SCIENTIFIC AMERICAN

MIRROR MAZE

$2.50

June 1986

re is full dictions.

you might deduce that the handling is an exhilarating experience. You would be correct.

The Accord DX Hatchback shown here comes with rear seatbacks that split vertically for versatility. It has thirty-seven percent more cargo area. And there is a rigid cargo area cover. The steering wheel adjusts. Panel switches work with a touch. And new front seats with adjustable headrests offer better support.

For your convenience, we have included a quartz digital clock. It will keep an accurate account of the good time you will have while driving the new Accord. Nothing contradictory about that.

HONDA

The New Accord

Established 1845 **SCIENTIFIC AMERICAN** June 1986 Volume 254 Number 6

ARTICLES

34 **BEYOND THE ERA OF MATERIALS, by Eric D. Larson, Marc H. Ross and Robert H. Williams** Changing patterns of materials consumption define the life cycle of an economic era.

42 **FIBRONECTINS, by Richard O. Hynes**
They hold cells in place, guide embryonic development and may play a role in tumor metastasis.

52 **IS NATURE SUPERSYMMETRIC? by Howard E. Haber and Gordon L. Kane**
A new theory may satisfy physicists' quest for fundamental laws that explain the structure of matter.

76 **EL NIÑO, by Colin S. Ramage**
An atmospheric-pressure seesaw warms the equatorial Pacific and brings global changes in weather.

84 **THE HEART AS A SUCTION PUMP, by Thomas F. Robinson, Stephen M. Factor and Edmund H. Sonnenblick** It expands actively, drawing blood from the veins into the ventricles.

92 **SEXUAL SELECTION IN BOWERBIRDS, by Gerald Borgia**
Why do females prefer certain males? Study of the bowerbird's ornate bowers may provide answers.

102 **THE PERCEPTION OF APPARENT MOTION, by Vilayanur S. Ramachandran and Stuart M. Anstis** How does the visual system perceive that an object seen at intervals is really moving?

110 **MODERN WINDMILLS, by Peter M. Moretti and Louis V. Divone**
They generate electricity by utilizing new materials and technology according to ancient principles.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
14 THE AUTHORS
16 MATHEMATICAL GAMES
24 BOOKS
62 SCIENCE AND THE CITIZEN
120 THE AMATEUR SCIENTIST
128 BIBLIOGRAPHY

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1986 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, 415 MADISON AVENUE, NEW YORK, N.Y. 10017

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, UT 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, MA 01970. Specify CCC Reference Number ISSN 0036-8733/84. $1.25 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper 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 photograph on the cover shows part of the Hall of Mirrors at Glacier Park in Lucerne. The mirrors create such a visually confounding maze that a visitor who is not familiar with it is likely to become quite disoriented (see "The Amateur Scientist," page 120). For example, what looks like a hallway to the left of the visitor is in fact only an apparent hallway resulting from the pattern of reflections from the mirrors. The mirrors are laid out on an array of equilateral triangles marked on the floor. Some triangles lack mirrors; others have one or at most two of them. Here and there in the maze are pieces of Moorish furniture. Ordinarily a person in the maze sees only a jumble of images or one of the apparent hallways.

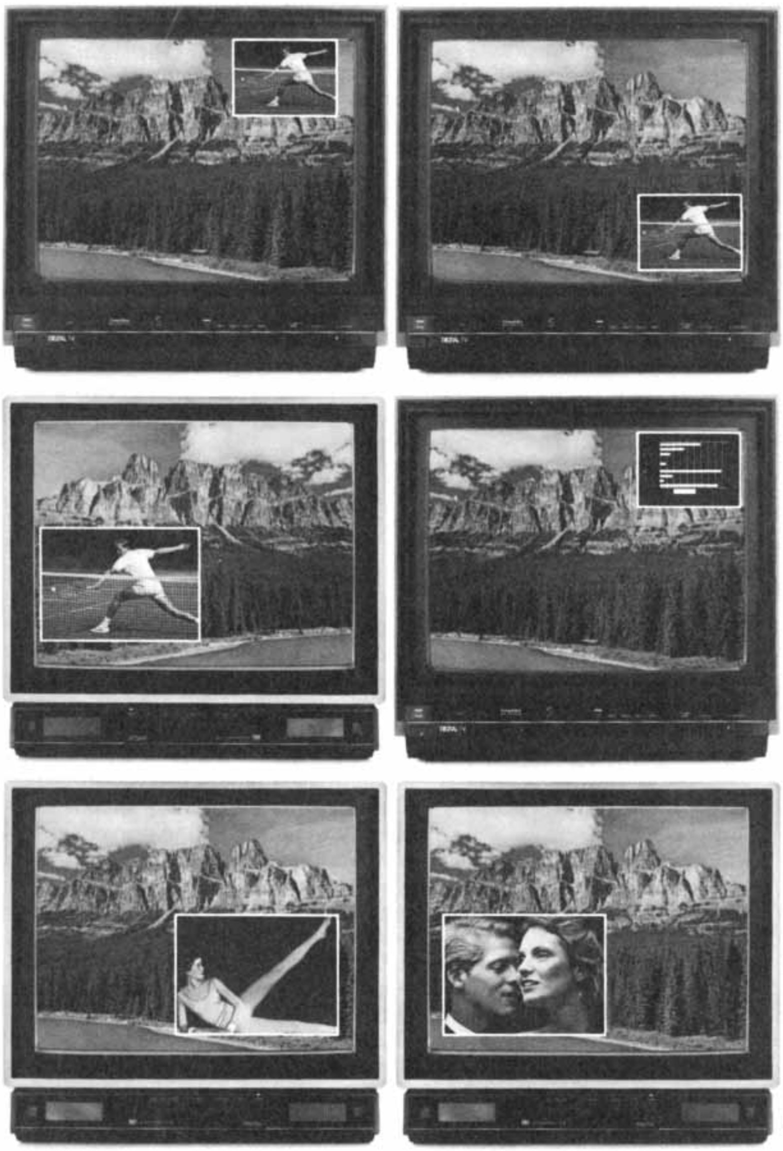
THE ILLUSTRATIONS

Cover photograph by Piergiorgio Sclarandis, Black Star

Page	Source
16–23	Joan Starwood
35–41	Andrew Christie
43	Richard O. Hynes, Massachusetts Institute of Technology
44	Richard O. Hynes, Massachusetts Institute of Technology (*top and bottom*); Kenneth M. Yamada (*middle*)
45	Jan Löfberg
46	Jean Paul Thiery (*top*), George V. Kelvin (*bottom*)
47–51	George V. Kelvin
53–60	Gabor Kiss
76–82	Ian Worpole
85–86	Carol Donner
87	Ilil Arbel
88–89	Carol Donner
90	Thomas F. Robinson, Albert Einstein College of Medicine of Yeshiva University
93	Hans and Judy Beste, Animals, Animals

Page	Source
94–98	Barbara L. Gibson
100	Hans and Judy Beste, Animals, Animals
103	The Granger Collection
104	Joan Starwood
105	Joan Starwood (*top*), Ronald M. Pickett (*bottom*)
106	Joan Starwood
107	Joan Starwood (*top*); Vilayanur S. Ramachandran, University of California at San Diego, and Patrick Cavanagh, University of Montreal (*bottom*)
108–109	Joan Starwood
111	Sandia National Laboratories
112–115	Alan D. Iselin
116	Solar Energy Research Institute
118	Louis V. Divone
120	Piergiorgio Sclarandis, Black Star
121–126	Michael Goodman

The hottest thing on TV this season will be the sets.



You're going to see them advertised everywhere.

Television sets that let you watch more than one thing at a time.

For example, you can take a peek at what your videotape recorder is recording, or see another channel. Without leaving the program that's on.

A picture within the picture lets you zoom in. Freeze the action. Someday, even see *nine* pictures at once.

This is digital television. And ITT pioneered it.

Lots of fine manufacturers are introducing digital TV in the U.S. (In Europe ITT had it first.)

And every set, on both sides of the Atlantic, starts with ITT microchips.

It's a good example of the innovative things we're doing in microchips. And of the strength of our semiconductor business.

What it all adds up to (says a leading newsweekly) is a "quantum leap in technology."

And a quantum opportunity for us.

Want to know more about ITT? Phone toll free 1-800-DIAL-ITT for a continuously updated message.

ITT

It's a different world today.

LETTERS

To the Editors:

"Athletic Clothing," by Chester R. Kyle [SCIENTIFIC AMERICAN, March], is a fascinating article about the role of athletic clothing in optimizing athletic performance and preventing injury. I believe an important issue has been overlooked, however.

