists of three protein subunits designated alpha, beta and gamma.

Having the purified transducin in hand enabled us to test our hypothesis about the overall flow of information in the activation cascade. We had suggested that the signal moved from activated rhodopsin to transducin (in its GTP form) to the phosphodiesterase. If that picture was correct, two things followed. First, it could be inferred that transducin could be converted into its GTP form in the absence of the phosphodiesterase. It also followed that the phosphodiesterase could be activated in the absence of photoexcited rhodopsin.

My colleagues and I set out to test both propositions. To test the first we assembled a synthetic-membrane system from which the phosphodiesterase was absent. Purified transducin in the GDP form was introduced into the artificial membranes. When rhodopsin was activated by light, we found that each molecule of rhodopsin catalyzed the uptake of 71 molecules of a GTP analogue to the membrane. Thus it appeared that each molecule of rhodopsin was activating transducin by catalyzing the exchange of GTP for GDP in many molecules of transducin. This was the first time an amplified effect of rhodopsin requiring only one other protein—transducin—had been found.

We then wanted to test the second proposition: that transducin in the GTP form can activate the phosphodiesterase in the absence of any photoexcited rhodopsin. The first step was to purify the active form of the enzyme: the transducin-GTP complex. When we did so, we got a surprise. We knew that in the inactive, GDP form transducin was complete; all three subunits were joined. In the active, GTP form, however, transducin came apart. The alpha subunit floated free of the joined beta and gamma units. GTP was bound to the alpha subunit.

We were now in a position to ask an even more precise question than the one we had intended to ask. We could now inquire whether it is the alpha subunit (with its attached GTP) or the beta-gamma subunit that stimulates the action of the phosphodiesterase. We obtained a decisive answer. The alpha subunit (with GTP) activates the phosphodiesterase; the joined beta-gamma unit has no effect. Moreover, the alpha unit was effective in the absence of rhodopsin, validating the idea that transducin can activate the phosphodiesterase without rhodopsin and confirming our overall picture of the information flow.



ACTIVATION OF TRANSDUCIN is accompanied by the binding of guanosine triphosphate (GTP). The horizontal axis shows the concentration of photoexcited rhodopsin (increasing to the right). The vertical axis shows the binding of GppNHp (a GTP analogue) in fragments of artificial membrane containing only rhodopsin and transducin. The slope of the curve indicates that each rhodopsin molecule activates 71 molecules of transducin in this system. (In intact rod cells 500 molecules of transducin are activated by each rhodopsin.) Activated transducin triggers the phosphodiesterase that cleaves cyclic GMP.

What is the mechanism by which transducin activates the specific phosphodiesterase? Bitensky's group found that in the dark the phosphodiesterase has a low activity because it is subject to an inhibitory constraint. When a small amount of trypsin (a protein-digesting substance) was added, the constraint was removed, activating the enzyme. It was known that the phosphodiesterase consists of three polypeptide chains; as in the case of transducin, they are designated alpha, beta and gamma. James B. Hurley, a postdoctoral fellow in my laboratory, found that trypsin degrades the gamma subunit but not the alpha or beta units. Taken together, Bitensky's and Hurley's results made it seem likely that the inhibitor of the phosphodiesterase is the gamma subunit.

That notion was confirmed by results from my laboratory and others. My co-workers and I purified the gamma subunit, added it to the active alpha-beta complex and found that gamma eliminated more than 99 percent of the catalytic activity. Further confirmation came from the fact that the rate of destruction of the gamma unit by trypsin closely resembled the rate of activation of the phosphodiesterase in the excitatory cascade. Finally, Marc Chabre and his colleagues at the Center for Nuclear Studies at Grenoble found that transducin in the GTP form can bind to the gamma subunit of the phosphodiesterase and form a complex with it.

The picture that emerges from these results is that after illumination the alpha subunit of transducin with its attached GTP binds to the phosphodiesterase and carries away the inhibitory gamma subunit. The departure of gamma unleashes the catalytic activity of the phosphodiesterase. That activity is powerful: each activated enzyme molecule can hydrolyze 4,200 molecules of cyclic GMP per second.

With these insights much of the activation cascade became clear. The first step is the activation of transducin by photoexcited rhodopsin. In that interaction rhodopsin causes the attached

GDP to break loose from transducin. A molecule of GTP takes its place, and the alpha subunit dissociates from the rest of the protein, taking the GTP with it. This process takes only about a millisecond, as Chabre showed with T. Minh Vuong, a graduate student, during a sabbatical spent in my laboratory. The production of hundreds of active alpha-transducin-GTP complexes by a single rhodopsin is the first stage of amplification in vision.

Alpha-transducin with its GTP then triggers the activity of the phosphodiesterase. At this stage there is no amplification: each alpha-transducin unit binds to and activates a single phosphodiesterase. The transducin-phosphodiesterase pair act as a single unit to provide the second amplified stage. The transducin remains associated with the phosphodiesterase as it does its work of cleaving cyclic GMP. As I described, each activated enzyme molecule can cleave several thousand cyclic GMP's. This amplification, along with the activation of many transducins by each rhodopsin, accounts for the remarkable magnification of a single photon into a palpable nerve impulse.

Yet if the organism is to be able to see more than once, this cycle must also be turned off. Transducin has a key role in deactivation as well as in activation. The alpha subunit has a built-in chemical timer that terminates the activated state by converting the bound GTP into GDP. The mechanism of the timer is not fully understood. It is known, however, that the hydrolysis of GTP into GDP in the deactivation phase has an important role in driving the entire cycle. The activation reactions are energetically favorable. Some of the deactivation reactions, on the other hand, are not; without the conversion of GTP into GDP the system could not be reset for future firing.

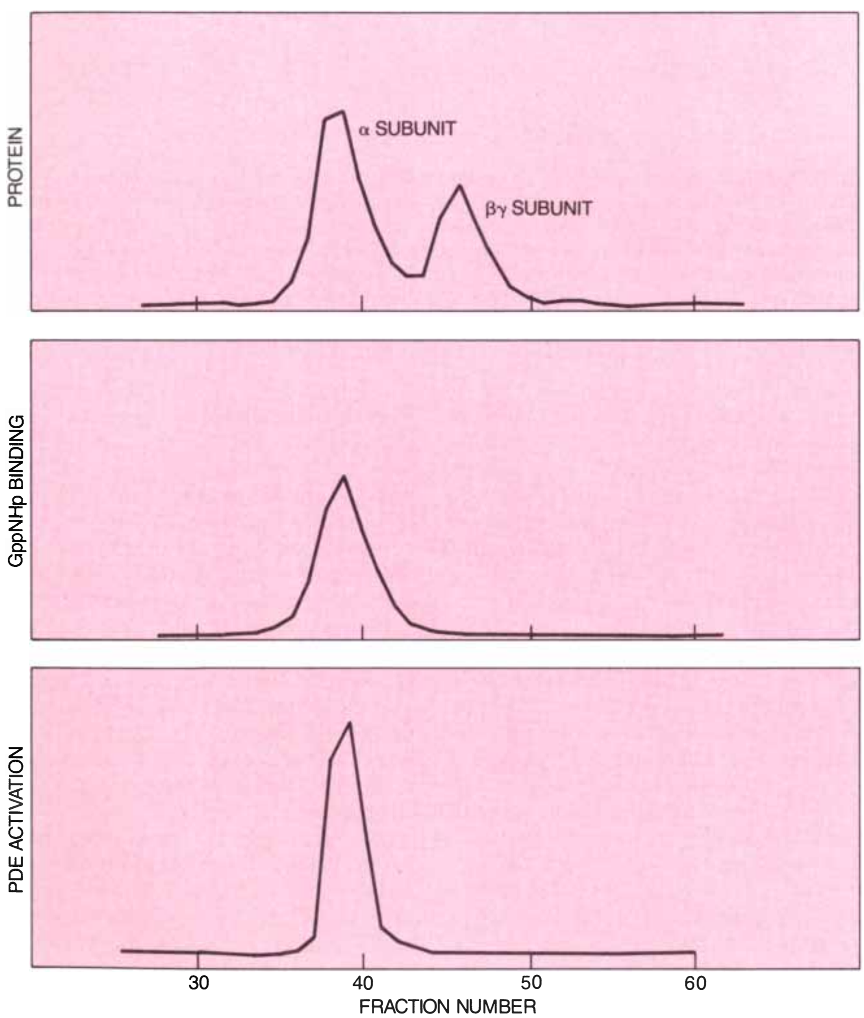
As GTP is cleaved to form GDP, the alpha unit of transducin releases the inhibitory gamma unit of the phosphodiesterase. The gamma unit returns to the phosphodiesterase, binds to it and restores it to the quiescent state. Transducin is then restored to its preactivation form by the rejoining of the alpha subunit and the beta-gamma unit. Rhodopsin is deactivated by an enzyme that recognizes its specific structure. That enzyme, called a kinase, attaches multiple phosphate groups to amino acids at one end of the opsin chain. As Kühn has shown, rhodopsin then forms a complex with a protein called arrestin, which blocks the binding of transducin and puts the system back in the dark state.

Much of the working out of the visual cascade that was done in the late 1970's and early 1980's proceeded on the assumption that cyclic GMP somehow opens the sodium channels in the outer membrane and that hydrolysis of cyclic GMP leads to their being closed. Little was known, however, about the mechanism by which this might take place. Was cyclic GMP acting directly on the channels, or did it act through intermediaries? A decisive answer was obtained in 1985 by Evgeniy E. Fesenko and his co-workers at the Institute of Biological Physics in Moscow.

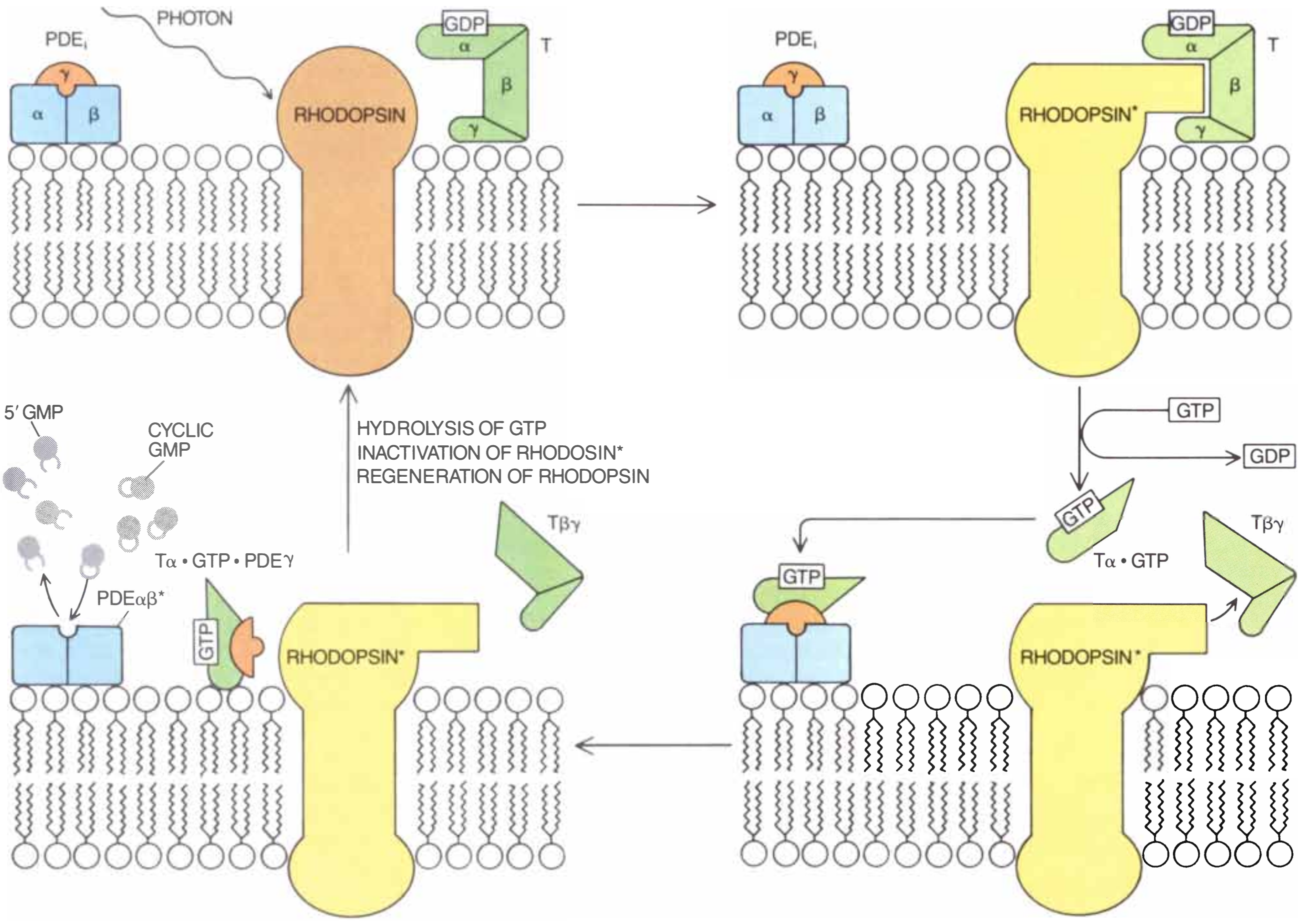
Fesenko employed a micropipette to pull off a small patch of the rod cell's plasma membrane. The patch adhered tightly to the end of the pipette, with the side that would normally be inside the cell facing out. This face of the membrane was then exposed to various solutions in order to test their effect on the sodium conductance. The results were unambiguous: the channels were directly opened by cyclic GMP but were not affected by other substances, including calcium ion.

Fesenko's incisive experiments gave the final blow to the notion that calcium ion might be the excitatory transmitter and established the last link in the excitatory cascade. The outline of the excitation pathway had been delineated; as we had hypothesized, the overall information flow is from rhodopsin to transducin to the phosphodiesterase and then to cyclic GMP.

Although working out the pathway



ALPHA SUBUNIT OF TRANSDUCIN is the information carrier in the activation of the phosphodiesterase. Transducin has three subunits, which are designated alpha, beta and gamma. The three graphs show the result of passing a solution containing transducin through a filtration column. The horizontal axis indicates the array of equal portions of fluid that were collected at the bottom of the column. The top panel shows that the protein in the solution has two components of differing size. The small component is the alpha subunit; the large one is the joined beta-gamma unit. The middle panel indicates that the GppNHp-binding activity is associated with the alpha subunit. The bottom panel indicates that phosphodiesterase activation is also associated with the alpha subunit.



EXCITATORY CASCADE forms part of a cycle mediated by transducin. The cycle begins with the absorption of a photon by rhodopsin (*upper left*). Activated rhodopsin (*rhodopsin**) then interacts with transducin (*T*), as shown at the upper right. That interaction causes GTP to replace GDP in the alpha subunit of transducin and causes the alpha unit to split off from the beta-gamma part of the enzyme (*lower right*). Transducin relieves the inhibition of the inactive phosphodiesterase (*PDE$_i$*), perhaps by carrying away its gamma unit. The activated phosphodiesterase begins cleaving many molecules of cyclic GMP (*lower left*). Before long a built-in timer in the alpha subunit of transducin cleaves GTP to GDP. The alpha subunit rejoins the beta-gamma unit; the phosphodiesterase is also reassembled. At the same time rhodopsin is inactivated and then regenerated in its preactivation form.

and mechanism of the excitatory cascade was very gratifying, several important questions have not yet been answered. One of them concerns how the response of the cascade is modulated. As I mentioned above, rod cells are much less sensitive in bright light than they are in the dark. The background lighting must somehow affect the degree of "gain" in the system: the total amount of amplification provided by the two amplified steps (from rhodopsin to transducin and from the phosphodiesterase to cyclic GMP). Much evidence suggests that calcium ion has a role in the process, but the details are not known.

Other questions involve the structure of the sodium channels and the mechanism that prevents the ultimate depletion of cyclic GMP in the cell. Liebman's group and that of Benjamin Kaupp of the Neurobiology Institute at the University of Osnabrück in West Germany have made important contributions to the first question by purifying cyclic-GMP-gated channels and reconstituting their function in model membranes. A key molecule in the answer to the second question must be guanylate cyclase, the enzyme that synthesizes cyclic GMP. Clearly there must be a feedback loop that restores cyclic GMP to its preillumination level; if there were not, the cell could fire only a few times before permanently exhausting its own capacity. The nature of the loop, however, is not known.

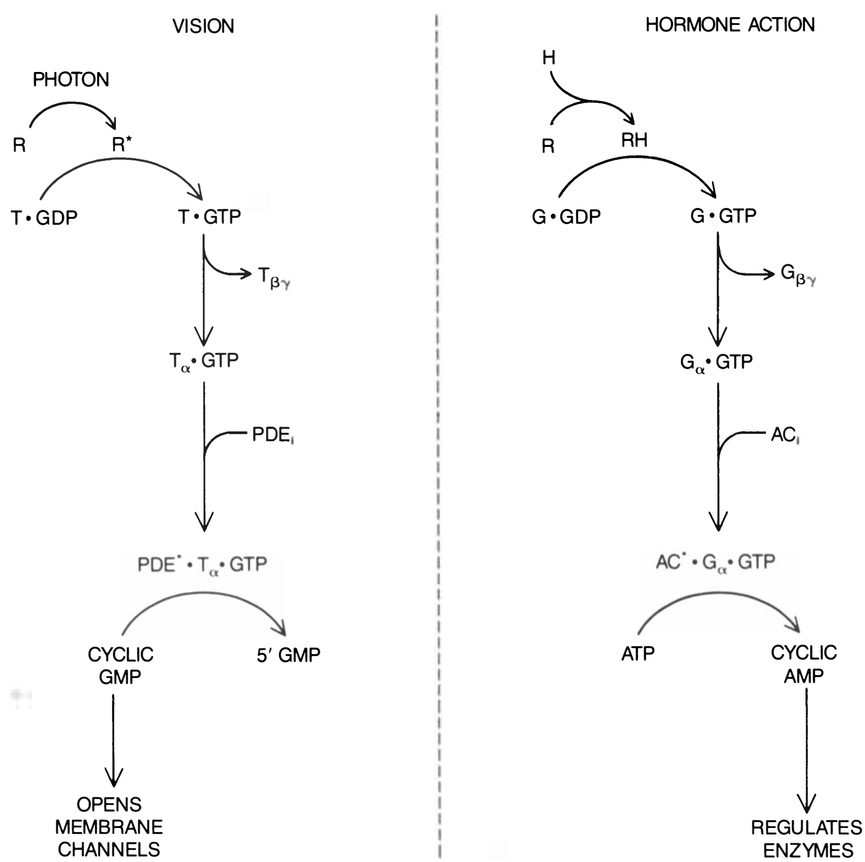
In addition to clearing up the remaining questions about the visual cascade in rod cells, current work is extending the recent findings beyond those cells. Cone cells have a transduction system resembling that of rods, as has been shown by Baylor as well as by King-Wai Yau of Johns Hopkins University. It has long been known that the cones contain three visual pigments (analogous to rhodopsin) that respond to red, green and blue light. All three pigments contain 11-*cis* retinal. Furthermore, molecular-genetic studies recently carried out by my colleagues Jeremy Nathans and David S. Hogness have shown that the three cone pigments have the same fundamental architecture as rhodopsin. Transducin, the phosphodiesterase and the cyclic-GMP-controlled channel in the cones are like their counterparts in the rods. Hence it may not be long before the transduction cycle there is understood in the same molecular detail as it is in the rod cells.

The cyclic-GMP cascade has a significance that extends even beyond vision. The excitatory cascade in the rod cells has a notable resemblance to the pathway by which certain hormones yield their effects. For example, the hormone epinephrine (adrenaline) acts by triggering the activation of an enzyme known as adenylate cyclase. Adenylate cyclase catalyzes the formation of cyclic adenosine monophosphate (cyclic AMP). Cyclic AMP is an intracellular messenger that mediates the action of many hormones [see "The Molecular Basis of Communication within the Cell," by Michael J. Berridge; SCIENTIFIC AMERICAN, October, 1985].

Alfred G. Gilman of the University of Texas Health Science Center at Dallas has worked out many of the facets of the regulation of adenylate cyclase, which turns out to have striking parallels to the excitatory cascade in the rod cell. Just as the excitatory cascade begins when rhodopsin absorbs a photon, the cascade of hormone action is initiated when a receptor on the cell's surface binds a specific hormone. The receptor-hormone complex then interacts with a *G* protein that resembles transducin. The same exchange of bound molecules that activates transducin—GTP for GDP—activates the *G* protein when the *G* protein interacts with the receptor-hormone unit.

The parallelism does not end there. Like transducin, the stimulatory *G* protein has three subunits. It is the alpha subunit that activates adenylate cyclase by relieving an inhibitory constraint. Again as in the case of transducin, the *G* protein's stimulatory action is turned off by a built-in timer that converts GTP into GDP.

The similarities between transducin and the *G* proteins (several have been identified) apply to their structure as well as to their activity. Transducin and the *G* proteins are members of the same overall family of signal-coupling proteins. All members of that family identified so far have nearly the same beta subunit. In addition, their alpha subunits have the same function, a similarity that has now been shown at the molecular level. The stretches of DNA encoding the alpha subunits of transducin and three *G* proteins have recently been sequenced in several laboratories. The information from the DNA indicates that about half of the amino acid sequences in the four proteins are identical or nearly so.



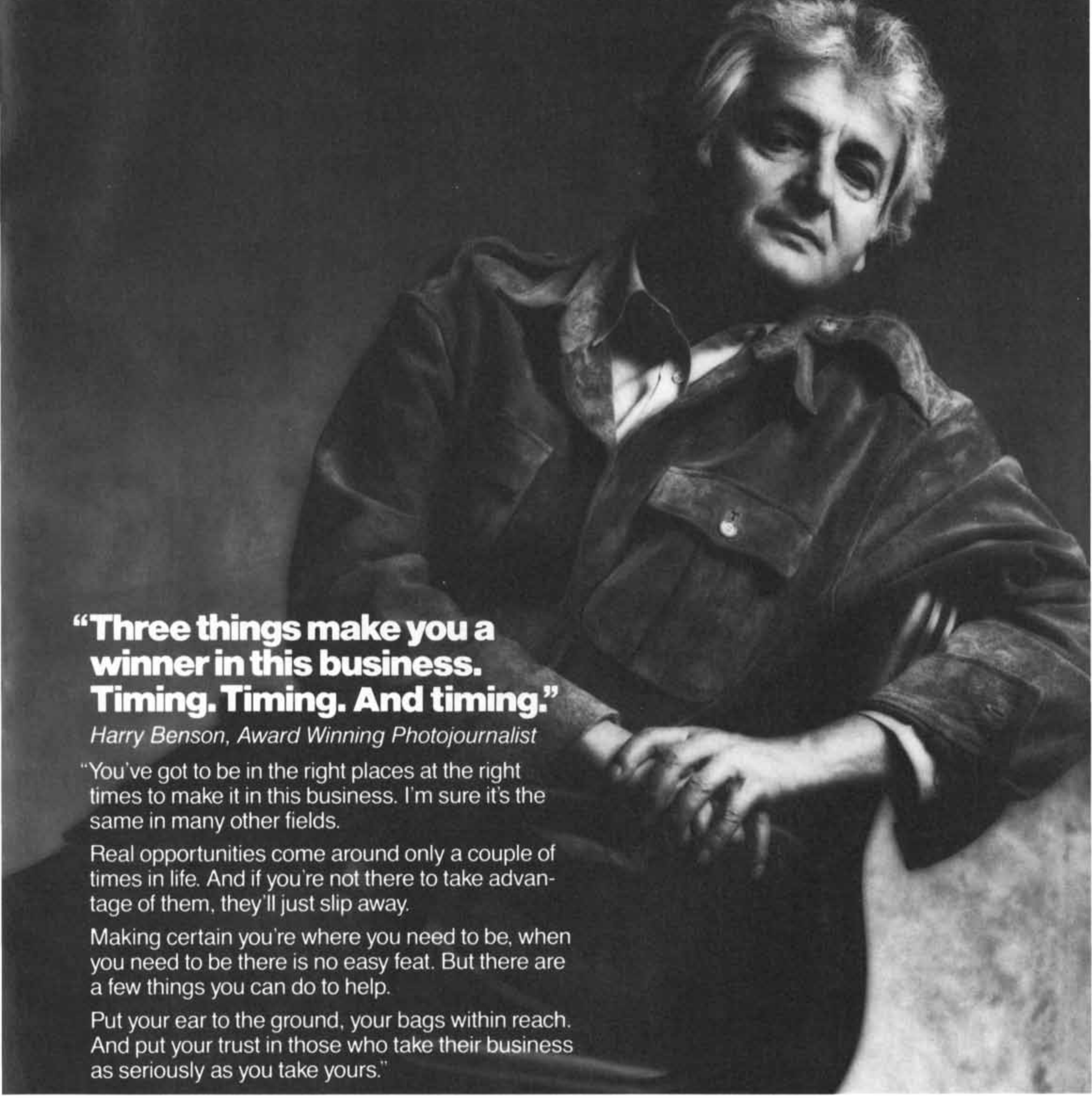
HORMONE ACTION AND VISION show striking parallels. The sequence at the left shows the excitatory cascade of vision; the one at the right shows the events by which a variety of hormones have their effects. The binding of a hormone (*H*) and its receptor (*R*) to form a hormone-receptor complex resembles the excitation of rhodopsin. As rhodopsin activates transducin, the hormone-receptor complex activates a stimulatory *G* protein. Like the activation of transducin, that of the *G* protein entails the binding of GTP by the alpha subunit. The *G* protein activates adenylate cyclase (*AC*), which is analogous to the phosphodiesterase. Adenylate cyclase catalyzes the conversion of adenosine triphosphate (*ATP*) to cyclic adenosine monophosphate (*cyclic AMP*), which regulates many enzymes.

When the genetic information is examined on an overall level, it is found that the proteins include some regions that have been conserved quite stringently through evolution as well as some that have diverged widely. Each protein contains three binding sites: one for guanyl nucleotides, one for the activated receptor (rhodopsin or a hormone-receptor complex) and one for the effector protein (phosphodiesterase or adenylate cyclase). As might well be expected—given its crucial function in the activation cascade—the binding site for GTP or GDP is the most highly conserved among the various proteins.

Indeed, the GTP-binding regions of these proteins resemble a region of a functionally very different protein called elongation factor *Tu*. *Tu* has a crucial role in the synthesis of proteins in certain bacteria; it delivers transfer RNA's (the "hooks" that carry specific amino acids and allow them to be added to a growing amino acid chain) to the ribosome, which is the organelle where proteins are synthesized. In its work *Tu* undergoes a cycle akin to the transducin cycle in visual activation. The cycle is powered by the cleaving of a molecule of GTP, and there is a site on *Tu* where the GTP binds. That site is quite close in amino acid sequence to the guanyl-nucleotide binding sites of transducin and the various *G* proteins.

Now protein synthesis is one of the most fundamental metabolic activities of any cell. Therefore it seems likely that elongation factor *Tu*, which takes part in that work, originated earlier in evolution than the *G* proteins or their relative, transducin. Indeed, *Tu* could be the ancestor of both transducin and the *G* proteins. The controlled uptake and release of proteins coupled to GTP-GDP exchange (and subsequent cleavage) was undoubtedly perfected early in evolution; elongation factor *Tu* may represent an early version of the cycle.

One of the many fascinating things about evolution is that mechanisms evolved for a particular function may later be modified and applied to different functions. That, I think, is what happened to the mechanism of *Tu*. After evolving to mediate protein synthesis, it was retained for billions of years and ultimately put to work in the transduction of hormonal and sensory stimuli. In the past few years one of those functions—the transducin cycle—has been worked out in great detail. The result has been extremely satisfying. For the first time we now understand at the molecular level one of nature's most precise sensory events: visual excitation.



No one is more serious about your international travel needs than Lufthansa. Our outstanding on-time arrival record will tell you that. So will our meticulously maintained fleet and our friendly, courteous service. That's why in 161 cities around the world, when people need to be in the right place at the right time, they fly 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.

The Rings of Uranus

Why are they so narrow and dark? Findings from the Voyager 2 *encounter suggest that the austere ring system may be only a fleeting stage in a continuing saga of creation and destruction*

by Jeffrey N. Cuzzi and Larry W. Esposito

In 1977 an unexpected experimental result led to the surprising but inescapable conclusion that the planet Uranus is surrounded by a family of narrow, dark rings unlike any seen before. Subsequent accumulation of more information about the rings raised several fundamental questions. Why are the Uranian rings so different from the broad, bright rings of Saturn? Why are some of them noncircular? Why do they have varying widths, sharp edges and vast reaches of empty space between them?

The *Voyager 2* encounter with Uranus in January, 1986, was designed to test dynamical models of ring structure by determining precisely the locations of the known rings, searching for new rings and observing the size and detailed distribution of particles within the individual rings. Perhaps most important, the spacecraft was also to search for small ring moons orbiting between the rings that are thought to shepherd the ring particles, or keep them in orbit, by exerting gravitational forces on them.

The cameras on *Voyager 2* discovered one new narrow ring and about 100 almost transparent bands that are invisible from the earth. Several partial ring arcs were also seen. The particles in the known rings are larger than had been expected, ranging between a softball and an automobile in size (between 10 centimeters and a few meters). The particles in the newly discovered dust bands, in contrast, are much smaller (roughly .02 millimeter) and their locations are not obviously related to the previously known rings. Ten small moons were discovered that lie within the orbit of Miranda, which is the closest to Uranus of the five main moons [see "The Moons of Uranus," by Torrence V. Johnson, Robert Hamilton Brown and Laurence A. Soderblom; SCIENTIFIC AMERICAN, April]. Two of the new moons appear to shepherd the particles in the outermost ring, but no ring moons of the sizes that had been expected were found anywhere near the other rings. The main rings and new moons are charcoal black.

The information *Voyager 2* transmitted across some three billion miles of space has already verified many aspects of the dynamical models of the Uranian rings. At the same time the new results have highlighted some difficulties with the models. The information has provided new perspectives on the rings of Uranus as well as the closely related ring systems of Saturn, Jupiter and Neptune.

The discovery of the Uranian rings was made during a stellar occultation, in which the brightness of a star is monitored as a planet passes between it and the earth. The main advantage of the technique, which was developed to study the atmosphere of planets, is its high resolution, because of the small apparent size of stars. Features as small as one kilometer to five kilometers can be distinguished from the earth during stellar occultations; in comparison, the best telescopic views of the outer planets have relatively poor resolutions of thousands or tens of thousands of kilometers.

On March 10, 1977, several groups of investigators observed the occultation of the star SAO 158687 by Uranus. To calibrate their instruments they started their observations well ahead of the predicted occultation and continued for as long as possible afterward. They were surprised to find that the starlight was sharply blocked during several short intervals both before and after Uranus passed in front of the star. At first the brief occultations were attributed to a belt of asteroids around the planet. The notion of rings was initially rejected because the occultations were so brief that any rings would have had to be much narrower than Saturn's. As the data from the independent observers were compared during the next few days, however, it became clear that each occultation on one side of Uranus was matched by another occultation on the other side. The observations could be explained only by the conclusion that the planet is surrounded by a family of threadlike rings.

During the next few years observations of more than 200 stellar occultations by Uranus revealed nine narrow rings, all of which lie within one planetary radius of the top of the Uranian atmosphere. Following the differing notation of the discoverers the rings are called (in order of increasing distance from Uranus) 6, 5, 4, alpha, beta, eta, gamma, delta and epsilon. Most of the more recent occultations were observed by James L. Elliot and Richard G. French of the Massachusetts Institute of Technology and independently by Philip D. Nicholson of Cornell University, along with their collaborators. Their work disclosed the properties of the ring orbits with an extraordinary degree of precision. In fact, even the tiny perturbations of Uranus itself by its major moons are detectable in the data.

The rings are not all circular, nor do they all lie in the plane of Uranus' equator. Their widths range from less than two kilometers to nearly 100 kilometers. In comparison, the broad, main rings of Saturn essentially fill a region that is about 60,000 kilometers wide. The epsilon ring of Uranus, which is the outermost one and the largest, ranges in width from 20 to 96 kilometers. This variation in width is not random: the width increases in proportion to the distance of the ring material from Uranus. In other words, at its nearest point to Uranus the epsilon ring is narrowest and almost opaque. At its farthest point it is five times wider and five times more transparent. The alpha and beta rings show similar behavior in that orbits with-

URANIAN RINGS are seen in an image obtained by *Voyager 2* from a distance of about a million kilometers as the spacecraft approached Uranus. The threadlike rings, which are for the most part densely packed with particles, are only a few kilometers wide. They are separated by hundreds of kilometers of virtually empty space. A new ring designated 1986U1R is barely visible between the outermost ring (epsilon) and the next bright ring (delta). The rings reflect only 1 percent of the incident light.

in each ring are nested together. The alignment of the long axes of the ring orbits varies randomly from one ring to the next.

The next three most prominent rings (eta, gamma and delta) have very narrow cores only a few kilometers wide, and eta and delta have broader and more transparent components as well. Although these rings are circular on the average, each of them shows apparently random variations in radius and width. The variations are even larger than the average width of the rings themselves.

The apparently random "wiggles" in the radius and width of the eta, gamma and delta rings are similar to the behavior of some narrow rings in the Saturnian system. Saturn's *F* ring and two ring arcs in a large gap called Encke's division are strands only a few kilometers wide and show highly irregular kinks and wiggles as well as a large fraction of microscopic dust. These properties have been attributed to vigorous ongoing collisions brought about by gravitational forces exerted by nearby moons.

Knowing the widths of the Uranian rings and how little light they reflect, investigators determined even before the *Voyager 2* encounter that the particles in the rings must be very dark; it appears that a flat surface of particle material would reflect less than 5 percent of sunlight falling directly on it. The five principal moons of Uranus, in contrast, have a reflectivity of about 30 percent; the earth's moon and Mars have a reflectivity of approximately 10 percent.

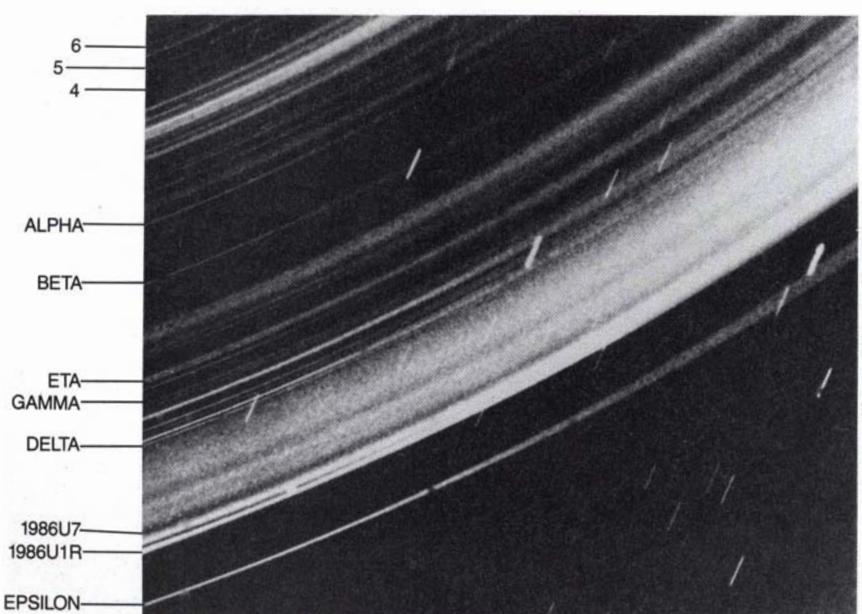
The *Voyager 2* encounter provided the first close view of the rings of Uranus, a view that now allows the Uranian rings to be compared with those of the other giant planets (Jupiter and Saturn) visited by the spacecraft. As *Voyager 2* was drawn ever closer to Uranus—a nearly featureless, gas-enshrouded planet 17 times as massive as the earth—its attention was divided among the Uranian atmosphere, rings and moons. Although the nine main rings were known to be too narrow and dark for the satellite's cameras to study in much detail, noticeable differences among the rings were nonetheless detected [*see illustration on preceding page*].

The quality of the images produced during the mission is particularly remarkable considering the difficulties that were encountered. Making a picture of the rings is comparable to making a picture of charcoal briquettes on a black background at the base of a Christmas tree lighted by a single bulb at the top, employing a camera loaded with film rated ASA 1. Such an image would require an exposure of 15 seconds. Unsmeared images of the rings would have been impossible to obtain without major improvements made painstakingly over several years by the *Voyager 2* spacecraft engineers [see "Engineering *Voyager 2*'s Encounter with Uranus," by Richard P. Laeser, William I. McLaughlin and Donna M. Wolff; SCIENTIFIC AMERICAN, November, 1986].

Additional images were obtained as *Voyager 2* crossed the plane of the rings and peered through them nearly sideways. After the spacecraft had streaked by the planet, and while the sun was hidden behind the planet, a single long-exposure image was made looking at only a small angle away from the sun. The image reveals about 100 very diffuse, nearly transparent bands of microscopic dust surrounding the known rings [*see illustration on this page*].

Like earthbound observatories, *Voyager 2* also made use of stellar occultations. The procedure was similar to the one that led to the discovery of the rings. In this case, however, the resolving power was much greater because the spacecraft passed so close to Uranus. The closeness of the approach overcame a fundamental limitation of all occultation studies: the washing out of shadows cast by the object of interest. Shadows cast by smaller objects or structural features in rings are easily degraded by diffraction and can be seen only at such close range. Finally, as *Voyager 2* passed behind the rings its radio signals to the earth were also occulted, providing another, independent means of making high-resolution observations.

The best *Voyager 2* images revealed quantities of microscopic dust widely spread throughout the rings of Uranus. Such dust is of special interest because tiny particles in and around planetary rings are rapidly removed as a result of erosion by micrometeoroids and radiation-belt electrons. The existence of the dust implies that a local, long-lived source consisting of more massive particles must exist.



DIFFUSE BANDS of microscopic dust were recorded in a single long-exposure image after *Voyager 2* had streaked by Uranus. Nearly 100 bands are evident. The outermost feature is the epsilon ring, which seems to be isolated from the bulk of the dust. Here the ring 1986U1R is the brightest feature, indicating that it has a large component of dust.

Relatively little of the dust (much less than 1 percent by fractional area) is found in any of the main rings, however. The dustiest feature, in fact, is one of the most prominent new rings, designated 1986U1R. The lack of dust in the main rings came as a surprise. Analogies with several of the narrow rings of Saturn had led workers to expect that at least the eta, gamma and delta rings of Uranus would contain noticeable dust. *Voyager 2* revealed, however, that the rarefied upper atmosphere of Uranus is considerably denser than had been thought. Apparently the atmosphere drags microscopic dust particles out of their orbits in only a few hundred years, much as the upper traces of the earth's atmosphere caused the demise of the much larger (but much closer) *Skylab* orbiter.

Comparing the transparency of the rings at the two radio wavelengths

MERKUR

SCORPIO

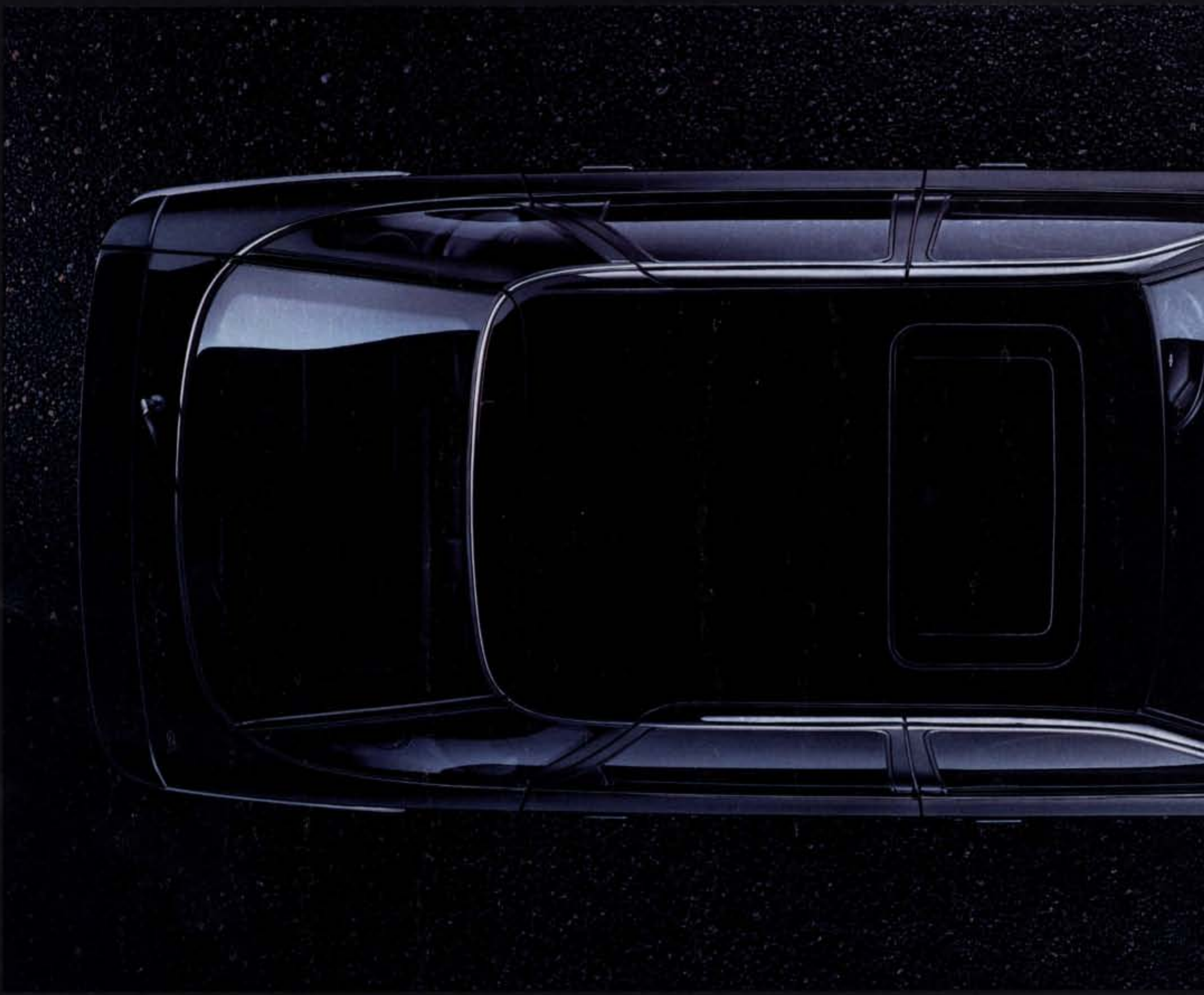
. . . AND ENJOY IT TOO.