Two years ago I injured my knee, apparently from running. The injury has healed, but the long recovery made me consider the many factors that might have contributed to the injury. The primary culprit, I have tentatively concluded, is the long-distance running shoe.

I do not run long distances. On the average I run two miles. Although this distance is, I believe, common among amateur runners, it seems as if nearly all off-track running shoes on the market are designed for the long-distance runner. Rarely do I find a shoe that does not have so much padding that one feels as though one is running on a sponge.

That is only part of the problem. I believe such running shoes are designed for an "average" mode of running that almost never occurs. Specifically, advertisements for Nike running shoes (which are mentioned in the article) claim that Nike engineers have uncovered the significant phenomena at work in the dynamics of running as the foot moves forward, strikes the ground and eventually leaves the ground. Nowhere do I see a reference to the contribution of the lateral running motions that take place when a runner changes direction to the overall dynamics. Kyle also fails to mention this aspect of running motion in his article.

Moreover, I have informally surveyed friends who run and have found that none of them has the luxury of running in a straight line on a flat surface. All roads are crowned, and when one travels on foot, it is apparent just how much the average surface is crowned. Sometimes cul-de-sacs have sidewalks that provide a long stretch of flat running surface. But how often does one have to twist and sidestep and change direction on such a road to avoid puddles and breaks in the pavement and to accommodate turns in the road? Running on a track avoids some of these problems, but tracks are curved.

Crowned road surfaces, like changes in direction, cause the foot to strike the ground at an angle. Since I am an engineer, I made some back-of-the-envelope calculations of the forces at work when the foot, shod in a typical long-distance running shoe, lands at an angle. As an approximation, ignoring forward motion, I modeled the foot as two lever arms, one above the other, connected by a pivot, with the lower lever connected by a second pivot to the ground. Each pivot allows one angle of freedom, with the axes of motion oriented in the direction of travel, giving a total of two degrees of freedom. The upper lever arm represents the tibia-fibula complex, connected by a pivot (the ankle) to the lower lever arm, which represents the rigid combination of the foot and the shoe. The contact point with the ground is represented by the second pivot. The foot rotates about the point while making contact with the ground. Since the foot is assumed to be landing at a lateral angle because of a crowned surface or a change of running direction, the shoe strikes the ground along either its inner or its outer edge.

A similar model of the bare foot serves for comparison, making it possible to judge the effects of wearing a running shoe when the foot lands at an angle. Using these models, I have reached two conclusions. The first is that the torque on the ankle is larger with the running shoe. The second is that the ankle must rotate to be stable, otherwise it is poised holding the sole of the shoe on edge. Both effects arise from the thick flared heel of most modern running shoes, which the article states is intended to reduce the rolling of the foot. I must conclude that the designers of such shoes (and Kyle) are considering only the case of running in a straight line on a flat surface; on surfaces that cause the foot to land at an angle the laws of physics indicate that a thick flared sole must cause a great deal more roll than the naturally rounded bare heel, and that the resulting torque transmitted to the ankle must be larger with such a sole. It is well known that foot roll is a primary cause of knee injury.

To make matters worse, Kyle suggests that a way to avoid foot roll is to form the inside of the heel sole of a harder material. This would indeed reduce roll in straight-line running, but by making the heel more "boardlike" such a sole would magnify the effects that take place when the runner turns.

I do not dispute the desirability of padding and pronation-control features; I simply contend that some of the mechanisms in today's long-distance running shoes work properly only under very limited conditions having little to do with the reality of running, and that these mechanisms may in fact increase the risk of injury for the average runner. Perhaps the heel sole itself ought to be rounded, like the ball of the foot, to accommodate turning. Nature could not have been so wrong in designing the foot this way.

CLIFFORD J. BERG

Burtonsville, Md.

To the Editors:

I have consulted with Peter R. Cavanagh of the Biomechanics Laboratory at Pennsylvania State University and Edward C. Frederick of Nike Research, who agree that Mr. Berg has cited a problem often mentioned by runners. The modern running shoe is not designed for rapid cutting or turning, as shoes for tennis or basketball are. In those activities the shear forces are much higher than they are in long-distance running, and a thinner sole with less heel lift gives better performance under those conditions. The running shoe is also not ideal for the rough ground encountered in trail running. A high-top shoe will help a trail runner to sense the position of the ankle on rough ground and make corrections to stabilize it, even though such a shoe does not do much to support the ankle.

Frederick emphasizes that research shows heel lifts of up to between half an inch and five-eighths of an inch actually stabilize the subtalar joint, the joint below the ankle at which pronation takes place. Higher heel lifts can cause instability and high ankle torques. On a crowned road the angle assumed by the foot is the same regardless of the thickness of the sole, provided the foot does not move excessively in the shoe. It is true, as Berg says, that if the shoe is soft enough to allow excessive foot motion, a thick sole or high heel lift will cause a greater torque on the ankle and could lead to injuries. Frederick feels, however, that the modern running shoe is sufficiently well designed to prevent excessive motion even on crowned roads and moderately uneven surfaces. He also thinks crowned roads will bother a runner no matter what his footgear.

Cavanagh points out that many runners consistently run against traffic, thereby stressing the legs unevenly if the road is cambered. Cavanagh runs equal distances both with and against traffic to balance the stress. Another problem is that runners tend to wear shoe soles or heels unevenly, causing the shoe itself to take on a camber. This can magnify problems caused by crowned roads or rapid turning. Hence shoes should not be worn too long after uneven wear is observed.

CHESTER R. KYLE


The Normalization Pairing

The Normalization Pairing

A scientist at the General Motors Research Laboratories has developed a new method for accurately determining the effectiveness of safety belts in preventing traffic fatalities. The approach may be used to answer a wide variety of questions using data bases that lack conventional measures of exposure.

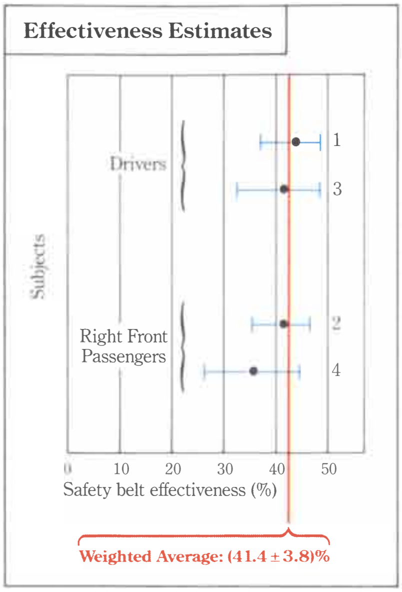
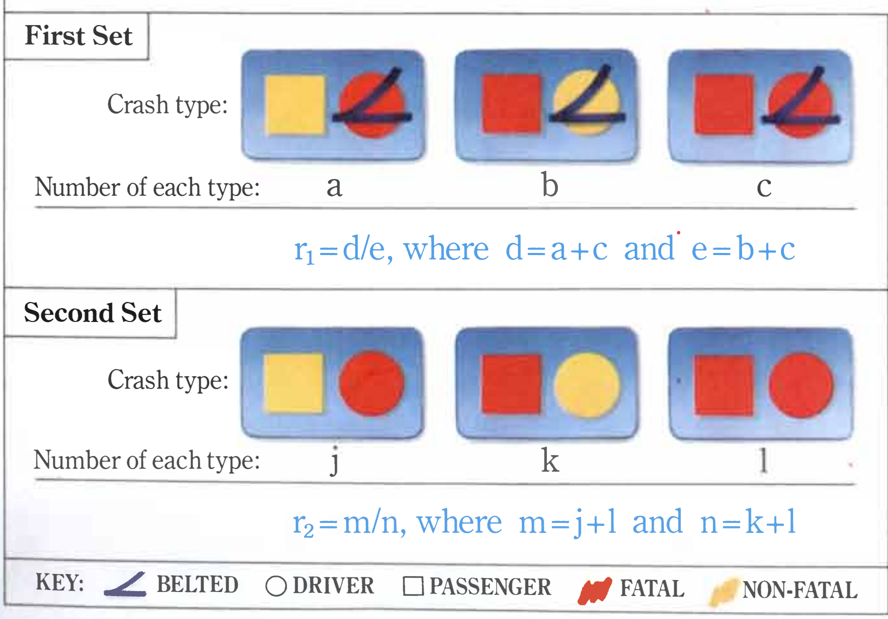


Figure 1: Weighted estimates of safety-belt effectiveness by subject, with standard error. Estimate 1 pairs subjects with right-front passengers; 2 pairs subjects with drivers; 3 and 4 pair subjects with occupants of all other seating positions.

Figure 2: Schematic representation of a sample double-pair comparison.



THERE IS A serious problem that researchers often encounter when trying to analyze large collections of information. It is the problem of measuring exposure. Though a collection of data may contain a large number of cases, and though the facts in each case may be highly detailed, there may be no way of comparing events selected for inclusion in the collection against the normal occurrence of similar events in the world at large.

One such data base is the Fatal Accident Reporting System (FARS) maintained by the U.S. Department of Transportation's National Highway Traffic Safety Administration. FARS details all fatal accidents in the U.S. since January 1, 1975—more than 300,000 crashes. However, it lacks an explicit measure of exposure.

FARS contains, for example, the number of fatalities classified by safety belt use. But fatalities among users depend on two considerations: first, the effectiveness of safety belts; and second, the crash involvement differences between users and nonusers—that is, the exposure of belt users to crash involvement. If crash involvement were independent of belt use, it would be a simple matter to calculate the effectiveness of safety belts in preventing fatalities. However, belted drivers have fewer crashes, and the crashes they do have tend to be of lower average severity than those of unbelted drivers.

Now a scientist at the General Motors Research Laboratories has developed an approach to drawing inferences from FARS using only the information contained in the file. Dr. Leonard Evans has designed a method for comparing the effects of isolated characteristics by using two sets of crashes. In each set, a *subject* occupant is paired with an *other* occupant. In the first set, the subject exhibits the characteristic to be studied; in the second, the subject does not. The *other* occupant is chosen to have similar characteristics in both sets of crashes (e.g. always unbelted), and thereby acts as a measure of exposure.

To illustrate the workings of the method of double-pair comparison, Dr. Evans first applied it to a study of the effects of safety belt use on fatality risk. He could define the effectiveness of safety belts in terms of the ratio:

$$R_{true} = \frac{N_b}{N_u} = \frac{N' \int q_{D,b}(s) f_u(s) ds}{N' \int q_{D,u}(s) f_u(s) ds}$$

where N' is the number of crashes per year by unbelted drivers, s is crash severity, $f_u(s)$ is the probability that a crash involving an unbelted driver has a severity s, $q_{D,u}(s)$ is the probability that an unbelted driver will become a fatality in a crash of severity s, and $q_{D,b}(s)$ is the probability that a belted driver will become a fatality in a crash of severity s. R_{true} is a ratio of new to old fatalities—assuming a formerly unbelted population became a belted population, with nothing else changing. But while N_u, the number of unbelted driver fatalities, can be determined from the FARS data, N_b, the number of these who would still have been fatalities had they been wearing safety belts, clearly is not coded in the data base.

Dr. Evans applied the double-pair comparison method to determine a quantity, R, that would, under plausible assumptions, accurately estimate R_{true}. Figure 2 shows the pattern of the first application. In it, one set of crashes paired belted drivers and accompanying unbelted front-seat passengers, generating a ratio, r_1, of belted driver fatalities per unbelted passenger fatality. The second set paired unbelted drivers with unbelted front-seat passengers, leading to a ratio, r_2, of unbelted driver fatalities per unbelted passenger fatality. This yields a value of $R = r_1/r_2$ as a measure of safety-belt effectiveness.

IN ADDITION to calculating R for driver *subjects* using front-seat passenger *others,* effectiveness was also calculated for right-front passenger *subjects* using driver *others.* Additional calculations were made pairing driver or right-front passenger *subjects* with passengers in any other seating position. Figure 1 reflects the synthesis of these estimates. Estimates 1 and 2 represent *subject* and *other* occupants disaggregated into three age categories and averaged. Estimates 3 and 4 represent pairings of *subjects* with occupants in other seating positions and averaged.

In all, Dr. Evans calculated 46 estimates of R. The weighted average of these gives a safety-belt effectiveness of $(41.4 \pm 3.8)\%$. This should be an accurate estimate whenever the distribution of severities is the same for both sets of crashes in each double-pair comparison.

Moreover, a formal analysis showed r_1/r_2 to be an accurate estimate of R_{true} under much less stringent restrictions. Even when the distributions of crash severity differ for belted and unbelted drivers, Dr. Evans concluded that the simple ratio $R = r_1/r_2 = nd/me$ is indeed an accurate estimate of safety-belt effectiveness.

Dr. Evans' confidence in the method rests on some key assumptions. But, as he points out: "One of the beauties of the method is its ability to remove the biasing effects of confounding interactions that may undermine those assumptions. It is necessary only to disaggregate occupants into different categories of the suspect variable.

"Because of bias elimination, and the ability to create a measure of exposure, the method of double-pair comparison lends itself to a broad range of investigations. We can estimate, for example, fatality risk as a function of helmet use by motorcyclists, or safety-belt effectiveness in different accident types. More broadly, we can estimate fatality risk as a function of age, sex, or alcohol use. We may even have revealed a trend in trauma response, in general, as a function of sex and age."

THE MAN BEHIND THE WORK

Dr. Leonard Evans is a Senior Staff Research Scientist in the Operating Systems Research Department at the General Motors Research Laboratories.

He received his undergraduate degree in physics from The Queen's University of Belfast, and holds a D. Phil. in the same discipline from Oxford University. He was a Post-Doctorate Fellow at the National Research Council of Canada in Ottawa.

Since joining GM in 1967, Dr. Evans has published research on such diverse topics as atomic physics and trauma analyses. His current area of concentration is traffic safety research.

He is a member of the Human Factors Society and is a Past President of the Society's Southeastern Michigan Chapter. In 1985, Dr. Evans received the Society's A. R. Lauer Award "for outstanding contributions to the human factors aspects of highway safety."

General Motors

MARK OF EXCELLENCE

A REPAIR SHOP LIKE NO PLACE ON EARTH.

In low orbits around the Earth, satellites gather, analyze and transmit critical information–for scientists, for corporations, for countries.

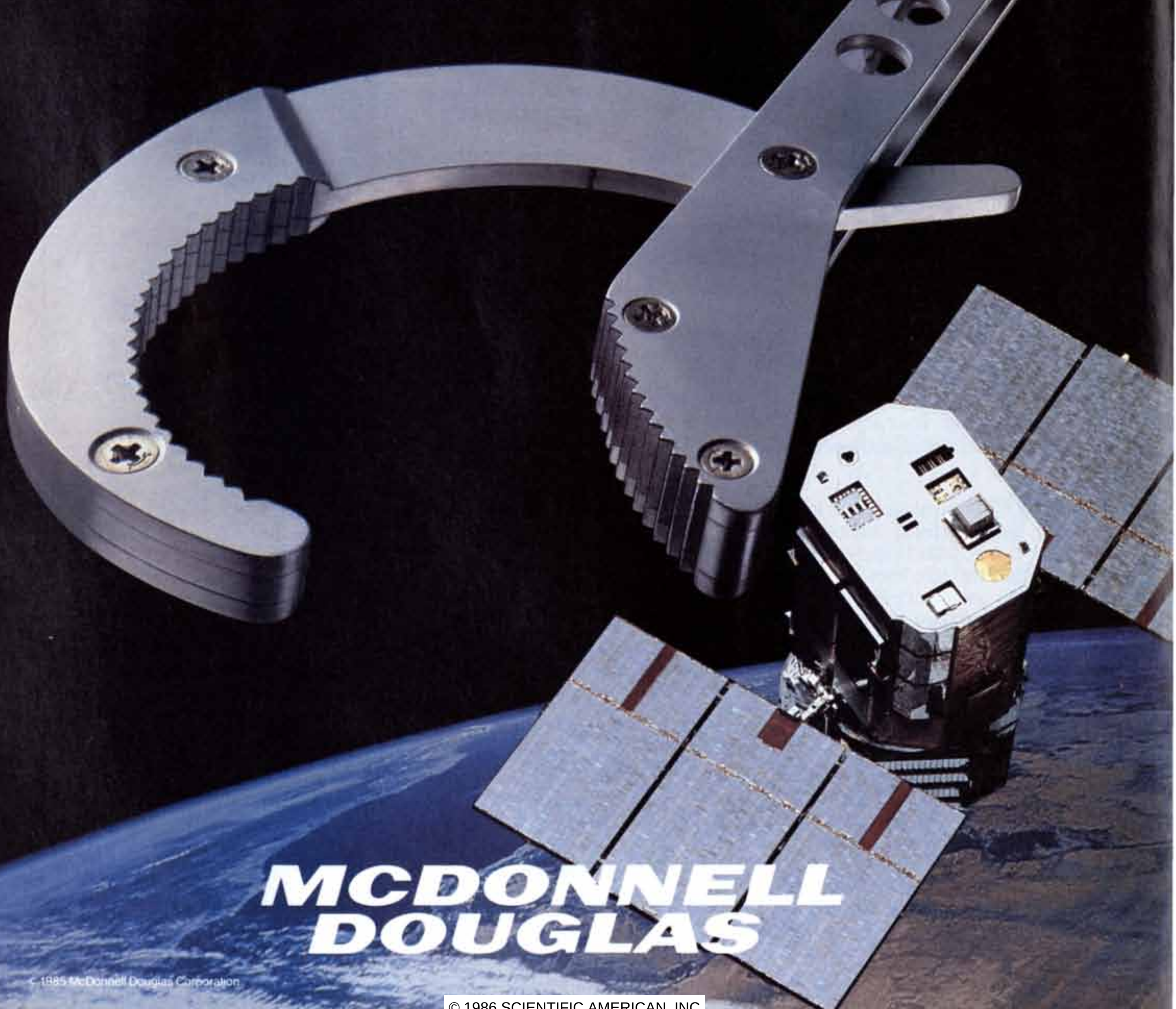
Because of the high cost of getting them up there in the first place, some of these satellites are designed to be repaired where they are if something goes wrong.