Scorpio explodes the myth that owning a well engineered German sedan requires sacrificing personal comfort. Nowhere is this more apparent than in the rear seating area where passengers enjoy more than three feet of legroom, electric reclining seats, and three-point safety harnesses. Up front, Scorpio is equipped with anatomically correct power bucket seats and a 4-way adjustable steering wheel.

THE AMENITIES OF STYLE. Scorpio has all the power conveniences: windows, door locks, seats, steering, brakes, and outside mirrors (heated). Its instrumentation is of precise analog design. No gimmicks. Just accurate information delivered in an effective manner. There's Electronic Automatic Climate Control and of course music. An electronic AM/FM stereo cassette with 12 station memory, Dolby,® a power amplifier and six speakers. An optional Touring Package includes leather upholstery, power moonroof

M E R K U R

SCORPIO



BRING SCORPIO CLOSER TO HOME.

It's a streak in the night called Scorpio, and you can have an insider's view, even before the American debut, by ordering a complimentary Scorpio Preview Collection. The Collection describes Scorpio's advanced components and uses extensive photography to show how Scorpio gives you the handling of a German touring sedan and uncompromising comfort.

It will introduce you to an exceptional automobile through these publications: A 38-page full-color catalog with complete details on this unique touring sedan. *Engineering Notes* to show design criteria that shaped Scorpio. The *Merkur Story* detailing the German heritage of Scorpio.

Also, when Scorpio reaches your Merkur dealership, you'll receive a *Driving Review* that can familiarize you with Scorpio's controls prior to your road test.

SCORPIO WAS THE MOST HONORED EUROPEAN CAR. FIND OUT WHY.

To receive your complimentary Scorpio Preview Collection,

FOLD

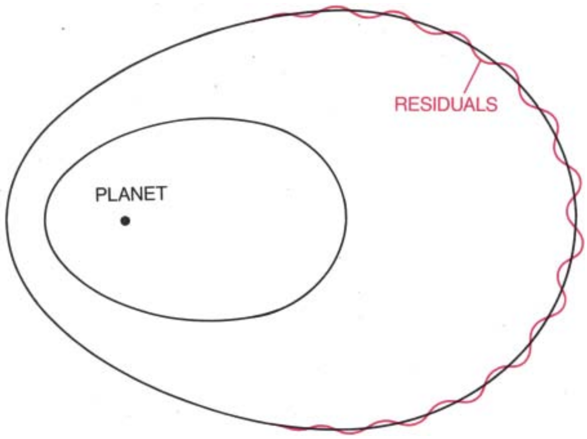
HERE

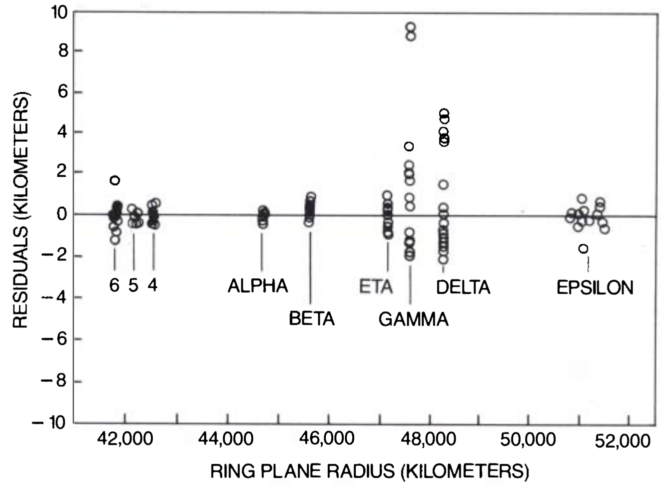
SIRCSA

MERKUR

FORD WERKE A.G. · KÖLN, GERMANY

MERKUR ADVANCING THE ART OF DRIVING





TYPICAL RING can be modeled by assuming its edges are defined by a combination of smooth elliptical orbits and random "wiggles" (*left*). The wiggles are inferred from residuals: differences between observation and smooth ring models. The residuals associated with the nine main rings of Uranus are as large as several kilometers (*right*). The various residuals of the eta, gamma and delta rings are so large that they are thought to have arisen from gravitational perturbations exerted by nearby moonlets.

transmitted by *Voyager 2* as it passed behind the rings led to firm knowledge of the ring particles. The sizes of the largest particles in the rings are on the order of a few meters, comparable to the sizes of particles in Saturn's rings. Strangely, the Uranian rings seem to contain very few millimeter- and centimeter-size particles, in great contrast with the situation at Saturn. Knowing the sizes of the particles and assuming that their density is about that of the most primitive meteoroids—1.5 grams per cubic centimeter—one obtains a mass density of several hundred grams per square centimeter for the rings in general. That is considerably larger than the mass density of even Saturn's imposing *B* ring.

Unfortunately the estimated density conflicts with the density predicted from theories of the dynamical behavior of the rings. Because Uranus is not perfectly spherical, the particles in the elliptical rings undergo precession, or slew around, so that the point of closest approach (the periapsis) to Uranus advances slowly in the orbital direction. Since the rate of precession varies with the distance of the orbit from the planet, the observed nested alignment of particle orbits should be destroyed by interparticle collisions in 100 years or so, causing the orbits to become circular. The most popular theory before *Voyager 2* was that the nested ellipses are preserved because the self-gravity of the particles themselves exactly counteracts the differential precession. According to these predictions, however, the required mass density of the rings should be more than an order of magnitude smaller than the observed density.

One alternative possibility is that the slight friction produced by variations in the orbital velocity of the gently colliding particles as their orbits fall out of phase may provide the force needed to prevent further disruption. The nature of the forces that sustain the nested elliptical orbits remains elusive.

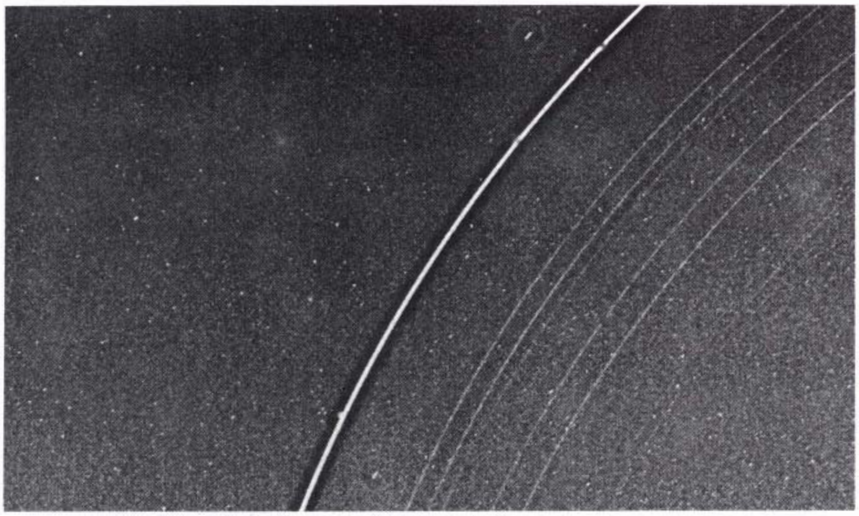
The darkness and the lack of color of the ring particles are similar to those of the 10 inner, small moons discovered by *Voyager 2*. One face of the outer, large moon of Saturn (Iapetus) appears similarly dark, but it has a reddish hue. Most asteroids, as well as the moons and rings of Jupiter, are also reddish.

Why are the inner moons and rings so dark? According to one hypothesis, their surfaces initially incorporated methane-rich ice. Energetic electrons in the surrounding radiation belts of Uranus struck the surfaces and expelled hydrogen atoms, leaving a dark carbon residue behind. (The five main moons of Uranus, which lie mostly outside the Uranian radiation belts, are considerably brighter than the rings or inner moons.) One problem with this scenario is that the radiation process, which has been modeled in terrestrial laboratories, usually produces reddening as well as darkening. Yet none of the rings or moons of Uranus is noticeably reddened.

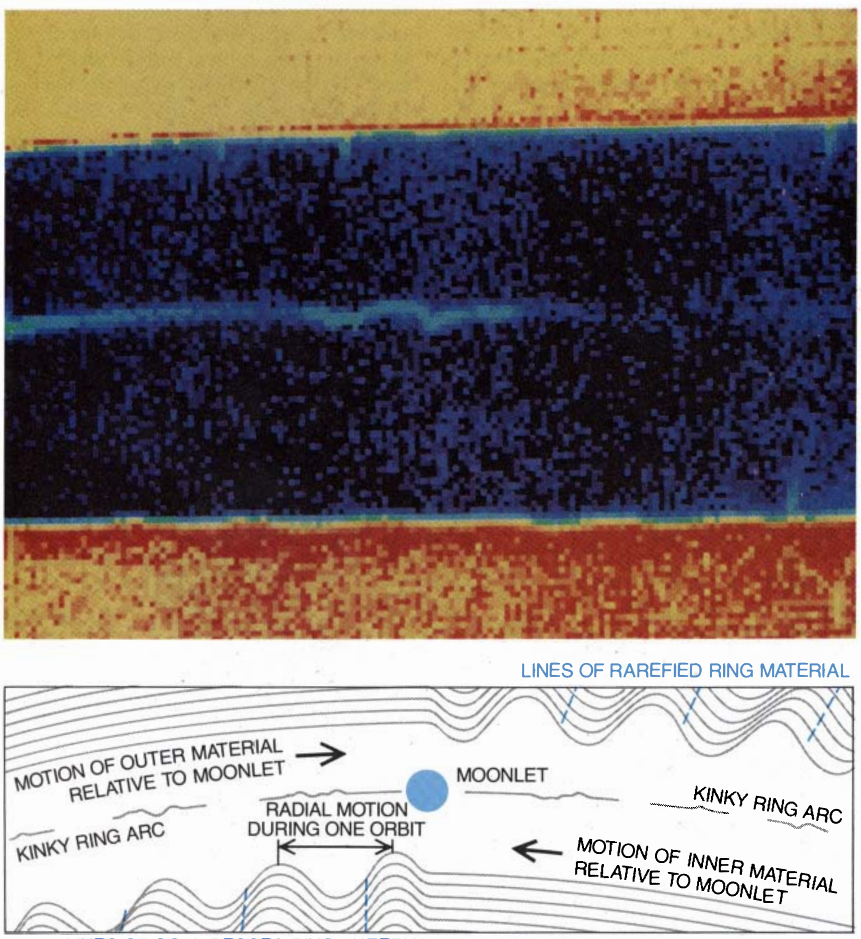
Alternatively the general darkness and lack of color in the ring and inner-moon system may attest to the presence of unaltered material with the properties of primitive meteoroids, called carbonaceous chondrites, which contain large amounts of carbon, opaque minerals and dark organic materials. A more speculative possibility is that the material may be a conglomerate, made up of atmospheric material that was thoroughly processed during an impact on Uranus of a body as large as the earth. Such an impact could also account for the fact that the pole of Uranus' rotation lies virtually in the planet's orbital plane [see "Uranus," by Andrew P. Ingersoll; SCIENTIFIC AMERICAN, January]. We look forward to comparing the Uranian ring properties with the brightness and color of the ring arcs and possible inner moons of Neptune, Uranus' sister planet, when *Voyager 2* encounters that planet in August, 1989.

Regardless of why the particles are dark, the structure of the rings must be strongly influenced by the moons of Uranus. Peter Goldreich and Scott D. Tremaine of the California Institute of Technology have proposed that the unusual narrowness and ellipticity of the rings arise from moons that confine the ring particles much as shepherds confine their flock. The essence of the proposed process is that initially circular orbits are perturbed into ellipses by gravitational tugs from passing (or overtaken) ring moons, or shepherds. By changing the eccentricity the shepherds change the angular momentum of the ring particles without changing their orbital energy. An eccentric orbit has more energy than a circular one of the same angular momentum. As excess energy is removed (for instance during subsequent collisions between particles) a circular orbit again results. The new circular orbit has a different angular momentum that places it farther from the moon.

As a consequence ring material recedes from, or seems to be repelled by, a ring moon over many cycles. The balance between the rate at which energy decays and the rate at which angular momentum is transferred through the ring may be responsible



SHEPHERDING MOONS on each side of the epsilon ring keep the ring particles in place by exerting gravitational forces on them. The moons, which appear as bright streaks, are designated 1986U8 (*outside the epsilon ring at top*) and 1986U7 (*inside the epsilon ring at bottom*). The dark band around the epsilon ring is an artifact of the data processing. *Voyager 2* did not detect shepherding moons near any of the other eight rings.



INDIRECT EVIDENCE indicates that unseen shepherding moons orbit in a 325-kilometer gap (*top*) in Saturn's rings called Encke's division. First, the inner (bottom), greenish edge has a wavy appearance compared with the outer (top) edge. Second, there are slanted intensity contours seen within the ring material at the bottom; the contours result from periodic compressions and rarefactions of ring material caused by the wavy edge. Third, there is a narrow, kinky ring arc in the center of the gap. The illustration at the bottom shows how compression and rarefaction arise, as seen from a shepherding moon.

for the overall eccentricity of these narrow shepherded rings.

It is known, for instance, that at least one gap in Saturn's rings, Encke's division, is cleared by an as yet unseen orbiting ring moon that is thought to have a radius of about 10 kilometers. The ring moon clears the gap by the same process of angular-momentum redistribution that causes the ring confinement described above. The small orbital eccentricities that transfer the momentum produce waves that are seen to form and then die out along the edges of the gap.

From the numerous high-resolution images obtained by *Voyager 2* it is clear that most of the 10 new moons lie just outside the main rings. The two innermost moons, 1986U7 and 1986U8 (which have been named Cordelia and Ophelia) straddle the epsilon (outermost) ring. Their diameters, between about 40 and 50 kilometers, are comparable to those of typical asteroids, and their orbital periods have been determined with good accuracy from *Voyager 2* image sequences.

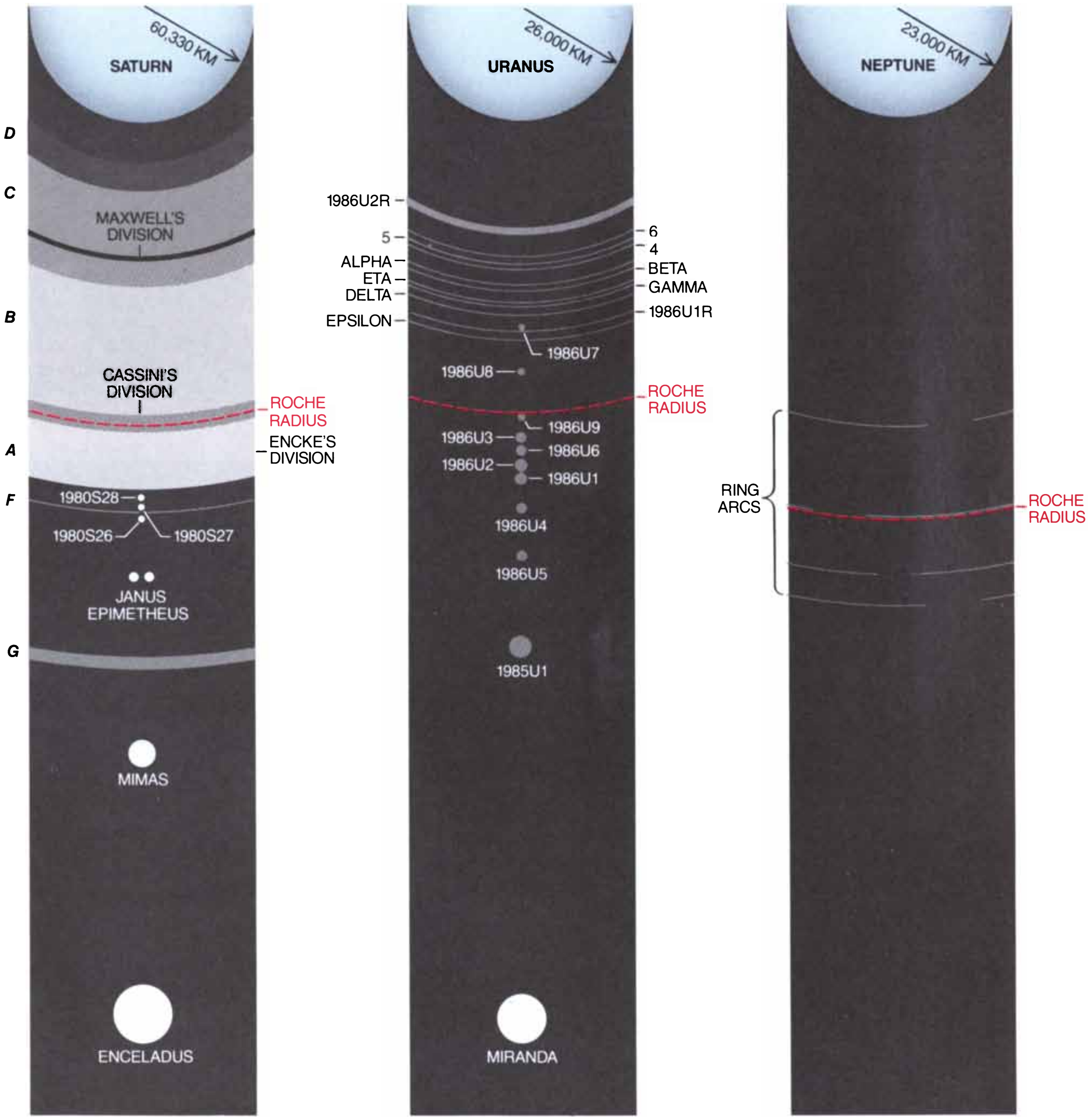
On the basis of such data, Carolyn C. Porco of the University of Arizona and Goldreich have shown that 1986U7 and 1986U8 exert "resonant" gravitational forces that keep the epsilon ring particles in their confined orbits. In a resonance relationship the gravitational tugs of the moon on the particles are timed so that they add constructively, and the orbital eccentricity will increase just as the amplitude of a swing increases if the swing is given pushes at precisely the right frequency. A resonance relationship exists between the outer edge of Saturn's *A* ring and the large moon Janus, for example; the ring particles complete seven orbits for every six orbits made by the moon. A resonance relationship also exists between the outer edge of Saturn's *B* ring and the moon Mimas; the orbital period of the moon is exactly twice that of the particles.

In the case of Uranus, particles on the inner edge of the epsilon ring make 24 orbits for every 25 made by ring moon 1986U7; particles on the outer edge make 14 orbits for every 13 made by ring moon 1986U8. In other words, material in the epsilon ring appears to be confined in much the same way as the outer edge of Saturn's *A* and *B* rings are maintained by Janus and Mimas. The process is able to shepherd ring material, but only at discrete resonance locations.

Does some kind of shepherding process also confine each of the eight other main rings of Uranus? A few other weak resonances do exist, but they cannot account for the observed orbits. Moreover, the *Voyager 2* cameras did not detect any other small ring moons. Of course, this does not mean that none exist. Limitations in the observations could allow for the existence of large numbers of small moonlets sprinkled among the rings that could in principle provide the confinement needed to keep the inner rings from spreading and to supply the required orbital eccentricity.

Another confinement problem concerns the several incomplete ring arcs around Uranus discovered by *Voyager 2*. Similar partial ring arcs have recently been discovered around Neptune, and two were seen within Encke's division during the *Voyager 1* and *Voyager 2* encounters with Saturn in 1980 and 1981. Although hypotheses of the dynamical confinement of arcs have been proposed, the conditions they must satisfy are not met at either Saturn or Uranus. The existence of partial ring arcs around Saturn and Uranus remains a puzzle.

Several insights can be drawn even from our first look at the data. A band of 10 successively smaller moons lies within the five main Uranian moons, merging with the ring system at its outer edge. Any plausible model of ring dynamics predicts that the narrow, dense rings would diffuse in width and become quite transparent in



RING AND MOON SYSTEMS of Saturn (*left*), Uranus (*middle*) and Neptune (*right*) show a striking similarity: a transition from a few large moons in outer regions to many small moons just outside ring-filled inner regions. The location of the transition coincides roughly with the Roche zone, the region in which planetary tidal forces are strong enough to fragment small objects. Recent data suggest that many smaller objects are intermingled with the rings and moons. Here the radius of each planet is set equal to 1.

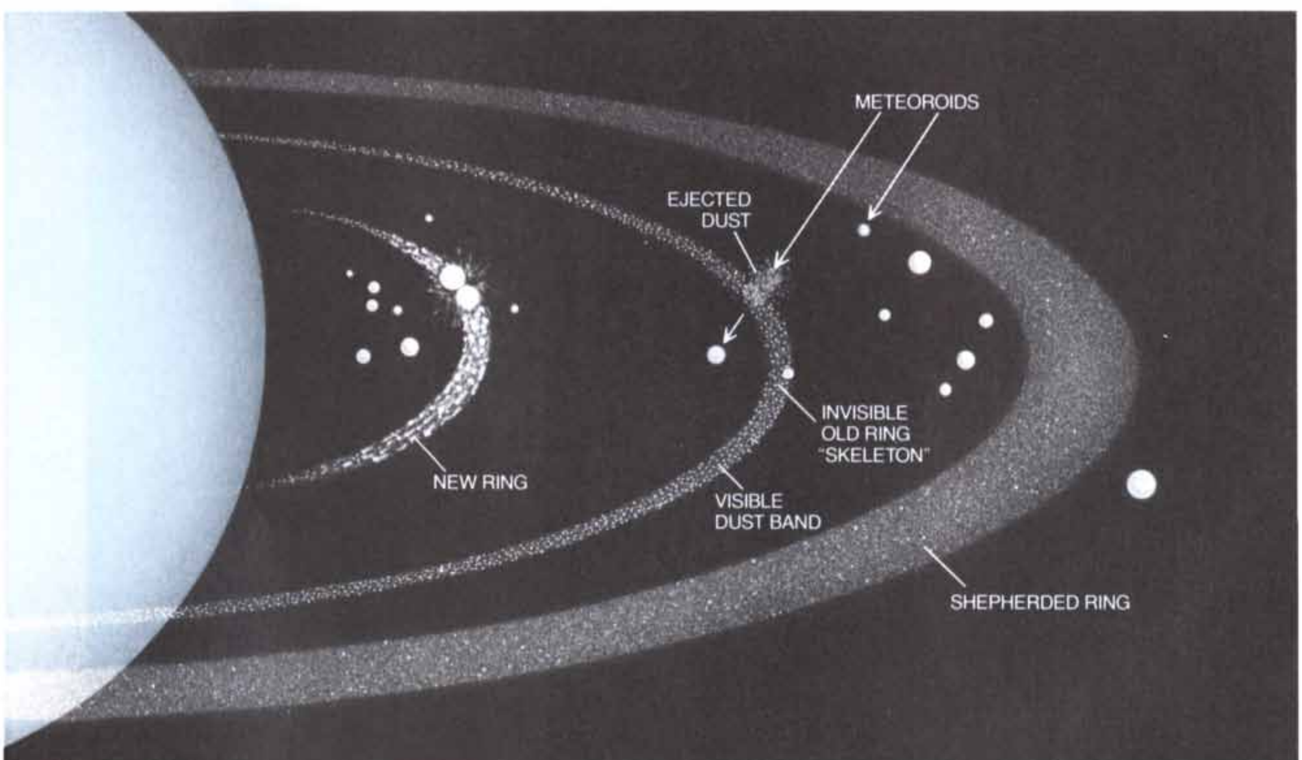
a period much shorter than the age of the solar system. As a consequence either the rings must be confined by shepherding moonlets or they must have been created recently. Incomplete ring arcs, such as those seen around Neptune and in Encke's division, have an even shorter life in the absence of some confinement process.

A number of observations suggest that the Uranian ring system, as well as other ring systems, has many other youthful features. All the small dust particles that are spread prominently, if thinly, through the ring system will be dragged out of orbit, perishing in a meteor flash within the Uranian atmosphere after only a few hundred years. Even the newly discovered moons may not have existed for more than a billion years (about a fifth of the age of the solar system). As Eugene M. Shoemaker of the U.S. Geological Survey has pointed out, any moons older than a billion years probably would have been destroyed even at the current rate of bombardment by cometary meteoroids. Uranus' inner moon Miranda, for example, has an unusual surface structure, which may attest to a geologically recent reassembly of the moon from collisional fragments. Smaller moons, which would have to be more numerous to provide the shepherding necessary to maintain the main rings, are even more susceptible to such catastrophic destruction.

It may be that some or even all of the narrow rings of Uranus were created, simultaneously or separately, during the catastrophic destruction of moons by cometary meteoroids. Most of the rings, even if the largest mass densities are assumed, contain only the mass of a one- or two-kilometer-radius object. The earliest precursors of the rings may have gone long ago, swept up by surviving ring moons or dragged into the planet itself. Similarly, the partial arcs seen by *Voyager 2* may be transient clumps of debris released during impacts involving smaller and as yet undetected members of a band of moonlets that resembles an asteroid belt. As such partial arcs spread they would become symmetrical, nearly transparent bands of debris. Even if the large-particle skeletons of such bands could not be seen, their continual erosion by micrometeoroids would provide a source for the observed bands of microscopic dust. Similar processes could be taking place in the ring and moon systems of Jupiter, Saturn and Neptune.

The underlying population of objects that could be ultimately responsible for such interactions might consist of literally thousands or even millions of moonlets whose sizes range from hundreds of meters to a few kilometers. The moonlets could be the extension of the inward trend toward more numerous, smaller moons seen in all three giant planet systems that *Voyager 2* has explored. A population of moonlets would provide a missing link between rings and major satellites. It might be no accident to find such a population near the so-called Roche zone: the region in which a planet's tidal forces prevent the growth of large satellites.

In all, we are beginning to suspect that the austere Uranian ring system may have had a violent and chaotic past. Uranus' present system—consisting of 10 narrow rings, numerous dusty bands, some narrow ring arcs and a bevy of moonlets—may be only a fragment of its former self and merely one more passing stage in an ongoing process of creation and loss from which future rings of Uranus will arise. The ring systems of Jupiter, Saturn and Neptune may have had a similar history. Indeed, some aspects of ring and moon systems may be as evanescent, and continually evolving, as the drifting continents on the earth are now known to be.



FORMATION OF RINGS and partial ring arcs may have taken place as moonlets collided with one another or were eroded or destroyed by incoming meteoroids. The collisions would have generated lumps of material; small differences in the velocity of the material would have caused it to spread rapidly into arcs. After a longer time each arc could have become a complete, if nearly transparent, ring (*left*). Such stable bands could then provide a source of target material that would be eroded by microscopic meteoroids, continuously generating short-lived dust (*middle*). Some rings could, of course, be sustained by shepherding moons (*right*).

Our Customers are the World's Smartest Buyers of Computer Products because they will SAVE $15 MILLION this year... Why not join them... at ELEK-TEK

DATA STORAGE

IOMEGA BERNOULLI BOX

For PC/XT/AT and most compatibles

10 meg (single) **$899**
20 meg (single) **1199**
20 meg (10 + 10) **1399**
40 meg (20 + 20) **1799**
Bernoulli Plus (80+20+20) **3299**

Interface priced separately

Bernoulli Cartridges

IOMEGA 10 meg—3 pak **155**
IOMEGA 20 meg—3 pak **239**
Dysan 10 meg—3 pak **140**
Dysan 20 meg—3 pak **209**
BB HEAD CLEANING KIT 59

MOUNTAIN

445818 20 MB Internal Bernoulli Drive .... **CALL**
445826 20 MB Cartridge for Internal Bernoulli Drive **90**

HARD DRIVE CARDS

Western Digital 20 MB **$489**
Plus Development 20 MB .... **CALL**
Western Digital 30 MB .... **CALL**

MULTI FUNCTION BOARD

AST

Advantage **$369**
I/O Mini II **99**
Six pak + with 384K **209**
Six pak premium with 512K **229**
Rampage PC **239**
Rampage 286 .... **409**
Preview I/O **150**

GENERIC Multi-function Board

(clock & calendar, 2 ser., 1 par.)
0K **99**
384K **175**

INTEL ABOVEBOARD

All products .... **CALL**

ORCHID TECHNOLOGY

All products .... **CALL**
Generic Compatibles of most of the above .... **CALL**

CHIPS

64K ram (150 or 200 NS) .... **CALL**
128K ram .... **CALL**
Samsung 256K RAM 150 NS . **CALL**
Major Brand 256K 120/150 NS **CALL**
NEC V-20 or V-30 .... **12^{00}**
Intel 8087 math chip (5MH) **110**
Intel 80287 math chip (6MH) .. 155

HARD DRIVES

SEAGATE

Internal Hard Disk Subsystems

20 Megabyte **$369**
30 Megabyte ½ Height **450**

Each system includes:

- Seagate Drive
- Controller
- Documentation
- 2 Cables
- 1 yr limited warranty
- For use in IBM PC, XT, AT&T, most compatibles

40 MGB Drive (only) for AT **699**
80 MGB Drive (only) for AT **999**

TAPE BACKUPS

Western Digital 60 meg **$699**
Archive XL 5240-XT 40 meg **465**
Archive XL 5540-AT 40 meg **499**
Tallgrass 1020i-XT **339**
Tallgrass 1020i-AT **339**

MODEMS

Hayes

Model 1200

HAYES *Internal*

(including software)
1200B 1200 Baud **$349**
2400B 2400 Baud **510**

HAYES *External*

(no software included)
1200 1200 Baud **380**
2400 2400 Baud **579**

OTHER MODEMS

Hayes compatible

Internal *(including software)*

Practical Peripherals 1200B **100**
Novation 1200B **125**
Generic 2400B **209**

External *(no software included)*

U.S. Robotics 1200 baud **199**
U.S. Robotics 2400 baud **399**
Practical Peripherals 1200 baud . **125**
Novation 1200 baud **175**

HEWLETT PACKARD
LaserJet Toner Cartridges
HP92285A for Original LaserJet
HP92295A for LaserJet Series II
Mfr. Sugg. Ret. $115 **$84**

PRINTERS

EPSON

$335
FX 86e

Epson's FX-86e is the industry's standard DOT MATRIX PRINTER. Features include 8K buffer, NLQ, 40/240 cps.

OTHER EPSON PRINTERS

FX 86e (New)200 CPS **$335**
FX 286e wide carriage (New) ... **475**
LX 86 120 CPS **189**
EX 800 300 CPS **399**
EX 1000 300 CPS **545**
LQ 800 NLQ (Parallel) **455**
LQ 1000 NLQ (Parallel) **650**
LQ 2500 NLQ 324 CPS **899**
CR 420i NLQ 420 CPS **699**

SAVE OVER 80%
EPSON SQ-2000
wide carriage Ink jet printer
Mfr. Sugg. Ret. $2395 **$399**

PRINT BUFFERS

PRACTICAL PERIPHERALS

64K parallel or serial **$129**
256K parallel or serial **159**

HANZON

64K universal **250**
64K memory expansion **56**

QUADRAM

Microfazers .... **CALL**

MONITORS

BEST BUY

Kimtron KC-1 13″ RGB **$275**
Samsung 1252G 12″ Amber TTL . **99**

NEC

Multisync **575**

AMDEK

300A amber composite **130**
310A (TTL) Amber **129**
410A, 410G, 410W .... **CALL**
600S RGB color **339**
722C **488**

PRINCETON

HX9 9″ RGB **425**
HX12 12″ RGB **449**
HX12E 12″ EGA **535**
MAX12 12″ amber **175**

Huge Discounts on

TOSHIBA NEC OKIDATA XEROX / Diablo

DICONIX 8213
Portable Printer
Perfect for Laptop Computers
Call for Price

SPECIAL PURCHASE!
Diablo D-625 Daisywheel 25CPS
Elek-Tek Price $399*
*Price includes IBM parallel interface

TRIPPLITE & PERMA POWER POWER PROTECTION

Surge protector strips

PermaPower 6 outlet **$28**
Isobar4-6 4 outlet **42**
Isobar8-15 8 outlet **50**

Line conditioners/stabilizers

TL1200 4 out 1200W **135**
TL1800 6 out 1800W **199**

Command consoles

TLCCI 612 .... **79**
PermaRS562 Power Commander **65**

Standby power systems

TLBC450 450 watt system **360**
BC675 675 watt system **535**
TLBC1000 1000 watt system ... **850**
TLBC2000 .... **1299**
SAFE400VA 400 volt system ... **350**
SAFE1200VA 1200 volt system .. **899**

PUSH BUTTON SWITCH BOXES

448010 2 pos. 25-pin AB **$29**
448052 2 pos. 36-pin AB **32**

ROTARY SWITCH BOXES

448028 3 pos. 25-pin ABC **$29**
448060 3 pos. 36-pin ABC **33**
448036 4 pos. 25-pin ABCD ... **33**
448078 4 pos. 36-pin ABCD ... **35**
448044 4 pos. 25-pin ABCD ... **38**
448086 4 pos. 36-pin ABCD ... **40**

VIDEO BOARDS

Hercules graphics PLUS **$199**
Generic compatible **80**
Hercules color **155**
Generic compatible **80**
Vega Deluxe **369**
Quadram EGA **330**
Generic EGA ½ card **179**
NEC GB1 EGA Card **359**

HUGE SAVINGS ON DISKETTES *Call for Quantity pricing for 10 boxes or more*

(*Packed in flip n' file) (**unformatted)	Dysan	maxell	3M	ELEK-TEK	SONY	MEMOREX
3½″ SSDD	14.00	13.00	13.00*	—	13.00	12.00
3½″ DSDD	22.00	20.00	20.00	—	20.00	16.00
5¼″ SSDD	13.00	8.00	8.50	—	—	7.00
5¼″ DSDD	18.00	10.00	10.00	9.00	—	8.00
5¼″ SSDD96TPI	20.00	18.00	17.00	—	—	—
5¼″ DSDD96TPI	22.00	22.00	22.00	—	—	10.00
5¼″ DSDDHD (for IBM AT)	12.50	22.00	22.00	12.50	—	17.00
8″ SSDD**	19.00	22.00	20.00	—	—	—
8″ DSDD**	23.00	23.00	23.00	—	—	—

3M DATA CARTRIDGES

3M	1	10	25
DC100A	$13.50	$13.00	$12.00
DC300A	17.50	17.00	16.00
DC300XL	20.50	20.00	19.00
DC300XL/P	21.50	21.00	20.00
DC600A	23.00	22.50	21.50
DC1000	14.50	14.00	13.50
DC1000 alpha*	16.50	16.00	15.00
DC1000 delta**	16.50	16.00	15.00
DC2000	19.00	18.50	18.00

* formatted 10 MB Cartridges
** formatted 20 MB Cartridges

ELEK-TEK

6557 North Lincoln Avenue, Chicago, Illinois 60645 312-677-7660 312-631-7800
Call Toll Free 800-621-1269—Except IL, AK • Canadian Toll Free 800-458-9133

Corp. Accts. Invited. Min. Ord. $15.00. Visa or MasterCard by Mail or Phone. Mail Cashier's Check, Mon. Ord., Personal Check Add $4.00 1st Item. (AK, HI, P.R., Canada add $10.00 first item) $1.00 ea. add'l shpg. & handl. Shipmentsto IL address add 7% tax. Prices subj. to change. WRITE forfree catalog. RETURN POLICY: Defectives Only: Most products replaced within 30 days of purchase with identical merchandise only. Computer and large peripherals replaced only when defective on arrival (within 3 work days of delivery). Other problems covered by mfr. warranty. ALL ELEK-TEK MERCHANDISE IS BRAND NEW, FIRST QUALITY AND COMPLETE. Delivery subject to availability. **DUNS #09-718-0517**

Beaches and Barrier Islands

Most manmade structures designed to protect these landforms along the Atlantic and Gulf coasts of the U.S. eventually fail. It might often be wiser not to interfere with the course of natural events

by Robert Dolan and Harry Lins

One of the world's most spectacular sand beaches runs from New England down the Atlantic coast, winding around Florida to reach along the northern edge of the Gulf of Mexico. Much of the 2,700-mile beach lies on the 295 "barrier islands" that stand between the sea and the mainland along the two coasts. Both the mainland beaches and the islands are under constant attack from the sea, which sometimes builds them up, sometimes tears them down and continuously reshapes them.

The early inhabitants of the shore zone, recognizing that the coast has always been a hazardous place for people, settled on the bay side of the barrier islands, as far inland from the beach as possible. Construction was also kept well back of the mainland beaches. Over the past several decades that pattern has been reversed. Construction now takes place as close as possible to the shoreline. Today such resorts as Atlantic City, Ocean City, Virginia Beach, Hilton Head, Jekyll Island, Miami Beach and Galveston Island occupy barrier islands, and summer homes crowd many of the beaches. Naturally pressure for public works to protect the islands and beaches is strong.

It is impossible for a northeaster or a hurricane to move along either coast without crossing or at least affecting a beach or a barrier island. Hurricanes are more powerful than the winter northeasters, but they come less often, particularly along the Atlantic coast. There, about 30 times a year, northeasters generate waves with sufficient force to erode the barrier-island beaches and frontal dunes, which run parallel to the shoreline. Every 100 years or so one of these winter storms has a devastating impact on the beaches and islands and the people who live on them. Along the Gulf Coast, on the other hand, hurricanes cause most of the damage. The question public officials face is whether works of engineering are effective in the long term or whether it would be better to let nature take its course along the two coasts, as it has for centuries.

A severe northeaster in 1962 (March 7, which was Ash Wednesday) now serves as the benchmark for workers in coastal geology and engineering. The Ash Wednesday storm resulted in many deaths and injuries and caused $300 million in damage. Several days passed before the full extent of the havoc wreaked on the beaches and barrier islands became apparent. Beaches were eroded back to the dune lines. Along sections of the Outer Banks of North Carolina erosion ate far enough back into dunes to expose the sand fences that had been emplaced 30 years earlier to enhance the natural processes of dune construction. Roads were buried by sediment from the overwashing ocean water, and lagoon channels were choked with sand and debris.

The storm of 1962 convinced many people that faulty practices had been followed in the development of Atlantic-coast beaches and barrier islands. It became apparent that manmade sand dikes and barrier dunes could not protect beachfront development under the force of a major storm. It also became clear that a better understanding of the natural dynamics of beaches and barrier islands would be fundamental to any effort to deal with the short- and long-term hazards of living on them.

Hence the storm was a turning point in coastal studies. As a result of work done since then it is now understood how beaches and barrier islands were created, how external forces change them and why they will continue to change in spite of human efforts to halt the natural processes.

Beaches and barrier islands are constructed of whatever sedimentary material is available to be transported by waves and near-shore currents. Beaches are among the most widely distributed of the coastal sedimentary environments: 33 percent of the North American shoreline is beach. The sources of the sediment that built the beaches are rivers, eroded cliffs, glacial deposits and biological material such as coral and shell fragments.

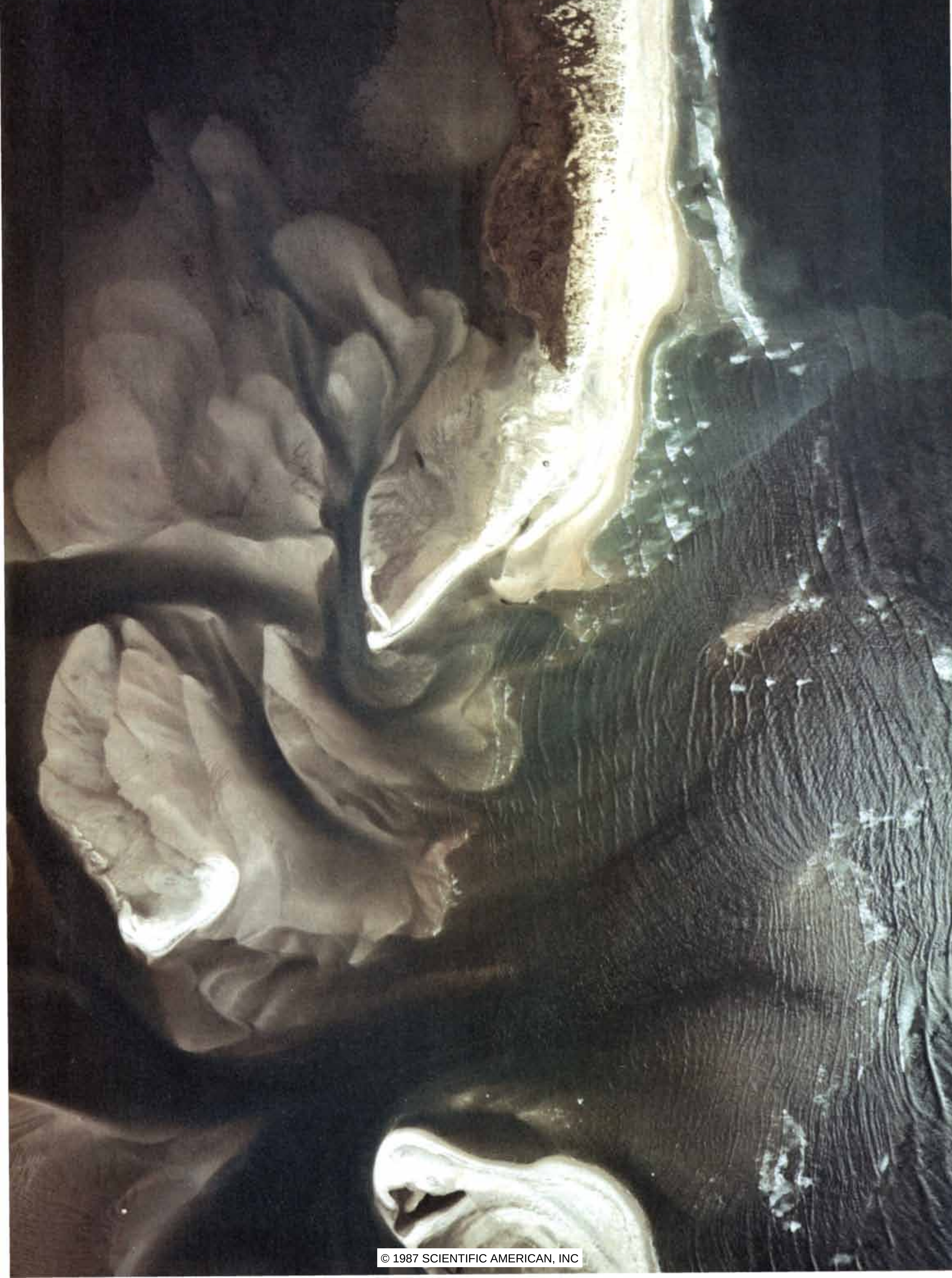
Most of the world's beaches consist primarily of quartz sand, together with fractions of heavy minerals and fragments of rocks and shells. On the West Coast and the coast of New England some beaches near rocky cliffs and headlands consist of larger material, including gravels and the naturally rounded stones called cobbles. A cross section of a beach reveals sedimentary layers made up of particles of various sizes that indicate their source and the processes responsible for depositing them. Bedded within the layers are units of well-sorted finer sands brought in by wind.