A manned space station could act as the neighborhood garage –the local spare parts and repair facility for these valuable satellites. That's one reason why McDonnell Douglas is working to put a space station in orbit by the 1990s.

Since 1960, when we did the first Manned Orbiting Research Laboratory studies, we've been a world leader in space systems. We built the original space station–Skylab. And we've been a pioneer in space systems operations, from launches to payload integration.

Now we're devoting our space experience, systems expertise, and technological ingenuity to taking America the next step into space–with a manned station.

Such a station could lower the cost of technological upkeep–for science, for industry. We think that's a very good reason to build a repair shop like no place on Earth.



50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JUNE, 1936: "For those people who revel in astronomical figures, a new ultracentrifuge at the Du Pont Experimental Station, at Wilmington, Delaware, should hold much fascination. The rotor is designed to operate for prolonged periods at 60,000 revolutions a minute, thus creating a centrifugal force of 250,000 times that of gravity. Under such conditions a mass of one pound would weigh 125 tons. The rotor has a peripheral velocity of more than 20 miles per minute, or approximately one and one-half times the muzzle velocity of an ordinary 22-caliber bullet. The machine has exciting potentialities in the field of colloid chemistry, which field includes the chemistry of giant molecules."

"Years of scientific research are behind the idea of the industrial use of farm crops as raw materials. Soy beans furnish the almost perfect example. In 1935 almost 21,000,000 pounds of American-produced soy-bean oil was used by industry. The American textile industry's annual imports of starch amount to hundreds of millions of pounds; only recently it was proved that a better starch is yielded by Southern sweet potatoes at a return of 40 dollars per acre to the farmer for the starch alone. Moreover, because of chemical conversion into cellulose, the uses of cotton have become literally hundreds."

"Nearly every insect that has been accidentally imported into the United States has become either a nuisance or a real menace. One need only witness the ravages of the European corn-borer, the dreaded Japanese beetle and a dozen others. It is therefore interesting to find that the oriental praying mantis is highly beneficial. Mantids possess enormous appetites for all other kinds of insects."

"Announcement has been made by the Forest Service of a new instrument called the Byram haze meter. Because of the important role which visibility conditions play in the detection of forest fires, the discovery of a simple and accurate means to measure the effect of haze on the distance fires can be seen marks a real advance. The meter is based on the discovery that a smoke column is just barely visible to lookouts with good eyesight against a background about 60 percent as bright as the sky at the horizon. The instrument provides a mechanical means for finding a point on the background or landscape that is 60 percent as bright as the horizon. The distance to that part of the landscape is scaled off on the map and becomes the visibility distance or the maximum distance at which the observer can expect to detect a smoke column of a certain size in that direction at that particular time."

"Reports coming from England indicate that the 'anti-horn' campaign inaugurated some time ago as a safety measure for motorists is highly successful; a similar experiment has been tried in New York City. 'Anti-horn' campaigns have two objectives. First, they are aimed at a much-to-be-desired reduction in the noise level of our more congested areas, and second, at the safety of the motorist who all too often is inclined to rely more on his horn and less on his brakes to keep himself out of trouble."

"The new German Zeppelin, *LZ-129,* has been christened the *Hindenburg* and has passed its flying tests with perfect success. If the wonderful record of its sister ship the *Graf Zeppelin* over the South Atlantic is borne in mind, there is little doubt that the same regularity and safety may be expected in the service of the new airship over the North Atlantic."

SCIENTIFIC AMERICAN

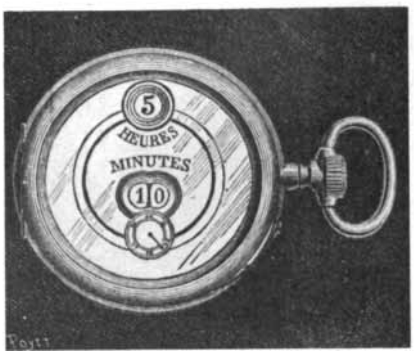
JUNE, 1886: "Prof. Alexander Graham Bell and his cousin, Dr. Chichester Bell, have recently discovered that a falling jet of water or a flame of gas burning in a room reproduces every word spoken and every sound uttered within a given distance. In this new apparatus the effect of sound waves on a jet of water is caught by instantaneous photography, and permanently recorded on a glass plate in the form of minute irregularities of surface. By suitable apparatus these elevations and depressions, which correspond to pulsations of air, are retranslated into air waves, and the voice is heard again. Prof. Bell regards this discovery to be quite as important as the telephone."

"For some months past, excavations have been carried on at Ghizeh, near Cairo, with the view of freeing the famous Egyptian Sphinx from the masses of sand which have gradually buried the monument. M. Gaston Maspero, the Director of the Boulak Museum, has superintended the operations, and in a recent letter he states: 'The result is beyond all my hopes. The face, raised approximately fifteen meters above the surface, is becoming expressive, in spite of the loss of the nose. The expression is serene and calm. The work now going on is in beds of sand, which have not been disturbed since the first centuries of our era.'"

"The largest balloon in the world, according to the San Francisco *Chronicle,* has recently been built in that city by Mr. A. P. Van Tassel. It has a capacity of 150,000 cubic feet of gas, and has been constructed for the specific purpose of enabling the well-known aeronaut to undertake a journey across the continent, from ocean to ocean. Mr. Van Tassel has had considerable experience in aerial traveling. A careful study of the aerial currents leads him to believe that by seeking the proper stratum of air he can be carried eastward at high speed, possibly 100 miles an hour."

"The Union Electric Company has recently been operating an experimental electric motor car in Philadelphia, and has met with very fair success. The estimated cost per day of running the electric car is $1.84, while that of operating a horse car is $4.74. Neither estimate includes salaries. The cost of ten miles of electric railway on this system, together with fifty cars, is stated to be $175,000."

"The Messrs. Schwob have recently brought a novelty into fashion, and that is a watch without hands. Its face, which has no divisions, is provided with two small apertures merely, one above the other. In the one above we read the hour, and in the one below, the minutes. The hour and minutes appear clearly in black upon a silver background."



Watch without hands


PEELING THE LASER BEAM DOWN TO ITS CORE.

AT&T Bell Laboratories scientists have generated a beam from a diode laser with a frequency spectrum 1,000,000 times narrower than that of today's most advanced commercial semiconductor lasers.

Part of a record-breaking coherent lightwave system, this laser 'peels away' a tremendous number of unwanted frequencies that can clutter up a beam—to create a lightwave so pure, its frequency variance is limited to 1 part in 40 billion. The equivalent of one drop of ink in a million-gallon, Olympic-size swimming pool.

A drop in 40,000,000,000

Purity Has Its Rewards

The new narrow-spectrum laser is the key element in a coherent lightwave communications system that increases the information-carrying capacity of an optical fiber, as well as the distance over which an unboosted laser beam can be received.

Increased capacity comes from dramatically increasing the number of individual laser beams that can be sent through a fiber's best transmission window.

Each laser in a coherent system produces an exceptionally stable, pure wavelength—allowing thousands of non-interfering wavelengths to travel side-by-side on a fiber. (Only a handful of beams can be combined using today's commercial semiconductor lasers.)

Dial 'M' For Movie

In the future, the capacity of coherent transmission could allow us to send 10 million conversations—or 10 thousand digital TV channels—simultaneously, on a single fiber.

Or, using the full capacity of a fiber, a coherent system could dump a movie like 'Gone with the Wind' into a home memory unit in one second flat. Or deliver Beethoven's '5th' in less than a 50th of a second.

Making A Little Go A Longer Way

Increasing capacity is important in an age of rapidly expanding information needs. But so is reducing costs—in this case, by nearly doubling the distance an unboosted signal can be received.

A newly developed AT&T coherent lightwave receiver contains its own narrow-spectrum laser. The beam from this laser reinforces the transmitted signal as it detects it—a technique only possible with two such pure beams.

Using this receiver, AT&T has achieved a laboratory transmission record of nearly 100 miles at a data rate of 1 billion pulses per second.

We Don't Keep The Future Waiting

Coherent lightwave transmission is just one of the ways AT&T is working toward the high-capacity, high-speed integrated networks of the future.

Meanwhile, we're bringing tomorrow closer with leading-edge lightwave systems we're building today.

AT&T this year introduced a commercial lightwave system—the FT Series G—designed to operate at up to 1.7 billion bits per second, a rate that permits the transmission of 24 thousand simultaneous calls on a single pair of fibers.

And by 1988, we'll have installed the first transatlantic and transpacific lightwave systems to Europe and the Far East—systems capable of transmitting 40 thousand simultaneous conversations on two pairs of fibers.

Clearly, whether on land or underseas, AT&T is lighting the way in lightwave. And peeling the laser beam is part of it.

AT&T

The right choice.

THE AUTHORS

ERIC D. LARSON, MARC H. ROSS and ROBERT H. WILLIAMS ("Beyond the Era of Materials") share an interest in issues related to energy use and the environment. Larson holds a postdoctoral research position at the Center for Energy and Environmental Studies at Princeton University. He got his Ph.D. in mechanical engineering in 1983 from the University of Minnesota and stayed on to take part in the Global Environmental Policy Project at the Hubert H. Humphrey Institute of Public Affairs. He moved to Princeton in 1984. Ross is professor of physics at the University of Michigan and senior scientist in the Energy and Environmental Systems Division of the Argonne National Laboratory. He earned his Ph.D. in physics from the University of Wisconsin at Madison and held several postdoctoral appointments before joining the faculty at Indiana University in 1955. He took up his position as full professor at Michigan in 1963. Williams is senior research physicist at the Center for Energy and Environmental Studies at Princeton. He studied at the University of California at Berkeley, getting his doctorate in 1967. In 1972 he became chief scientist of the Ford Foundation's Energy Policy Project, and he moved to Princeton in 1975.

RICHARD O. HYNES ("Fibronectins") is professor of biology at the Massachusetts Institute of Technology, where he is associated with both the department of biology and M.I.T.'s Center for Cancer Research. He was graduated in 1966 from the University of Cambridge with a bachelor's degree in biochemistry and received a Ph.D. in biology from M.I.T. in 1971. He did postdoctoral research at the Imperial Cancer Research Fund in London until 1975, when he returned to M.I.T. to join its faculty.

HOWARD E. HABER and GORDON L. KANE ("Is Nature Supersymmetric?") have collaborated in the study of very-high-energy phenomena in particle physics for the past several years. Haber is adjunct assistant professor at the University of California at Santa Cruz. He got his Ph.D. from the University of Michigan in 1978, and he did postdoctoral work at the Lawrence Berkeley Laboratory of the University of California at Berkeley and at the University of Pennsylvania before joining the faculty at Santa Cruz. Kane, who is professor of physics at Michigan, received his doctorate from the University of Illinois at Urbana-Champaign. He did postdoctoral research for two years at Johns Hopkins University before joining the faculty at Michigan in 1965. He is currently on leave at CERN, the European laboratory for particle physics.

COLIN S. RAMAGE ("El Niño") is professor of meteorology at the University of Hawaii at Manoa. He has been engaged in meteorological work since 1936, when he joined the New Zealand Meteorological Service. He received a degree in physics at Victoria University College in Wellington in 1940 and later became a meteorological officer in the Royal New Zealand Air Force. After his discharge in 1946 he went to the Royal Observatory in Hong Kong as a scientific officer. Since 1956 he has been at the University of Hawaii. Ramage is currently taking a two-year leave of absence to serve as science adviser to the director of the National Oceanic and Atmospheric Administration's Environmental Research Laboratories.

THOMAS F. ROBINSON, STEPHEN M. FACTOR and EDMUND H. SONNENBLICK ("The Heart as a Suction Pump") are respectively associate professor of medicine and of physiology and biophysics, professor of pathology and associate professor of cardiology, and Olson professor of medicine at the Albert Einstein College of Medicine of Yeshiva University. Robinson has a B.S. in chemistry and an M.S. in physics from the State University of New York at Albany. His Ph.D. in biophysics is from the Rensselaer Polytechnic Institute. In 1978 he moved to Einstein. Factor got his bachelor's degree from Queens College of the City University of New York and his M.D. from Einstein. From 1971 to 1973 he was an officer in the U.S. Army, stationed at Martin Army Hospital at Fort Benning in Georgia. Factor has been a member of the faculty at Einstein since 1975. Sonnenblick holds a B.A. (1954) from Wesleyan University and an M.D. (1958) from the Harvard Medical School. After completing his residency in 1963 he took a position at the National Heart Institute, remaining there until 1967. He then joined the faculty at the Harvard Medical School and took on clinical and research duties at Peter Bent Brigham Hospital in Boston. In 1975 Sonnenblick moved to Einstein, where he has been director of the cardiovascular center since 1984.

GERALD BORGIA ("Sexual Selection in Bowerbirds") is assistant professor of zoology at the University of Maryland at College Park. He got a bachelor's degree from the University of California at Berkeley in 1970 and did graduate work at the University of Michigan, where he received a master's degree and a doctorate respectively in 1973 and 1978. After postdoctoral work at the University of Chicago and the University of Melbourne he became assistant professor at the University of Maryland in 1980.

VILAYANUR S. RAMACHANDRAN and STUART M. ANSTIS ("The Perception of Apparent Motion") share an interest in the neuropsychology of vision. Ramachandran is professor of psychology at the University of California at San Diego and holds a joint appointment as a visiting associate in biology at the California Institute of Technology. He was trained as a physician at the University of Madras and went on to earn a Ph.D. in neurophysiology from the University of Cambridge. After a series of postdoctoral appointments he joined the faculty at San Diego. Ramachandran welcomes inquiries from those who would like to obtain a tape or a computer disk reproducing the illusions described in the article. Anstis is professor of psychology at York University in Toronto. He got an M.A. and a Ph.D. from the University of Cambridge, where he remained for two more years as a postdoctoral fellow. He taught at the University of Bristol before moving to York.

PETER M. MORETTI and LOUIS V. DIVONE ("Modern Windmills") have worked together in the U.S. wind-energy research program. Moretti is professor of mechanical and aerospace engineering at Oklahoma State University. He is a graduate of the California Institute of Technology and Stanford University, where he received his Ph.D. in 1965. After working in private industry for several years, he joined the faculty at Oklahoma State in 1970. Moretti took a year's leave from the university in 1977 to serve in the U.S. Department of Energy, where he worked with Divone. Divone is director of the office of solar electric technologies in the U.S. Department of Energy. He holds degrees in aeronautical engineering from the Polytechnic Institute of Brooklyn and the Massachusetts Institute of Technology. Divone headed the Federal wind-energy program from its inception under the National Science Foundation to its current incarnation in the Department of Energy.

MATHEMATICAL GAMES

Casting a net on a checkerboard and other puzzles of the forest

by Martin Gardner

No tree in all the grove but has
its charms,
Though each its hue peculiar.
—WILLIAM COWPER, *The Task, Book 1: The Sofa*

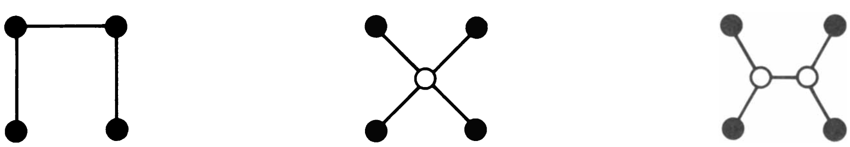
In graph theory, which is the study of the structures formed by joining points with lines, a tree is a connected network of line segments that includes no circuits. A circuit is a closed path that allows one to travel along a connected network from a given point back to itself without retracing any lines. It follows that any two points on a tree are joined by a unique path. Trees are extremely important in graph theory, and they have endless applications in other branches of mathematics, particularly probability theory, operations research and artificial intelligence.