Formative Processes

On a sandy coast almost any variation in wave action or sea level alters the beach. Sand transported by waves and inshore currents moves offshore, onshore and in the direction of prevailing longshore currents. Beaches constantly adjust in response to different tide, wave and current conditions.

The Atlantic and Gulf coastal plains are relatively flat sedimentary formations that slope gently seaward to wide continental shelves. The sand beaches

DRUM INLET on Core Banks on the North Carolina coast shows how the tides, waves and currents move sand. During the incoming tide inshore currents carry sand through the inlet and deposit it as a fan-shaped shoal (*left*). To some extent the reverse takes place when the tide goes out, as is evident from the less distinct shoal outside the inlet.



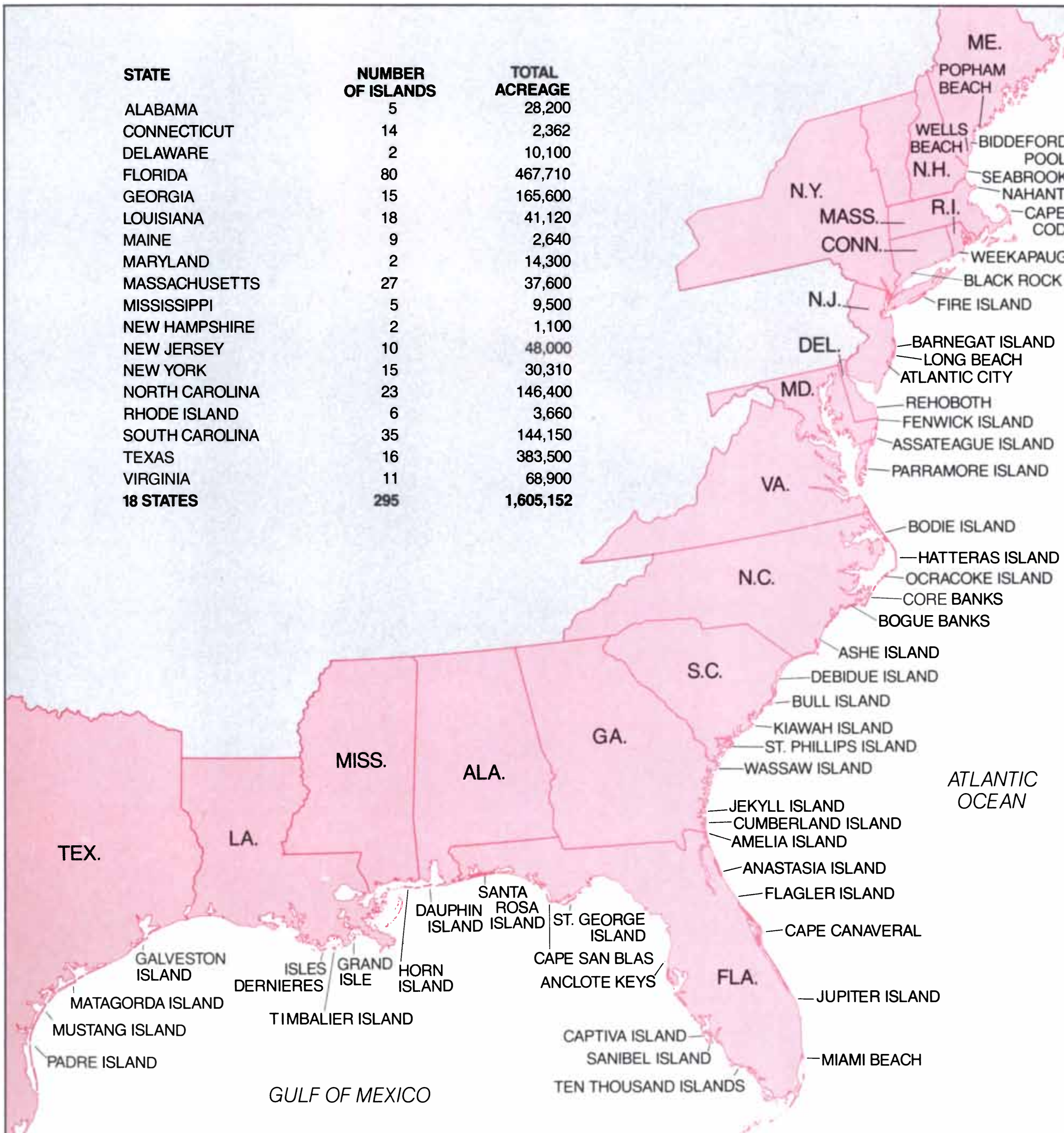
and barrier islands that mainly constitute the shore zones of these plains are the product of marine processes working and reworking the seaward margins of the coastal plains.

The processes that form barrier islands have been debated among earth scientists for many years. There is, however, strong evidence that most of the barrier islands along the Atlantic and Gulf coasts formed about 4,000 to 6,000 years ago. They have undergone change since then, and they are still changing. Peat and tree stumps—remnants of forests that once stood on the bay side of the islands—are now being found on ocean beaches. In other words, the islands are being moved toward the mainland. It is also possible to record short-term recessions of beaches and barrier islands by comparing recent aerial photographs with older ones or with old maps.

Over the long term the changes in beaches and barrier islands result from several factors, including the nature of the sediment and the magnitude of the local processes at work. The major factor, however, has been periodic change in the sea level.

Oscillations of Sea Level

The global sea level has oscillated several times in the past half-million years. During the warmer intergla-



STATE	NUMBER OF ISLANDS	TOTAL ACREAGE
ALABAMA	5	28,200
CONNECTICUT	14	2,362
DELAWARE	2	10,100
FLORIDA	80	467,710
GEORGIA	15	165,600
LOUISIANA	18	41,120
MAINE	9	2,640
MARYLAND	2	14,300
MASSACHUSETTS	27	37,600
MISSISSIPPI	5	9,500
NEW HAMPSHIRE	2	1,100
NEW JERSEY	10	48,000
NEW YORK	15	30,310
NORTH CAROLINA	23	146,400
RHODE ISLAND	6	3,660
SOUTH CAROLINA	35	144,150
TEXAS	16	383,500
VIRGINIA	11	68,900
18 STATES	**295**	**1,605,152**

MAJOR BARRIER ISLANDS along the Atlantic and Gulf coasts of the U.S. are identified. In all there are 295 such islands, which take their name from the fact that they stand between the sea and the mainland. The action of the sea built them from sediments.

cial periods continental ice melted, the oceans expanded and shorelines moved inland. During the cooler glacial periods, as water was withdrawn from the oceans and stored on the continents in the form of glacial ice, the shorelines moved seaward. These processes involved large quantities of water, enough to advance the shoreline great distances across flat coastal plains and continental shelves.

When the last period of glaciation (the Wisconsin) came to an end between 12,000 and 14,000 years ago, the sea level was some 350 feet lower than it is today, and the shorelines of the Atlantic and Gulf coasts were from 20 to 75 miles seaward of their positions now. With the change from glacial to interglacial conditions the world's oceans and seas began to rise. They continued to do so for 8,000 years, reaching within a few feet of the present level between 4,000 and 6,000 years ago. As the sea level rose and the shoreline progressed across the continental shelf, large masses of sand were moved with the migrating shore zone and deposited as beaches. Sediment that had been deposited as floodplains and deltas in the coastal river systems was also disturbed by wave action and moved along the shore.

Once the sea level had become fairly stable the waves, currents and winds worked together on the sand to form the beaches and barrier islands. As long as the inshore beach system contained surplus sediment, the beaches continued to build seaward until equilibrium was established. Equilibrium in this case was a function of the balance among storm and wave energy, sea level and the amount of sediment in the transport system.

All the evidence suggests that this equilibrium was reached at about the time the sea approached its present level. At that time the barrier islands were probably much wider than they are now. As time passed, the complex landscape of the barrier islands evolved. In the narrow stretches surging water from storms frequently breached the islands, creating new inlets. Long spits connected the wider, more stable stretches where sequences of beach and dune ridges developed. In this way long chains of barrier islands evolved. Exceptions to the pattern are the "sea islands" off Georgia and South Carolina. They originated as segments that became detached from the mainland and then were built up by beach material deposited on their seaward side.

Since the time of equilibrium the sea level has continued to rise slowly (a little more than one foot over the past 100 years). The result has been the further recession of shorelines and the enlargement of bays and sounds behind the barrier islands. The recession rate of beaches and barrier islands has varied over the past several thousand years as the rate in the rise of the sea level changed, as the supply of sand dwindled and as the slope of the bottom of the inshore zone evolved in response to storms and waves. Some of the eroded material was deposited in large offshore sediment sinks, such as Diamond Shoals off Cape Hatteras, but much of it has remained within the barrier-island sediment budget and has contributed to the growth of spits, the filling of inlets and the building of dunes, as well as to the deposits that get carried onto land by storms.

STORM DAMAGE is a hazard for buildings near the shoreline. This is how the shoreline of Fire Island, N.Y., looked not long after the Ash Wednesday (March 7) storm of 1962, a particularly severe northeaster. Such a major storm occurs about once per century.

Migrating Islands

The two most important factors in the landward movement of beaches and barrier islands are the overwash of water and sediment and the formation of inlets. Storms cause both phenomena. During severe storms the beach zone and the nearby dunes are overtopped by high water and waves. As the sediment-charged mass of water spills across the beach and flows toward the bays and sounds on the inland margins of the barrier islands, a layer of sediment is removed from the beach and added to the island's interior. This process transforms the shapes and positions of the islands but conserves their total sediment mass.

Overwash and inlet formation are common along the Atlantic coast, particularly south of Cape Cod. Temporary inlets are formed during storms when the narrower reaches of islands are overwashed and breached, creating openings to the lagoons and bays behind the beaches. While the inlet is open sand moves through it and is deposited on the inside of the island as a large, fan-shaped shoal. Sand is also carried out during ebb tides, so that a similar delta may be created in the ocean. The inlet shoals are exposed at low tide and eventually become a new substrate for the formation of salt marshes. Shoals below the low-tide level support underwater grass beds.

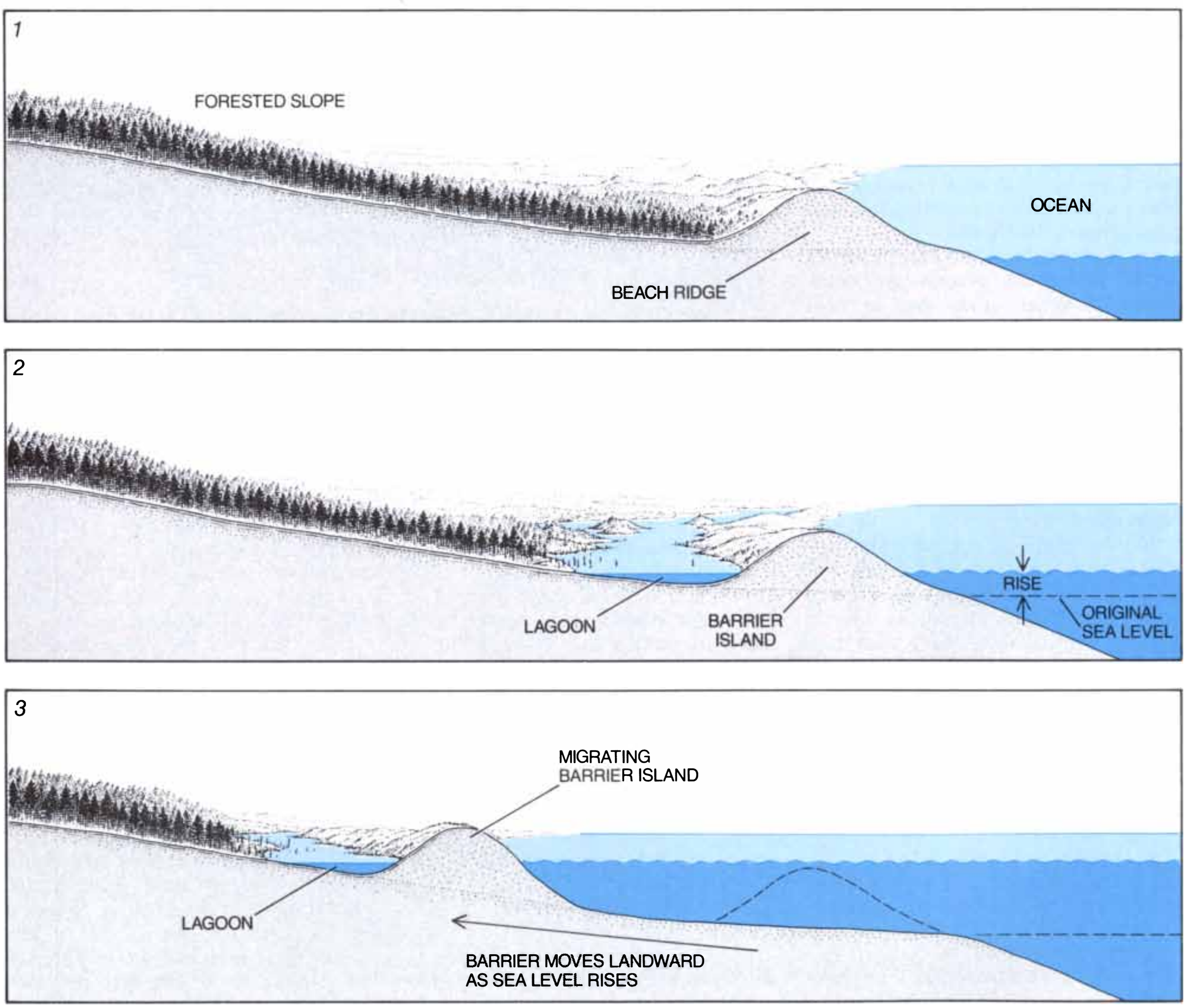
Overwash is commoner along the mid-Atlantic coast than along the other sections. The mid-Atlantic section is close to the track that most of the winter northeasters follow as they move offshore, and the tide range is small. As a result high storm surges are frequent. To the north and south the tide range is higher and the frequency of overwash decreases. The Gulf Coast gets few of the winter storms but is often assailed by hurricanes. Each region, then, has special beach and barrier-island characteristics reflecting the frequency and magnitude of storms and the range of the tide.

Studies of beach erosion are usually based on data obtained from ground surveys, maps, charts and a sampling of the changes recorded by repetitive aerial photography. Over the past 10 years we have developed at the University of Virginia a common-scale mapping program of shoreline movements and changes in storm-surge penetration that now covers more than 750 miles of the Atlantic and Gulf coasts. The results show that the overall shoreline-erosion rate along the mid-Atlantic coast is from 1.5 to 4.5 feet per year. These mean rates encompass considerable variation from island to island and obscure the fact that some short stretches of beach may even be growing larger.

The rates of change are a function of the position and direction of storm tracks, the direction from which waves approach the beach and the orientation of the shoreline. Exceptionally high rates of erosion are found on the barrier islands off the Virginia coast and the beaches along the deltaic coast of Louisiana. There the erosion rate is as high as 25 feet per year. The Virginia and Louisiana islands are shorter and lower than many of the other barrier islands and consist of a thin veneer of fine sand and shell overlying a silty clay substrate. Isolation, vulnerability to storm surge and a relative lack of fresh water made the Virginia islands less attractive during the decade of rapid development along the Atlantic seaboard in the 1960's. As a result the Nature Conservancy, an organization of people interested in protecting and preserving threatened natural resources, was able to buy 13 of the islands, which are being left in a natural state.

Question of Patterns

It is noteworthy that some homes built on Atlantic-coast beaches and barrier islands in the 1950's are still



EVOLUTION OF A BARRIER ISLAND is depicted, beginning with the situation 15,000 years ago (*1*), when the sea level was 350 feet or more lower than it is now. Beach ridges, or dunes, were formed along the continental shelf by waves and winds. In time the sea level rose (*2*), breaking through the ridge and flooding the low area behind it to form a lagoon. As a result the former dune area became an island. Constant action by waves and the continuing rise in the sea level caused the island to migrate landward as sand was removed from the beach and deposited inland. Its present position (*3*) will change for the same reasons.

in place whereas newer houses built nearby have long since succumbed to storm damage. This contrast brings into focus the question of whether shoreline changes at a particular place along a sedimentary coast are simply a matter of chance or are naturally recurring and therefore predictable patterns reflecting some organization of the processes that cause the changes. The results of 20 years of research suggest that, even at scales of hundreds of feet, shoreline processes and shorezone landforms assume systematic patterns. This finding departs from the common belief that change on a beach or a barrier island is a random event.

Even a cursory inspection of photographs of the Atlantic coast made from space reveals a repetitive pattern of crescents along shorelines. The crescents appear on both large and small scales. Some of them result from variations in the intensity of shoreline processes along the coast. For example, the rate of shoreline erosion along the barrier islands of Virginia varies with the configuration of the shoreline. Erosion rates are highest where the shoreline faces northeast and lowest where it faces southeast. The result appears as a series of crescentic landforms. The crescents are concave at some places and convex at others.

The largest of the crescentic landforms along the Atlantic coast are the broad arcs of the Outer Banks of North Carolina, which span distances of from 70 to 100 miles. Smaller crescentic forms appear within these large arcs, including beach cusps (from 30 to 100 feet), giant beach cusps (from 300 to 600 feet) and even larger forms a mile or more long. Inshore bars and troughs may also assume crescentic and rhythmic configurations in response to sea states, tides and sea level. The smaller forms appear and disappear and may migrate along the shore; the larger ones establish the pattern for shoreline erosion and storm overwash in severe storms.

The pattern of storm-surge deposits along the Delaware coast following the Ash Wednesday storm of 1962 shows that although all sections of the beaches and barrier islands were overwashed, the distance the water and sand penetrated inland varied markedly. Analysis of these patterns and of the 40-year averages for the position of the shoreline suggests that along the coast regular, recurring patterns exist for both the long-term average of shoreline erosion and the penetration of surging water in a single storm. These periodicities play an important role in determining the patterns of erosion and storm damage.

CRESCENT PATTERNS along the shore are evident in photographs made from spacecraft and airplanes. Some cusps are quite large and others are fairly small. They suggest that the action of waves, currents, tides and winds is repetitive rather than random over long periods of time. Data on where the effects are likely to be severe can help in the planning for new structures near the shore. This photograph was made over Cape Cod.

The protection and restoration of beaches are expensive measures that are generally beyond the means of the individual property owner. Although the owner's best solution is to plan carefully before buying or building on shoreline property, one still sees structures being built in stretches where older buildings are falling into the sea. An understanding of the relation between beach characteristics and potential surges of water in storms is needed so that buildings can be constructed with a knowledge of the probability of wave damage within a given number of years.

On a larger scale the Government enters the scene. The Coastal Engineering Research Center of the U.S. Army Corps of Engineers investigates shoreline problems for the purpose of designing structures to prevent erosion and to protect coastal property from storm surge. In 1946 Congress first authorized the expenditure of Federal funds to build such structures. Under that authority the Corps of Engineers has built more than 100 projects. They encompass both "hard" structures and "soft" engineering.

Engineering Approaches

The hard structures include breakwaters, groins, seawalls and revetments. A breakwater is designed to intercept waves that normally break on a beach and move sedimentary material. The purpose of a groin is to provide an obstruction that retains a beach, or at least retards its erosion, by jutting

It's like having 256,000 in one box.

Back in the dark ages of personal computing, the world was ruled by numbers and words. Graphics were a nicety, but rarely a necessity.

Welcome to the Renaissance.

And to the new IBM® Personal System/2.™ Its talent for graphics is dazzling.

Each Personal System/2 can paint up to 256 colors on the screen at once, drawing from an incredible palette of over 256,000. And not one of those colors costs a penny extra.

Even in monochrome, things aren't monotonous. There can be up to 64 shades of gray for new dimension and contrast.

And the images themselves are greatly improved. The tiny "pixels" that create the image can now be tinier, and there can be lots more of them. Even the space between them seems to have disappeared. So pictures are sharp and clearly defined.

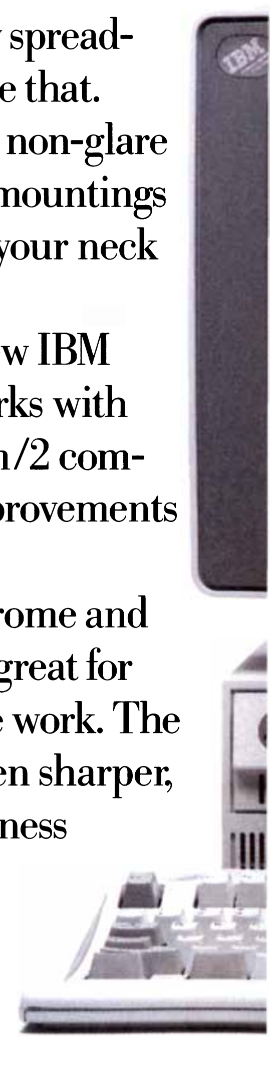
Better letters.

Equally important, letters and numbers are clean-edged and precise, looking more like they're printed than projected. After a few hours with your trusty spreadsheet, you'll appreciate that.

You'll also like the non-glare viewing surface, and mountings that tilt and swivel so your neck doesn't have to.

There are four new IBM displays, and each works with every Personal System/2 computer, all showing improvements in price.

The 12″ monochrome and 14″ color displays are great for most general-purpose work. The 12″ color display is even sharper, ideal for detailed business



crayons

graphics. And for design work, there's the big 16″ color display with even higher resolving power.

Your favorite programs.

Just about any program you can run on the IBM Personal System/2 will look better, and will likely be more pleasant to spend time with. Many other programs are being reworked just to take advantage of the new graphics.

But the future holds real surprises. The screens of the Personal System/2 are like a brand new kind of canvas. How the artists will use them should be something to see.

For a graphic demonstration, call your IBM Marketing Representative, or visit an IBM authorized dealer.

For the dealer nearest you, call

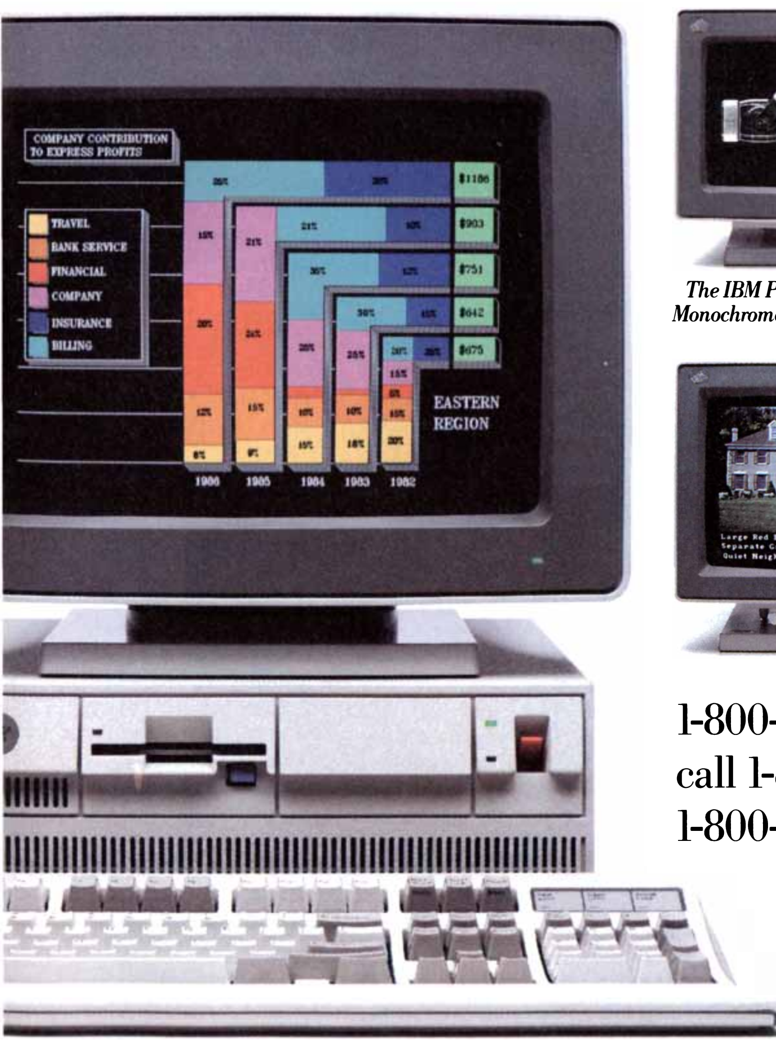
The IBM Personal System/2 Monochrome Display 8503.

The IBM Personal System/2 Color Display 8513.

The IBM Personal System/2 Color Display 8512.

All screens are actual and unretouched.

1-800-447-4700, ext. 9. (In Alaska call 1-800-447-0890, in Canada 1-800-465-6600.) IBM®



The IBM Personal System/2 Model 50 and the IBM Personal System/2 Color Display 8514.

out from the beach and intercepting the transport of sand along the shore. Sand accumulates on the side facing the movement and erodes on the other side. A seawall is designed to keep the sea from directly eroding a beach by separating the two. A revetment is a structure made of a material such as stone and emplaced on the beach or shoreline to prevent waves from removing sand.

The trend in coastal engineering today is to take the soft approach. This technique entails adding sand to a beach to replace material lost by natural erosion. The sand is generally brought in from an offshore area by a dredge and pumped onto the beach or into the water near the shore. Periodic replenishment of sand often costs less than building massive hard structures. It also has the advantage of providing a beach for recreation while protecting the area behind the beach.

Few standard guidelines or rules of thumb are available for planning and managing the use of land on barrier islands or for developing beachfront property. Every coastal site is different. Many costly failures have resulted when a solution that worked on one beach was tried on another. Hence each site should be treated as a unique problem calling for specific plans.

The Human Impact

Advanced development or urbanization has come to about 70 of the 295 barrier islands that rim the Atlantic and Gulf coasts of the U.S. About 80 other islands have been bought by state and local governments for recreation areas and nature preserves, and 120 more are privately owned and largely undeveloped. Fifteen of the largest islands have been acquired by the Federal Government for wildlife refuges and national seashores.

The development of North Carolina's Outer Banks of typifies what has happened on many of the islands. Even though the dynamic nature of the beaches and dunes has always been part of the aesthetic and recreational appeal of the Outer Banks, the islands remained remote and were seldom visited until the first bridges from the mainland were built in the 1930's.

Soon thereafter a plan was developed to build a road running the length of the Outer Banks with a barrier dune to prevent storm surge and overwash. Until then the islands were overwashed several times a year in places, which precluded a permanent road system. Beginning in 1936 the Civilian Conservation Corps and the Works Projects Administration, under the direction of the National Park Service, erected almost 3.3 million feet of sand fences (which act much like snow fences) to create a continuous barrier dune 70 miles long along Hatteras, Pea and Bodie islands.

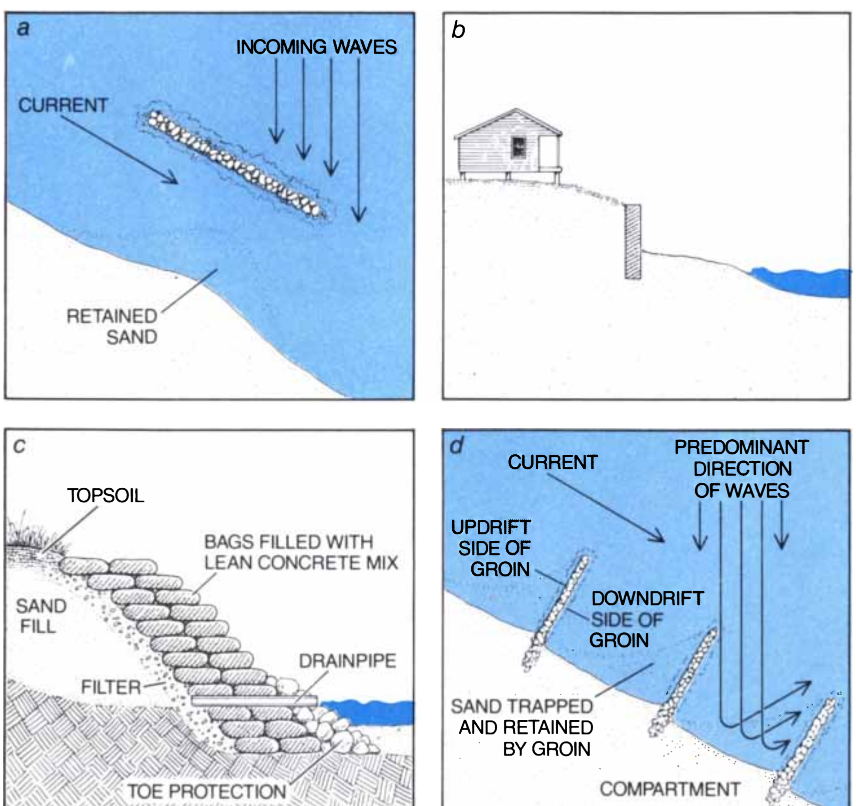
Most of the dune construction took place in the open sand flats behind the original low, irregular beach dunes, about 300 to 500 feet inland from the shoreline. The sand that collected along the fencing was stabilized further with some 1.4 million trees and shrubs and enough grass to protect more than 3,000 acres. The Park Service resumed dune construction in the late 1950's; the result was an almost continuous mass of dunes and vegetation within the Cape Hatteras National Seashore.

Thirty years of artificial dune stabilization have altered the ecology and geology of the Outer Banks. A comparison of a cross section of Hatteras Island and Core Banks, representing the altered and natural states of barrier islands, shows how stabilization has changed the morphology and ecology of the beaches, dunes and marshes.

Viewed from the air the most striking contrast between the natural and altered barrier islands, other than the artificial barrier dune, is a marked difference in beach width. The unaltered islands have beaches from 350 to 600 feet wide, whereas on Hatteras Island the beach has been reduced to 100 feet or less. The paradox suggests that manmade structures do not merely fail to protect beaches but actually work to destroy them.

One major difference between natural and stabilized barrier islands is the role played by the artificial dune in altering the normal vegetation sequence. The stabilized dune, as high as 30 feet in places, prevents overwash and obstructs salt spray; hence plants, which usually grow farther inland from the natural beach, survive close to the ocean on the lee side of the barrier dune. The plants moving seaward because of dune stabilization are shrub communities that now form impenetrable thickets from 10 to 15 feet high. Today the U.S. Fish and Wildlife Service and the National Park Service attempt to check the further spread of the shrubs with controlled fires.

This manipulation of the natural processes affecting beaches and barrier islands became the policy of the Park Service not long after the Outer



PROTECTIVE STRUCTURES represent the "hard" approach to preventing or reducing erosion of a beach. A breakwater (*a*) intercepts waves before they reach the beach. A seawall (*b*) separates the sea from the beach. A revetment (*c*) is built on the beach to prevent waves from removing sand. A groin (*d*) juts out from the beach and retains sand on the side facing the current. An alternative to hard structures is the "soft" engineering technique of replenishing the sand on a beach periodically. The sand comes from offshore.

EFFORTS AT STABILIZATION of barrier islands by building or enhancing dunes and planting grass and shrubs result eventually in beaches that are narrower than those on unstabilized islands. Two islands in the Outer Banks of North Carolina are examples: Hatteras Island (*left*), scene of much stabilization work since the 1930's, and Core Banks (*right*), which is in its natural state.

Banks became the first of the national seashores in 1953. Yet as early as 1938 geologists of the Park Service asserted in reports describing the new seashore area that the low, open nature of the Outer Banks was due to natural processes, not to deforestation as a result of human activities. Between 1970 and 1973 Paul J. Godfrey of the University of Massachusetts and one of us (Dolan) presented to the Park Service the results of research indicating that barrier islands are intrinsically unstable. Dolan and Godfrey advocated the termination of large-scale dune-stabilization programs, arguing that such programs lead to major modifications of the beaches and barrier islands that cause long-term adjustments of geologic and ecological processes. In 1973 the Park Service adopted the policy of letting nature take its course on all the national seashores.

The Prospects

In 1978 Richard A. Frank, then administrator of the National Oceanic and Atmospheric Administration, warned that "a hurricane will kill hundreds, if not thousands, of Americans and cause billions of dollars of property damage sometime soon." The storm has not struck yet, but when it does, much of the damage and some of the deaths will probably take place on one or more of the beaches and barrier islands of the Atlantic and Gulf coasts. Today millions of people live in those vulnerable areas. Most of the people have never experienced a severe hurricane and have little or no idea of its destructive potential.

Many coastal specialists believe much of this development would have been held in check if the only force at work had been knowledge of the high risk associated with beaches and barrier islands. They contend that the Federal Government, by offering various subsidies for public development, is responsible for most of the problems. Of particular concern has been Federal support for the construction of bridges—the first step toward the urbanization of barrier islands—and the Federal Flood Insurance Program.

Not all the blame, however, can be laid at the Government's door. These are affluent times, and many people are investing in second homes. Federal programs such as flood insurance, disaster relief, tax shelters and support for water and utility projects clearly minimize the risk to individuals. On the other hand, many of the barrier islands were well on the way to development before the programs were initiated, and much development would probably continue even if the Government programs were withdrawn.

Over the past several decades the sea level has risen at a rate of one foot per century, and there is little evidence to suggest that the trend will reverse in the foreseeable future. The continuing rise in sea level and encroachment on the shoreline will challenge engineers seeking to maintain the resort communities built on beaches and barrier islands. The beaches and barrier islands are, in effect, migrating out from under coastal development, making each building in turn a beachfront property. Many communities that reduced the risks by building dunes, groins, seawalls and other structures during the 1940's and 1950's now face rapidly increasing maintenance costs for those structures—and perhaps the total loss of their sand beaches no matter what protective efforts are mounted.

If the predicted increase in atmospheric carbon dioxide from the burning of fossil fuels brings about a global warming and increased melting of polar ice, the sea level will rise even faster and shoreline recession will accelerate. Estimates of the rise in sea level over the next century range from two to 10 feet. Even the lower rise would stress the coastal engineering works that now help to maintain the beaches in many of the large resort cities; the higher one would be disastrous for most of the world's developed beaches and barrier islands.

Living on beaches and barrier islands will continue to be hazardous; with a rising sea level storm surges will increasingly cause beach erosion, property damage and sometimes loss of life. The alternatives are limited: to build hard structures and add sand as the sea encroaches or to retreat landward as the sea engulfs the developed beaches and barrier islands. The ones that are undeveloped will adjust, remaining places of great beauty and constant change.

Lyme Disease

A bacterium transmitted to human beings by the bite of a deer tick causes this hazard of summertime. Interleukin-1, an immune-system regulator, may mediate its potentially serious arthritis-like symptoms

by Gail S. Habicht, Gregory Beck and Jorge L. Benach

Lyme disease is an affliction of summer. It is a tickborne bacterial disease that is most likely to be contracted during the months of June through September, when youngsters and adults are outdoors, walking barelegged in woods and long grass.

It is spreading rapidly and is now the most frequently diagnosed tick-transmitted illness in the U.S., if not in the world. In 1975, 59 cases were recorded in Connecticut; in 1985 the number had climbed to 863 cases. Moreover, Lyme disease has now spread to three regions of the U.S.: the Northeast (in coastal areas), the northern Middle West (Minnesota and Wisconsin) and the West (parts of California, Oregon, Utah and Nevada). The disease is also found throughout Europe and has been recorded in Australia, the Soviet Union, China, Japan and Africa.

Because its symptoms can be severe, ranging from acute headache to neurological impairment and manifestations resembling rheumatoid arthritis, the disease has elicited concern in the 12 years since it was first described. The discovery of this disease, from its recognition as a clinical entity to the identification of its causative agent, is a triumph of modern medical research and a tribute to the collaborative efforts of a great many scientists.

Lyme disease was first reported in November of 1975, when the Connecticut State Health Department received telephone calls from two mothers whose children had just been diagnosed as having juvenile rheumatoid arthritis. The condition is a devastating one that can lead to lifelong suffering and physical debilitation, and it is not surprising that these mothers were concerned. What alarmed Health Department officials, however, was the news that these cases were not isolated ones: according to the women who telephoned, a number of adults and children in the town of Lyme had recently been diagnosed as having rheumatoid arthritis. Health officials concluded that this was more than a regional anomaly and might represent something very serious: either the presence of an environmental toxin or possibly the beginning of an epidemic.

They contacted Allen C. Steere, who was then a postdoctoral fellow in rheumatology at the Yale University School of Medicine. He had just completed training with the epidemic-intelligence service of the U.S. Centers for Disease Control in Atlanta. Intrigued by this bizarre outbreak of arthritis, he agreed to undertake an epidemiological investigation.

Steere and his colleagues discovered that the disease was limited to three townships in eastern Connecticut, Old Lyme, Lyme and East Haddam, which are adjacent communities on the east bank of the Connecticut River. Juvenile rheumatoid arthritis is normally a rare condition, affecting only one in 100,000 children, and yet out of a total population of 12,000 in these three towns, 39 children (and 12 adults) had been diagnosed as having the condition: 100 times the normal occurrence.

Moreover, distinct cluster patterns emerged within each town. Most victims lived in heavily wooded areas and only a few of them in town centers. In Old Lyme and East Haddam half of the affected individuals lived on just four roads, and here the frequency in children was 10,000 times higher than normal: one in 10 children as opposed to an expected one in 100,000. This was clearly no ordinary form of rheumatoid arthritis, and there were few clues to guide Steere in his investigation. Nevertheless, he made several important findings.

One was that the disease is not particularly contagious: individuals in the same family often contracted it in different years. Another was that the majority of cases, regardless of the year, first presented symptoms during the summer months, June through September. A third finding was that 25 percent of the patients Steere interviewed remembered having a strange skin rash from one to several weeks before the onset of arthritis-like symptoms. The descriptions of the rashes were remarkably similar: they had started as a red papule, or small bump, and gradually expanded to form a bull's-eye from 10 to 50 centimeters in diameter. The occurrence of the rash on the chest, abdomen, back or buttocks of most patients suggested that most likely a crawling, rather than a flying, insect or an arachnid had transmitted the disease, although no one could clearly remember being bitten.

Steere concluded on the basis of these findings that he was dealing with a previously unrecognized disease probably caused by a virus and transmitted by an unknown arthropod (the group to which insects, spiders and ticks belong). He named it Lyme arthritis or Lyme disease for the town in which it was first observed. In 1975–76 he began testing sera from victims of Lyme disease for the presence of specific antibodies against 38 known tick-transmitted diseases and 178 other arthropod-transmitted viruses. Not a single test result was positive.

As he carried out research on the disease and its probable causes, however, Steere came across some interesting information. In 1909 a similar phenomenon had been described in Europe. A Swedish physician, Arvid Afzelius, described an expanding red skin rash in patients who had been bitten by the tick *Ixodes ricinus*. Afzelius named the rash erythema chronicum migrans (ECM), which literally means "chronic migrating red rash."

ECM sounded remarkably like the bull's-eye rash that had been observed in Lyme disease patients. Although it lacked the arthritis-like symptoms characteristic of Lyme disease, Steere concluded that ECM and Lyme disease might be closely related and have similar modes of transmission.

European physicians had successfully treated ECM with penicillin, indicating that the most likely agent of infection was not a virus but a bacterium. Yet when fluid was removed from the joints of Lyme disease patients and cultured, no microorganisms could be found. Meanwhile the number of cases of Lyme disease continued to climb.

Finally, in 1977, nine patients affected by the ECM rash that year remembered having been bitten by a tick at the site of the rash. One of them had removed the tick and saved it, and was able to give it to Steere for identification. The tick, barely larger than a pinhead, was a dark brown, hard-bodied animal that might easily be mistaken for a scab or a piece of dirt. It was not surprising that it had taken Steere and his group almost two years to locate it.

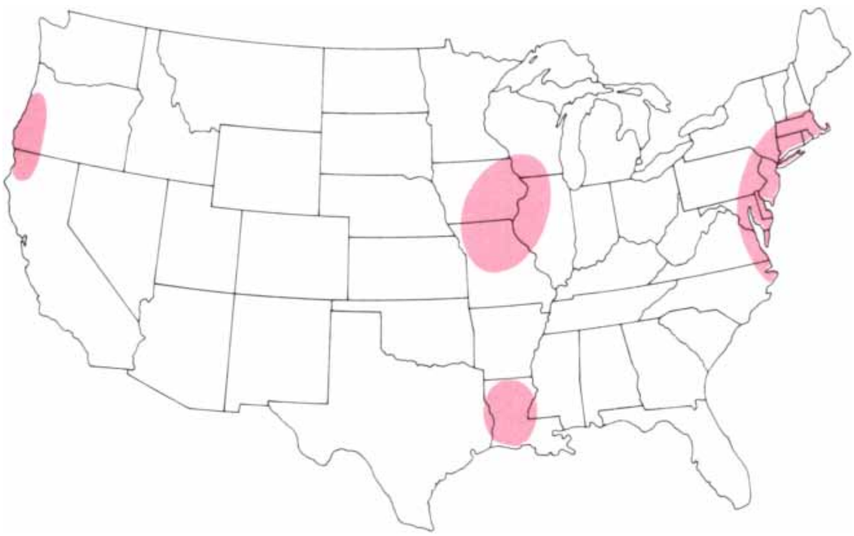
The tick was identified by Andrew Spielman of the Harvard School of Public Health as *Ixodes dammini,* a species closely related to *I. ricinus,* the tick responsible for European ECM. Now that *I. dammini* was identified, investigators working on Lyme disease hoped to isolate the actual agent of infection. First they had to be certain that the tick was indeed the vector for Lyme disease. If the distribution of *I. dammini* in the wild corresponded to the outbreak of Lyme disease, the circumstantial evidence linking the two would be strong indeed.

Biologists at Yale set out animal traps on both sides of the Connecticut River in order to map the distribution of *Ixodes* along the river and at the same time find out on which mammalian species it was feeding. The distribution of ticks was just as they hoped: the dog tick *Dermacentor variabilis* was equally common on both sides of the river, but *I. dammini* was 12 times more abundant on the east side—near Lyme, Old Lyme and East Haddam, where Lyme disease was by that time known to be endemic. The workers were convinced that *I. dammini* must be the primary vector in the transmission of Lyme disease.

Still the agent responsible for both ECM and Lyme disease remained elusive. Repeated cell cultures and microscopic examinations of the tick's internal organs failed to reveal the presence of a bacterium or any other pathogen. Then, in the fall of 1981, a fatal case of Rocky Mountain spotted fever, a rickettsial disease transmitted by the dog tick, was reported on Shelter Island, off the coast of eastern Long Island. The New York State Department of Health sent a team of biologists to the island to collect live ticks. Because the normal vector, *Dermacentor variabilis,* is not found in the fall, adult *Ixodes dammini* were collected instead and were sent to the Rocky Mountain Laboratories in Hamilton, Mont., for study.