Suppose a finite set of n points are randomly scattered about in the plane. How can they be joined by a network of straight lines that has the shortest possible total length? The solution to the problem has practical applications in the construction of such networks as roads, power lines, pipelines and electrical circuits. If no new points are allowed to be added to the original set, the shortest network connecting them is called a minimal spanning tree. It is easy to see the network must be a tree: if it included a circuit, one could shorten it at once by removing a line from the circuit.

There are many ways to construct a minimal spanning tree. The simplest is known as a greedy algorithm, because at each step it bites off the most desirable piece. It was published in 1956 by Joseph B. Kruskal, now at the AT&T Bell Laboratories. First find two points that are closer together than any other two and join them. If more than one pair of points are equally close, choose any such pair. Repeat the procedure with the remaining points in such a way that joining a pair never completes a circuit. The final result is a spanning tree of minimal length.

A minimal spanning tree is not necessarily the shortest network spanning the original set of points. In most cases a shorter network can be found if one is allowed to add more points. For example, suppose you want to join the three points defining the corners of an equilateral triangle. Two sides of the triangle make up a minimal spanning tree. The spanning tree can be shortened by more than 13 percent by adding an extra point at the center and then making connections only between the center point and each corner [*see upper illustration on this page*]. Each angle at the center is 120 degrees.

How a minimal spanning tree **(left)** *is shortened if an extra point is allowed* **(right)**



Minimal spanning tree of a square, shortened by adding one or two extra points

A less obvious example is the minimal network spanning the four corners of a square. You might suppose one extra point in the center would give the minimal network, but it does not. The shortest network requires two extra points [*see lower illustration on this page*]. Again all the angles around the extra points in the network are 120 degrees. The network with one extra point in the center has length $2\sqrt{2}$, or about 2.828. The network with two extra points reduces the total length to $1 + \sqrt{3}$, or about 2.732.

One of the first mathematicians to investigate such networks was Jakob Steiner, an eminent Swiss geometer who died in 1863. The extraneous points that minimize the length of the network locally are now called Steiner points. (I shall describe what is meant by "locally" below.) It has been proved that all Steiner points are junctions of three lines forming three 120-degree angles. A tree with Steiner points is called a Steiner tree. Although adding Steiner points can reduce the length of the spanning tree, a Steiner tree is not always the shortest network spanning the original set of points. When it is, it is called a minimal Steiner tree.

Minimal Steiner trees are almost always shorter than minimal spanning trees, but the reduction in length may depend on the length of the original spanning tree. It has been conjectured that for any given set of points in the plane, the length of the minimal Steiner tree cannot be less than $\sqrt{3}/2$, or about .866, times the length of the minimal spanning tree; the result has been proved, however, only for three, four and five points. Just as a set of points can have more than one minimal spanning tree, so it can have more than one minimal Steiner tree, although of course all minimal Steiner trees for a given set of points have the same length. A Steiner tree can have at most $n - 2$ Steiner points, where n is the number of points in the original set.

Many simple Steiner trees can be found empirically by a simple analog device you can build. Two parallel sheets of Plexiglas are joined by perpendicular rods that correspond to the

---

A. K. Dewdney is on vacation. His "Computer Recreations" column will appear again next month.

All we've ever asked of Businessland is that they be in 197 places at once.

Norman Coulson, President and CEO of Glendale Federal Savings and Loan.

It wasn't all that long ago that Mr. Coulson asked us to install a few computers in Glendale Federal's main office.

And then a few more in their Torrance office.

And then a few more in their Ft. Lauderdale office.

Before we knew it, the nation's fifth largest savings and loan was enlisting Businessland's services throughout their branch network.

We must be doing something right.

And they must be doing something right: this fiscal year was the most profitable in Glendale Federal's history.

Partly due to their ability to aggressively manage assets and liabilities.

Which is partly due to our ability to provide them with sophisticated networking solutions.

Hardware and software evaluation.

On-site service, installation support, and more.

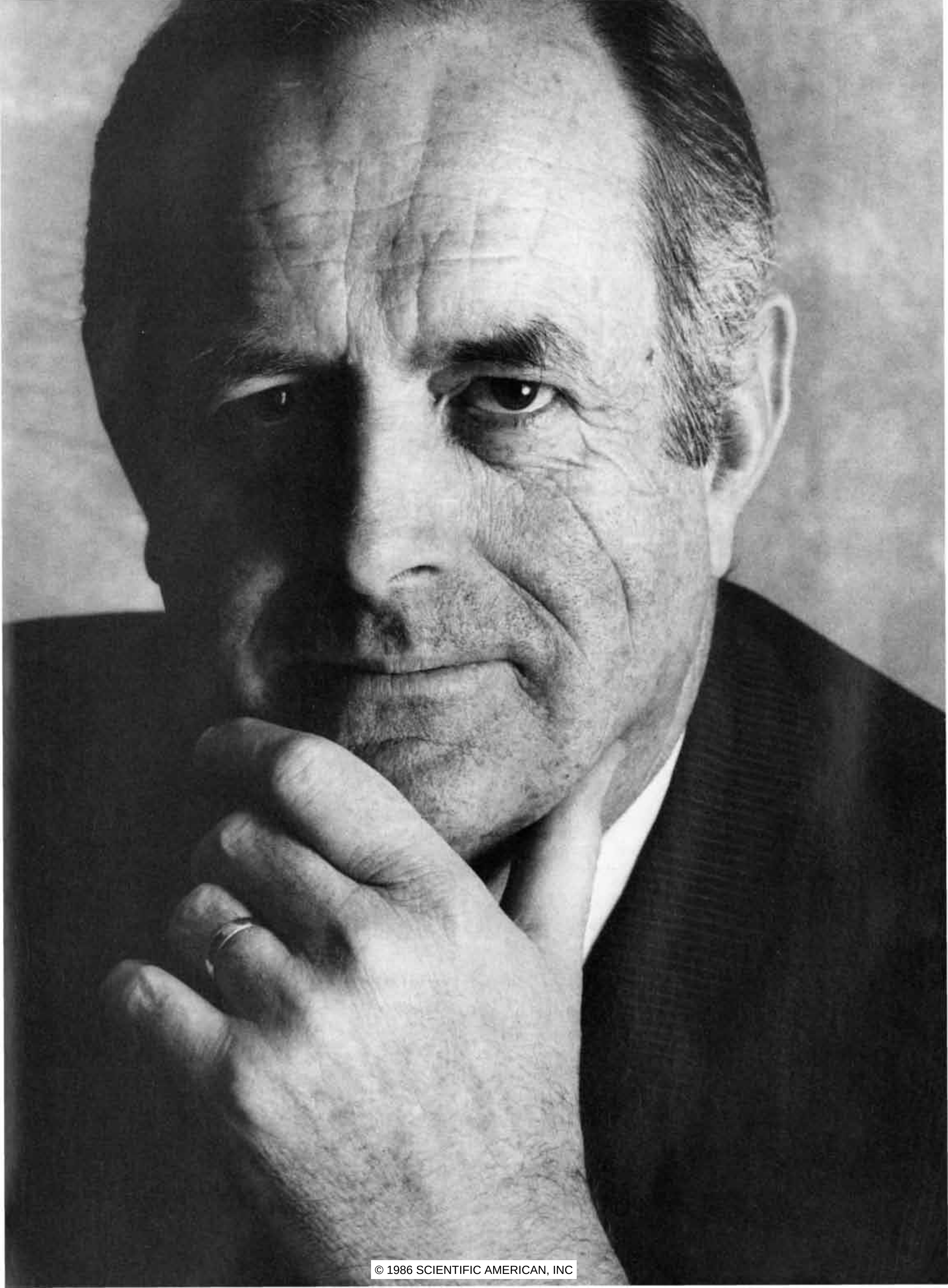
And because we're company-owned and nationwide, we've been able to consistently coordinate the needs of Glendale Federal's offices all across the country.