There Willy Burgdorfer, an international authority on tickborne diseases, squashed the digestive tract of the *Ixodes* tick and examined it by dark-field microscopy. To his surprise he found the gut teeming not with the rickettsiae that cause Rocky Mountain spotted fever but with long, irregularly shaped spirochete bacteria. Burgdorfer knew that *I. dammini* had been implicated as the probable vector for Lyme disease, and he also knew that the spirochetes were not the infectious agent of Rocky Mountain spotted fever. He wondered whether these bacteria could be the cause of Lyme disease. Fortunately Alan G. Barbour, then at the Rocky Mountain Laboratories, was able to grow the spirochetes in pure culture and obtain them in sufficient quantities for experimentation.

HEALTH WARNING posted near the Atlantic coast in Amagansett, N.Y., is one of many such notices distributed throughout tick-infested areas on Long Island. Lyme disease is rapidly increasing in wooded areas where the mammals on which the ticks feed are abundant. People who develop symptoms characteristic of Lyme disease are advised to seek prompt treatment from a physician.



LYME DISEASE continues to spread throughout the U.S.; it has been reported in 25 states. Most cases are found along the North Atlantic coast from Massachusetts to North Carolina, in Minnesota and Wisconsin, in Texas and along the Pacific coast in California and Oregon. Scattered cases, however, have also been reported from Arkansas, Florida, Georgia, Indiana, Kentucky, Maine, Michigan, Montana, Nevada, New Hampshire, Ohio, Tennessee, Utah and Vermont. The areas of heaviest incidence are indicated in color.

Because patients exposed to an infectious agent have antibodies in their serum that react to the agent, an antibody test can be a good indicator of infection. Serum samples from New York patients infected with Lyme disease were sent to Burgdorfer, who tested them for the presence of antibodies against the spirochetes. Unlike the earlier series of tests conducted by Steere, the results this time were positive: the sera showed a pronounced antibody response to the bacteria, indicating that the patients had in fact been infected by the spirochete.

A more direct test of the pathogenicity of the spirochete was carried out on rabbits. Spirochete-infected ticks were placed on the shaved skin of albino rabbits, where they could be observed feeding on the blood of their hosts. After several weeks lesions similar to the ECM rash appeared, and microscopic examination of the skin at the site of tick attachment revealed the presence of live spirochetes.

From this point investigations proceeded rapidly. By the summer of 1982 spirochetes had been isolated from the blood, skin and cerebrospinal fluid of Lyme disease victims by investigators at the New York State Department of Health and at Yale. Russell C. Johnson and his colleagues at the University of Minnesota Medical School studied the Lyme disease spirochete and determined, on the basis of its DNA, that it was a new species in the genus *Borrelia*. In 1984, to honor its discoverer, Burgdorfer, they named it *Borrelia burgdorferi*.

B. burgdorferi is a typical spirochete: it is a unicellular, loosely coiled, left-handed helix (that is, it coils in a counterclockwise direction). Its length varies but averages 30 micrometers (thousandths of a millimeter), with seven turns of the coil. Like most spirochetes, it is small and difficult to detect: the diameter of the cell ranges from .18 to .25 micrometer, allowing it to pass through many filters designed to retain bacteria.

Once *B. burgdorferi* had been conclusively identified as the agent of Lyme disease it was possible to track its distribution in nature. Edward M. Bosler of the New York State Department of Health found the spirochete in the tissues of several mammals, including field mice, voles and deer, as well as in all stages of *I. dammini*.

Detection of the spirochete in mammalian tissue is difficult. Not only is the spirochete extremely small but also it is normally present in very low numbers. The preferred method of detection therefore depends on fluorescein-labeled antibodies specific for *B. burgdorferi*. These bind to the spirochetes and fluoresce on illumination with ultraviolet light, making it possible to detect the presence of even a few spirochetes. Such studies indicate that *Borrelia* travels widely once it enters the bloodstream: it has been detected in the eyes, kidneys, spleen, liver, testes and brain of nonhuman mammalian hosts, and also in several species of passerine birds. (The geographic distribution of Lyme disease suggests that *Borrelia* spreads when ticks infested with the bacterium attach themselves to migratory birds.)

Borrelia burgdorferi can be detected in the gut of *I. dammini* by dark-field microscopy or by removing the contents of the tick's gut and growing the spirochetes in culture. Surveys along the North Atlantic coastline indicate that in highly endemic areas from 80 to 90 percent of *Ixodes* ticks have *B. burgdorferi* in their gastrointestinal systems. In contrast, only 3 percent of all *Ixodes* tested on the West Coast harbor the spirochete, a finding that correlates well with the much lower incidence of Lyme disease there.

The life cycle of *I. dammini* normally spans two years. Eggs are deposited in the spring and hatch into free-living larvae a month later. During the first summer the larva feeds once (for a period of two days) on the blood of a host and then enters a resting stage coincident with the onset of cold weather in the fall. The following spring the larva molts, enters a second immature stage called the nymphal stage and again attaches itself to an animal host, this time to feed for three or four days. Although the larvae and nymphs attack a variety of vertebrates, the majority of the ticks in these age cohorts are found on the white-footed mouse, *Peromyscus leucopus*. It is at this stage that ticks are most likely to attach themselves to humans.

At the end of the summer nymphs molt into the adult stage. They can be found in brush about one meter above the ground, where they can easily attach themselves to larger mammals. Like the immature ticks, the adults feed on a variety of mammalian hosts, but in the northeastern U.S. they are found predominately on the white-tailed deer, *Odocoileus virginianus*. The adult ticks mate on the host soon after the female attaches herself to it. Only the females overwinter; the males die soon after mating. It is not known where the eggs are deposited, but they hatch in the spring—and the entire cycle is repeated.

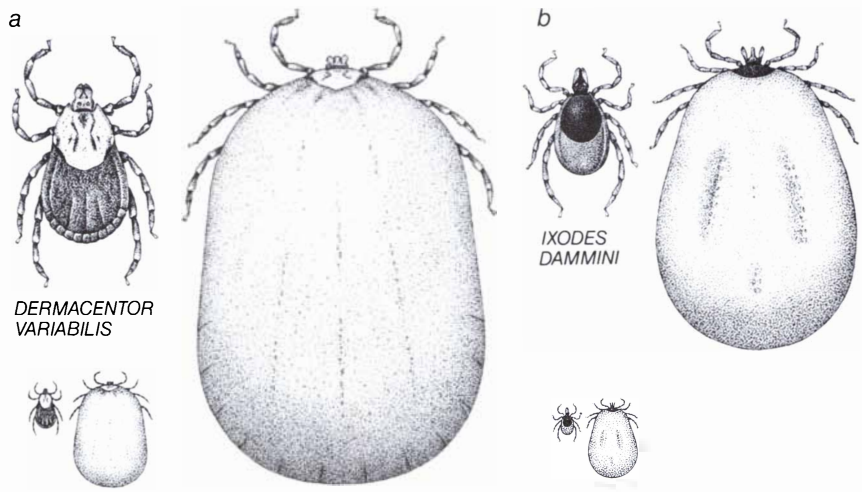
Anyone who lives in or visits an area where Lyme disease is endemic is susceptible to the condition. Lyme disease is indiscriminate: it affects both sexes and all age groups. Although disproportionate numbers of children have been affected, this may simply reflect the fact that children spend more time playing in wooded areas than adults do. In addition, people who have outdoor animals are known to be at greater risk for contracting Lyme disease, but it is not clear whether this reflects a more outdoor lifestyle or the fact that the humans are bitten by ticks attached to the fur of their domestic animals. Lyme disease is also rapidly becoming a veterinary problem: dogs and horses in endemic areas have developed debilitating joint problems that veterinarians believe are caused by *Borrelia burgdorferi.*

Clinically Lyme disease can be divided into three stages. The first and most obvious stage is characterized by the erythema chronicum migrans rash, which develops from two to 30 days after an individual has been bitten. The rash is frequently accompanied by profound fatigue, fever, chills, headache and backache. In some patients, however, these symptoms, including ECM, fail to appear. In from 25 to 50 percent of cases secondary lesions appear at various sites on the body. Because these lack distinct red papules at their center, they probably reflect the spreading of spirochetes by way of the blood, rather than additional tick bites.

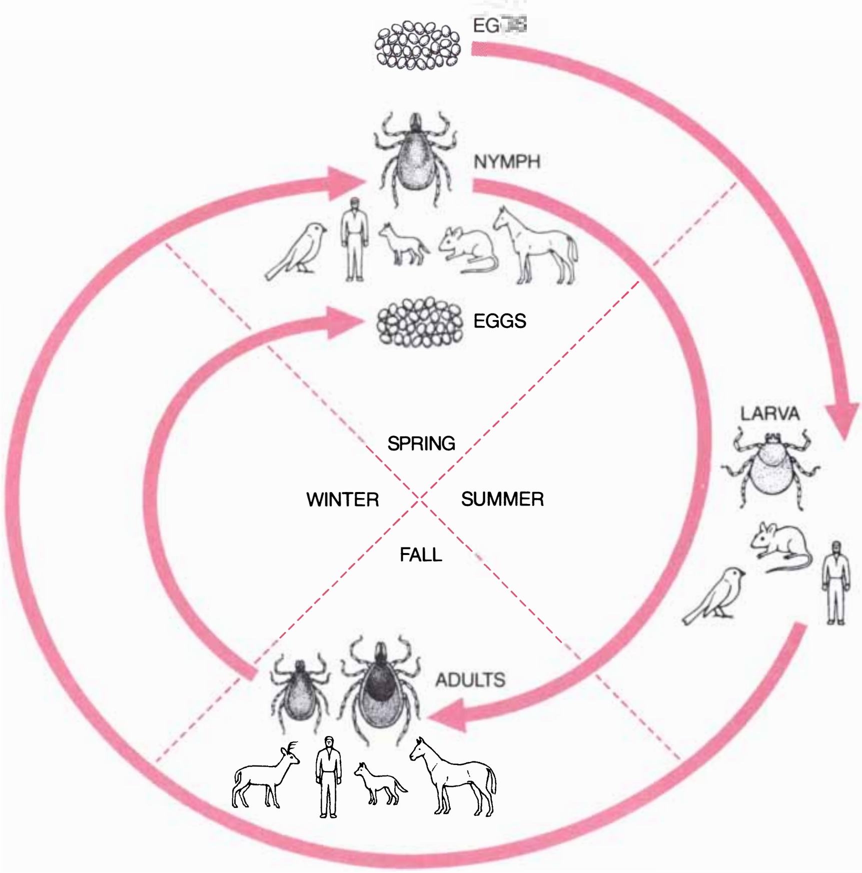
The second stage (also not always expressed) is marked by neurological complications and migratory musculoskeletal pain. Approximately 5 percent of patients develop cardiac difficulties lasting for from three days to six weeks. These patients experience palpitations, dizziness or shortness of breath associated with irregular electrical impulses to the heart (atrioventricular block), and some may require temporary pacemakers.

The third stage typically involves the onset of arthritis. Joint problems characteristic of rheumatoid arthritis occur in about 60 percent of Lyme disease patients who have not been treated, generally within several months but not more than two years after the onset of ECM. Attacks of arthritis usually last from a few days to a few weeks at a time and primarily affect the knees (which can lead to difficulty in walking) and other large joints.

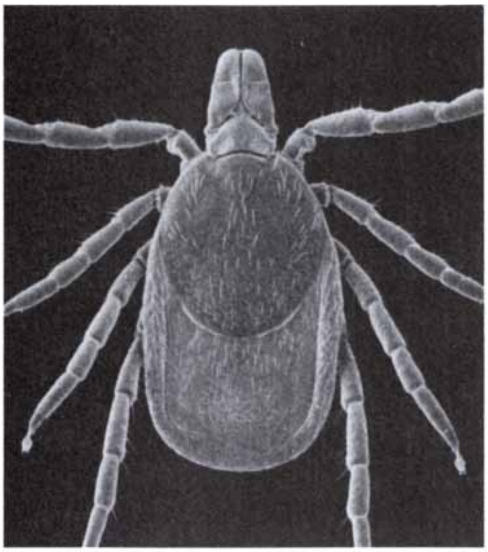
The nervous system may be involved in all stages of Lyme disease: live spirochetes have been detected in the cerebrospinal fluid and brain tissue of patients diagnosed as having the disease. The episodic, excruciating headaches and neck pain experienced during the first stage are believed to result from irritation of the meninges, the membranes surrounding the brain. During the second stage 15 percent of patients develop more severe neurological complications, including meningitis, inflamed nerve roots in the neck and Bell's palsy: a paralysis of the seventh cranial nerve, which controls many facial muscles. Some patients experience increased sensitivity of the

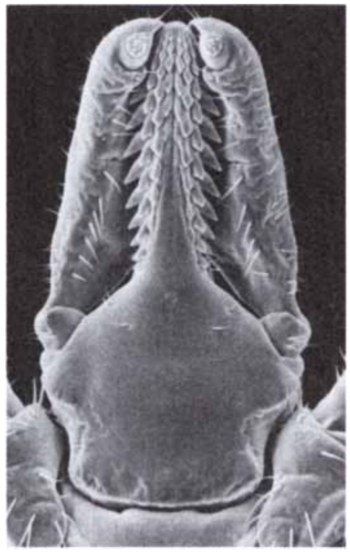


COMMON DISEASE-CARRYING ticks in the U.S. are the dog tick *Dermacentor variabilis* (*a*), which transmits the bacterium that causes Rocky Mountain spotted fever, and the deer tick *Ixodes dammini* (*b*), which transmits the Lyme disease spirochete. Both tick species feed on the blood of mammalian hosts, including human beings, and may triple in size following a meal. The bottom drawings show the actual size of each tick species and compare ticks that have not fed recently (*left*) with ticks that are engorged by a meal (*right*). The top drawings (enlarged four times) show greater anatomical detail.



LIFE CYCLE of *I. dammini* in New England lasts for two years. Eggs are deposited in the spring and the larvae (which have only six legs and are practically invisible to the unaided eye) emerge several weeks later. They feed once during the summer, usually on the blood of small mammals such as mice. Larvae molt the following spring into slightly larger eight-legged nymphs, which also feed once during the summer—on mice or larger mammals such as dogs, deer or human beings—before molting into adults in the fall. Adults attach themselves to a host, usually the white-tailed deer, where they mate. The males die shortly thereafter, but the females continue to feed to obtain the protein necessary for egg development. Females lay their eggs and die, and so the cycle is repeated.





SCANNING ELECTRON MICROGRAPHS of a female *I. dammini* show the animal, enlarged 26 diameters, in dorsal view (*left*) and a close-up of the head, enlarged 120 diameters, in ventral view (*right*). The long, segmented legs have claws and adhesive pads at their tips that help the tick cling to its host while feeding. The head consists of a small cranium and a large proboscis, called the hypostome, surrounded by sensory palps. The hypostome drills through the skin of a host until it reaches a capillary and then draws blood out of the host and into the tick. The serrations on the hypostome's surface help to anchor it in place once it has pierced a blood vessel. The palps (one on each side) are sensory organs, covered with hairs and other receptors that enable the tick to locate a suitable warm-blooded victim. The photographs were made by Stanley F. Hayes, Willy Burgdorfer and M. D. Corwin at the Rocky Mountain Laboratories in Hamilton, Mont.

FIRST STAGE of Lyme disease, observed in 60 percent of the individuals bitten by infected *Ixodes* ticks, is very pronounced. It appears as a large bull's-eye rash, which expands radially from the site of the tick bite (seen here as a small red papule) and is noticeably swollen at its center. The rash on this patient's back measures 14 centimeters in diameter. The photograph was provided by the New York State Department of Health.

skin to touch or changes in temperature. During the third stage a small percentage of patients also suffer from somnolence, loss of memory, mood swings and an inability to concentrate.

Fortunately Lyme disease can be treated successfully at any stage with broad-spectrum antibiotics administered orally, including penicillin, tetracycline and erythromycin. Current studies suggest that cephalosporin antibiotics are also effective. Treatment during the first stage greatly reduces the likelihood of developing neurological, cardiac or arthritic complications. Even if it is left untreated until the third stage, Lyme disease can still be eradicated in most patients by antibiotic therapy, although hospitalization and intravenous administration of the antibiotics may be necessary at this stage.

Physicians who treat patients with Lyme disease have observed an unusual phenomenon. Immediately following antibiotic therapy there is a temporary exacerbation of symptoms. This phenomenon, known as the Jarisch-Herxheimer reaction, was first seen in syphilis patients who were treated with mercury ointments in the 16th century. Syphilis is also caused by a spirochete, *Treponema pallidum,* and it shares many symptoms with Lyme disease including rashes, joint involvement and neurological complications. The Jarisch-Herxheimer reaction has also been observed following treatment of other spirochete infections, such as relapsing fever. The reaction has given us a major clue for elucidating the pathogenesis of Lyme disease.

It is interesting that Lyme disease patients experience an extensive array of symptoms in spite of the presence of only a small number of spirochetes. Two theories of Lyme disease pathogenesis have been advanced to explain this fact; both involve the immune system and both appear to be operative. The first theory holds that immune complexes, which consist of antigens from the spirochete and antibodies and complement from the human host, accumulate in a patient's joints. This buildup in turn attracts neutrophils (phagocytic white blood cells), which release a variety of enzymes that attack the antigen-antibody complexes. According to this hypothesis, it is the enzymes released by the neutrophils that attack the joint and erode bone and cartilage tissue to cause arthritis-like symptoms.

Work done in our laboratory at the State University of New York at Stony Brook suggests a second hypothesis. We believe the pathological effects of spirochetes are amplified not only by neutrophil-secreted enzymes but also by the immune-system mediator called interleukin-1 (IL-1).

IL-1 is a protein with a molecular weight of 17,000 daltons that is synthesized primarily by the phagocytic white blood cells called macrophages. It is a regulator of the body's immune response and acts as the molecular orchestrator of nonspecific defense mechanisms against a variety of environmental insults. It coordinates the body's reaction to bacterial infection and trauma by regulating the onset of fever, the release of neutrophils from bone marrow and the proliferation of fibroblasts (connective-tissue cells).

One of the most powerful stimuli for the release of IL-1 is a lipopolysaccharide (LPS), a complex of sugar and lipid molecules, that is found in the outer envelope of the cell wall of all gram-negative bacteria. Because *Borrelia burgdorferi* is a gram-negative bacterium, we speculated that it might contain LPS that could trigger the release of IL-1, which in turn would exert powerful local and systemic effects on the human body.

We approached our hypothesis in several ways. First we needed to demonstrate that the cell wall of *B. burgdorferi* does indeed contain LPS. To do this we cultured the spirochetes in a special growth medium and then harvested large numbers of them, which we tested for the presence of LPS. Once we confirmed that LPS was present, we applied a chemical extraction method to isolate the LPS. At the end of the procedure we had obtained pure extracts of *Borrelia* LPS with which to test our theory.

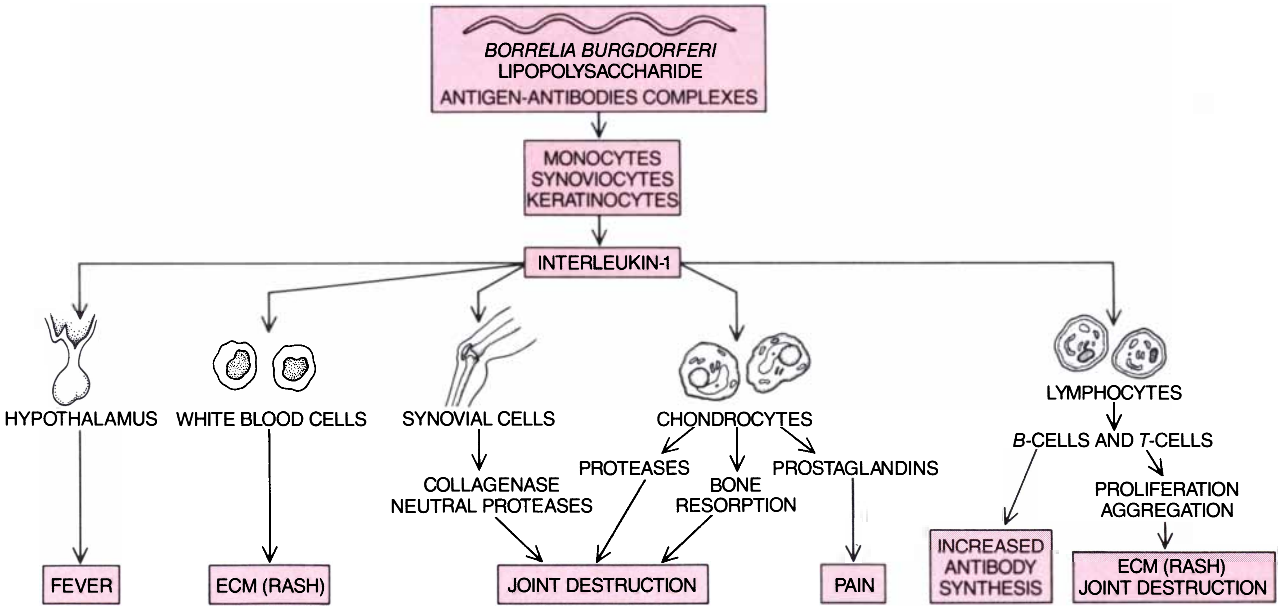
We carried out two series of experiments. In the first series we injected both humans and rabbits with pure LPS. The results were striking: rabbits that received the LPS intravenously became feverish within a few hours, and rabbits and humans who received injections of LPS under the skin developed an ECM-like rash.

In our second series of experiments we wanted to observe both the in vivo and the in vitro responses of human macrophages (which synthesize IL-1) to *B. burgdorferi*. We cultured the two cell types together and found that the macrophages secrete large quantities of IL-1 in the presence of the spirochete. Then we examined thin sections of rabbit skin that had been injected intradermally with either IL-1, *B. burgdorferi* or LPS. All were capable of inducing an acute inflammatory response in the skin. We therefore believe that both the skin lesions and the fever can be explained by the release of IL-1.

We also believe that IL-1 released in response to *Borrelia* is directly responsible for the arthritis characterizing the third stage of Lyme disease. When IL-1 is put into culture with cells from the synovium, the inner lining of a human joint such as the knee, it stimulates production of two compounds: the enzyme collagenase and a substance called prostaglandin. Both contribute to arthritis. Collagenase does so by degrading collagen, the primary component of the connective tissue in joints; the degradation leads to a pattern of erosion much like the one seen in severe cases of Lyme arthritis. Prostaglandin contributes to arthritis by promoting pain. Our experiments also show that synovial cells from Lyme disease patients release IL-1 when the cells are exposed to either *Borrelia* or LPS extracts from *Borrelia*.

The Jarisch-Herxheimer reaction experienced by some Lyme disease patients is consistent with our theory: antibiotic treatment kills large numbers of the spirochetes at the same time, releasing large quantities of LPS into the bloodstream and triggering the production of IL-1.

Much remains to be learned about Lyme disease. Currently there are no effective control programs; the best protection against it is caution in tick-infested areas and sensitivity to the early warning signs of the disease. Although our research on the role of IL-1 may not contribute to the control of the disease, we believe our work may have significance beyond this single illness. By examining the physiological response to infection we have shown how a single infectious agent can set off a chain reaction, and how the localized production of a powerful biological mediator such as IL-1 may account for the inflammatory changes characteristic of Lyme arthritis. We hope this will be a useful model for other arthritic diseases of unknown origin.



LIPOPOLYSACCHARIDES from the spirochete *Borrelia burgdorferi* trigger the release of interleukin-1, which plays a major role in the pathogenesis of Lyme disease. The IL-1 acts on various organs and cells to produce symptoms such as the rash, fever and arthritis that characterize stages one, two and three of Lyme disease. The spirochete (*top*) is enlarged roughly 1,000 diameters.

Cold Nuclear Fusion

The electronlike particles called muons can catalyze nuclear fusion reactions, eliminating the need for powerful lasers or high-temperature plasmas. The process may one day become a commercial energy source

by Johann Rafelski and Steven E. Jones

Ordinarily the mention of nuclear fusion calls forth images of enormous magnets, powerful lasers and plasmas hotter than the center of the sun. These create the extreme conditions in which pairs of hydrogen atoms fuse, or join, creating helium and giving off energy that could be used to produce electric power. A less familiar but perhaps more promising kind of fusion, known as muon-catalyzed fusion, has a very different aspect: it circumvents the need for high temperatures entirely.

Muon-catalyzed fusion, which is also called cold fusion, can take place rapidly at room temperature in a simple chamber containing certain kinds of hydrogen known as deuterium and tritium. Particles known as negative muons are introduced into the chamber, and they form unusually tight bonds between the nuclei of some of the hydrogen atoms. The muon-bonded hydrogen nuclei then fuse, ejecting the muons, which can go on to catalyze other fusion reactions. The other atoms in the gas are essentially unaffected, except that each fusion increases the temperature of the gas as a whole. The heat from muon-catalyzed reactions might someday drive turbines for generating electricity.

Muon-catalyzed fusion is not limited to room temperature. The process has been made to work in liquid and solid forms of hydrogen at temperatures as low as 13 degrees Kelvin (degrees Celsius above absolute zero) and in gases at temperatures as high as 530 degrees C. Recent research suggests the reactions involved should be most efficient at about 900 degrees C.

The entities at the heart of the process, muons, are short-lived elementary particles. They are found in nature in secondary cosmic rays, which are produced when primary cosmic rays collide with the upper atmosphere. Muons can be made artificially by causing a fast-moving beam of ions (electrically charged atoms) from a particle accelerator to collide with a sample of ordinary matter such as carbon. The collisions produce particles called pions, which quickly decay to make muons in a process much like the one that takes place when primary cosmic rays strike the atmosphere.

Muons can carry positive or negative electric charge. A negative muon has properties quite similar to those of the electron but is about 207 times more massive. As we shall discuss, it is the muon's large mass that enables it to catalyze fusion reactions.

It is not yet possible to build a cold-fusion reactor that produces more energy than is required to operate it. A major stumbling block results from the short lifetime of muons, which decay, on the average, in about two microseconds (millionths of a second) after they are created. In that short time each muon must catalyze enough reactions so that the reactor as a whole can power the accelerator that generates muons. Until recently that goal seemed remote. In the past few years, however, theoretical and experimental advances have shown that under the proper conditions a muon may catalyze hundreds of times more reactions before it decays than had seemed possible. It is now conceivable that cold fusion may become an economically viable method of generating energy.

The possibility that negative muons could catalyze nuclear fusion was suggested on theoretical grounds by F. C. Frank and Andrei D. Sakharov in the late 1940's. The first experimental observation of muon-catalyzed fusion came by chance a decade later. Luis W. Alvarez and his colleagues from the University of California at Berkeley, analyzing the results of an unrelated experiment, noticed certain unusual patterns on films that recorded the tracks of particles through a bubble chamber. The Berkeley workers had not heard of the earlier suggestions by Frank and Sakharov, but, assisted by Edward Teller, they deduced that the patterns recorded the products of muon-catalyzed fusion reactions.

The discovery caused great excitement at first. As Alvarez remarked in his Nobel-prize acceptance speech in 1968, "We had a short but exhilarating experience when we thought we had solved all of the fuel problems of mankind for the rest of time." Unfortunately later calculations showed the reactions he had observed were too slow to generate useful energy: the average muon had time to catalyze at most only a single fusion before it decayed, producing too little energy to supply muons for later reactions. Most investigators went on pursuing other methods of stimulating fusion.

Nevertheless, some workers continued to study muon-catalyzed fusion. Investigators found that muons can catalyze fusion through several processes other than those first observed. The Alvarez group had seen reactions involving only deuterium and ordinary hydrogen, but it became clear that muons could catalyze reactions involving deuterium and tritium much more quickly. V. P. Dzelepov and his colleagues at the Joint Institute for Nuclear Research in Dubna found experimentally that the rate of one such process depended strongly on temperature, yielding more fusions per muon at higher temperatures.

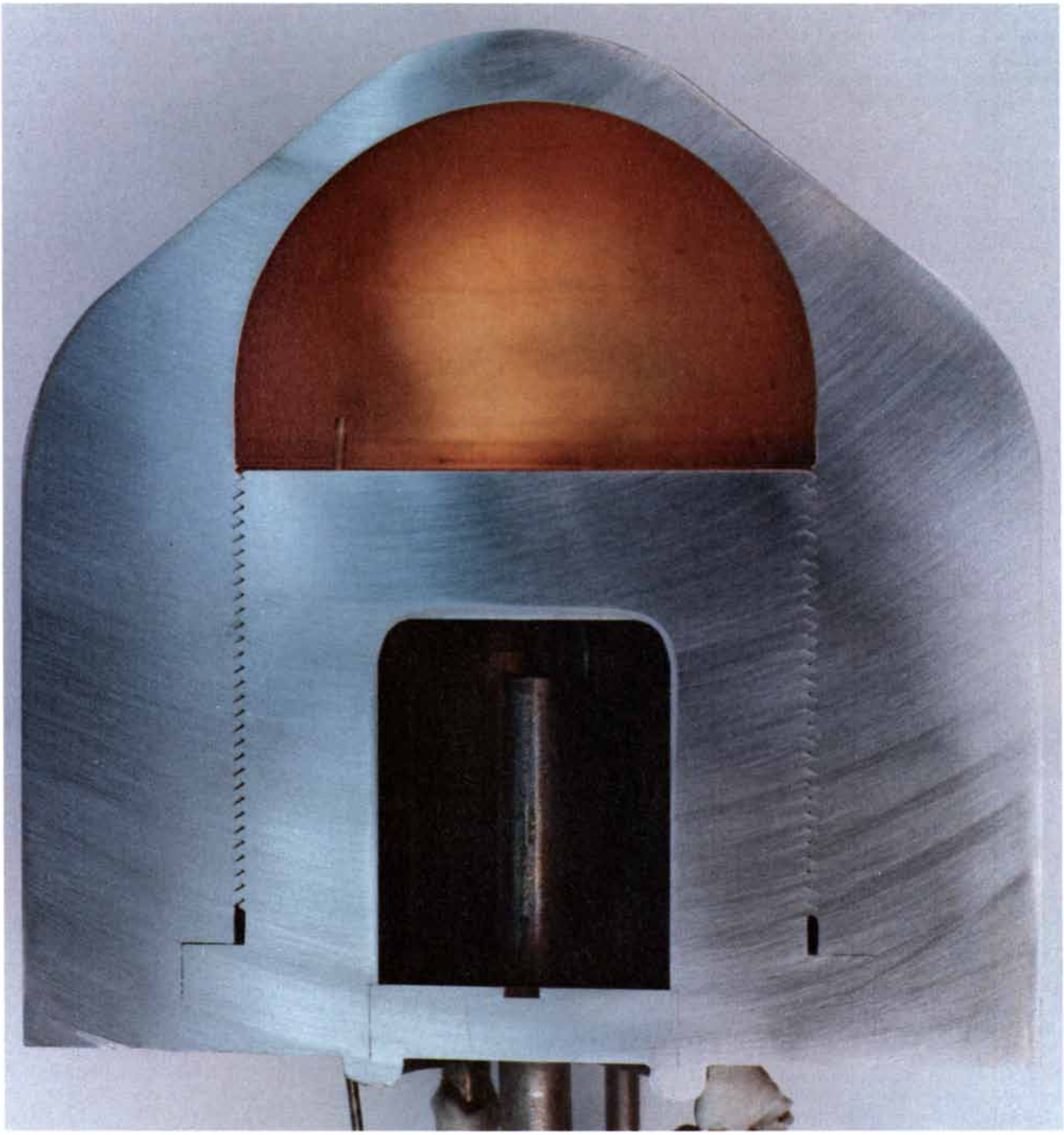
A model that explained the results was developed by E. A. Vesman of the Estonian Academy of Sciences in 1967. In 1977 S. S. Gershtein, L. I. Ponomarev and their co-workers in Dubna built on Vesman's model to predict that at certain temperatures, and in certain high-density mixtures of deuterium and tritium, muon-catalyzed fusion could occur much more rapidly—perhaps thousands of times more rapidly—than the processes first seen. The prediction led one of us

(Jones) and his colleagues to do experimental studies of muon-catalyzed fusion in deuterium and tritium.

What happens when a beam of negative muons enters a vessel of deuterium and tritium? To understand what ensues, it is first necessary to understand how deuterium and tritium differ from ordinary hydrogen.

An ordinary hydrogen nucleus is a proton: a massive, positively charged particle. In deuterium the nucleus is a proton bound to a neutron, which is slightly more massive than the proton and has no electric charge. In tritium the nucleus is a proton and two neutrons. In each kind of hydrogen atom the nucleus is orbited by an electron (which is negatively charged).

Like ordinary hydrogen atoms, deu-



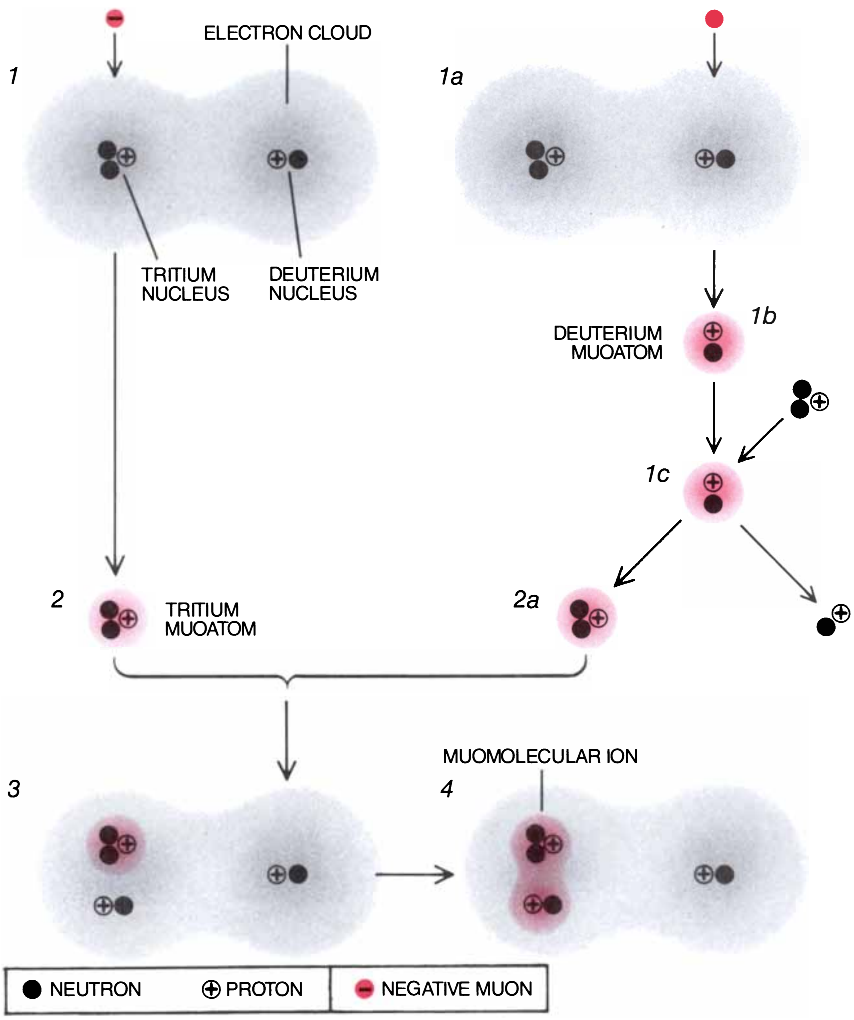
EXPERIMENTAL REACTION VESSEL in which cold nuclear fusion reactions take place is shown in cross section. The reactions occur in a gold-lined hemisphere (*top*) measuring two inches in diameter. A thin pipe fills the vessel with deuterium and tritium (forms of hydrogen). A beam of negatively charged muons is aimed from above at the vessel's cone-shaped top; it enters the reaction chamber by passing through the vessel's stainless-steel casing. The muons act as a catalyst, causing deuterium and tritium atoms to fuse, releasing energy and producing helium and neutrons. Liquefied gases are fed through a thick pipe into the hollow component screwed into the base of the reaction chamber (*bottom*). They cool the vessel for studies of how the rate of muon-catalyzed fusion reactions varies with temperature. Similar reaction vessels, built on a larger scale, could someday produce energy in a muon-catalyzed-fusion power plant. The vessel shown here was developed by Augustine J. Caffrey and Kenneth D. Watts of the Idaho National Engineering Laboratory, with help from Michael A. Paciotti and H. Richard Maltrud of the Los Alamos National Laboratory.

terium and tritium atoms combine in pairs to form molecules. In each molecule two nuclei are bound together by electrons, which form a unifying "cloud" between and around the positively charged nuclei. In a mixture of deuterium and tritium some molecules consist of two deuterium atoms, some of two tritium atoms and some of a deuterium atom bound to a tritium atom. Nuclei bound together in a molecule are relatively far apart; the distance between them is about 30,000 times greater than the radius of the nucleus itself.

A negatively charged muon traveling at high speed through a mixture of deuterium and tritium is slowed by collisions with the electrons that bind the molecules. The collisions usually knock the electrons out of their molecular orbits. Soon the muon is moving slowly enough so that as it displaces an electron it is captured into an orbit similar to the electron's. Almost immediately, however, the muon tumbles into a very tight orbit around one of the nuclei in the molecule.

It is because of the muon's large mass that it can orbit the nucleus so closely. In general, negatively charged particles such as electrons and muons can follow only certain stable orbits around nuclei. According to the laws of atomic physics, the size of a particle's smallest stable orbit is roughly proportional to the inverse of its mass. Because the muon is about 200 times more massive than the electron, it can follow an orbit about 200 times smaller than the electron's.

When the muon enters the tight orbit, the molecule it has invaded breaks apart; the muon and the nucleus it orbits—a "muoatom"—emerge with a small velocity. The muoatom can include a nucleus of deuterium or tritium. Because a tritium nucleus is more massive, it can bind the muon more strongly than a deuterium nucleus can. For this reason muons that originally bind to deuterium nuclei usually transfer, during some later collision, to a tritium nucleus. The events leading to the formation of a tritium muoatom—the slowing of the muon, its binding to a nucleus and any transfer from a deuterium to a tritium nucleus—can be made to take place in less than a thousandth of the muon's lifetime.



INITIAL STEPS of muon-catalyzed fusion bring a deuterium nucleus and a tritium nucleus into unusually close proximity. First a muon invades a molecule consisting of two nuclei held together by an electron cloud. The muon may collide with a tritium nucleus (*1*). Then, as a quantum-mechanical consequence of the muon's high mass, the muon falls into a very tight orbit around the nucleus, forming what is called a tritium muoatom (*2*). Alternatively (*1a*), the muon may collide with a deuterium nucleus, forming a deuterium muoatom (*1b*). In a later collision (*1c*) the muon can be transferred to a tritium nucleus, forming a tritium muoatom (*2a*). The tritium muoatom then penetrates the electron cloud of another molecule and collides with a deuterium nucleus (*3*). The tritium nucleus, the deuterium nucleus and the muon then combine within the molecule to form a muomolecular ion, in which the muon holds the nuclei together much as an electron binds nuclei in an ordinary molecule (*4*). Because of the muon's high mass, the nuclei in a muomolecular ion are approximately 200 times closer together than the nuclei in an ordinary molecule.

The events that take place in this first part of the muon-catalyzed fusion process are governed by the laws of atomic physics. The next steps in the sequence are governed by the laws of molecular physics.

The tritium muoatom is a small, electrically neutral object that wanders freely through the gas, easily penetrating the electron clouds of gas molecules. When it comes close to a deuterium nucleus in an ordinary molecule, it can combine with the nucleus to form what is called a muomolecular ion. In a muomolecular ion the muon binds the two nuclei (in this case a tritium nucleus and a deuterium nucleus) much as electrons bind the nuclei of a normal molecule. The muon can pull the two nuclei much closer together, however, because it is so much more massive than the electron. The distance between the nuclei in a muomolecule is usually about 200 times less than the distance between the nuclei in an ordinary molecule.

The small muomolecular ion can nest within the electromolecule (that is, the ordinary molecule). Because the muomolecular ion has a net positive charge (two positively charged protons and only one negatively charged muon), it takes the place of one of the positively charged nuclei in the electromolecule. Then the muomolecular ion and the remaining deuterium or tritium nucleus are bound to each other by electrons, just as the two nuclei in the muomolecular ion are bound together by a muon.

The speed with which the muomolecular ion forms is one of the key determinants of how many reactions each muon can catalyze. The ordinary molecule that hosts the muomolecular ion is critically important for its rapid formation.

For many years it was thought that the formation of the muomolecular ion is an inherently slow process, for reasons that concern the muomolecular ion's binding energy: the amount of energy the pair of nuclei gives off when it forms a muomolecular ion (which is equivalent to the amount of energy that would have to be added to split the nuclei apart). The muomolecular ion cannot form unless there is some mechanism for carrying away the binding energy.

An electron could serve the purpose; that is, the binding energy might be absorbed by an electron belonging to the electromolecule in which the muomolecular ion forms. The electron would then be ejected from the molecule at high speed. Unfortunately this simple mechanism is also a slow one. According to quantum mechanics, the electron can absorb such a large amount of energy only under certain relatively rare conditions.

There is a much faster mechanism, however: the one, in fact, that was suggested by Vesman in 1967. The mechanism depends on a resonance effect. The molecule hosting the muomolecular ion can vibrate in space. The molecule's vibrational states are quantized: only certain amounts of vibrational energy are allowed. In a sense the molecule is like a xylophone that has only a fixed set of tones. In Vesman's mechanism the energy given off by the formation of the muomolecular ion "rings" one of the vibrational states of the molecule—the molecule as a whole absorbs the binding energy and vibrates as a result.

The binding energies of muomolecular ions are usually about 100 times larger than the energies of normal molecules. The muomolecular ion involved in muon-catalyzed fusion, however, could not form by resonance unless it had a much lower binding energy, in order to match a vibrational energy level of the larger molecule. The muomolecular ion that forms by resonance is therefore quite loosely bound: it has a very low binding energy in relation to typical muomolecular energies. It was Ponomarev's achievement to show, by a series of detailed calculations, that such an unusual, loosely bound state can exist.

The resonance mechanism calls for precise tuning: the energy absorbed by the molecule must be equal to the energy of its vibrational state. The binding energy of the loosely bound muomolecular ion turns out to be slightly less than the energy of a vibrational state in the larger molecule. The muoatom and the deuterium nucleus can supply the extra energy if they have the right amount of kinetic energy when they collide to form the muomolecular ion. The kinetic energy of the colliding particles can be tuned by adjusting the temperature of the gas.

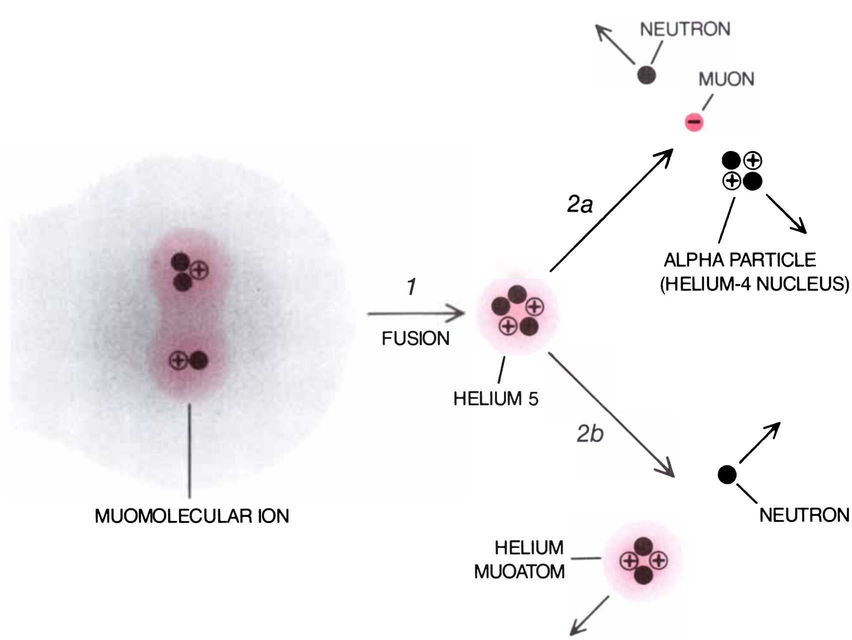
According to the resonance model, then, the temperature of the gas has a great effect on the rate at which muomolecular ions are formed. In 1982 one of us (Jones) and his fellow workers initiated a new experimental program in muon-catalyzed fusion at the Los Alamos Meson Physics Facility (LAMPF). Under the aegis of Ryszard Gajewski of the U.S. Department of Energy, we tested the predictions of the resonance model and found that the rate of muon-catalyzed fusion reactions does indeed vary with temperature in much the way that the theory predicts. To our surprise, we also found that the overall reaction rates were much more rapid than had been predicted; in one case the fast rates allowed about 150 fusions per muon, with the potential for still higher yields. These exciting discoveries helped to rekindle widespread interest in muon-catalyzed fusion.

After the muomolecular ion forms, it falls from its loosely bound state into a more strongly bound one, giving off energy that might be carried away by an electron. The nuclei in the strongly bound muomolecular ion are confined together in a very small volume. They repel each other because they both bear positive charges, but the muon repeatedly draws them together. Eventually the nuclei are united by a quantum-mechanical phenomenon known as tunneling: they pass through the "barrier" of their mutual repulsion until they are so close that the strong nuclear force (the force that binds protons and neutrons in nuclei) asserts itself. Now the laws of nuclear, rather than molecular, physics begin to govern events.

The nuclei fuse to form a single nucleus of helium 5, which has two protons and three neutrons. Soon afterward, this nucleus breaks up into an alpha particle, which is an ordinary helium nucleus (two protons and two neutrons), and a free neutron.

The reaction releases energy, which takes the form of kinetic energy: the alpha particle and the neutron move away from each other at high speed. The muon is usually left behind, and so it is free to repeat the cycle. It acts as a true catalyst for nuclear fusion.

Sometimes, however, the alpha particle created in the fusion reaction captures the negatively charged muon by virtue of its own positive charge, there-



FUSION occurs as a tritium nucleus and a deuterium nucleus in a muomolecular ion combine, forming a helium nucleus and a free neutron and releasing energy. Because the tritium and deuterium nuclei in the muomolecular ion are held close together by a muon, the strong nuclear force can cause them to fuse (*1*) into a helium-5 nucleus orbited by a muon. The helium-5 nucleus breaks up almost immediately. Usually (*2a*) it breaks into a neutron, an alpha particle, or helium-4 nucleus, and a muon. Sometimes, however (*2b*), the positively charged alpha particle captures the negatively charged muon, forming a helium muoatom and preventing the muon from catalyzing another reaction. In any case, the fusion releases kinetic energy: the neutron and the alpha particle move away at high speed.

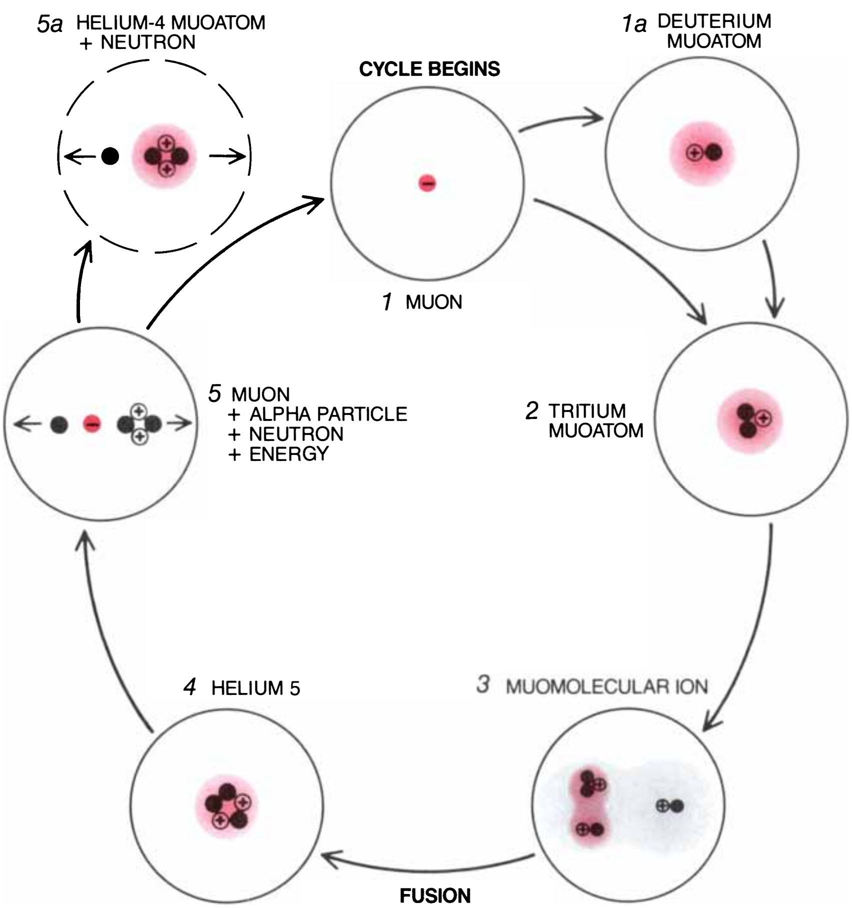
by preventing the muon from going on to catalyze another fusion reaction. Even if the muon is captured, there is still some probability that it will be stripped away by a collision, because the alpha particle is initially moving rapidly through the dense gas in the reaction vessel. In 1981 Gershtein and his collaborators and, independently, Giovanni Fiorentini and Luciano Bracci of the University of Pisa found that about 25 percent of the muons that stick to alpha particles are eventually stripped free. More recently James S. Cohen of the Los Alamos National Laboratory has suggested that perhaps as many as 40 percent are stripped free. There is still much room for further understanding of this process. If the muon sticks to the alpha particle until the pair comes to rest, it will remain bound until it decays.

The frequency with which a catalyzing muon sticks to an alpha particle, breaking a chain of fusion reactions, is the primary obstacle to muon-catalyzed fusion. Unfortunately it is difficult to calculate the sticking probability accurately because the multibody reactions involved are so complex. In 1957 J. David Jackson, then at Princeton University, first recognized that the catalyzing muon could be carried away by the alpha particle in about 1 percent of fusion reactions. He therefore postulated that no more than 100 fusion reactions per muon could ever be achieved.

The LAMPF experiments, which saw well over 100 fusions per muon, led to theoretical reconsideration of the sticking probability as well as to additional experiments. Data from the LAMPF experiments, as analyzed by Allen N. Anderson of Idaho Research, Inc., showed that the alpha-muon sticking probability is about .4 percent or less under certain conditions—less than half the long-standing theoretical value. These surprising results have since been confirmed by experiments at the Swiss Institute for Nuclear Research (SIN), led by Wolfgang H. Breunlich of the Austrian Academy of Sciences and Jean-Claude Petitjean of SIN, as well as by experiments at the Japanese high-energy physics laboratory under the direction of Kanetada Nagamine.

Theoretical readjustments, due in part to David M. Ceperley and Berni J. Alder of the Lawrence Livermore National Laboratory and to Michael Danos of the National Bureau of Standards, Berndt Müller of the University of Frankfurt and one of us (Rafelski), have removed much of the disagreement between the theory of muon-alpha sticking and the experimental results. Important questions are still open, however. In an effort to find answers a major international effort is beginning at the Rutherford Appleton Laboratory under the leadership of Jones and John P. Davies of the University of Birmingham. The investigators will measure the sticking probability directly by using conventional particle-physics techniques to record all the products of fusion reactions.

In parallel with the experimental effort, a major theoretical program has been launched under the leadership of Hendrik J. Monkhorst and Krzysztof Szalewicz of the University of Florida and Lawrence C. Biedenharn, Jr., of Duke University. The workers hope to gain further understanding of the intricate resonance phenomena.



REACTION CYCLE of a cold-fusion reaction starts and ends with a free muon (*1*). A tritium muoatom forms (*2*), sometimes by way of a deuterium muoatom (*1a*). The muoatom combines with a deuterium nucleus to form a muomolecular ion (*3*), which fuses to make a helium-5 nucleus (*4*). The helium 5 splits into an alpha particle and a neutron, releasing energy (*5*). If the muon sticks to the alpha particle (*5a*), the cycle is interrupted because the catalyst has been removed; sticking is a primary obstacle to the development of muon-catalyzed fusion. If the muon does not stick, it is free to begin another cycle (*1*).

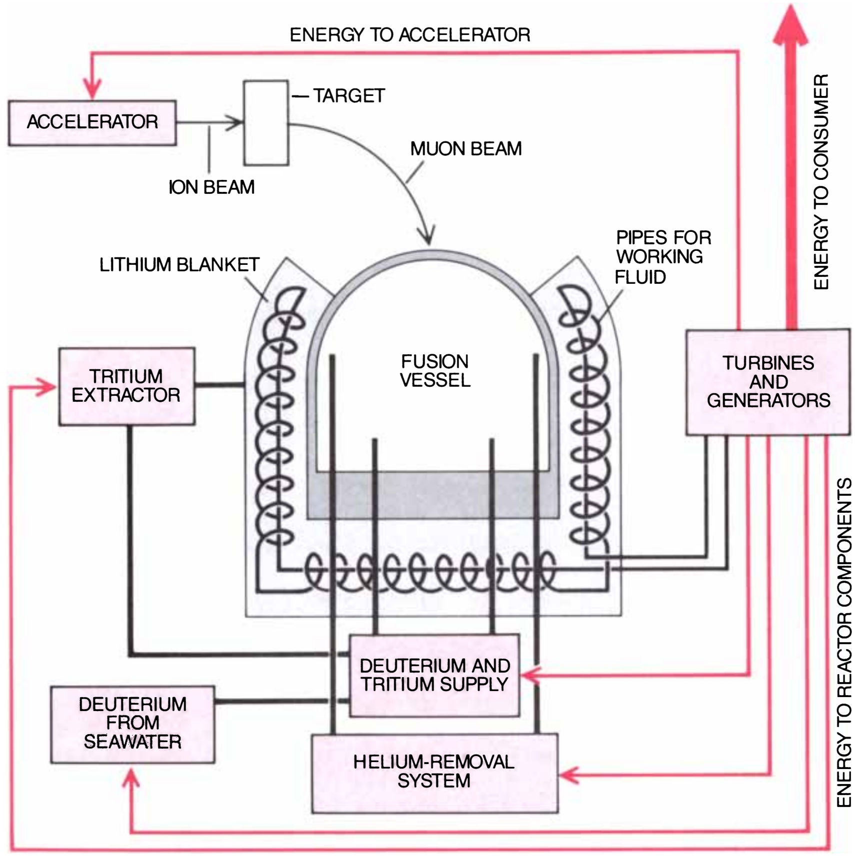
How might muon-catalyzed fusion serve as a practical source of energy? Although there has been little time to evaluate the implications of the recent discoveries, a number of possibilities have begun to emerge. In addition to the number of fusions per muon, the efficiency with which muons can be generated in the first place may determine which scheme can become viable. The energy equivalent of the muon's mass is roughly equal to the energy gained from six fusion reactions; in other words, an ideal, perfectly efficient muon-generating machine could create one muon for every six muon-catalyzed fusions taking place in the separate reaction vessel. Unfortunately the actual energy cost of creating a muon is at least 20 times greater. Advances in the technology of particle accelerators may well bring this cost down.

The cost of producing muons has been explored at CERN, the European laboratory for particle physics, by Magnus Jändel of the University of Uppsala and one of us (Rafelski). Our study shows that a single muon could be created by aiming an ion beam at a vessel of deuterium and tritium at an energy cost equivalent to between 100 and 500 fusion reactions. As the LAMPF experiments have shown, the muon thus produced could go on to catalyze well over 100 fusions.

One concept for generating energy by muon-catalyzed fusion was proposed in 1980 by Yu. V. Petrov of the Leningrad Institute of Nuclear Physics, and was based on the assumption that no more than 100 fusions per muon were possible. Petrov suggested a hybrid of muon-catalyzed fusion and nuclear fission. In addition to producing steam for driving turbines and dynamos directly, the fusion chamber would also serve as a source of neutrons for breeding nuclear fuel to be used in a fission-power plant.

A similar fusion-fission concept was proposed by Marshall N. Rosenbluth, Shalom Eliezer and Toshiki Tajima, then working at the University of Texas at Austin. In this scheme muons for the fusion process would be produced by aiming a high-energy beam of ions directly at the chamber of deuterium and tritium where the fusion reactions are to take place. The muons would then not need to be transported from a separate accelerator to the reaction vessel. As in Petrov's scheme, the muon-catalyzed fusion would provide neutrons for breeding fission fuel.

In addition to these concepts, a pure fusion reactor may also be possible. Such a reactor would have several advantages over fission reactors and fusion-fission hybrids. For one, the "ash" inevitably produced by the fusion process is helium, a harmless gas, rather than radioactive waste. For another, the materials required for fueling fusion reactions are abundantly available in seawater (which contains plentiful deuterium, as well as lithium that can be used to make tritium).



COMMERCIAL COLD-FUSION REACTOR could be built with existing technology. To make the required muons, a particle accelerator directs a beam of ions (charged atoms) at a target made of a substance such as deuterium or lithium. The resulting muon beam is guided into a reaction vessel supplied with deuterium (which can be collected from seawater) and tritium. There the muons catalyze fusion reactions. A purifier removes the helium produced by the reactions. Neutrons from the fusion reactions collide with a blanket of lithium, producing tritium and helium. The tritium is channeled into the fusion vessel and the helium is removed. Heat from the fusion reactions vaporizes a working fluid that is piped through the lithium blanket. The vapor spins a set of high-pressure turbines that run electric-power generators. Some of the electricity powers the particle accelerator, pumps and other components of the reactor. The rest is sold to consumers.

The zoo of elementary particles is a large one, and it is conceivable that some particle other than the muon might also be able to catalyze fusion. As George Zweig of the California Institute of Technology has suggested, free quarks (if they are ever found) could catalyze fusion reactions. Nevertheless, the muon suits the role remarkably well. Each muon can catalyze a long series of reactions, both because the resonance mechanism makes it possible for muomolecular ions to form quickly and because each muon has a good chance of being freed after a fusion event to catalyze another reaction. It also happens that the temperature at which the resonance mechanism should work best—about 900 degrees C., according to Melvin Leon of LAMPF—is also near the temperature at which many energy-generating systems, such as high-pressure steam turbines, are most efficient.

At much higher temperatures the resonance mechanism would ensure that fusion would proceed more slowly: the increased kinetic energy of the colliding muoatom and molecule, when it is added to the binding energy given up by the muomolecular ion, would be more energy than the larger molecule could readily absorb. This has two important implications. First, it means that a muon-catalyzed fusion reactor would not be susceptible to runaway reactions or meltdown. Second, it implies that muon-catalyzed fusion cannot be used as the basis for thermonuclear weapons.

Historically, pioneering research in physics tends to precede applications beneficial to society by one or two generations. The physics of exotic particles is now entering a stage at which applications are emerging; muon-catalyzed fusion is a prominent example.

In the case of muon-catalyzed fusion, the applications that are beneficial to society may themselves produce spin-offs that aid basic research. For instance, in the fusion reactor only negative muons would be needed, but a large number of positively charged muons would also be produced by the muon-generating mechanism and would therefore be available for the purposes of pure research. Ready availability of a large number of positive muons could open the door to fuller understanding of the enigmatic muon itself. Muon-catalyzed fusion thus could foster a close symbiosis of fundamental research and advanced technological applications.

Aside from its possible technological applications, research on muon-catalyzed fusion touches on many areas of modern physics. The processes involved depend on the laws of molecular, atomic, nuclear and particle physics. Research on muon-catalyzed fusion challenges our ability to combine concepts from these diverse fields and deepens our knowledge of each.

Aerodynamics of Wind Pollination

Many plants are almost perfectly engineered to capture pollen from the wind. Cones, flower clusters and other structures channel the airflow—and sperm-producing pollen—toward reproductive surfaces

by Karl J. Niklas

The gusts of pollen that seasonally vex the sinuses of hay-fever sufferers are a by-product of the reproductive strategy of many plant species. Such plants loft showers of pollen grains (sperm-producing spores) into the air. If all goes well, wind currents then convey the grains to other plants of the same species. Pollen that lands on ovules (specialized egg-producing bodies), or on structures that encase them, fertilizes the eggs and enables the ovules to develop into seeds.

Biologists have traditionally considered wind dispersal, the oldest surviving method of pollination, to be more wasteful than the strategies of many other plants, which depend on insects, birds or other animals to transfer pollen. Perhaps one in 1,000 grains of pollen entrusted to the wind reaches the female organ of a target plant. Most of the rest collide with a variety of unreceptive objects, such as leaves, branches, telephone poles or human nasal passages.

Wind pollination might be more wasteful than other methods, but I believe it is less extravagant than it might seem. After all, if the method were truly inefficient, thousands of plant species would not have retained it over the course of evolutionary history or—what is more striking—have recently adopted it. Many gymnosperms (nonflowering seed plants), including the conifers and certain of the palmlike plants known as cycads, are pollinated by wind. So are certain members of the other main group of seed plants: the angiosperms (flowering plants), which evolved some 200 million years after the gymnosperms. Grasses, one of the largest and most successful families of flowering plants, are almost all wind-pollinated. In addition solid evidence now shows that pollen capture is often far from random. My colleagues and I at Cornell University, in collaboration with workers at the University of Arizona and the Fairchild Tropical Garden in Coral Gables, have found that many plants are aerodynamically designed to filter large amounts of pollen from the air.

An understanding of our findings requires some knowledge of the reproductive machinery of seed plants. In gymnosperms the ovule lies unenclosed on modified branches or leaves. ("Gymnosperm" literally means naked seed.) Pollen needs only to land on or near the opening of the ovule—the micropyle—for the pollen grain to mature and ultimately release sperm that can fertilize an egg. In angiosperms, in contrast, the ovule is enclosed within the bulbous part of a flask-shaped body known as a carpel. Pollination is complete when the grain lands on the stigma, the carpel's upper tip. The pollen then extends a long tube in order to reach the egg for fertilization.

In some species of seed plants the sexes are divided: pollen-bearing and ovulate organs grow on different individuals. Most species, however, are bisexual; even so, many do not pollinate their own eggs. (There is a reasonable explanation for this apparent inefficiency: cross-pollination preserves genetic diversity.)

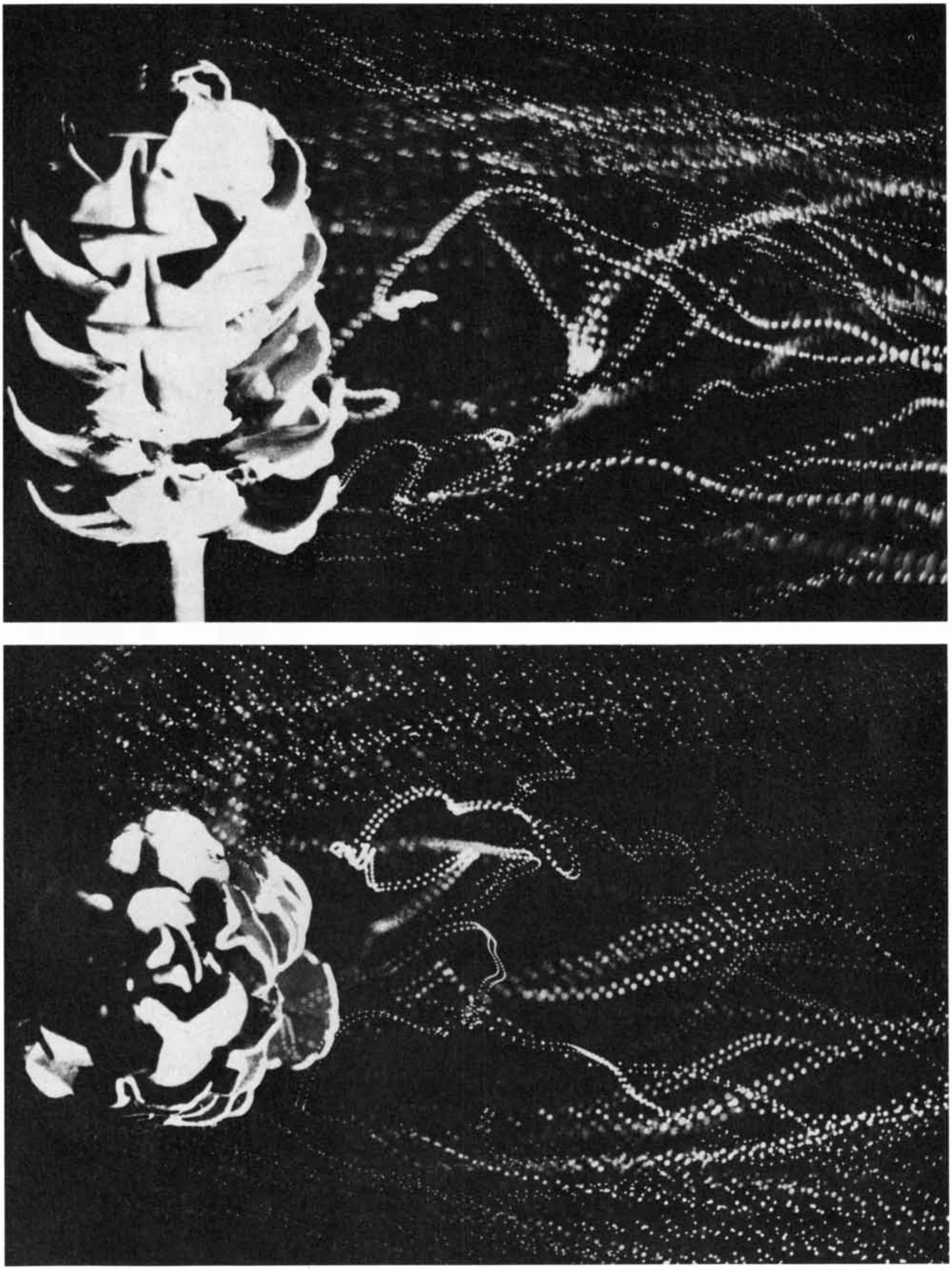
Pine trees, which are bisexual as adults yet rely on cross-pollination to fertilize 95 percent of their eggs, provide an excellent illustration of the influence of a plant's aerodynamic design on its ability to snare pollen from the wind. The male reproductive organs of these conifers are small cones that usually grow in clusters. When a cone is completely grown, its pollen-producing chambers rupture, releasing their contents into the air. The female cones, on the other hand, are larger and can grow singly. Their scales, which have small leaves known as bracts at the base, bear two ovules on their upper surface. When the cone is ready to be pollinated, the scale-bract complexes separate slightly from one another, giving airborne pollen access to the micropyle.

When my colleagues and I began to study female pinecones, we were particularly intrigued by a puzzling structural feature: the ovules are buried close to the axis (the cone's stemlike core), and their micropyles open toward the axis and not the external environment. This seemed a poor design for fostering contact between pollen and micropyle. Could it be, we wondered, that the scales and bracts somehow obstruct the flow of air in such a way as to deflect pollen toward the oddly placed ovules?

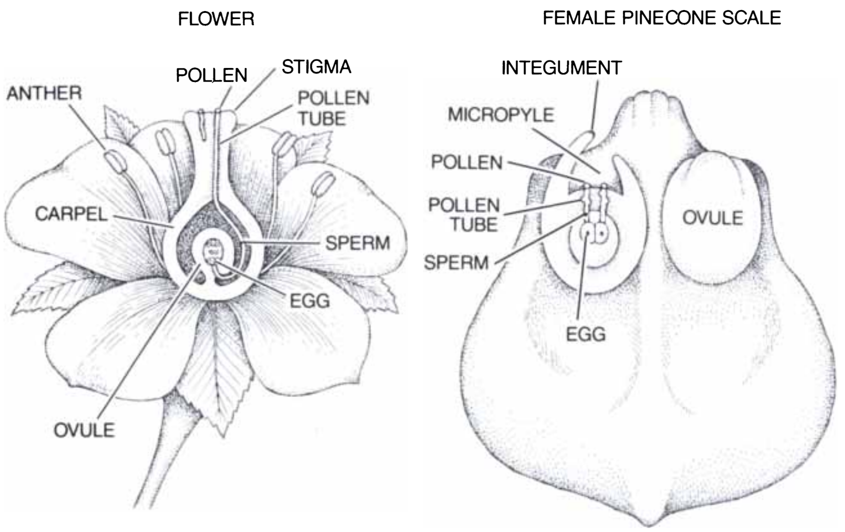
We sought the answer by placing a larger-than-life model of a pinecone in a wind tunnel. We visualized the airflow disturbances around the model by releasing helium-filled bubbles into the wind; such bubbles trace airflow with excellent precision and can pass through tight spaces at a high speed without bursting. We then recorded the trajectories of these bubbles with stroboscopic photography: a camera "shot" the bubbles as a strobe light flashed at preset intervals. A computer analyzed the images to yield data on the speed and direction of the wind in various "cells," or tiny regions, in the microenvironment around the cone.

The computer indicated that the model disturbed the linear movement of the wind in three basic ways [*see bottom illustration on page 92*]. First, it deflected the wind into the core of the cone. There the air circled the axis, washing over the upper (ovule-bearing) surface of the scale-bract complexes. Second, wind passing over each scale dropped toward the base and swirled chaotically near the micropyles. Third, the cone as a whole, with its many protrusions, resulted in significant turbulence along the leeward (downwind) side. There the wind was sucked downward and back (upwind), striking the leeward scales.

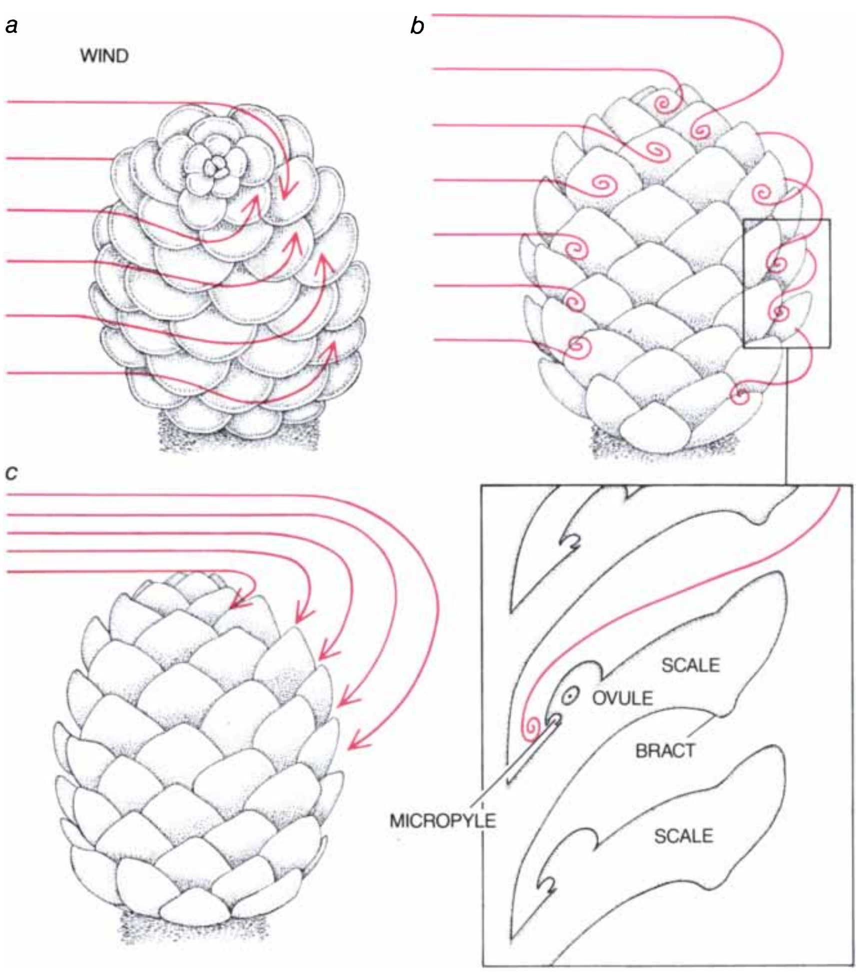
All three of the patterns suggested



HELIUM-FILLED BUBBLES in a wind tunnel reveal the turbulence created by a model of a female pinecone, which is seen from the side (*top*) and from above (*bottom*). The wind is blowing from left to right. Such turbulence fosters collisions between airborne pollen grains (sperm-producing spores) and a cone's ovules: the egg-bearing structures on the upper surface of each scale. The airflow around the model was photographed by a stroboscopic technique in which the bubbles were illuminated at preset intervals.



FERTILIZATION of an egg follows pollination: the delivery of pollen to the stigma of a flower (*left*) or, in a nonflowering plant, to the micropyle (opening) of an ovule (*right*). The pollen then matures, if necessary extending a tube to reach the egg. Eventually the pollen releases sperm, fertilizing the egg and causing the ovule to develop into a seed. A pine scale bears two ovules, each of which holds two or three eggs; only one egg becomes an embryo, however. (The left ovule is rotated 90 degrees to show the inner structure.)



THREE PATTERNS of airflow are typically generated around a female pinecone that is ready to be pollinated. Wind is deflected into the core of the cone (*a*), where it circles the axis, or central stem, brushing over the upper (ovule-bearing) surface of the scales. The air also swirls chaotically above each scale (*b*), near the micropyle (*inset*). On the leeward (downwind) side of the cone air is drawn down and is pulled toward the scales (*c*).

that pollen grains suspended in the wind brushing over an actual cone have a good chance of colliding with the micropyles. Other experiments involving real cones and pollen confirm the suggestion. The same basic turbulence patterns were evident, and pollen moving around ovulate cones roughly followed the course of the wind. We also determined that a large fraction of the spores settled on the inward-facing micropyles.

Knowing that female organs of conifers can deflect airborne pollen toward their ovules, we confronted another question: What prevents the pollen grains of one plant species from being "wasted" on the ovules of another? In plants, as in animals, sperm can generally fertilize eggs from the same species only.

Our studies reveal that the unique shape of the cone produced by each plant species results in idiosyncratic modifications of the airflow patterns described above. The specific patterns are influenced by such factors as the diameter and length of the cone, the number of scales attached to the central axis, their shapes, the angle at which each scale meets the axis and the speed of the wind. Similarly, each type of pollen has a distinctive size, shape and density, causing the pollen to interact with the turbulence in a unique way. Pollen grains from one species, for example, may be too dense to follow precisely the currents generated by a cone from another species; instead particles on the leeward side may settle out of the air currents, falling to the ground before they can be pulled back toward the scales.

Apparently many varieties of cones generate wind-flow patterns that best suit the pollen of their species; most of the cones we studied filtered their "own" pollen from the air but not that of other species. This filtering ability has obvious benefits. In addition to capturing the appropriate pollen, female pinecones also deflect undesirable particles, such as fungal spores, that can damage the ovules.

The cone itself is not the only part of the conifer to exert an aerodynamic influence on pollination; the leaf clusters that typically surround a female cone also have an effect. These leaves act much like a snow fence: a slatted fence that screens a roadway from blowing snow. The spaces between the slats create eddies that slow the wind and therefore the flakes it carries. As a result the flakes drift to the ground on the downwind side of the fence instead of streaming onto the roadway. Similarly, the leaves around a pinecone can decrease the speed of the air, thereby showering the cone with pollen grains, which are then sucked into eddies above and around the scales. Because the leaves are arranged symmetrically around the cone, they can trap wind-borne particles coming at them from any direction.

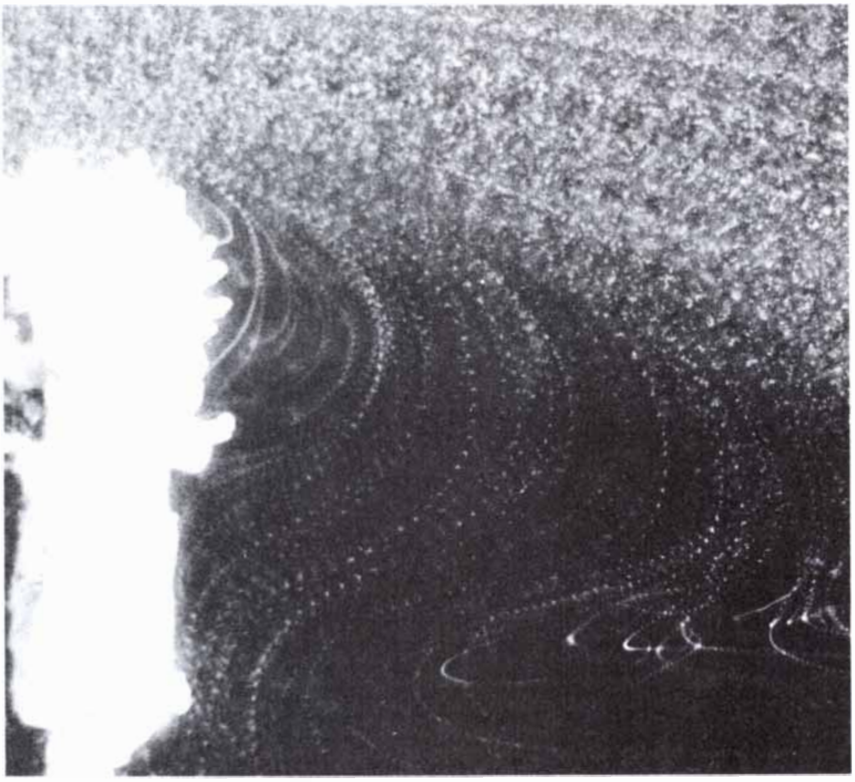
One additional characteristic of pine trees also facilitates pollen capture. Ovulate organs generally grow at branch tips. As the wind blows it can cause branches to undulate in a nearly circular path, enabling the cones to harvest spores from various levels of the air.

Like pinecones, the reproductive organs of grasses also generate unique turbulence patterns. Some grass inflorescences, or clusters of flowers, are borne at the tips of flexible stems and consist of numerous flowers arranged around a central stalk, much as scales are arranged around a pinecone. Not surprisingly, the flow of air around these "tight" inflorescences resembles the flow around a female pinecone. Floral surfaces deflect pollen into the spaces between flowers, and the inflorescence as a whole creates an aerodynamic sink for other grains on its leeward side. When the stem sways in the breeze, it plunges the flowers into this sink, increasing their uptake of pollen.

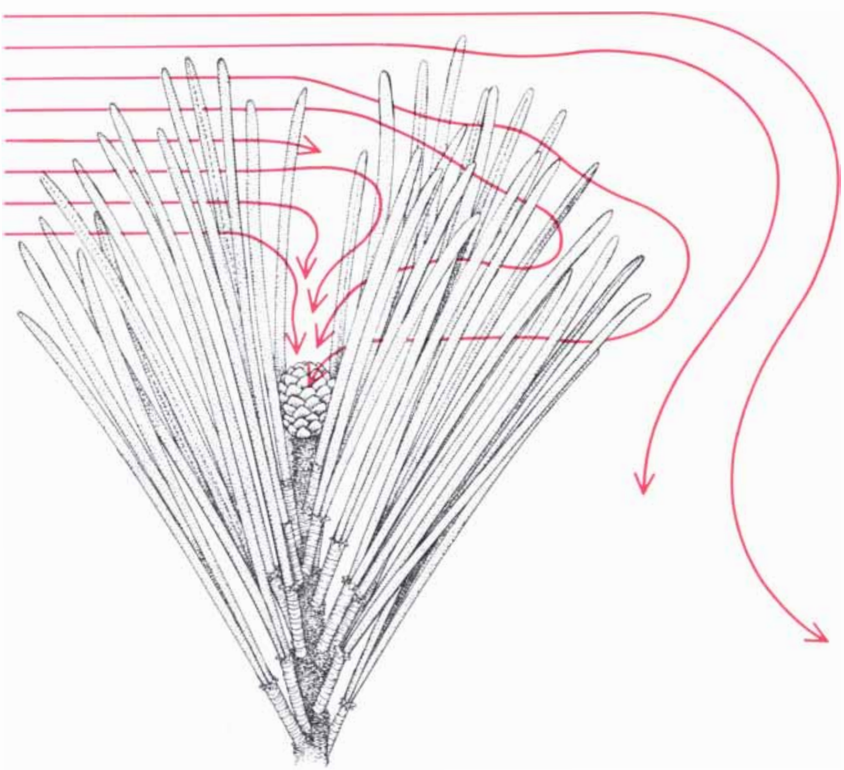
The inflorescences of certain grass species consist not of dense aggregates but of a main stem bearing several small stems, each of which supports one or more flowers at the tip. These "loose" inflorescences create little local turbulence, but they compensate in part by traversing a broad area when the wind stirs them. As the main stem is set into harmonic motion, the smaller stems follow the same trajectory and also oscillate independently, creating a highly complex motion that allows the flowers to intercept much of the pollen in their vicinity.

The studies discussed thus far have been primarily descriptive, but research into the physics of pollination has also extended to prediction. Computer models make it possible to estimate the likelihood that a particular ovulate organ will entrap pollen grains having a hypothetical structure. One goal of such modeling is to determine whether the spore produced by a given plant is aerodynamically ideal. Even if an ovulate organ is known to take up its own species of pollen in preference to other types, the possibility remains that the plant could improve its rate of pollination if the grain had slightly different features.

As in our descriptive studies, we set up a cone or an inflorescence in the wind tunnel and determined its characteristic patterns of airflow. We then

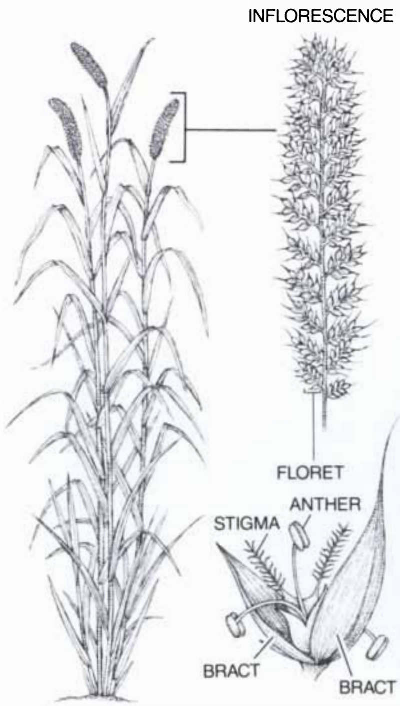


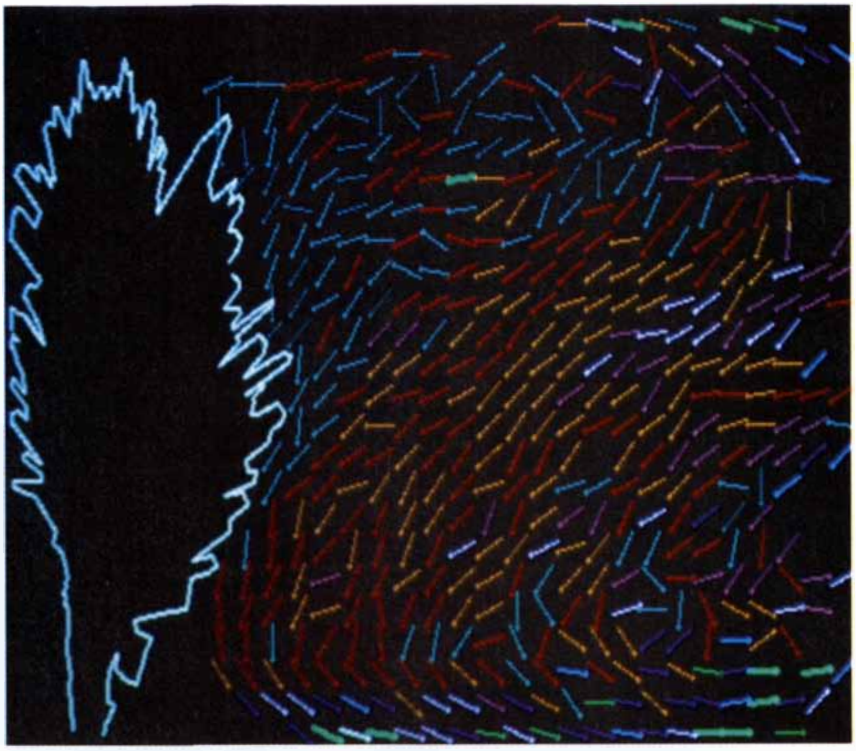
AIRBORNE POLLEN blowing past a female cone of the species *Pinus taeda* is shown by stroboscopic photography; the flow is from left to right. Many grains hit windward surfaces or continue beyond the cone, but others fall back and collide with leeward scales.



LEAVES around a female pinecone act much like a snow fence: a slatted barrier set parallel to and at a distance from a roadway to keep the road free of snow. The spaces between the slats of such a fence hamper the airflow, causing snowflakes to settle to the ground immediately downwind of the fence. Similarly, the leaves upwind of a pinecone cause airborne pollen to settle on their downwind side, showering the cone with spores.