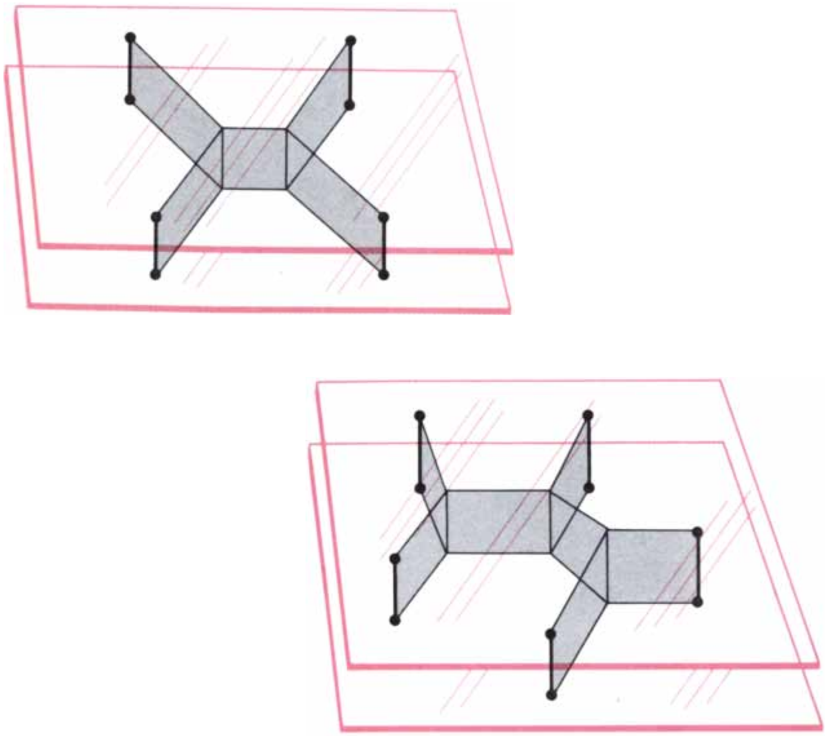
We can do the same for your company. Call (800) 228-7463 for the Businessland nearest you. And if you only have a couple offices, that's okay.

We can be in two places at once.

BUSINESSLAND®

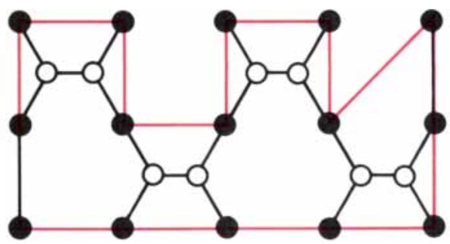
Where business people are going to buy computers.



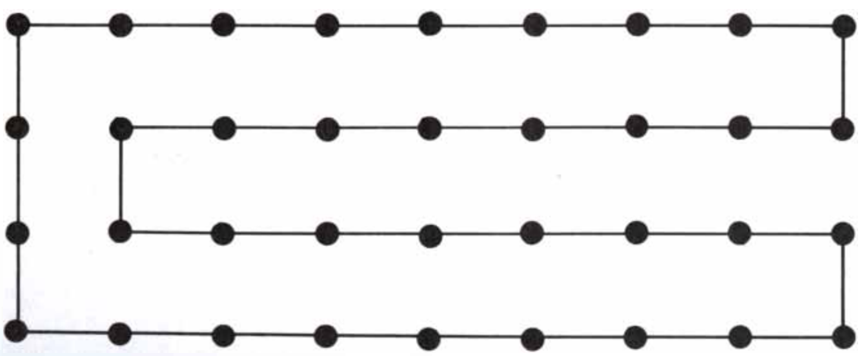


Steiner problems for the vertexes of a square and a regular pentagon, solved by soap film

Two stable soap-film trees for four points on a rectangle. Only the left tree is minimal



Path of a traveling salesman (color) *and a Steiner tree* (black) *on a rectangular lattice*



Another traveling salesman's path. Is there a Steiner tree shorter than 32.095... units?

points to be spanned in a given network. Drill holes in the sheets, insert the rods and immerse the entire assembly in a soap solution of the kind used for making bubbles. When the assembly is lifted out of the solution, a soap film forms surfaces that span the rods. Because such surfaces shrink to minimal area, the pattern formed by the film when it is viewed from above is a Steiner tree.

Such a device can find the minimal Steiner tree for the corners of a square [*see top illustration at left*]. The tree can take either of two forms, one of them a 90-degree rotation of the other. By blowing on the film you can make it jump from one pattern to its rotated form. Similarly, the device can model the minimal Steiner tree for the five points at the corners of a regular pentagon. For the six corners of a regular hexagon (and all higher regular polygons) extra Steiner points are of no help. The minimal spanning network is simply the perimeter of the polygon with one edge removed.

Even in these simple cases, however, one must be wary of the soap-film computer. For example, if the four points in the given network mark the corners of a rectangle a trifle wider than it is high, the film can stabilize in one of two patterns [*see second illustration at left*]. Both are Steiner trees, but only the one at the left is minimal. As the rectangle widens, the vertical line *AB* in the nonminimal pattern on the right becomes shorter. The line shrinks to a point when the vertical side of the rectangle is 1 and the base is $\sqrt{3}$, and for all wider rectangles only the minimal Steiner tree is stable. The tree at the right is said to be locally minimal. In other words, if you think of the lines as being elastic bands anchored at their ends to the four corner pegs, any slight shifting of the extra points will increase the length of the tree.

Given the simplicity of Kruskal's greedy algorithm for the construction of minimal spanning trees, one might suppose there would be correspondingly simple algorithms for finding minimal Steiner trees. Such, alas, is not the case. The task is part of a special class of "hard" problems known in computer science as NP-complete. When the number of points in a network is small, there are known algorithms for finding Steiner trees in a reasonably short time. As the number of points grows, however, the computing time needed increases at a rapidly accelerating pace. Even for a relatively small number of points it can be thousands or even millions of years. Most mathematicians believe no efficient al-

"Not now, I've got a headache."

—Computer backup excuse #14

It's not a very good excuse.

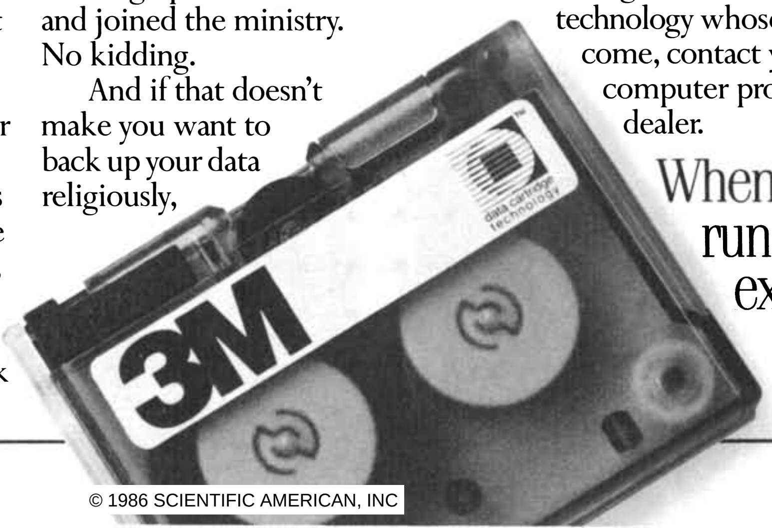
But when it comes to doing computer backup, any excuse will do. Because backing up, let's be brutally honest, is somewhat short of the thrill-a-minute category.

Not backing up, on the other hand, can really be exciting. Involving some interesting things like travel to other cities (while you're looking for another job), and the fun of even entering a whole new line of work.

After one big failure on disk the night before a big meeting, the programmer (who didn't back up) was so crushed that he hung up his business suit and joined the ministry. No kidding.

And if that doesn't make you want to back up your data religiously, we don't know what will.

To learn more about backup and other applications of the data cartridge, a 3M developed technology whose time has come, contact your local computer products dealer.



When you run out of excuses.™

gorithm exists for constructing minimal Steiner trees that connect arbitrary points in the plane.

Imagine, however, that the points are arranged in a regular lattice of unit squares, like the points at the corners of the cells of a checkerboard. Is there a "good" algorithm for finding a minimal Steiner tree spanning the points of such regular patterns?

The question occurred to me several years ago when I thought of the following problem. What is the length of the minimal Steiner tree that joins the 81 points at the corners of a standard checkerboard? Henry Ernest Dudeney, England's greatest puzzle maker, and his American counterpart Sam Loyd were both fond of puzzles based on checkerboard patterns. I checked all their books carefully, but they had not considered the problem. Indeed, I could find no evidence it had ever been posed before, let alone solved.

When I tried to solve the problem, I was surprised by its complexity. Although I could not prove it, it seemed obvious that the minimal Steiner tree would be constructed by joining many replicas of the regular four-point tree. The four-point tree has no name; let us call it *X* because in working on Steiner-tree problems for rectangular lattices, an *X* is easier to draw than the full tree. The difficulty in solving such problems is that it is hard to know where to place the *X*'s. It is easy to place them so as to make a Steiner tree, but it is not so easy to make the tree minimal.