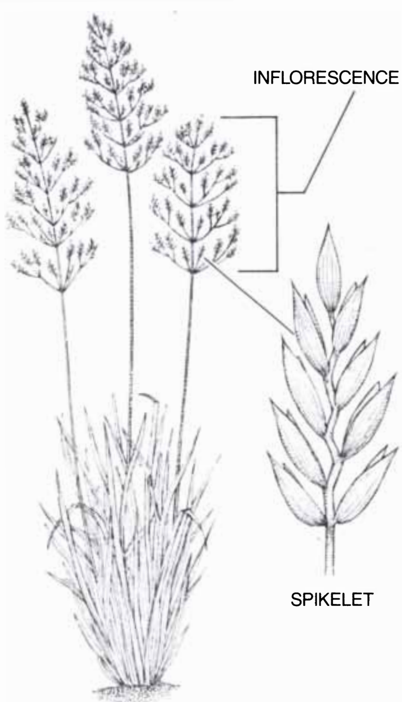
MILLET (*SETARIA*)

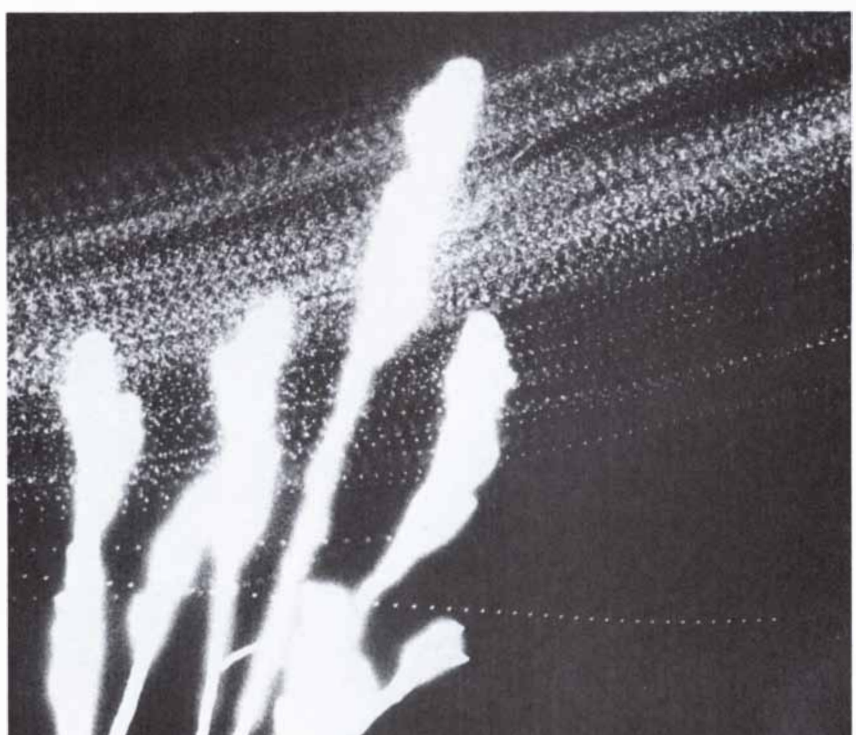




INFLORESCENCE, or flower cluster, of *Setaria geniculata,* a species of millet (*left*), produces leeward turbulence similar to that of a pinecone. A computer analysis (*right*) reveals that pollen grains carried by the wind follow a Z-shaped trajectory and slow as they drift back toward the inflorescence. The computer determined the pattern by calculating the average direction (*round-headed arrows*) and speed (*colors*) of all the pollen in every cell of an imaginary grid superposed on the air around the inflorescence. Green represents the highest speed and is followed in decreasing order by dark blue, white, purple, yellow, red and light blue. The turbulent, slow-moving pollen constitutes a virtual reservoir into which the inflorescence may plunge when it sways in the breeze.

BENT GRASS (*AGROSTIS*)





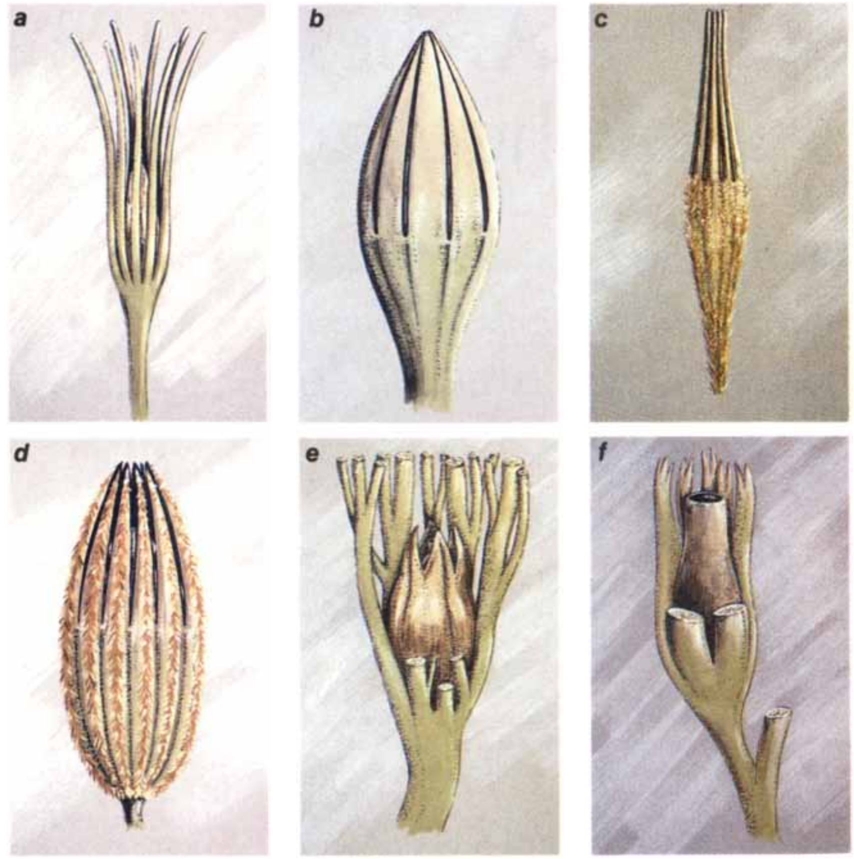
GRASS *Agrostis hiemalis* (*left*) produces "loose" inflorescences: each has a main stem that branches into several smaller stems bearing flowers at the tip. Pollen brushing across such an inflorescence (*right*) does not become trapped in turbulence as it would when passing over a more compact cluster of flowers. Loose inflorescences do have one aerodynamic advantage, however. When the wind blows, their parts move in complex patterns that enable them to capture pollen scattered throughout a large volume of air.

specified the size, shape and density of a hypothetical pollen grain. On the basis of the known air patterns and basic aerodynamic principles, the computer estimated the number of spores that would be likely to reach an ovule (or a stigma).

We have now evaluated approximately 20 plant species that are pollinated by the wind. Most of these have evolved a type of pollen that well suits the aerodynamic environment of their female cones or inflorescences. Certain of the remaining species, however, produce grains whose physical properties appear to be far from optimal. In some of those instances it turns out that the plants do not rely solely on the wind to deliver pollen. A few species are pollinated by insects as well; others reproduce by sprouting cloned offshoots. These plants apparently survive with a "sloppy" match between aerodynamic design and pollen structure because they have reproductive alternatives.

The fact that plants with no alternative methods tend to have the most effective systems of wind pollination suggests aerodynamic design may be an evolutionary adaptation to the need for capturing pollen. It is possible, however, that the architecture of the plants is only a fortunate by-product of adaptations to other demands in the environment. For instance, the tight clustering of ovules in a pinecone or in a grass inflorescence could result solely from the selective pressure to package numerous reproductive units into a small volume. In a similar vein, the arrangement of scales and flowers could be the result of several different pressures, all of which happen to have the same design solution.

This issue is extremely difficult to resolve, but studies of plant fossils carried out by my colleagues and me do support the notion that the need to filter pollen from the air may well have played a role in shaping wind-pollinated species. In particular we evaluated the pollen-capturing abilities of the fossil ovule, often the only part of ancient plants that has been preserved.



FOSSIL OVULES became aerodynamically more efficient as they evolved. Some of the oldest varieties (*a–d*), dating from more than 300 million years ago, consisted of an inner, barely visible egg chamber surrounded by a truss, or ring, of branches. The turbulence created by these protrusions was not focused at the open tip of the chamber and so was inefficient at delivering pollen to the egg. In later ovules (*e, f*), shown within a different set of branches (a cupule), the truss appears to have melded and become shorter. At the rim of the new design an eddy formed a sink for airborne pollen. Modern ovules resemble the inner part of figure *f;* the egg chamber is encased by tissue known as the integument.

Ovule-bearing plants first appeared some 350 million years ago. Although no one is certain of their pollination method, there is a good chance that they depended on the wind. Working under this assumption, my group made life-size models of fossil ovules, put them in our wind tunnel and released various types of pollen around them. A simple conclusion emerged: Among the many ovule shapes in the history of seed plants, the ones that resulted in the most efficient trapping of pollen from the wind were those most similar to the shapes of modern ovules. Nature seems to have culled the least efficient structures.

One such inefficient ovule (or, more properly, ovule precursor) consisted of a short, central egg chamber surrounded by a truss, or ring, of branchlike protrusions; pollen grains reached the egg by falling into a hole at the tip of the chamber. In our studies the protrusions acted like a snow fence, slowing the airflow, but they generated little turbulence over the hole. As a result much of the pollen fell onto the outside of the central body, missing the entryway. In contrast, the egg chambers of modern ovules are encased not by branches but by a continuous layer of tissue known as the integument, which is open only at the rim (the micropyle). In many plants with naked ovules, air that strikes the integument is channeled toward the micropyle and into an eddy above it. This turbulence forms a reservoir for pollen grains, many of which drop into the ovule.

The fossil record suggests that the integument is a reduced and melded form of the ancient protrusions, and our findings indicate that the need to improve pollen capture influenced the alteration. It may well be that plants with the most streamlined ovules were favored by natural selection because they had a higher probability of being pollinated successfully and hence of producing a high yield of seeds that could grow into new plants.

Regardless of the selective forces responsible, it is clear that many wind-pollinated plants leave as little as possible to chance. Botanists have long known that some species exude sticky droplets or extend tentaclelike outgrowths to trap pollen. Most wind-pollinated plants also grow in stands, or clusters, thereby limiting the distance that pollen must travel. Now my colleagues and I have discovered that certain plant species also employ an additional stratagem: by channeling the local airflow, they harness the wind for reproduction.

Arches and Vaults in the Ancient Near East

Working with sun-dried mud brick and mud mortar, masons in ancient Egypt, Mesopotamia and the Levant built arches and vaults that were graceful as well as durable. Their methods are worth studying today

by Gus W. Van Beek

It is almost a truism among students of architectural history that the use of the arch and the vault began with the Romans. And indeed the Romans exploited those structures to the fullest, building them not only often but also on a monumental scale, in bridges and aqueducts, triumphal gates and amphitheaters that are still scattered through southern Europe. Yet the Romans did not invent the arch; nor were they the first to combine arches into vaults. By the time the Colosseum was erected, arches and vaults had been built in the Near East for about 3,000 years.

The few archaeologists who have studied the question are unanimous in the view that arches and vaults originated in the marshlands of Lower Egypt or Mesopotamia. The prototype was a structure built of bundles of reeds, which were placed upright in the ground, bent inward and tied together at the top to form a roof. Early Egyptian drawings, including hieroglyphs, depict reed vaults over sanctuaries, boat huts and other structures. Although no ancient reed buildings have survived, the technique has—in southern Iraq, at the confluence of the Tigris and the Euphrates, where a people called the Marsh Arabs still construct enormous vaulted buildings of reeds.

The outer surfaces of some of these buildings are covered with mud plaster. This type of construction, known as wattle and daub, is probably a relic of an intermediate stage in the evolution of the vault. Eventually most Near Eastern builders dispensed entirely with reeds (which in any case would have been available only in marshy areas) and came to rely on a more durable and massive construction material: sun-dried mud brick. Nearly all the surviving Near Eastern arches are made of mud brick, or adobe (a Spanish loanword from the Arabic *aṭ-ṭūb*). Even after fired brick was introduced most buildings in the Near East continued to be built of the sun-dried kind.

The reasons are easy to understand. Because mud brick is made of widely available ingredients—soil, water and chopped straw, with occasional additions of dung or sand—and because it is baked in the sun rather than in a kiln, it is one of the cheapest of all building materials; it is labor-intensive rather than energy-intensive. Furthermore, it is a poor conductor of heat, and so it is particularly suitable for arid regions where the daily variation in temperature is high. When the temperature climbs above 90 degrees Fahrenheit on a summer day, the temperature inside a mud-brick house stays in the 70's; in contrast, in a non-air-conditioned house built of prefabricated concrete, it can rise above 100 degrees.

The same arid climate that makes mud brick the ideal building material in the Near East makes vaulting the perfect technique for constructing a ceiling or a roof. A vaulted ceiling allows hot air to rise higher than a low flat ceiling does and thus helps to keep the living space even cooler. More important, in many parts of the Near East there are no forests, and consequently the timber needed to support flat ceilings is scarce. A mud-brick vault requires no wood beams for support. Not only is it practical and economical but also it is a singularly graceful way to cover a building.

Three types of mud-brick arching and vaulting have been unearthed at Near Eastern archaeological sites. The first type, which probably looks most familiar to the modern eye, is the radial arch [*see illustration on page 98*]. In a radial arch the first layer of bricks was canted inward by setting small stones or potsherds in mud mortar along the outer edges of the side walls. The bricks were laid face down, and the long edge was aligned with the wall. (Typically the bricks were between 30 and 40 centimeters long and between nine and 12 centimeters thick; they have been found in both rectangular and square form.) The procedure was then repeated, and the layers of brick were canted at progressively steeper angles until the arch was closed at the top with bricks in a vertical or near-vertical position. In arched doorways and short vaulted passages the bricks in successive layers were usually laid directly on top of one another, but in longer vaulted rooms the layers were bonded, or staggered, as they are in the construction of a vertical wall.

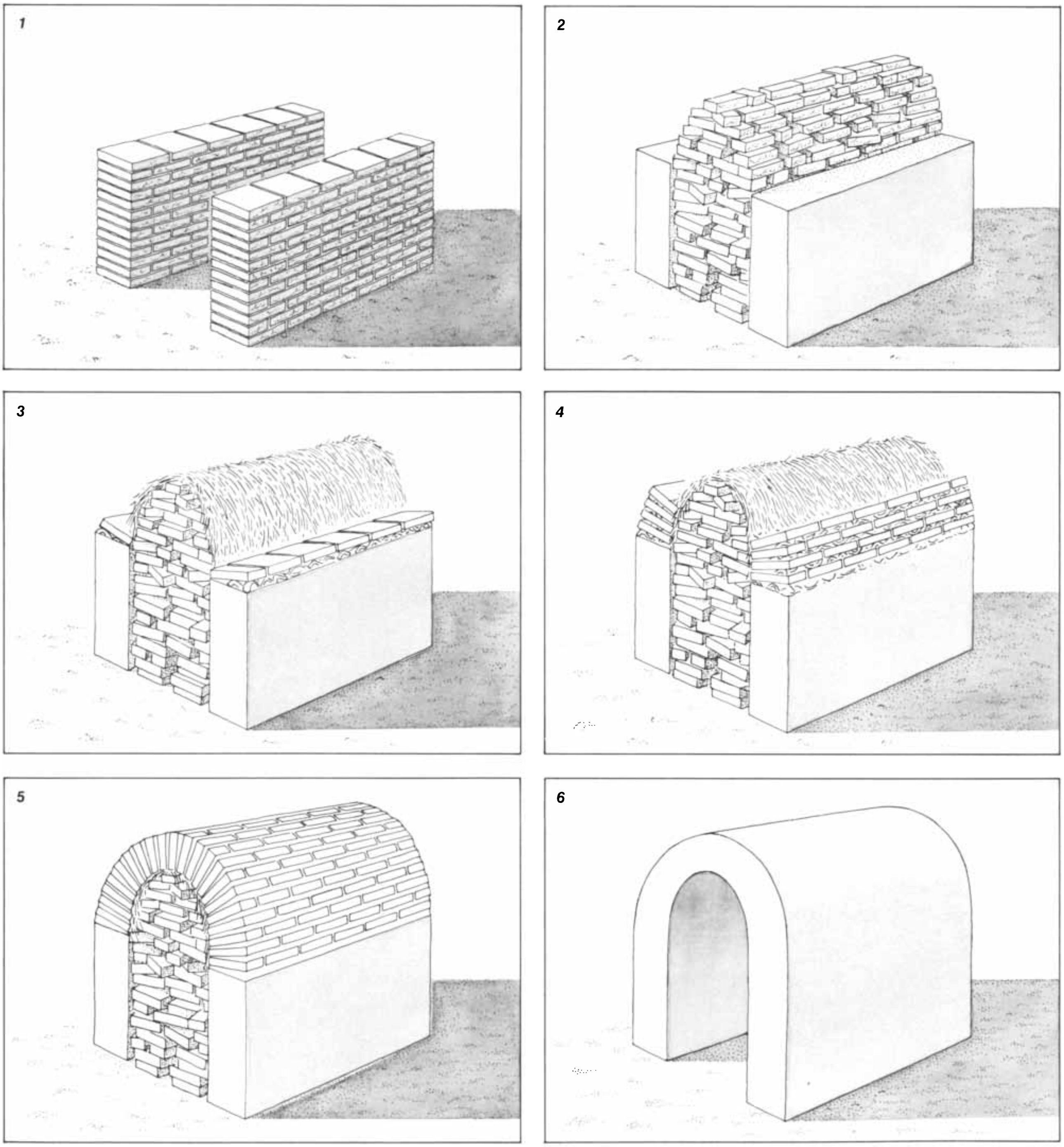
It seems likely that the radial arch and vault evolved from corbeling, a technique that from the earliest times was common throughout the Near East and the Mediterranean region. In corbeling the successive layers of brick are cantilevered rather than canted inward; that is, each layer is laid horizontally but in such a way that it projects over the edge of the layer below. At Tell Razuk in Iraq a building dating from about 2900 B.C. displays what may be transitional evolutionary forms: vaulted roofs in which the successive layers are both corbeled and canted inward.

In Egypt radial arches and vaults were built sporadically in most periods of Pharaonic history, primarily in tombs and in monumental gateways. The earliest example known has been found at Helwan, in a tomb dating from late in the First Dynasty (about 3000 B.C.). A somewhat later but particularly instructive example is the arched gateway of a mastaba (a bench-

MUD-BRICK VAULTS support the storehouses (*top*) in the mortuary complex of Ramses II, the great Pharaoh who led the Egyptian empire to its final peak in the 13th century B.C. The complex, called the Ramesseum, is across the Nile from Luxor. Each storehouse is surmounted by a pitched-brick vault: the bricks stand upright on the side walls and lean toward one end wall. (The end walls have not survived.) Each vault consists of four courses, or layers, of brick; the photograph of the top of a side wall (*bottom*) shows how successive courses of vault brick lean in opposite directions. The storehouses held provisions for the Pharaoh's afterlife.

like tomb with inclined sides) at Giza that dates from the Fourth Dynasty (between 2680 and 2560 B.C.) and that belonged to the nobleman Neferi. The bricks on the inside of the arch were shaped to resemble arching reed bundles, with two half-round moldings on their interior surface; furthermore, the inside of the arch was painted a bright, rich red, the traditional color of dried reeds in Egyptian art. This unmistakable imitation of reeds is strong evidence that early mud-brick arches were modeled on a reed prototype.

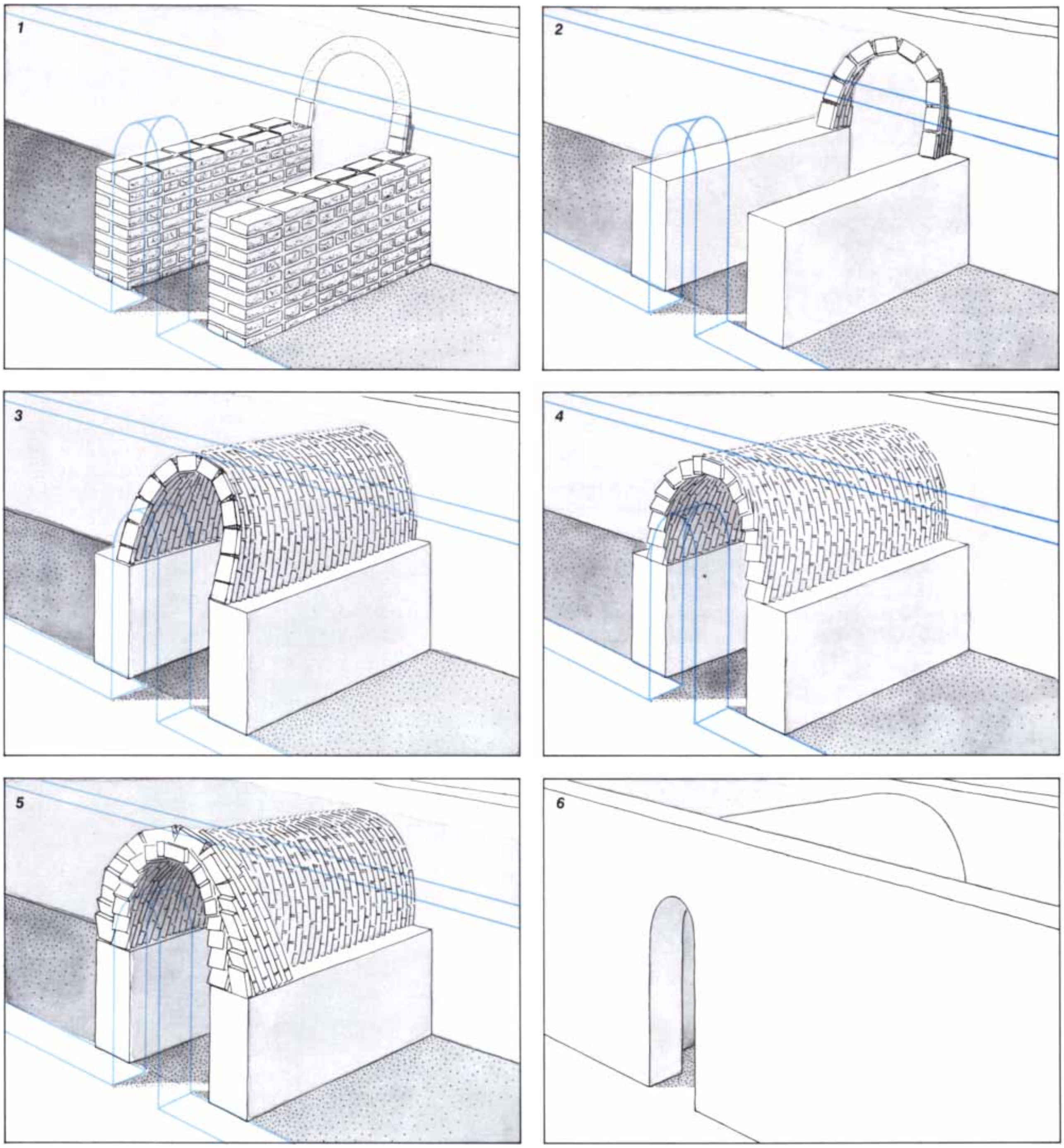
The possibilities of the radial arch and vault were more fully exploited in Mesopotamia, where they probably evolved independently and at roughly the same time as in Egypt. The earliest example known is in a hall built at Tepe Gawra late in the fourth millen-



RADIAL VAULTS were common in the Near East from the fourth millennium B.C. well into the first millennium A.D. The first step in building a radial vault was to raise the side walls to their full height (*1*). Mud bricks were then piled without mortar between the walls to serve as centering, or temporary support, for the vault (*2*). Mud mortar was spread on top of the walls, and small stones or potsherds were placed on the outer edges to cant the first vault bricks inward (*3*). The bricks were laid face down; the ones shown here are typical Mesopotamian square bricks measuring 35 centimeters across. Successive layers of brick were staggered and were also canted inward with stones and sherds (*4*) until the vault was closed at the top (*5*). Within a few days, when the mortar had dried, the support was removed; the vault was covered with mud plaster to help protect it from wind and rain (*6*).

nium B.C., at the time of the emergence of the Sumerian civilization. At Ur, the most famous Sumerian site, the tombs of King Abargi and Queen Shubad (built in about 2500 B.C.) had radially vaulted roofs and radially arched doorways.

The most impressive Mesopotamian radial arches and vaults are at Tell al Rimah, which dates from the end of the third and the first half of the second millennium B.C. In a temple complex at that site, vaulted ceilings span rooms that are as much as 3.8 meters wide; a stairway is supported by eight radial arches of progressively increasing height, and three vaulted passageways cut through a monumental entrance ramp to connect one terrace with another. There are also arched doorways, a vaulted burial chamber



PITCHED-BRICK VAULTS did not require temporary support during construction. The vault was outlined on an end wall with a thick layer of mortar, and the first brick, standing upright on the side wall and canted inward by stones and sherds, was leaned against the mortar (*1*). (Egyptian rectangular bricks are shown here.) Mortar was spread on the face of the first brick and a second brick was leaned against it; a third brick was placed on top of the second brick, canted inward and leaned against the end wall, and so on until the first arc was closed at the top (*2*). Successive arcs were laid against the first one until the opposite end wall was reached (*3*). The triangular space there was filled with smaller arcs and finally with stones, sherds and mortar (*4*). Often a second course, pitched in the opposite direction, was laid on top of the first one (*5*). The finished vault was covered with plaster (*6*).

and a series of radial arches on two levels that supported an overlying terrace or perhaps even a building.

Recent excavations in Israel have brought to light the first examples of mud-brick radial arches in the Levant. The oldest is an 18th-century-B.C. gateway at Tell Dan that consists of three concentric courses, or layers in thickness, of arch brick. At Tell Jemmeh, a site I excavated, four radial arches supported the floor separating the firebox from the baking chamber in a large pottery kiln built by the Philistines in the 12th century B.C. Elsewhere at the same site, in a cross wall that carried the roof of a large circular granary from the third century B.C., I found a mud-brick radial arch with the unusually large span of 4.25 meters.

Taken together, the various archaeological excavations have shown that radial arches and vaults were widely and diversely employed in the ancient Near East, from southern Egypt to western Persia and from late in the fourth millennium B.C. well into the first millennium A.D. The radial design, however, had a serious drawback: the arch or vault required centering, or temporary support, while it dried. Given the scarcity of timber, the commonest type of centering was probably just a dry, unmortared pile of mud bricks, perhaps covered with a bed of straw. Filling a doorway or an entire room with bricks and then removing them once the arch or vault had dried would have taken a lot of time and effort.

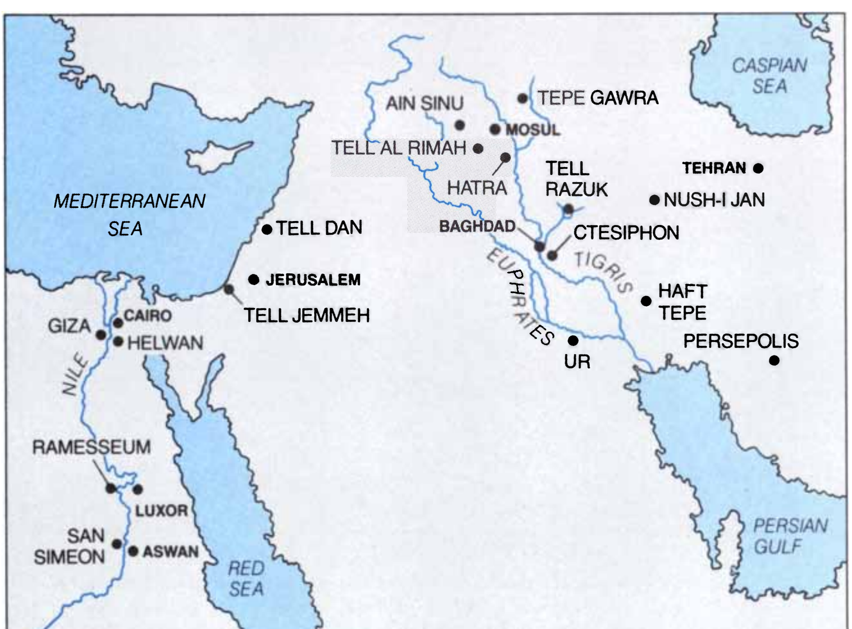
The second method of vault construction, the so-called pitched-brick method, avoided the need for centering. The bricks in pitched-brick vaults were lighter than those in radial vaults because they were generally smaller and thinner and also because extra straw was mixed with the mud. More significant than the weight of the bricks, though, was the ingenious way they were laid: they were stood on end and pitched, or leaned, against one of the end walls of the vault [*see illustration on preceding page*]. At the same time they were canted inward, like the bricks in a radial vault, by stones or potsherds placed under their outside edges. The finished vault consisted of a series of inclined arcs of brick; the hole left at the opposite end wall was filled with smaller arcs, brick fragments, sherds, stones and mortar.

Because each arc supported the next one, the construction of a pitched-brick vault could be stopped at any time and the vault would remain upright without interior support. Mud mortar spread thick between the arcs kept them from slipping. The suction of the wet mortar was augmented by an interesting feature of the bricks themselves: on one face (and in Egypt sometimes on both faces) each brick had a series of deep grooves gouged by the mason's fingers while it was still wet. The grooves allowed the mortar to act on a greater surface area, and the increased suction made for a stronger bond between bricks.

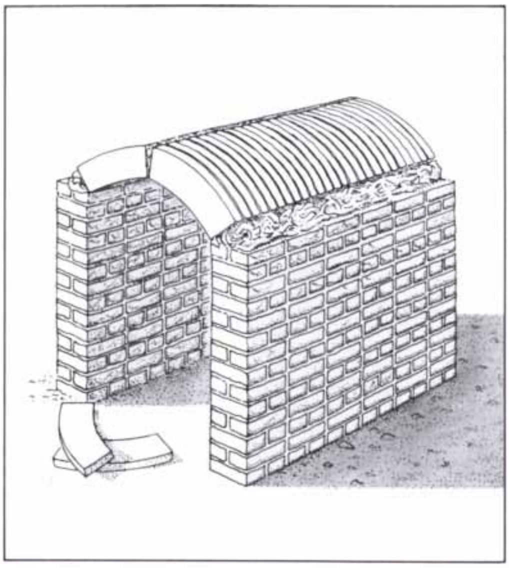
A number of variations on the basic pitched-brick technique have been observed. Occasionally a vault was built from both end walls; this was not particularly clever, because instead of having to fill a small triangular space at one end, the masons had to fill a larger ellipsoidal space at the center, where the oppositely inclined arcs met. Often, after a room had been covered with one vaulted layer of brick, the thickness of the vault was increased by laying additional courses over the first one. In Mesopotamia, where square bricks were the norm, one additional course was enough to achieve the desired thickness. In Egypt, where narrow, rectangular bricks were favored, vaults were often four or more courses thick. Alternating courses were inclined in opposite directions: the first course leaned against one end wall, the second course leaned against the other end wall and so on.

The earliest example known of a pitched-brick vault is at Helwan, in the same late First Dynasty (about 3000 B.C.) tomb that contains the oldest Egyptian radial vault. The mature design of the pitched-brick vault suggests, however, that the technique had been employed for some time and that older specimens either have not been preserved or have not yet been discovered. In the many vaults dating from the Fourth Dynasty some experimentation, mostly with the shape of the bricks, is still evident. One of the most interesting experiments is in the mastaba of the priest Sabef in the Giza necropolis. The bricks in that vault have triangular projections at both ends that made it possible to interlock them with neighboring bricks in the same arc. Theoretically such a design would yield a stronger vault. Yet shrinkage of the bricks as they dried might often have resulted in a poor fit, and the triangular projections would also have tended to break during construction. Presumably such problems explain why no other mud-brick vaults with interlocking bricks have been found.

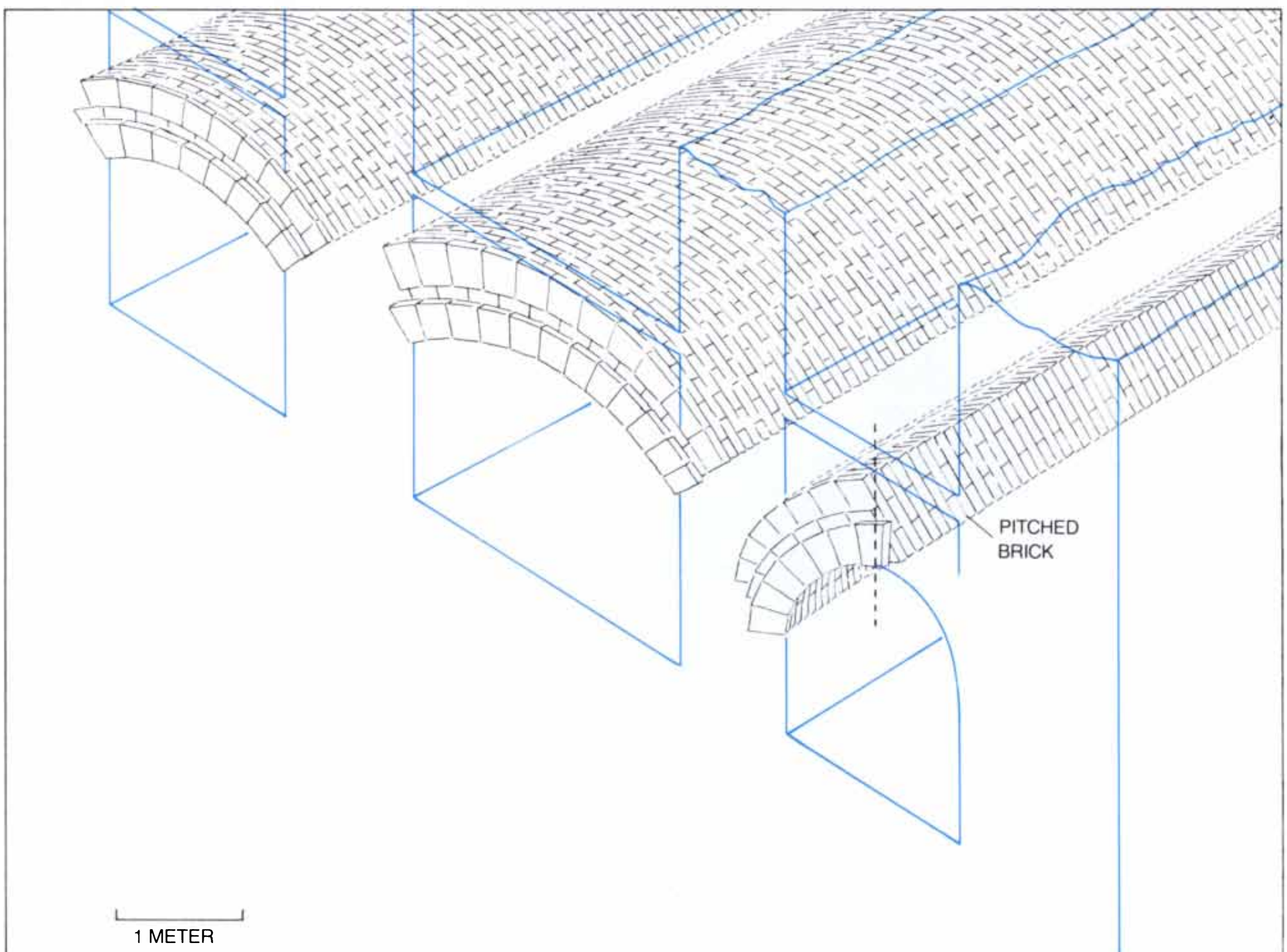
In contrast, barrel vaults built of rectangular pitched brick appear in every period of Egyptian history from ancient to modern, and in a wide range of applications: in burial chambers and associated offering rooms; as supports for ceilings and roofs in ordinary houses; in cupola-shaped ventilators on flat roofs; as supports for stairways and as covers for them, and in storehouses, churches and monasteries. Eventually pitched-brick vaults largely supplanted radial vaults. The radial design remained the best option for doorways or open-ended vaults, which lack the end wall required by the pitched-brick design, but in other



ARCHAEOLOGICAL SITES where brick arches and vaults have been found are shown on this map of the Near East. They are in Egypt, Mesopotamia (the valley of the Tigris and Euphrates), the Levant (the region bordering the eastern Mediterranean) and Iran.



RIBBED VAULTS were built with pairs of long, slightly curved mud bricks that leaned against one another over the center of the room. The spaces between the bricks and under their ends on the side walls were filled with mortar and potsherds. At Nush-i Jan in Iran a ribbed vault supported the floor of an upper room in a temple constructed between 750 and 600 B.C. (*right*). The bricks in the vault are 120 centimeters long. The site was excavated by David B. Stronach, now at the University of California at Berkeley.

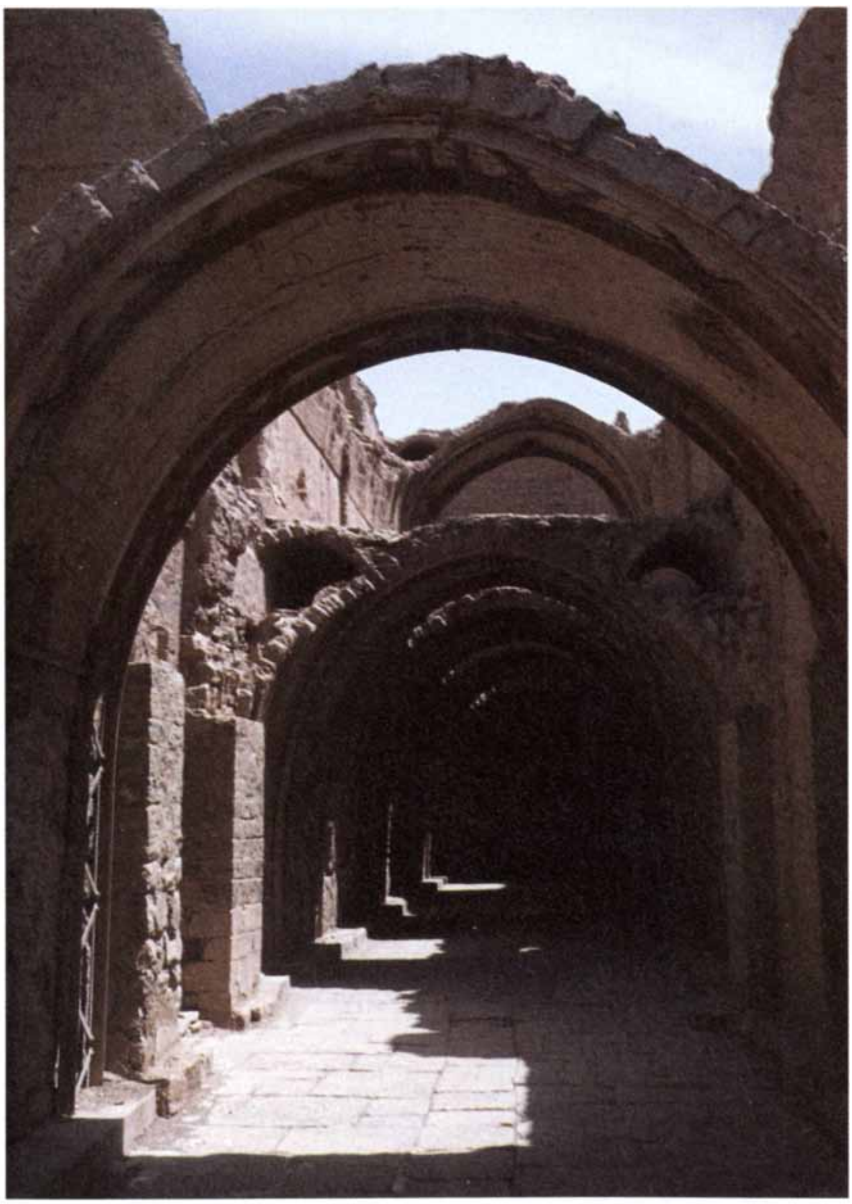


WEDGE-SHAPED BRICKS, or voussoirs, resulted in stronger arches and vaults. The oldest voussoir vaults are in an Assyrian building of the seventh century B.C. at Tell Jemmeh in Israel. The three rooms in the basement differ in width, but the ceiling vaults were built to the same height by tailoring the bricks' shape and by making the vaults over the wider rooms flatter. The vaults were one and three-quarters of a brick thick: in each of the two courses arcs of full-size bricks alternated with arcs of three-quarter bricks.

applications it survived primarily for reasons of tradition.

The finest vaults from the time of the Pharaohs are at Luxor, in the storehouses constructed by Ramses II (who reigned from about 1290 to about 1224 B.C.) in his mortuary complex, the Ramesseum. The long, narrow storehouses held provisions for the sovereign's afterlife. In each building the side walls are topped by four courses of corbeled brick; the corbeling creates broad platforms from which the thick vaults spring. Each vault consists of at least four courses of pitched brick, with alternate courses inclined toward opposite end walls.

The extraordinary adaptability of pitched-brick vaulting in both form and function can be seen in the Coptic monastery of San Simeon at Aswan, which was built in the seventh century A.D., nearly 2,000 years after the Ramesseum. The vaults that stretched the length of the church (and are now preserved only on the sides of the chancel) were not simple barrels: originally they were intersected by cross vaults in the nave. In the building housing the monks' quarters the pitched-brick vaulting over the ground-floor corridor was also an ingenious ventilation system. Secondary vaults occupying the space formed by the curves of the primary vault and the side walls served both to lighten the load on the primary vault and to channel cool night air into each monk's cell.



MONASTERY OF SAN SIMEON, built by the Copts at Aswan in Egypt in the seventh century A.D., contains striking examples of pitched-brick vaulting. The photograph shows a corridor in the monks' quarters. The vaults, which have since been reinforced with steel, supported the floor of another vaulted corridor on the second story. Secondary vaulting between the primary vaults and the corridor walls served to channel air into each cell.

In Mesopotamia the oldest pitched-brick vaults are at Tell al Rimah; they were built in about 2000 B.C., roughly 1,000 years after the appearance of the technique in Egypt. At Tell al Rimah there is in fact a series of vaults covering the entire period from 2000 to 1350 B.C. Some of them exhibit interesting variations in technique. For example, several vaults that helped to support a terrace or a building have a flattened, domical shape. In order to erect them the end walls were built to the same height as the side walls and a fan of brick (a pendentive) was laid across each corner in such a way that it projected up and into the room. The resulting octagonal opening was vaulted with arcs that leaned against the pendentives and the end walls.

Other Mesopotamian pitched-brick vaults have been uncovered in a temple from the 18th century B.C. at Tell Taya, not far from Tell al Rimah, and in the roofs of two 15th-century-B.C. tomb chambers at Haft Tepe in Iran. None has been found from the first millennium B.C., but that is probably just an accident of preservation and archaeological discovery. Pitched-brick vaults reappear in the first millennium A.D. at various sites. Indeed, the acme of pitched-brick vaulting is the Taq Kisra, the great hall of the palace at Ctesiphon, south of Baghdad, which was built sometime between the third and sixth centuries A.D. (There is no general agreement on the date.) The Taq Kisra was made of fired brick rather than of sun-dried brick. It still stands 28.4 meters high and spans 25.5 meters, making it the largest single-span vault of unreinforced brick in the world [*see cover of this issue*].

Judging from the archaeological record, the greatest technological advance in pitched-brick vaulting came not in Egypt or in Mesopotamia but in the Levant. It consisted in the introduction of voussoirs, or wedge-shaped bricks, which eliminated the need for inserting potsherds and stones under the outer edges of the bricks. Vaults built of voussoirs are inevitably stronger, because the bricks are in full contact with one another and the mortar is compressed in very narrow joints. Wedge-shaped bricks make it possible to flatten the vault and give it something other than a semicircular arc. This flexibility in turn enables an architect to design a building with vaulted rooms of different widths and still keep the ceiling or an upper floor at a uniform height.

MODERN MUD-BRICK VAULTS built in the traditional way support the buildings occupied by the Dar Al Islam educational foundation at Abiquiu in New Mexico. The complex was designed by the Egyptian architect Hassan Fathy. The structure projecting to the right is a mosque. The remainder of the complex is a school; the unfinished structure at the upper left will house the cafeteria.

Such was the strategy adopted by an Assyrian architect in about 675 B.C. when he designed the basement of a large building at Tell Jemmeh, where I found the oldest vaults known built of wedge-shaped bricks. The building is three rooms wide. Each room in the basement has a different width and a pitched-brick vault with a different arc: flattened in the two wide rooms and more semicircular in the narrow room. The average width of an individual brick in the vaults is 30 centimeters at the top and 24.5 centimeters at the bottom, but of course the bricks in the flatter vaults taper less than the bricks in the more semicircular one.

Ironically, in the same building the doorways between rooms were covered by a third type of vaulting, rib vaulting, that seems considerably less sophisticated than either the pitched-brick or the radial design. Six pairs of long bricks arched over each door; one end of each brick rested on a side wall, and the other end was propped against the opposite brick over the center of the doorway. The spaces between the brick ends and between the bricks and the side walls were filled with stones, sherds and mortar.

Rib vaulting appears to have been confined primarily to Iran; the technique may have been brought to Tell Jemmeh by a Median builder in the service of the Assyrian imperial forces. It was employed extensively at Nush-i Jan, a site in Iran that dates from between 750 and 600 B.C. The vault bricks there are huge: about 1.2 meters long. One would expect such long bricks to be fragile, but actually they were strong enough to support the floor of an upper room in the central temple. Nevertheless, a rib vault could not have been as strong as a pitched-brick or radial vault, and except for a certain simplicity of construction it probably had few advantages.

Voussoir brick, on the other hand, had distinct advantages, and yet it too did not receive wide application in the ancient Near East. Other than at Tell Jemmeh, voussoir vaulting has been discovered only at the Roman frontier post of Ain Sinu in northern Iraq. More examples may be discovered someday; if they are not, one can only conclude that the advantages of the technique were not considered worth the extra effort required to "custom make" bricks for each vault. Just as puzzling as the apparent failure of voussoir to catch on is the question of why it did not appear until the seventh century B.C.—why the Near Eastern peoples who invented writing, law, mathematics, astronomy and, more to the point, cities should have failed for more than 2,000 years to think of building their arches and vaults with wedge-shaped brick.

The invention of mud-brick arches and vaults, however, was in itself quite an achievement. Arch building never died out entirely in the Near East, and today it is enjoying a limited renaissance. The chief figure in that movement has been the Egyptian architect Hassan Fathy, whose brilliant and imaginative designs incorporating pitched-brick vaults have attracted a following not only in his own country but also in the U.S. As the advantages of mud brick are rediscovered by an energy-conscious world, the practical value of arches and vaults is becoming equally apparent.

Yet I do not want to stress only their practicality. Mud-brick arches and vaults are also beautiful. In a visual world too often dominated by rectangles and squares, they enable an architect to relieve the harshness and monotony of the straight line by employing nature's commonest form, the curve, in an almost infinite number of variations.

THE AMATEUR SCIENTIST

Why a fluid flows faster when the tube is pinched

by Jearl Walker

When I water the lawn with a garden hose, I sometimes partially block the end of the hose with a finger in order to make the water go farther. The action increases the speed of the water by forcing it to pass through a smaller opening.

Does the partial blockage increase the volume of water issuing from the hose per unit of time? No, the flow rate (as it is called) actually decreases, because my finger creates turbulence that removes some of the energy in the flow. Decreasing the size of the opening even more might further increase the speed of the water, but it would also decrease the flow rate. This experience is just as commonplace in a laboratory. When I want to decrease the flow of water or any other liquid from a tube, I tighten a clamp on the tube, narrowing the space through which the liquid must pass. The result has become something laboratory workers expect.

Werner J. Gatzek, a physician in Waynesboro, Va., recently surprised me with a related demonstration that ruffled my certainty. He was trying to model the flow of liquid through a human kidney because some properties of the flow had confounded medical researchers. In his experiment water drained through a garden hose from an elevated reservoir that was kept full. The lower end of the hose was connected to a rubber tube that included a loop along its course to simulate the arrangement of the input and output vessels of the kidney. The loop was held in place by a short corset consisting of three sections cut from the end of the tube; Gatzek forced the tube through the corset by stretching the sections with two needle-nose pliers. A screw clamp was mounted at the free end of the tube to simulate the variable resistance experienced by the output vessel of the kidney. The tube lay flat on a workbench, so that there were no variations in height along its length to affect the flow.

Gatzek measured the rate of flow through the system by collecting water in a small container for a measured time and then weighing the water on a dietary scale. Initially the flow was a trickle because of the constriction from the corset. While continuing to measure the flow rate Gatzek gradually tightened the clamp by repeatedly turning the screw through a measured angle. Initially the flow rate remained constant. Then, when the tube had been narrowed appreciably, the rate suddenly and dramatically increased. Eventually it reached a level more than five times higher than it was when the clamp was fully open. Further tightening of the clamp reduced the rate, which reached zero when the clamp was fully closed.

How could closing the clamp increase the flow rate? Surely the tightening of the clamp must introduce turbulence, thereby reducing the energy of the flow and the flow rate. Nevertheless, constraining the flow somehow increased it.

Once the clamp was fully closed Gatzek started opening it by specific increments, again monitoring the flow rate. This time a wider opening of the clamp was necessary to achieve a peak rate, which was higher than the peak rate attained while he was closing the clamp. When the peak was reached, further opening of the clamp suddenly decreased the flow rate until it had again become a trickle. When Gatzek plotted his measurements of flow rate versus clamp aperture, the graph revealed a hysteresis (a lag) in the pattern. Why did the system behave in a different way when the clamp was being closed?

The details of Gatzek's experiment offered no easy solution to the puzzle. The water reservoir was 5.2 meters above the workbench. The tube, made of amber latex rubber, had a thin wall and an outside diameter of five millimeters. (The characteristics of the tube are not critical, but it should be thin-walled and expandable.) The loop was 22 centimeters long, and the sections of the tube that served as input to and output from the loop were respectively nine and 10 centimeters long. The three sections of the corset were each one millimeter wide.

To measure the flow rate Gatzek collected samples of water for 30 seconds. After measuring the weight of a sample on the scale he converted the reading into milliliters and calculated the flow rate. (He could have measured the time the flow took to fill a container marked in milliliters.) The screw of the clamp was rotated five degrees between measurements. The tube showed no signs of fatigue under the water pressure.

Another experiment done by Gatzek compounded the puzzle. In that experiment he removed the clamp and gradually raised the height of the reservoir. At first, as intuition would suggest, the flow rate increased. After the reservoir passed a certain height, however, raising it farther either left the flow rate unchanged or decreased it as much as 30 percent. When Gatzek continued raising the reservoir beyond that region, the rate of flow began to increase again.

Gatzek originally did his experiments with a rotary pump that pushed pulses of water through the looped and clamped tube. He thought the pulsation might be responsible for the flow's perplexing properties. In 1986, however, Steven C. Wells, then a graduate student at Old Dominion University in Virginia, discovered that the paradox can be demonstrated with a steady flow. Since then Gatzek has substituted the elevated reservoir for the pump.

In the experiment in which the clamp is tightened Gatzek believes the sudden increase in the flow rate results from competition for space in the corset. The section of the input tube within the corset is under pressure from the water. When the clamp is open, the pressure causes that section to bulge against the adjacent section of the output tube within the corset, reducing the flow through that part of the output tube to a trickle.

When the clamp is gradually closed, the narrowing introduces a resistance to the flow, increasing the pressure in the region between the corset and the clamp and in the output tube within the corset. Eventually the pressure in the output tube is sufficient to counteract the pressure from the input tube. The output tube expands within the corset. Although it does so at the expense of the input tube, the result is a much greater flow rate. When the peak rate is reached, further tightening of the clamp introduces so much resistance that the flow gradually decreases and then stops.

I took another approach to the puzzle by focusing on the energy. The point that intrigued me was that although tightening the clamp is certain to remove energy from the flow, the flow actually then becomes more energetic. To follow my steps in solving this paradox you must first understand several general features of water flow. The flow rate through a tube is equal to the product of the speed of the flow and the cross-sectional area of the tube. Along the tube the flow rate must be constant, and so there can be no magical appearance or disappearance of water. If the tube narrows, the water must flow faster; if it widens, the water must flow slower. If the tube does not change in width, the speed of flow will not change.

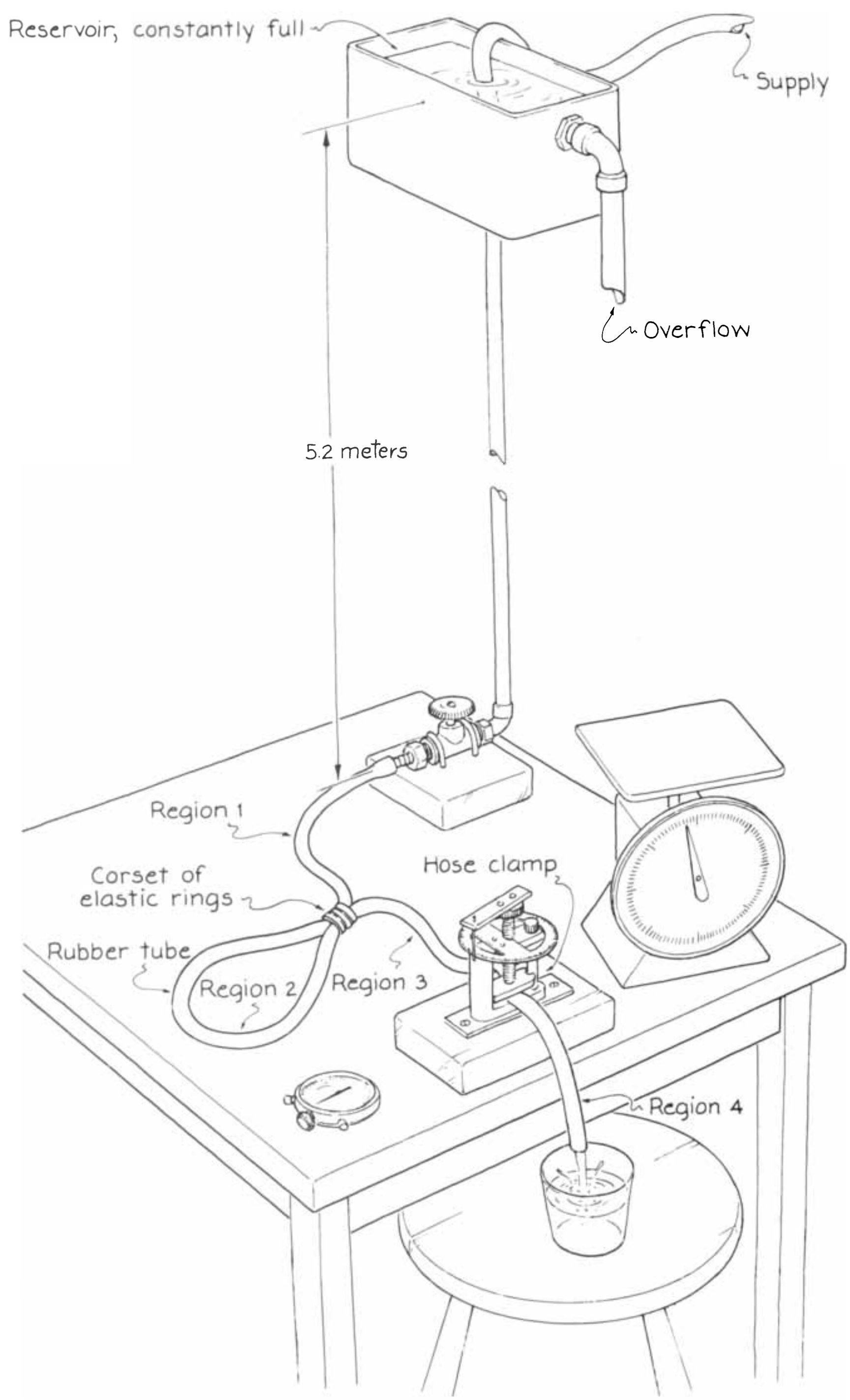
When the flow lacks turbulence, it is said to be laminar. In such a state it can be traced by streamlines that show where individual parcels of water travel. Every time a new parcel travels through any particular point in the system, it thereafter follows the path taken by all the parcels that preceded it through that point.

Flowing water has two types of pressure: static and dynamic. Static pressure exists in a conduit even if the water is stationary; only where the water is in direct contact with the atmosphere, as at the end of a hose, is there no static pressure. Dynamic pressure is proportional to the square of the flow speed. The pressures are related to energy. The static pressure is a form of potential energy per unit of volume; the dynamic pressure is the kinetic energy per unit of volume. If the flow changes in height, the gravitational potential energy per unit of volume must also be considered.

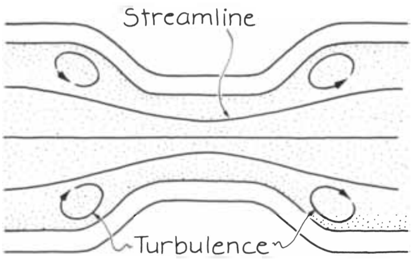
The Bernoulli equation, named after the 18th-century Swiss mathematician Daniel Bernoulli, states that in laminar flow the sum of those energies must be constant along a streamline. Suppose the water flows through a horizontal tube that narrows. As water enters the narrow section its speed and kinetic energy (dynamic pressure) increase because energy is transferred from the static pressure. If the tube returns to its former width, the energy is transferred back to static pressure. At all times the sum of the static and dynamic pressures is unchanged.

Suppose the tube descends without any change in width. As water flows through the tube, gravitational potential energy is converted into static pressure. Again the sum of the energies is unchanged.

The Bernoulli equation and the requirement that the flow rate be constant through a system are not sufficient to explain Gatzek's experiment. If the flow were laminar, the flow rate through the corset would be the same as it is when the corset is removed. The fact that the flow is only a trickle indicates the corset introduces enough turbulence to remove nearly all the energy in the flow. I believe that when



Werner J. Gatzek's apparatus for studying flow



Turbulence at a constriction

the clamp is tightened, the flow rate increases because the loss of energy in the output tube within the corset is greatly reduced.

To develop my explanation I began with a simple system of laminar flow in which water drains through a horizontal tube of constant width. It is attached near the bottom of a container, which is kept full. Since the water that issues from the tube is directly exposed to the atmosphere, the static pressure at the open end must be zero. Since the tube's radius is constant, the speed of the flow must be the same throughout its length. Therefore, according to the Bernoulli equation, the static pressure must be zero at any point in the tube.

Suppose you are looking at a parcel of water in the container near the drain. The pressure on the drain side is lower than the pressure on the side where the tube leaves the drain. The pressure difference accelerates the water parcel toward the drain, giving the parcel the speed of the rest of the water flowing through the tube.

I next worked with a drain tube that has a vertical section connecting two horizontal sections. Again the static pressure in the final horizontal section must be zero. Moreover, the flow through the tube must have a constant speed, even in the vertical section, because the radius is constant. What differs in this setup from the preceding arrangement is that when the water flows through the vertical section, gravitational potential energy is converted into static pressure. At the top of the vertical section the static pressure is negative.

Focus again on a parcel of water about to enter the drain. Since this time the parcel has a negative static pressure on one side, it accelerates more and enters the drain at a higher speed than the parcel in the preceding example. With the faster flow the flow rate is higher.

Note that the kinetic energy and the speed of the water are derived from the static pressure and the gravitational potential energy of the water in the container. Suppose a constriction in the final horizontal section of the tube introduced turbulence that removed energy from the flow. Then the kinetic energy of the water leaving the end of the tube would be lower and the water speed throughout the system would be reduced. In Gatzek's experiment the corset introduces two constrictions and the tightened clamp gives rise to a third.

To review Gatzek's experiment I worked with a schematic representation of the tube that included two short, narrow regions and a clamp in the final horizontal section [*see upper illustration on opposite page*]. The narrow regions represent the input and output tubes at the point where they pass through the corset. The static pressure at the tube's open end and in the regions numbered 3 and 4 is zero. In region 1 the static pressure is high, collapsing the output section of the tube in the corset and allowing only a trickle to flow through the system. Region 2 represents the loop.

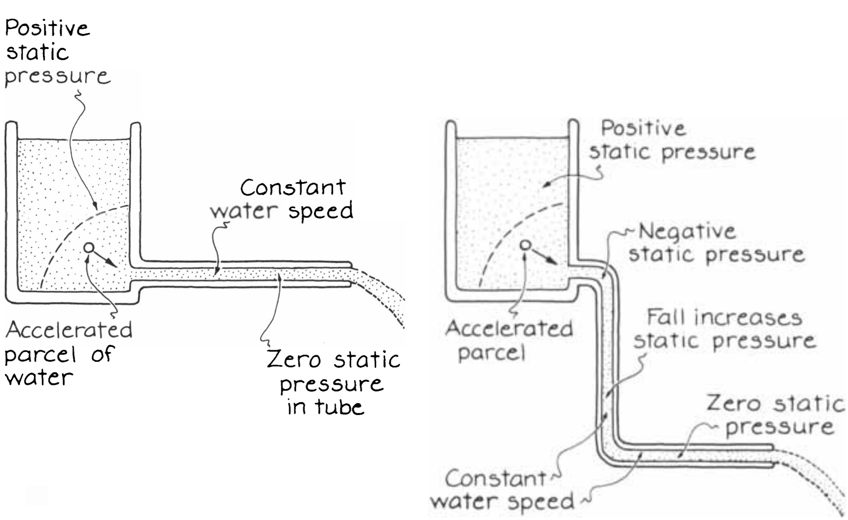
Examine the graph below the diagram of the schematic tube. The top line represents the sum of the pressures at points along the length of the tube. The line below it represents the static pressure. The distance between the two lines represents the dynamic pressure, or kinetic energy, of the flow. Since the water speed is constant and small, the distance between the lines is also constant and small.

In the vertical section of the tube the static pressure increases at the expense of the gravitational potential energy. Both lines in the graph rise, but the kinetic energy of the flow does not increase. Both the input and the output tube within the corset introduce turbulence that removes energy, reducing the static pressure. Since the input tube is relatively wide, the energy loss there is small. The energy loss in the partially collapsed output tube is considerably larger, reducing the static pressure to zero.

Next I drew a schematic diagram and a graph representing the maximum flow rate. Again the static pressure in region 4 is zero. Since the size of the tube in region 4 is unchanged, the fact that the flow rate is now greater means that the water speed is higher. Hence the two pressure lines in the graph are more widely spaced than they are in the preceding example.

Note the energy losses at the simulated input and output of the corset. Although the loss at the input is somewhat greater than it was before, the loss at the output is considerably reduced owing to the increased size of the output tube within the corset. The combined loss of energy at the corset and at the clamp is less than it was when the clamp was fully open. The reduced loss shows up as the added kinetic energy in the flow.

When does the corset allow maximum flow? Gatzek suggested that the flow rate is maximum when the input and output tubes within the corset are equal in size. To check his suggestion I assumed that the energy loss from a constriction in the tube is proportional to the ratio of the tube's normal cross-sectional area and its cross-sectional area within the constriction. Next I added the losses at the input and output points of the corset—under the requirement that the sum of the areas of the constricted tubes had to be the cross-sectional area of the corset. I



Two principles of drainage

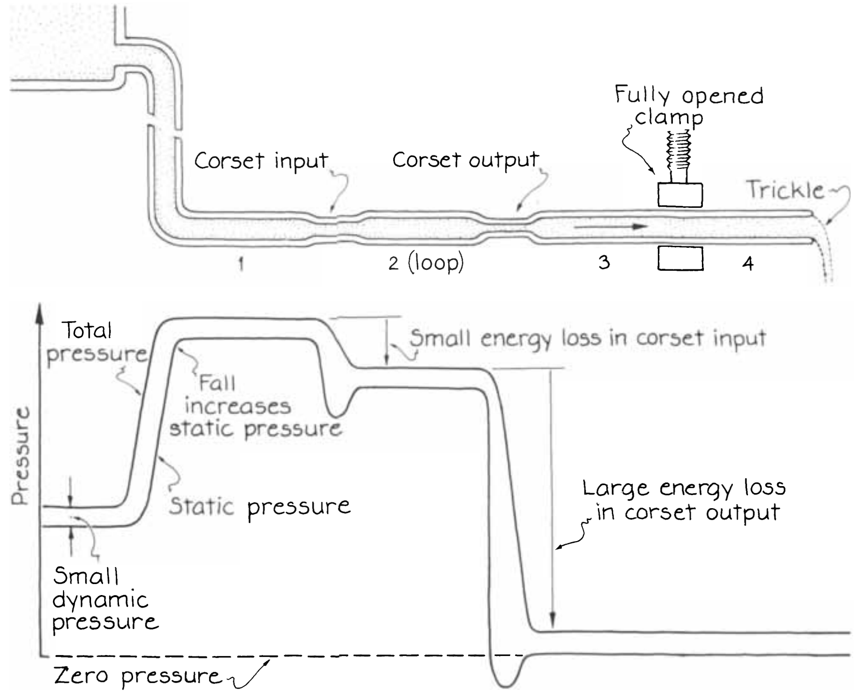
then checked for the case in which the combined loss is least. Indeed it is least when the input and output tubes within the corset are of equal size. Thus part of the puzzle is solved. When the clamp is tightened, the increased static pressure in the ouput tube within the corset widens the output tube, greatly reducing the energy loss there and allowing more of the energy of the flow to end up as kinetic energy.

Why is the transition to the peak flow rate sudden? I believe the suddenness arises from the shape of the walls where the input and output tubes touch in the corset. When the clamp is being tightened, the wall of the input tube is convex and the wall of the output tube is concave, leaving the output tube at a disadvantage. As pressure builds in that tube it suddenly becomes high enough to overcome the concave shape of the wall, whereupon the wall flexes outward. Then the walls of the input and output tubes are flat against each other and the flow rate suddenly increases to its peak value.

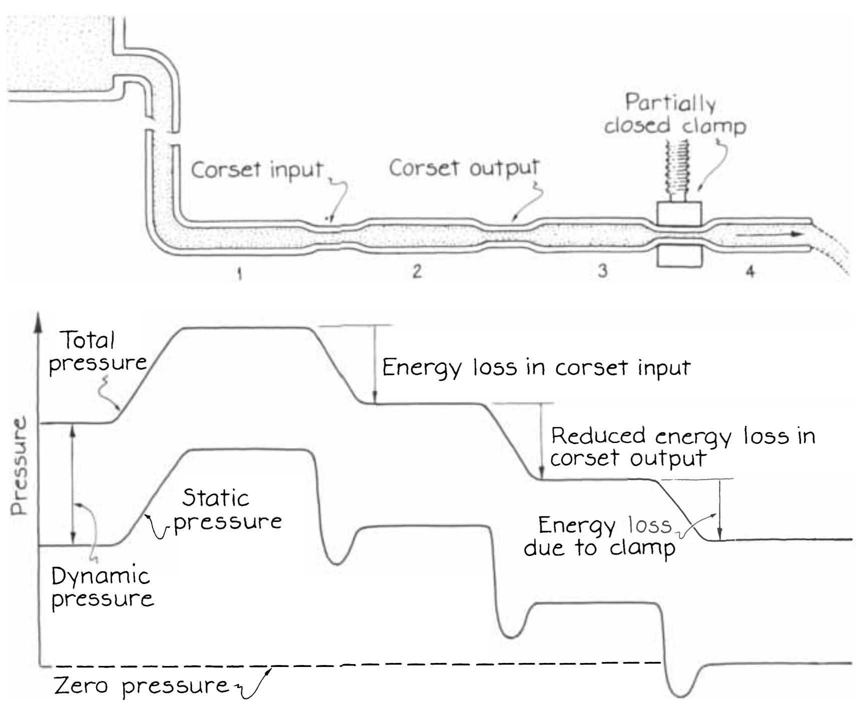
Why does the system exhibit hysteresis? Suppose the clamp has been tightened to the point where the flow rate is maximum. Since the walls of the input and output tubes are then flat against each other, the energy loss at the corset is either at the minimum or nearly so. If the clamp is opened somewhat, the walls of the input and output tubes remain flat and the energy loss in the corset is unchanged. Since the energy loss at the clamp is reduced, however, the combined energy loss in the system is also reduced. For this reason more of the energy in the flow is kinetic energy and the flow rate is at a higher peak value.

One more piece of the puzzle remains. What accounts for the behavior of the flow rate when the clamp is left open and the reservoir is gradually raised? Gatzek's explanation serves well. Initially the flow rate increases because the added height of the reservoir pushes water through the system faster. Soon, however, the increased pressure in the input tube of the corset begins to collapse the output tube. Once the reservoir is raised beyond a certain height, the collapse either holds the flow rate constant or decreases it. If the reservoir is raised even more, the rate begins to increase again because the high pressure in the input tube expands the corset.

Gatzek's experiments may help to explain some characteristics of the flow through the glomerulus in the kidney. The afferent vessel (the input arteriole) and the efferent vessel (the output arteriole) pass through the same opening in Bowman's capsule and may compete for space there. The efferent vessel does not have a clamp, but it is subject to a vasoconstricting hormone called angiotensin, a derivative of the renin produced by the kidney. In some physiological conditions the kidney may produce angiotensin in order to constrict the efferent vessel, increasing its resistance to flow. The increased resistance may then widen the vessel in the opening in Bowman's capsule, leading to a noteworthy increase in the flow rate.



Energy losses with the clamp open



Energy losses with the clamp partially closed

COMPUTER RECREATIONS

Probing the strange attractions of chaos

by A. K. Dewdney

Chaos has strange attractions for the mind that can see patterns therein. Some physical systems that exhibit chaotic behavior do so because they are in a sense attracted to such patterns. As a bonus, the patterns themselves are strangely attractive. Some readers may already be aware that the geometric forms underlying chaos are called strange, or chaotic, attractors [see "Chaos," by James P. Crutchfield, J. Doyne Farmer, Norman H. Packard and Robert S. Shaw; SCIENTIFIC AMERICAN, December, 1986]. Strange attractors can be generated with a home computer.

Before setting off with me, readers must be equipped with a protective coating of physical intuition. In particular, what is an attractor? Roughly speaking, an attractor is a generalization of the notion of equilibrium; an attractor is what the behavior of a system settles down to, or is attracted to. The pendulum is a simple physical system that illustrates the concept of an attractor. Suppose an ordinary pendulum moves under frictional forces that slow it eventually to a standstill. One can describe the pendulum's motion by means of a so-called phase, or state, diagram in which the angle the pendulum makes with the vertical is graphed against the rate at which the angle changes. The swinging motion of the pendulum is represented by a point circling the origin in the phase diagram; as the pendulum loses energy, the point spirals into the origin, where it ultimately comes to rest. In this case the origin is called an attractor because it seems to attract the moving point in the phase diagram. Readers would be correct in thinking there is nothing strange about an attractor consisting of a single point.

A slightly more complicated attractor underlies the motion of a grandfather clock. Here an escapement mechanism feeds energy to a pendulum to keep it from slowing down. If one starts the clock with an overly energetic push of the pendulum, it slows down to the tempo prescribed by the escapement but thereafter slows no further. If the clock is started with a push that is too gentle, however, the pendulum behaves like an ordinary one: it slows to a standstill as before. In the case of the overly energetic push, the pendulum's motion in a phase diagram is a spiral that winds ever more tightly about a circular orbit. Here the attractor is a circular loop. In this context a circle is no stranger than a point.

An ordinary pendulum can be made to show chaotic behavior by introducing a vertical, vibratory motion: if the point of support is moved up and down in a sinusoidal manner by an electric motor, the pendulum may begin to swing crazily, exhibiting no evidence of periodic behavior whatever.

To introduce chaos, however, I have selected a different physical system. Imagine an arrangement of three amplifiers in which the first amplifier outputs a signal x that is fed to the other two. The second amplifier outputs the signal $1 - x$ in response to x. The third amplifier takes the two signals, x and $1 - x$, as input. It generates the product, $x(1 - x)$, of the two signals and feeds it back to the first amplifier, which also receives a control voltage, r, as input. One additional component, a device that samples its input and delivers the same voltage as output for a short time, completes the circuit; it is inserted in the output line from the first amplifier. The three-amplifier circuit does a voltage dance that becomes more hectic as the control voltage r is gradually increased.

When r is less than 3 and x initially has some nonzero setting, the circuit oscillates briefly before it settles down to a specific value of x that remains the same thereafter. This value constitutes a single point attractor. If the control voltage, r, is now raised to a level just above 3, the circuit flutters between two values of x. At this level of r the circuit is said to be bistable and the attractor consists of two points. As r is increased further, the circuit oscillates among four points; yet another increase yields an eight-point attractor. The pattern of doubling and redoubling goes on as the knob controlling r is turned to higher values, until at a setting roughly midway between 3 and 4 the circuit suddenly goes crazy. It hunts endlessly at electronic speed for the simple recurring patterns that marked its earlier existence. Its behavior is now governed by a strange attractor that has a potential infinity of values. The result is chaos.

Electronically literate readers may be tempted to build such a circuit. Others may simulate it on a computer of any size, viewing the dance with great clarity on the display screen. To do so, they merely need to write a simple program that computes the iterated equation $x \leftarrow rx(1 - x)$. The program, which I call CHAOS1, has a core that consists of six instructions:

```
x ← .3
for i ← 1 to 200
    x ← rx(1 − x)
for i ← 1 to 300
    x ← rx(1 − x)
      plot(200x, 100)
```

The variable x starts at the value .3. CHAOS1 then enters a loop that iterates the basic equation 200 times to allow transients to die away. The transients are inherent in the equation itself, not in imprecise arithmetic. The reasons for this will be made clear in geometric terms below. The program then enters a new loop that iterates the equation 300 more times, plotting the value of x on each occasion.

The number 100 used in the plot instruction above is more generic than specific; here the screen has dimensions 200 by 200. The horizontal coordinate, $200x$, spreads the various values computed for x (always between 0 and 1) across one row of the screen, which is set at a height of 100—halfway up the hypothetical screen.

Depending on the setting for the control variable r, the core program will either plot a single point 300 times or several points fewer than 300 times each. It may even try to capture chaos by plotting 300 different points of a strange attractor. If the iteration limit is increased, more of the strange attractor will be seen. In all cases, once the iteration process has settled down, the x values jump in a systematic way from one point of the attractor to another. The attractors are also called orbits, regardless of whether they have a finite or an infinite number of points.

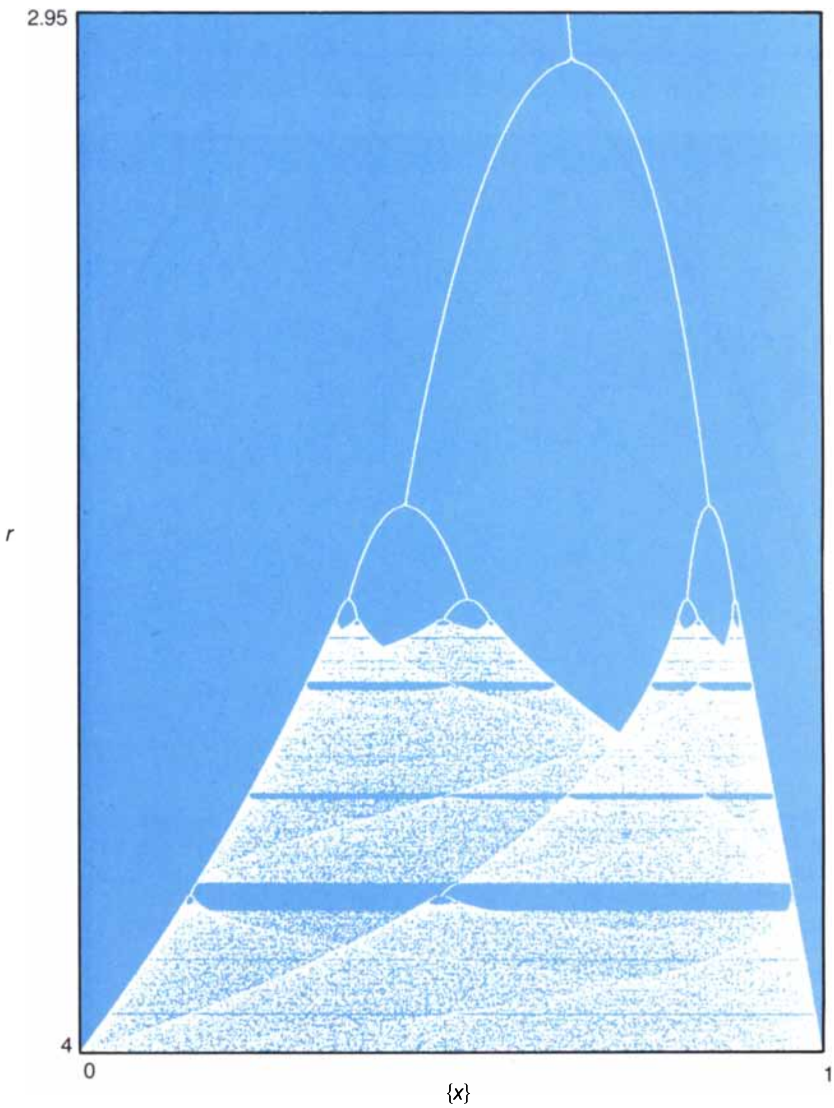
A complete picture of the behavior of the simple amplifier circuit emerges if the program computes a raft of plots, each plot below the last one [*see illustration at right*]. The plots result from a succession of r values that run from 2.9 to 4.0 in, say, 200 steps from the top of the screen to the bottom. A more elegant picture emerges if more steps are used, say 4,000, but in this case the diagram will not fit on the screen and it must be plotted to be seen as a whole.

For values of r less than 3.56 (the more precise value is 3.56994571869) the attractors of the simple dynamical system embodied in the iterated equation $x \leftarrow rx(1 - x)$ consist of a few points. These points, which represent nonchaotic behavior, are arranged in three large bands and an infinity of smaller ones. The attractors become strange as r approaches 3.56. Here chaos sets in as the hitherto smoothly bifurcating lines suddenly fall into a pepper-and-salt madness. Strangely enough, the chaotic regime vanishes from time to time as r continues its inexorable march to 4.

The entire plot is called a bifurcation diagram. When it is viewed sideways, it resembles the spectrum of chaos from a star named x. The plot is embellished by curves and attractively shaded folds. The reasons for the ornamental details are mysteries that can be explained only by the theory of chaos. I shall delve more into that topic below. For the present there is a mystery closer to home in the minds of most readers: Why does the innocent-looking equation behave so strangely?

The equation's behavior for nonchaotic values of r can be simulated geometrically by drawing a parabola described by the equation $y = rx(1 - x)$, where x is the horizontal variable and y is the vertical variable. Now superpose on the parabola the diagonal line $y = x$. Such a procedure has been followed in the top illustration on the next page, where r has been set at 3.3, a value at which the system's attractor consists of two points. To show how the system behaves, an initial value of x is chosen. I have picked .3, although almost any other value will do as well.

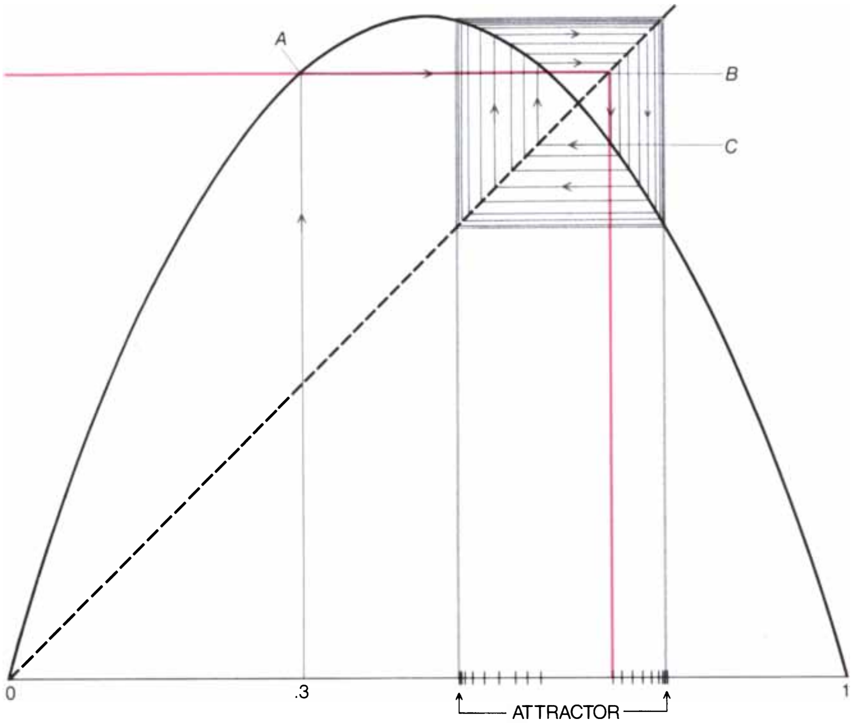
The first iteration of the equation is simulated by drawing a vertical line beginning at the point $x = .3$ at the bottom of the graph and continuing it upward until it hits the parabola. I have labeled the point where the line hits the parabola as A. The height of the intersection determines the corresponding value of y. In the second iteration that value of y is fed back into the equation as the x variable. Graphically the procedure corresponds to measuring the height of the intersection, marking it on the horizontal axis and drawing another vertical line from that mark until it hits the parabola. Here a short cut is employed by drawing a horizontal line from point A to the diagonal line $y = x$; I call the new point of intersection B. Note that point B and the origin lie at diagonally opposite corners of a square whose sides have a length equal to the value of y determined in the first iteration. As a consequence the y value can be fed back into the system by drawing a vertical line from B until it hits the parabola (point C). By continuously repeating the procedure of moving vertically until the parabola is hit and moving horizontally until the diagonal line is hit one produces a rectangular path that spirals into a square.



A bifurcation diagram shows the transition to chaos

The geometric recipe mimics the core procedure within CHAOS1. The two places where the resultant square intersects the parabola correspond to the two-point attractor. Enterprising programmers might undertake the interesting project of generating such figures by computer. In doing so armchair investigators could gain insights into the "simple" iterative equation under study. Specifically, what do the figures look like when chaos sets in? Are the random-looking numbers generated by values of r that produce chaos truly random?

I owe the idea for an excursion into chaos to a number of readers who wrote in. Among them was James P. Crutchfield, one of the authors of the article "Chaos" referred to above. Crutchfield and his coauthors explain that "the key to understanding chaotic behavior lies in understanding a simple stretching and folding operation, which takes place in the state space." In the case of the simple amplifier system the state space is a line segment that contains the attractor points and



A two-point attractor appears in a geometric simulation of a simple system

the point representing the current value of x. Where do the stretching and folding come in?

Iterating the equation $x \leftarrow rx(1 - x)$ amounts to mapping the points between 0 and 1 into a parabolic curve. Points that are close together on the unit interval, particularly those close to 0, end up farther apart when they are mapped into the parabolic curve. This happens, of course, when the number $rx(1 - x)$ replaces x. The folding operation comes about because of the bilateral symmetry of the parabola; except at the apex of the curve there are always two points on the unit interval that map into the same value $rx(1 - x)$. Those points are of course x and $1 - x$.

Much of the structure of the bifurcation diagram has been analyzed by chaos theorists. The boundaries of the chaotic regions are set by the minimum and maximum values of the iterates of $x = .5$. The curves followed by the minima and maxima, as well as those followed by the "veils" that hang so strangely down in the chaotic regions, are all simple polynomials in r. At the places where the shading is densest one finds the highest concentration of points in the strange attractors that cross them. In the empty bands mentioned above chaos gives way to order. Theory tells us that for every whole number there is a band (however narrow) with orbits of precisely that size. Finally, it will come as no surprise to readers familiar with chaos that strange attractors, even in the humble system just explored, have a fractal nature; an infinite number of points show interesting detail at all levels of magnification, like the Mandelbrot set described in this column in August, 1985.

More complicated dynamical systems are embodied in the equations named after Michel Hénon, a French mathematician. The so-called Hénon mappings not only describe physical systems such as moving asteroids and dripping faucets but also generate beautiful images in the process. A Hénon mapping consists of not one equation but two. Here is an example:

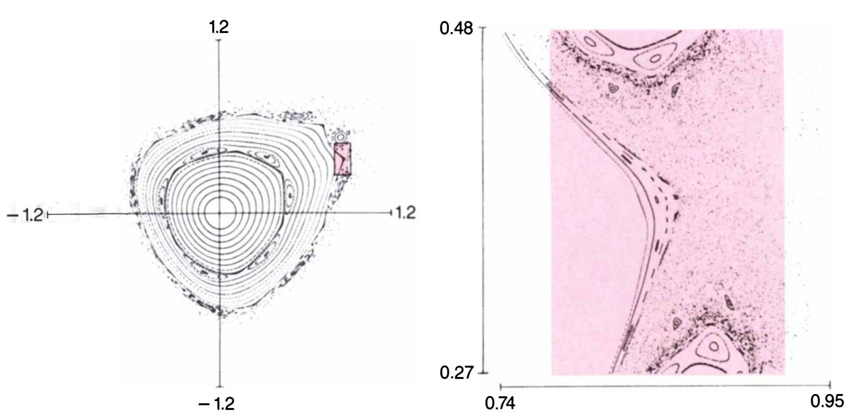
$$x \leftarrow x\cos(a) - (y - x^2)\sin(a)$$
$$y \leftarrow x\sin(a) + (y - x^2)\cos(a)$$

Current values of two variables x and y are used in the right-hand sides of both equations to produce new values (also symbolized by x and y) in the left-hand sides.

A program called CHAOS2 exploits the two equations to produce images of the order and chaos inherent in a wide class of dynamical systems. CHAOS2 has a core program that is similar to the core program of CHAOS1:

```
input x and y
for i ← 1 to 1,000
  xx ← xcos(a) − (y − x²)sin(a)
   y ← xsin(a) + (y − x²)cos(a)
   x ← xx
   plot (100x, 100y)
```

The differences between the two core programs stem from two sources: CHAOS2 has two iterated variables instead of one, and the system described by the Hénon mapping is conservative rather than dissipative. The presence of two variables forces one to employ a temporary variable xx for the new value of x while the current value of x is still being used in the second equation. The fact that the underlying dynamical system is conservative means the primary iteration loop for eliminating transient values can be removed. There are no losses of energy due to friction or other dissipative leaks. Consequently there are no attractors as such. One might say, however, that every orbit computed by the system is its own attractor. In any event, strangeness (and chaos) is certainly present in the Hénon mapping. Finally, for each setting of the parameter a the resulting system has a multitude of orbits and, owing to conservatism, any initial pair of values for x and y will represent a point that is already on one of the orbits; the attraction is instant, so to speak. For these



Successive orbits of a Hénon mapping **(left)** *degenerate into chaos* **(right)**

The Hénon mapping generates different figures for a = *.264* (left) *and* a = *1.5732* (right)

reasons the core of CHAOS2 does not use standard initial values for its iteration variables. They must be input by the computer programmer.

CHAOS2 is complete when its core is preceded by an input statement that allows the programmer to select the value of *a*. As in CHAOS1, each new value of *a* leads to a new system. But because the system is two-dimensional, a sampling of orbital plots takes up all available room; one cannot systematically vary the control parameter *a* without invoking chaos of an unwanted kind.

The user of CHAOS2 therefore specifies an initial orbit by typing in the coordinates of a point on it. Sitting back, he or she watches in fascination as the orbit is plotted. It might turn out to be a curve (traced not continuously but intermittently) or it might turn out to be something a little stranger. For example, the bottom illustration on the opposite page displays a succession of 38 orbits in a Hénon mapping with the value of *a* set at 1.111. From the center of the plot outward the orbits form a nest of closed curves until the sudden appearance of small "islands": individual orbits wedged between the larger nested ones. Farther out the nested orbits continue until the onset of chaos. In the outer reaches of the phase plot more islands appear, along with a random sprinkling of points that denotes the onset of chaos. One of the chaotic areas (outlined by a rectangle in the illustration) is shown in magnified form. Readers who want to magnify Hénon diagrams are warned to use the most precise arithmetic available on their machines.

As I have mentioned, Hénon mappings represent a great variety of conservative systems, such as asteroids orbiting the sun. Unfortunately the orbits in the diagrams are not the orbits of the asteroids but phase plots of those orbits. In the diagram just described the horizontal axis may represent the position of an asteroid in terms of its distance from the sun. The vertical axis may represent the radial velocity, or the rate of change, in this distance. Each point on the orbit computed by the Hénon mapping represents the radial distance and velocity of an asteroid at a specific angular position with respect to the sun, that is, when the asteroid passes through a vertical plane making that angle with the sun. Successive points computed by the mapping represent the asteroid's successive reappearances on the plane. The islands mentioned above are resonance bands due to perturbations in the asteroid's orbit by larger bodies in the solar system such as Jupiter. In the chaotic regions the radial position and velocity of an asteroid will vary in an essentially random way every time an asteroid revisits the specified plane. Its motion is unpredictable. Almost anything can happen.

On the aesthetic side it is worth looking at some other plots generated by Hénon mappings; quite apart from physical interpretations, there is a whimsical quality present, as is seen in the illustration above. They resemble strange, aquatic creatures.

Readers wanting to learn more about Hénon mappings should get a copy of the December 1986 issue of *Byte* magazine. There Gordon Hughes, a professor of mathematics at California State University, has engagingly described some of the relevant physics and mathematics underlying Hénon mappings. PASCAL programs are also listed.

A reader in Holland, Peter de Jong of Leiden, has already suggested some other iteration formulas that produce bizarre shapes and images. He recommends the four-parameter iterations $x \leftarrow \sin(ay) - \cos(bx)$ and $y \leftarrow \sin(cx) - \cos(dy)$. Begin with x and y both set equal to 0. Then, to get the figure de Jong calls "chicken legs," try $a = 2.01$, $b = -2.53$, $c = 1.61$ and $d = -.33$. The respective values -2.7, $-.09$, $-.86$ and -2.2 yield a "dot launcher," and the values -2.24, $.43$, $-.65$ and -2.43 produce a "self-decorating Easter egg."

Readers are free, like de Jong, to invent their own iteration formulas and to experiment with them. Anyone finding particularly attractive (or puzzling) chaos is hereby invited to send it to me in care of this magazine. Crutchfield has kindly agreed to correspond directly with those readers whose puzzlement I am not likely to satisfy. He can be reached at the Department of Physics, University of California, Berkeley, Calif. 94720.

Readers with one-track minds have by now undoubtedly solved last month's problem involving the interchange of two cars across a weak bridge. The cars are on a circular track, and an engine occupies another track connected to the circular one by a switch. The bridge is strong enough to hold one car but not the engine. How can the engine switch the cars?

The engine enters the circular track, goes to car *A* and pushes it onto the bridge. Then the engine backs around the track to car *B*, couples with it, pushes it to the edge of the bridge and couples *B* with *A*. Chugging back to the switch, the engine and two cars back onto the straight track, where *A* is uncoupled. Next the engine takes *B* back to the bridge, leaving it uncoupled there. Finally the engine circles the track, pulls *B* off the bridge to its new position and then retrieves *A*.

In the April column on computer music I left readers to ponder how to obtain long, nonrepeating sequences of notes by selecting numbers modulo *m*. The method of selection involved starting with a seed number and from then on continually multiplying by a number *a*, adding another number *b* and reducing the result by taking the remainder on division by *m*. If the numbers *a* and *m* are relatively prime (have no common factor larger than 1), the sequence will be the longest one possible. It will also produce the strangest music.

Peter de Jong notes that he has created strange music through chaos. Readers can create similar sounds by converting the numbers generated by CHAOS1 into musical notes. Outside chaotic zones there will be simple, repetitive musical phrases; inside the zones will be the very sounds of chaos.

BOOKS

The global dynamo, the earth's lights at night, Chinese fire drug and Third World agronomy

by Philip Morrison

THE EARTH'S ELECTRICAL ENVIRONMENT, by the Geophysics Study Committee, National Research Council. National Academy Press ($28.95).

Ben Franklin was lucky, all right: the key he touched to test the charge of his kite aloft under the thunderhead yielded only some hissing blue sparks. The French investigators who now fly small rockets trailing a grounded wire up into a New Mexico thunderstorm trigger full-grown lightning discharges two times out of three. A nighttime photograph catches a bolt branching upward from a twisted, glowing column. The column itself had struck upward from the lofted wire, which exploded under a sudden current load of a few tens of kiloamperes.

This book is a lively and comprehensive review of the present state of what we know about the electrical phenomena of the earth at all scales. Nearly every one of the 16 expert, readable chapters presents data and outlines rich understandings; maps, graphs and diagrams far outnumber equations and numerical tables. The authors, too many to praise by name, have made a first-rate review that is a model of its kind.

Dazzled awake by a midnight flash, you may take comfort in the thought that a watchful Defense Meteorological Satellite Program satellite has recorded it optically. The satellite flash count is mapped here worldwide for a number of months at various times of day. It is no surprise that the lightning peak moves with summer, north and south. It is interesting that lightning, like most people, avoids hot deserts and cold tundra and is a confirmed landlubber as well; although there are flashes over all the oceans, the frequency at sea is far below that on land.

The census is approximate. Midnight flashes amount to about 100 per second globally, judged by optical sensors that miss some of them. A satellite that listens for high-frequency radio noise certainly overcounts lightning; it reports about 300 flashes per second worldwide. Yet these statistics are much more reliable than the old studies from routine weather reports, when each station at day's end noted whether or not thunder had been heard.

Radio static has long served to pinpoint lightning. Nowadays, in the interest of forest-fire control, a net of automatic direction-finding wideband VHF magnetic receivers all over North America notes the location and particulars of each flash. An East Coast network reports each flash from Florida to Maine to a single computer at the State University of New York at Albany. On its busiest day, in June of 1984, it recorded 50,836 flashes around the clock, mostly from storms in three nearby states.

Our weather is a parade of flattened flows: persistent, slowly turning disks a couple of hundred miles across and only 10 miles deep. Thunderstorms are transient, smaller, chunkier: they are about as wide as they are deep, 10 miles each way. They are driven by local convective flow: thermal columns carry ground-heated moist air aloft. Finally they die out or are sheared apart. A couple of thousand of them are active worldwide at any one time. The typical lightning center is a ball of negative charge perhaps two miles across, precipitation furious within it, conspicuous to the weather radars. The flashes—of which about half remain in the clouds, half strike to ground—predominantly carry negative charge down to the earth. The separated positive charge usually towers high above in the thunderhead anvil.

Within the past decade it became clear that during severe winter storms in northern Japan the more intense flashes were reversed in charge; they brought positive charge to the earth, often with 10 times the energy of the typical summertime negative flash. By now positive flashes have been found generally. The hint that some novel charging mechanism was at work has not been borne out; it seems the cause may be simple geometry. Strong wind shear can drag the positively charged cloud 10 miles or more downwind from the central negative charge. The unusually big flash now carries positive charge from on high all the way down to ground. Although they are much less frequent, these positive flashes bear watching. They are the ones most damaging to power lines and perhaps to aircraft; quite possibly they are the chief fire setters.

The ability to observe and study thunderstorms has increased greatly within the past decade or two. It seems likely that strong electrification requires strong convective motion, precipitation and the formation of ice (although exceptional ice-free storms are reported). The negative cores of charge are found where the air temperature drops to well below zero degrees Fahrenheit, high aloft in summer, much lower in winter.

What separates charge in the first place? No crisp answer is in the offing. A fine table lists and characterizes eight microscopic mechanisms proposed between 1913 and 1977. All of them might occur in the hail stage of the thunderhead, some only in the immature cloud, among tiny cloud droplets or in the rain. The density of charge we are looking for is about a coulomb in a cubic kilometer; a couple of dozen such volumes are joined into the typical charged core. It isn't that much. The electrical energy in an entire cubic kilometer is a dime's worth or so; the kinetic energy of convection and the heat release available within that million tons of air and water are certainly larger by 100 million times. Yet the electricity-generating systems fashioned by human beings produce a flow of energy greater by several orders of magnitude than the output of the global atmosphere.

All the models of the day, now quite elaborated into three dimensions, begin with some charging at microscale, usually by cloud-particle collisions of one kind or another (bits of ice and water). The second stage is large-scale charge separation by gravitational settling or the much faster process of convection. One grand experiment in 1984 that reversed the polarity of a thundercloud by feeding it negative ions from a long suspended cable held at 100 kilovolts is a little unsettling for most model makers: it suggests a larger role for ions that are not attached to any cloud particles. That mechanism was proposed some 30 years ago by Bernard Vonnegut; he was among the workers who reversed the charge of the subject cloud.

Other chapters survey the overall processes in lower, middle and higher atmospheric layers. There are plenty of puzzles. The global electrical circuit has long been known to be closed through the conducting ionosphere. Thunderstorms are the generators that maintain a potential difference of approximately 300 kilovolts across the insulating atmosphere, with a total current flow of 1,000 amperes to the conducting earth. The input current is patched and peaky; the return current is the smooth flow from air to ground observed everywhere in fair weather.

Quite different magnitudes are observed in the three other geoelectromagnetic flows. These are the solar wind and its encounter in space with the earth's magnetic field, the winds that can deform the conducting ionosphere to set up horizontal voltages there and the majestic geodynamo currents in the earth's core. We see their effects in the aurora and in both the steady and the transient behavior of the compass needle. Perhaps one way or another they weakly couple our weather and even our seismic earth to the distant plasma-producing sun, as the lightning system is coupled to the galaxy at large through the cosmic rays that maintain the conductivity of the air at altitude.

There is more than can be summarized, of course. Thunder is now a way to map the lightning bolt by acoustic transit time; it would be fine to hear the computer-simulated thunder rolls and claps recently made by inverting the acoustic model used to map out real lightning. Lightning is a chemical process too; nowadays it contributes only some 1 percent of the total nitrogen fixed by bacteria and by humanity, but once perhaps it was the mainstay of early life. A final chapter treats earth currents induced in seawater and in all long conductors by primary currents high in the ionosphere. Predictions have it that every few years 1,000 direct-current amperes will flow along the Alaska pipeline. The phenomenon was first observed in the early days of telegraphy: the Northern Lights were bright at the end of August, 1859, and a few days later the incredulous telegraphers of New England "for more than one hour...held communication over the wires with the aid of celestial batteries alone."

EARTH AT NIGHT, by Woodruff T. Sullivan, III. A poster map, 23 by 35 inches, Catalog Number 34–5, Hansen Planetarium, 1098 South 200 West, Salt Lake City, Utah 84101 ($6, shipping charge $1.50). **NIGHT AS FRONTIER: COLONIZING THE WORLD AFTER DARK**, by Murray Melbin. The Free Press ($19.95).

The dark night sky is a sphere whose surface the lights of the stars define. Within that sphere is another, also marked in light: the surface of the earth at night. This unique poster map presents the night earth—not the night sky—on a Mercator projection, as a mosaic of carefully assembled actual nighttime scans. The images were made by two Defense Meteorological Satellite Program birds that orbit pole to pole, one near the shadow terminator around dawn and dusk, the other near the noon-midnight line, both 500 miles up.

Woody Sullivan is a radio astronomer with a keen interest in the idea of viewing our planet (and others) from afar. He has laid out these images into a world map with positional errors of only a few percent. Transmitted from orbit as digitized images of strips 1,800 miles by two miles, the scans were made by a photodiode sensitive to the entire visible range of color and the near infrared. They are reconstructed strip by strip into photographs by laser scan. The DMSP system and its striking yield were described in this magazine [see "Nighttime Images of the Earth from Space," by Thomas A. Croft; SCIENTIFIC AMERICAN, July, 1978]. The poster is the coherent composition of 40 DMSP photographs into a world map, adjusting for scale, exposure, moonlight and various small distortions. The poster is darkly beautiful, glittering in form and in meaning. To separate black land from black seas Sullivan calls on the clever graphic device of marking the empty oceans with a matte pencil striping that ends sharply at continental outlines.

The continents and islands blaze most unevenly with light. Three kinds of light are shown, all the work of our cunning species. The most familiar white areas are the populous cities whose denizens wake by night. Recognizably patterned areas shine from the Rhine-Ruhr, the south of England, the American Northeast, the Punjab, Rio de Janeiro and São Paulo, Japan east and west. Small maps in the margin neatly index about 180 identifiable cities large and small, from Fairbanks across the world to far Christchurch.

Then there is the light of the rural poor; these are not the dim lamps of the villages but the burning fields of shifting cultivators, the fires of herdsmen clearing their dry grasslands, the woodland set on fire to gain new fields. They all appear here in a distinct texture, less solidly overexposed. They glow from the Malagasy Republic to the Sahel, and from the highlands of Burma, Thailand and Laos. In years of scan some places never were freed enough from cloud cover to see; southwestern Africa and South China are thus abnormally dark, and many islands are missing.

The third kind of light is flame seriously overexposed. Its nature can be surmised at once from the bright patches that fill the desert shores of the Persian Gulf. There no one farms wide fields and few cattle graze, but the light from the Gulf rivals the brightest metropolitan centers on the earth: we are seeing the tall fiery flares of waste natural gas. These luminous genies stalk the Surgut gas fields along the Ob north of Omsk, the tropical island of Sulawesi, the delta of the Niger and the North Sea beyond the Shetlands.

Finally, strings and alignments mark out particular features such as the Nile, the Ohio, the Yellow River and the Trans-Siberian railroad. A dozen spots in a straight line not explained by the usual atlases join Madras to the bright patch of Delhi. All these lights must be centers of population that are arrayed along rivers or rail lines or perhaps are simply brightened by the megawatt bounty of a high-power line nearby. A single densely packed page of text comes with the poster to suggest still more that the eye may find and the imagination grasp in this marvelous (and inexpensive) view of the constellations of humanity.

Who is it that wakes by night among us five billion diurnal primates? City dwellers have long pushed back the night through the provision of artificial lights, which clear no stubble but make possible the activities of the urbanized. Murray Melbin is a sociologist at Boston University who documents his insight in a charming and persuasive book. Once the lands have been occupied and the spatial frontier comes to an end, human activities grow to fill a new temporal regime. Midnight and beyond are still a frontier. Few Americans dwell there, although urban life has long ago advanced well past sunset; prime time has been thickly settled and is no longer an adventuresome territory.

The cartography is tentative. A national census was taken on the American night frontier in May, 1980. About 30 million people were up and about after midnight; by the small hours between 3:00 and 5:00 A.M. the count had dropped to 10 million. That scant 5 percent represents the core of incessancy, the worldwide spread of purposeful social wakefulness that never stops. Fewer than four million were wage earners at regular work. More than four million were patrons of public establishments, stopping, say, for meals or fuel. Most of these seemed to have been on the way somewhere, perhaps home; a million and a half more were in transit by private automobile. It is a restless frontier.

Three schedules of the world of

work and play are clear. The first one is ruled by daylight. The second is the spread of everyday life over the boundaries of natural light. Fishermen, bakers and wholesale markets traditionally start early to prepare for the morning; at night the metabolism of every city now turns largely rejuvenative—to cleaning, repair, waste removal, maintenance, revelry and recreation. The third schedule is newer: it is the rule of the incessant organization explicitly at work around the clock, partly in response to telecommunications that can relate the distant sunlit world to local night.

Incessancy is still an open frontier. The White House, the chemical plants, the powerhouses, the international airports, the newspapers, the hospitals and the shops to serve their people work around the clock. Yet the Marine guards at the U.S. Embassy in Moscow and the operators at Bhopal and Chernobyl do not seem to have included the heads of those enterprises. Leadership is still largely single and diurnal; the night duty officer is junior. But the trend is clear. At Three Mile Island the staff knew what to do in their predawn crisis: they awakened the plant manager himself. "Incessance had arrived."

On the streets of Boston in 1974 the researchers asked for directions, sought open shops to count friendly comments and proposed interviews to a large sample right around the clock. The night people unmistakably treated one another in a friendlier way than is common among day dwellers. This is the hospitality of the frontier; people there are bound by a sense of common vulnerability to random risk. The people of the night see themselves as suited to a special niche, as pioneers. "In a well-established society there is less adversity and less good will than at the precarious edge of human settlement."

We seem slowly to be becoming more wakeful. "Reared in the changed environment, our children will take incessance for granted," and the midnight satellites of the future will record more light everywhere. Or perhaps not; a prosperous future will have to waste less fuel than we do.

SCIENCE AND CIVILISATION IN CHINA, VOLUME 5, PART 7. MILITARY TECHNOLOGY: THE GUNPOWDER EPIC, by Joseph Needham, with the collaboration of Ho Ping-Yü, Lu Gwei-Djen and Wang Ling. Cambridge University Press ($99.50).

"This volume has been 43 years in the gestating" since one summer evening when the talk turned to the history of gunpowder and a very old printed passage was found and copied out. Dr. Needham and a few senior Chinese historians had been evacuated after some adventures to a "delightful little town" in Szechuan. The young Wang Ling was there too, and he made the matter his lifelong study. The first draft of this sparkling fire fountain of learning was lighted by Dr. Ho; by now all four authors have had a hand in the book.

This is a story of power, power both as swift release of energy and as enforcement of the will of the state. On page 4 a chart diagrams the entire development, a current that flows from Byzantium to Cape Canaveral. The medieval Chinese fire drug would come to depose the feudal lords of distant Europe; in 1449 the artillery train of the king of France toured the castles held by the English in Normandy to batter them down one by one "at the rate of five a month." In China the same stuff, employed in the widest variety of weapons for five centuries, had essentially no effect on the age-old bureaucratic structure, which held no castles and fielded no armored cavalry. Like printing in China, gunpowder made no revolution there; like the Chinese compass, gunpowder in the end equipped the overseas barbarians who would for a while humiliate the descendants of its early inventors.

In summary of this thick volume we address mainly the engineering history of the first chemical explosive. Fire flies with the arrows of war out of the deepest archaic past. The natural seepage of petroleum was a major resource, found in many places and used widely for military incendiaries. The coming of the first big change, the formidable Greek fire itself, can be dated with some precision. It depended on the discovery of distillation to produce a low-boiling hydrocarbon fraction: naphtha, a kind of kerosene or petrol—free-flowing, volatile, most inflammable. The blazing stuff was pumped over enemy craft in the sea fight that saved Byzantium from Arab attack in the late seventh century. The defending ships were called siphon bearers: behind iron shields their Greek seamen worked rumbling bronze flamethrower pumps. The technique made its way to China, along with the fire oil itself, through the sea trade of the Arab merchants into southeastern Asia. By the year 1000 or so flamethrower pumps were government issue in the Chinese armies.

Meanwhile the Taoist alchemists, seeking not death but the elixir of life, had from the third century or so received sulfur and of course carbonaceous substances into their masteries of the furnace. But the singular key to gunpowder was of course its internal oxidant saltpeter, or potassium nitrate. There is a long history of Chinese alchemical experimentation with the natural decomposition products of organic material, nitrates recent or ancient, and a slow rise of the recognition and preparation of increasingly pure saltpeter. By 850 or 900 the chemistry and its product were well understood. After 1200 the Arab chemists would call the stuff "Chinese snow."

The first loud, clear report of the sudden deflagration of the triple mix, saltpeter, sulfur and carbon, is heard in an entirely nonmilitary context. Among dozens of procedures against which the responsible Taoist adept is warned, a famous book of the Tao canon, probably from before 900, records that some people who had heated together sulfur, realgar and saltpeter with honey had been burned in hands and face "and even the whole house burned down." The internal oxidant was at work.

The weapons makers enter slowly. The usual Chinese term relates gunpowder to medicine or alchemy; it is called fire drug. In English the name is military from the start, powder for guns. Contemporary texts—not fully clear to a reader—suggest its first military use as a slow-match igniter for Greek fire in China in about 920. Pretty soon it is certainly used on its own in incendiary bolts and arrows, wrapped up and lighted behind the arrowhead to set a fire that is hard to extinguish. There is a description and a picture, although much later, of this "fiery pomegranate shot from a bow." Next comes the gunpowder flamethrower, its burning gases erupting from a bamboo tube—a fire lance. Such arms are still clearly seen in photographs of passenger junks of the 1930's, mounted for defense against pirates in the South China seas.

If gases flow fiercely out, why not projectiles along with them? Now the fire tubes hold sand and broken porcelain, and even human feces and other materials hopefully deemed to be toxic, all of which spurt out with the rushing flame. Next, the true explosive. (Of course the powder recipe steadily changes; the texts confirm that.) By the 1100's they have made a higher-nitrate mixture, enclosed in a weak container of bamboo; a loud bang is its main effect. A picture shows this "thunderclap bomb." Progress is swift; by 1230 or so the Chin Tatars stoutly defended Khaifêng from the powerful Mongols. The only weapons the Mongol soldiers feared were the flying-fire spears (big fire lances) and the thunder-crash bombs. Those bombs, the oldest true explosive devices, were gunpowder-loaded cast-iron shells two inches thick, shaped like bottle gourds. It was a famous defense, although it could not save the city or the dynasty.

Then comes the true propellant. For a long time the fiery gases drag out only ill-fitted scraps of all kinds. But in the end they propel well-made balls and arrows with gaskets (wads) that begin to occlude the opening; the true gun was born, first of all in bamboo and finally in metal. The evolution is intricate, but by about 1290 the species has clearly appeared. An eight-pound bronze hand bombard has been excavated in Manchuria, without a date but with excellent supporting ages derived from the archaeology and from local texts. The oldest gun, it must go back to just before 1290.

The first illustration of a gun in Europe is dated 1327. The picture seems clear: judging by the comments of Roger Bacon and the response of the arsenals, Europe eagerly seized on the villainous saltpeter between about 1270 and 1310 or so, as soon as the long Chinese development reached the artillery stage.

How was gunpowder brought to far-off Europe? There is a closing piece full of surmises bolstered by what evidence there is: yellow-haired Norsemen from Novgorod, Arab military engineers in China, learned friars and Nestorians, an Italian merchant colony in Tabriz? By the start of the 1300's "the bell had rung...and the Western world was set upon the fateful road" that led to small arms, artillery, heat engines and space travel. The mandarinate, however experienced, soon proved no match for the fervor of Europe, of its investors, its inventors, its kings and its sects. Before 1650 the busiest and most celebrated of the cannon founders in Peking itself were two Jesuit fathers.

INDIGENOUS AGRICULTURAL REVOLUTION: ECOLOGY AND FOOD PRODUCTION IN WEST AFRICA, by Paul Richards. Westview Press ($34; paperbound, $24.95).

Among the Mende farmers of Sierra Leone, it seems, there has been an experimental approach to agriculture. They harvest not with a sickle but with a knife, and every head of grain comes to the palm of the hand and under a quick eye. They will mark a small plot, plant and harvest it and carefully measure with the same calabash the seed planted and the seed harvested. In 1983, the author recalls, "I came across three or four cases where farmers were...undertaking trials involving selections" among materials rogued from earlier harvests. The author has put a photograph of a rice head selected by the farmers on the cover of his small, tightly argued and subtle book. The grain has long awns and stiff outer glumes—"jaw-boned rice" in the local language—which deter birds. Bird damage is a major problem locally for the cultivator of quick rices. The textbooks advise selection to get rid of those very characteristics.

In that too small nutshell the nub of Richards' argument is held. On the positive side, the inventive smallholders of West Africa have long demonstrated both their degree of understanding of their world and their penchant for sensible innovation. Unlike those in many parts of the world of traditional agriculture, these farmers, men and women, have made their crops in a remarkably diverse environment where land was plentiful and labor short, where wilderness and cultivation—forest and savanna—adjoin, where rain is abundant but seasonal. They generally know more from life than the new expert is likely to know from books. On the negative side, an approximate if general theory tightened by experience of distant conditions is a poor guide. "No one expects a pilot to captain a plane on the basis of textbook knowledge alone." Yet the adviser without experience often expects to take the controls of the farm from someone whose life depends on his own crops. On-farm trials in the "full sceptical gaze of local agricultural opinion" offer a better model.

The tale is not new. Chocolate (cocoa) was brought by the Portuguese to the islands off West Africa as a plantation tree crop but was rigorously kept from the mainland. Under British colonial rule the major cash crop of several West African dependencies became cocoa, but it was not introduced by the colonial administrations. The economic historian Polly Hill in 1963 demonstrated that cocoa was brought to Ghana surreptitiously by a migrant worker who returned from the plantation, and that its thriving culture was almost wholly dependent on local initiative. In a similar way the Yoruban freed slaves who returned after emancipation in Brazil brought the technique for making the dried cassava meal called *gari*. That was an innovation of real importance, since it made possible widespread use of high-yielding "bitter" varieties of the root crop, which were otherwise rejected because of their high content of cyanide.

In India the famines that followed crop failure early induced the agriculturists of the Raj to place emphasis on peasant-farmer crops and crop systems; cash crops came second. In West Africa the official model was Caribbean: cash crops for export were everything, the food supply would take care of itself. There is an interesting case of an Indian expert brought in to advise on how the farmers in the mangrove swamps of the Scarcies estuary could be taught to grow more rice. He found their yields per acre were higher than those of Madras, and he recommended that some experienced farmers be dispatched for extension work elsewhere. The administrators were truculent, however: they thought the high yields only showed how immensely fruitful the land was even in the hands of negligent cultivators.

Why did the African farmers seem so shiftless? They worked hard, but differently. They did not plow: simple economics. Plowing saves labor in cultivation—but weeding is their main work; plows do not pay, particularly with no animal traction. They practiced bush-fallow rotation, burning up all that fine green manure—but nitrogen is not their need; plenty of cyanobacteria and legumes are about. It is minerals the land needs, and the ashes remain. The farmers plant a variety of crops on the same land; 80 percent of the land is intercropped. Some diversity is for nutrition and good eating, but some crops in fact help each other.

This is a book rooted in and admiring of the farmers' garden plots of Sierra Leone, but it is far from narrow and provincial. Its burden is accompanied by a good deal of cogent critique about grand theories from Spencer to Marx and Lenin, about how we know, about preconceptions and trials and about agrarian populists from Kansas to Russia. A "people's science" is worth pursuing in West Africa above all because it is good science.

The author himself argues from two deep defining points of theory that are hard to fault. Agricultural development, he says, is both applied biology and a social science. Literacy is not a necessary condition of successful agricultural experiment. The agricultural revolution in Britain was not made by trained agronomists from the universities. An 18th-century experiment in irrigation was recently described in the right tone; one Mr. Bakewell's "proof pieces" of land conveyed "complete conviction to the mind of every person who views them."

It is no use to be sentimental either. Village life is full of conflict and error, like life anywhere. Self-help R&D, like imported R&D, is not a magic formula. The farmers certainly need what science can bring them, while the rice breeders and the extension agents need the energy and the thought of experienced farmers. The options are there to use. The book has something important to say to everyone who is interested in development.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

CAN ADVANCED TECHNOLOGY SAVE THE U.S. STEEL INDUSTRY?

TECHNOLOGY AND STEEL INDUSTRY COMPETITIVENESS. Office of Technology Assessment. U.S. Government Printing Office, 1980.

RADICALLY INNOVATIVE STEELMAKING TECHNOLOGIES. Julian Szekely in *Metallurgical Transactions B,* Vol. 11B, No. 3, pages 353–371; September, 1980.

THE COMPETITIVE STATUS OF THE U.S. STEEL INDUSTRY: A STUDY OF THE INFLUENCES OF TECHNOLOGY IN DETERMINING INTERNATIONAL INDUSTRIAL COMPETITIVE ADVANTAGE. Bruce S. Old, chairman. National Academy Press, 1985.

THE MOLECULES OF VISUAL EXCITATION

CONTROL OF ROD DISK MEMBRANE PHOSPHODIESTERASE AND A MODEL FOR VISUAL TRANSDUCTION. P. A. Liebman and E. N. Pugh, Jr., in *Current Topics in Membranes and Transport,* Vol. 15, pages 157–170; 1981.

FLOW OF INFORMATION IN THE LIGHT-TRIGGERED CYCLIC NUCLEOTIDE CASCADE OF VISION. Bernard K.-K. Fung, James B. Hurley and Lubert Stryer in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 78, No. 1, pages 152–156; January, 1981.

MILLISECOND ACTIVATION OF TRANSDUCIN IN THE CYCLIC NUCLEOTIDE CASCADE OF VISION. T. M. Vuong, M. Chabre and L. Stryer in *Nature,* Vol. 311, No. 5987, pages 659–661; October 18, 1981.

CYCLIC GMP CASCADE OF VISION. Lubert Stryer in *Annual Review of Neuroscience,* Vol. 9, pages 87–119; 1986.

THE RINGS OF URANUS

RINGS: DISCOVERIES FROM GALILEO TO VOYAGER. James Elliot and Richard Kerr. The MIT Press, 1984.

THE RINGS OF URANUS. J. L. Elliot and P. D. Nicholson in *Planetary Rings,* edited by Richard Greenberg and Andre Brahic. University of Arizona Press, 1984.

RINGED PLANETS: STILL MYSTERIOUS. Jeffrey N. Cuzzi in *Sky and Telescope,* Vol. 68, No. 6, pages 511–515, December, 1984; Vol. 69, No. 1, pages 19–23, January, 1985.

VOYAGER 2 MISSION TO URANUS. Reports in *Science,* Vol. 233, No. 4759; July 4, 1986.

BEACHES AND BARRIER ISLANDS

MAN'S IMPACT ON THE BARRIER ISLANDS OF NORTH CAROLINA. Robert Dolan, Paul B. Godfrey and William E. Odum in *American Scientist,* Vol. 61, No. 2, pages 152–162; March–April, 1973.

BARRIER ISLANDS. Robert Dolan, Bruce Hayden and Harry Lins in *American Scientist,* Vol. 68, No. 1, pages 16–25; January–February, 1980.

THE OUTER BANKS OF NORTH CAROLINA. Robert Dolan and Harry Lins. *U.S. Geological Survey Professional Paper No. 1177-B,* 1986.

LYME DISEASE

LYME ARTHRITIS: AN EPIDEMIC OF OLIGOARTICULAR ARTHRITIS IN CHILDREN AND ADULTS IN THREE CONNECTICUT COMMUNITIES. Allen C. Steere, Stephen E. Malawista, David R. Snydman, Robert E. Shope, Warren A. Andiman, Martin R. Ross and Francis M. Steele in *Arthritis and Rheumatism,* Vol. 20, No. 1, pages 7–17; January–February, 1977.

LYME DISEASE: A TICK-BORNE SPIROCHETOSIS? Willy Burgdorfer, Alan G. Barbour, Stanley F. Hayes, Jorge L. Benach, Edgar Grunwald and Jeffrey P. Davis in *Science,* Vol. 216, No. 4552, pages 1317–1319; June 18, 1982.

LYME DISEASE SPIROCHETES INDUCE HUMAN AND MURINE INTERLEUKIN 1 PRODUCTION. Gail S. Habicht, Gregory Beck, Jorge L. Benach, James L. Coleman and Kimberly D. Leichtling in *The Journal of Immunology,* Vol. 134, No. 5, pages 3147–3154; May, 1985.

CHEMICAL AND BIOLOGIC CHARACTERIZATION OF A LIPOPOLYSACCHARIDE EXTRACTED FROM THE LYME DISEASE SPIROCHETE (*BORRELIA BURGDORFERI*). Gregory Beck, Gail S. Habicht, Jorge L. Benach and James L. Coleman in *The Journal of Infectious Diseases,* Vol. 152, No. 1, pages 108–117; July, 1985.

COLD NUCLEAR FUSION

MESIC MOLECULES AND MUON CATALYSED FUSION. L. Bracci and G. Fiorentini in *Physics Reports,* Vol. 152, No. 4, pages 169–216; 1982.

OBSERVATION OF UNEXPECTED DENSITY EFFECTS IN MUON-CATALYZED *D-T* FUSION. S. E. Jones, A. N. Anderson, A. J. Caffrey, C. DeW. Van Siclen, K. D. Watts, J. N. Bradbury, J. S. Cohen, P. A. M. Gram, M. Leon, H. R. Maltrud and M. A. Paciotti in *Physical Review Letters,* Vol. 56, No. 6, pages 588–591; February 10, 1986.

DETUNING REDUCTION OF MUON STICKING IN RESONANT MUON-CATALYZED *D-T* FUSION. Michael Danos, Berndt Müller and Johann Rafelski in *Physical Review A,* Vol. 34, No. 5, pages 3642–3645; November, 1986.

MUON-CATALYSED FUSION REVISITED. Steven Earl Jones in *Nature,* Vol. 321, No. 6066, pages 127–133; May 8, 1987.

AERODYNAMICS OF WIND POLLINATION

WIND POLLINATION IN THE ANGIOSPERMS: EVOLUTIONARY AND ENVIRONMENTAL CONSIDERATIONS. David R. Whitehead in *Evolution,* Vol. 23, No. 1, pages 28–35; March, 1969.

LIFE IN MOVING FLUIDS: THE PHYSICAL BIOLOGY OF FLOW. Steven Vogel. Princeton University Press, 1983.

PALEOBOTANY AND THE EVOLUTION OF PLANTS. Wilson N. Stewart. Cambridge University Press, 1983.

THE MOTION OF WINDBORNE POLLEN GRAINS AROUND CONIFER OVULATE CONES: IMPLICATIONS ON WIND POLLINATION. Karl J. Niklas in *American Journal of Botany,* Vol. 71, No. 3, pages 356–374; March, 1984.

ARCHES AND VAULTS IN THE ANCIENT NEAR EAST

ANCIENT EGYPTIAN MASONRY. Somers Clarke and R. Engelbach. Oxford University Press, 1930.

ARCHITECTURE FOR THE POOR. Hassan Fathy. University of Chicago Press, 1973.

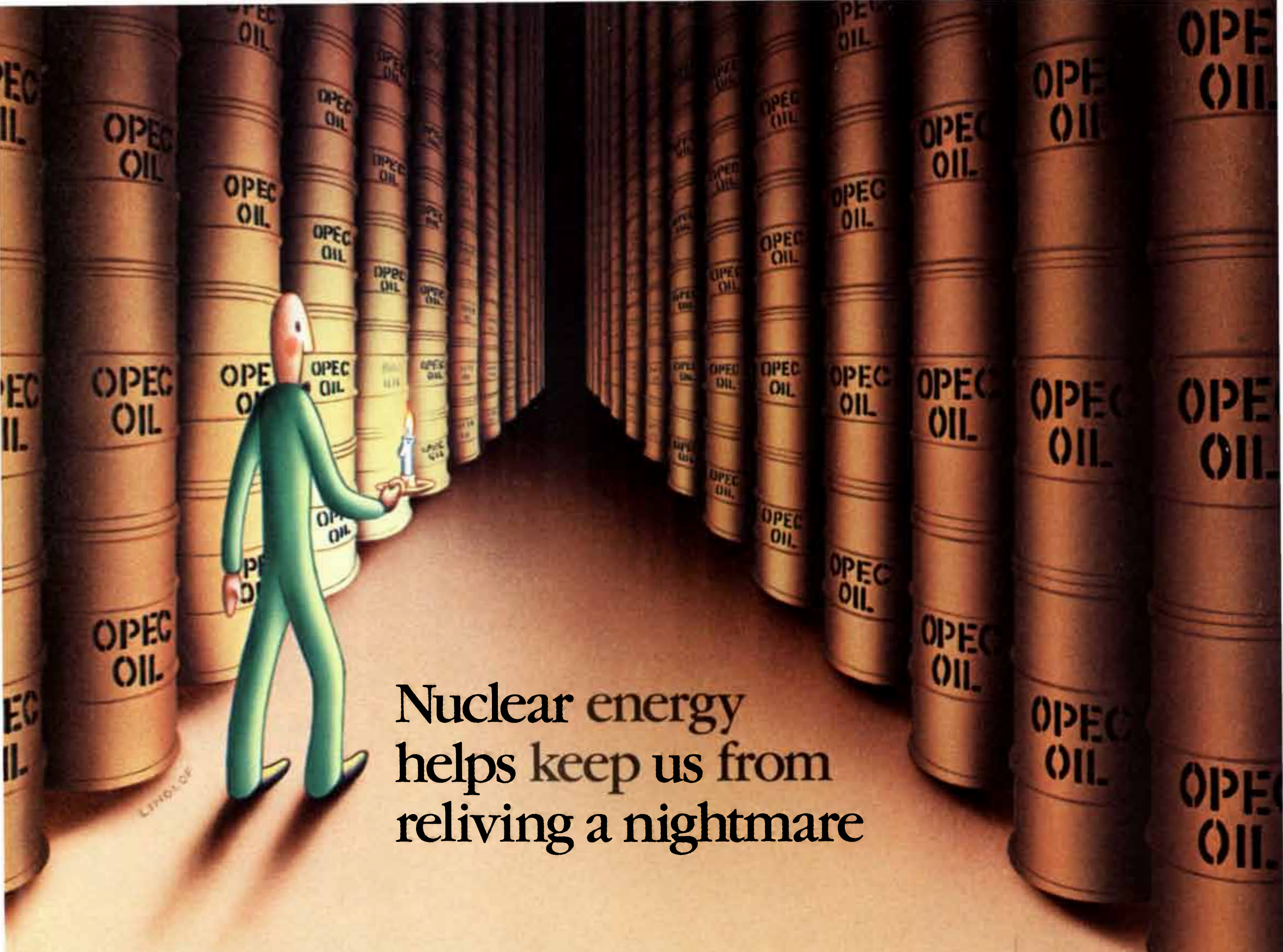
DIGGING UP TELL JEMMEH. Gus W. Van Beek in *Archaeology,* Vol. 36, No. 1, pages 12–19; January/February, 1983.

THE AMATEUR SCIENTIST

A NEW CONCEPT OF RENAL HEMODYNAMICS. W. J. Gatzek in *Medical Hypotheses,* Vol. 4, No. 3, pages 221–230; May–June, 1978.

COMPUTER RECREATIONS

HENON MAPPING WITH PASCAL. Gordon Hughes in *Byte,* Vol. 11, No. 13, pages 161–192; December, 1986.



Nuclear energy helps keep us from reliving a nightmare

The 1973 Arab oil crisis is a haunting reminder of the darker side of foreign oil dependence. Since then, America has turned more to electricity from nuclear energy and coal to help restore our energy security. As a result, these are now our leading sources of electricity and a strong defense against an increasing oil dependence that again threatens America's national energy security.

A dangerous foreign oil dependence

America imported four million barrels of oil a day in 1985. Last year that increased by another 800,000 barrels a day. The danger? Most of these new barrels come directly from OPEC. And the U.S. Department of Energy estimates that by year-end 1987, oil imports will be 30% higher than the 1985 level—an ominous trend.

U.S. Interior Secretary Donald Hodel recently warned that "OPEC is most assuredly getting back into the driver's seat" and our increasing dependence will be "detrimental to the country's economic and national security and its financial well-being."

Nuclear electricity's contribution

America's electric utilities have helped diminish OPEC's impact. Today, over 100 nuclear plants make nuclear energy our second largest electricity source, behind coal.

And nuclear energy has helped cut foreign oil demand. It's saved America over two billion barrels of oil since 1973, and our nuclear plants continue to cut oil use. The energy analysts at Science Concepts, Inc. estimate that by the year 2000, nuclear energy will have saved us between seven and twelve billion barrels of oil.

Nuclear energy for a secure future

Nuclear energy is not just helping here in America. According to OPEC, nuclear energy has permanently displaced about six million barrels of oil a day in world markets.

The lessons we learned in 1973 are lessons we can't afford to forget. Nuclear energy and coal can't offer us guarantees against another oil crisis. But the more we hear about the return of OPEC dominance, the more we need to remember the critical role played by electricity from coal and nuclear energy in fueling America's economy and protecting our future.

For a free booklet on energy independence, write to the U.S. Committee for Energy Awareness, P.O. Box 1537 (OP20), Ridgely, MD 21681. Please allow 4-6 weeks for delivery.

Information about energy America can count on
U.S. COMMITTEE FOR ENERGY AWARENESS