I finally convinced myself that the checkerboard puzzle has a unique answer, although I could not prove it [*see bottom illustration on these two pages*]. I call it the conjectured solution for the order-9 array, where the order is the number of points on the side of the square. Because the length of the line segments that make up each *X* is $1 + \sqrt{3}$, it is easy to determine the total length of the tree: $26\sqrt{3} + 28$, or about 73.033. Although it seemed I had found a new puzzle, I suspected that in the growing mathematical literature on Steiner trees there must surely be a paper describing a simple algorithm for finding minimal Steiner trees on rectangular lattices. I was encouraged by knowing that many problems involving paths through points in the plane, which are hard when the points are arbitrary, become trivial when the points form regular lattices.

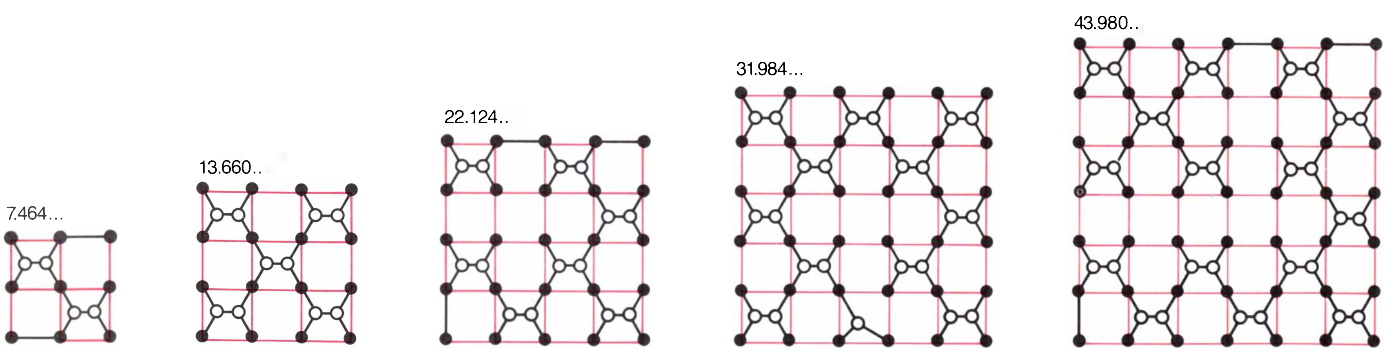
The traveling salesman problem is a notorious example. What is the shortest path allowing a salesman to visit each of *n* towns once and only once and return to the starting town? When the points are arbitrary, the task is NP-complete, and no efficient algorithm for solving it is known. But when the points are placed at the corners of squares and packed into a rectangular lattice, the problem is absurdly easy. If a rectangular array of *m*-by-*n* points includes an even number of points, the minimal path has length $m \times n$. If the array includes an odd number of points, the path has length $m \times n + \sqrt{2} - 1$ [*see third and bottom illustrations on page 20*]. I fully expected that the task of spanning points in such arrays by minimal Steiner trees would be equally trivial. I could not have been more wrong.

My first step was to send the checkerboard problem to my friend Ronald L. Graham, a distinguished mathematician at Bell Laboratories. I also asked him to direct me to a paper that might answer such questions. To my amazement, it turned out that the only relevant paper was one Graham himself had coauthored in 1978 with Fan R. K. Chung, also of Bell Laboratories. Titled "Steiner Trees for Ladders," it showed how to construct minimal Steiner trees for 2-by-*n* rectangular arrays of points, as well as for other kinds of 2-by-*n* "ladders." Aside from these special cases, nothing seemed to be known about how to find minimal Steiner trees for rectangular arrays when the number of points on each side is greater than 2.

The more Graham and Chung considered the matter, the more it intrigued them. On and off for better than a year they have been seeking an algorithm for the general case, but without success. Chung has recently been lecturing on the topic, and she and Graham plan eventually to write a paper on their progress.

Their best results are shown along with my checkerboard solution at the bottom of these two pages. Some of the trees have more than one minimal solution. Incredibly, only the pattern for the order-2 square lattice has been proved to be minimal. (There is a proof in Problem 73 of the book *100 Problems in Elementary Mathematics,* by Hugo Steinhaus.) Even the seemingly trivial order-3 pattern has eluded proof, although it would yield to brute-force methods carried out by computer. Graham and Chung firmly believe all their trees are minimal, but in the absence of proofs there may still be room for improvements.

It would be interesting to know whether soap film will solve the square lattices of order 3 and order 4. If it does, how far up the scale will soap film continue to find minimal trees?



A forest of conjectured minimal Steiner trees and their lengths for low-order square lattices.

What happens when Plexiglas sheets, joined by 81 rods in the checkerboard pattern, are dipped into the soap solution and then lifted out? Will the film generate Steiner trees spanning all 81 rods? If it does, what is the probability the tree will be minimal? Perhaps some venturesome readers will carry out these experiments.

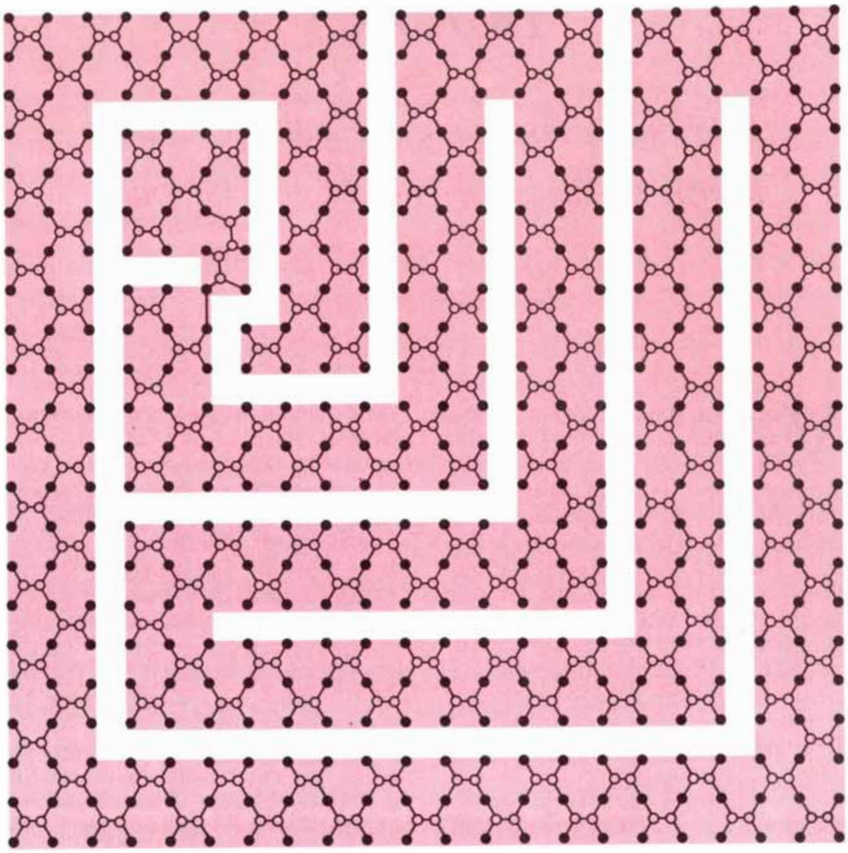
The order-6 square lattice is the smallest one from which an unexpected solution springs. When I worked on this forest of trees (sets of disconnected trees are known as forests to graph theorists), my order-6 pattern had length $11\sqrt{3} + 13$, or about 32.053. I almost fell out of my chair when I saw the shorter tree found by Graham and Chung. The little three-point tree in their pattern has length $(1 + \sqrt{3})/\sqrt{2}$, and so the total length of their network is $[(1 + \sqrt{3})/\sqrt{2}] + [11 \times (1 + \sqrt{3})]$, or about 31.984. It beautifully illustrates the kind of surprises—the "hue peculiar" of Cowper's epigraph—that lie in wait for anyone who tries to climb the ladder of square arrays in search of minimal solutions.

If you look closely, you will note that only squares of orders that are powers of 2 (2, 4, 8 and so on) have trees made entirely of *X*'s. Graham and Chung have proved an even more general result: a rectangular array can be spanned by a Steiner tree made up entirely of *X*'s if and only if the array is a square and the order of the square is a power of 2. Their clever proof, based on mathematical induction, is still unpublished. The unique spanning pattern generalizes in an obvious way to all squares whose order is a higher power of 2.

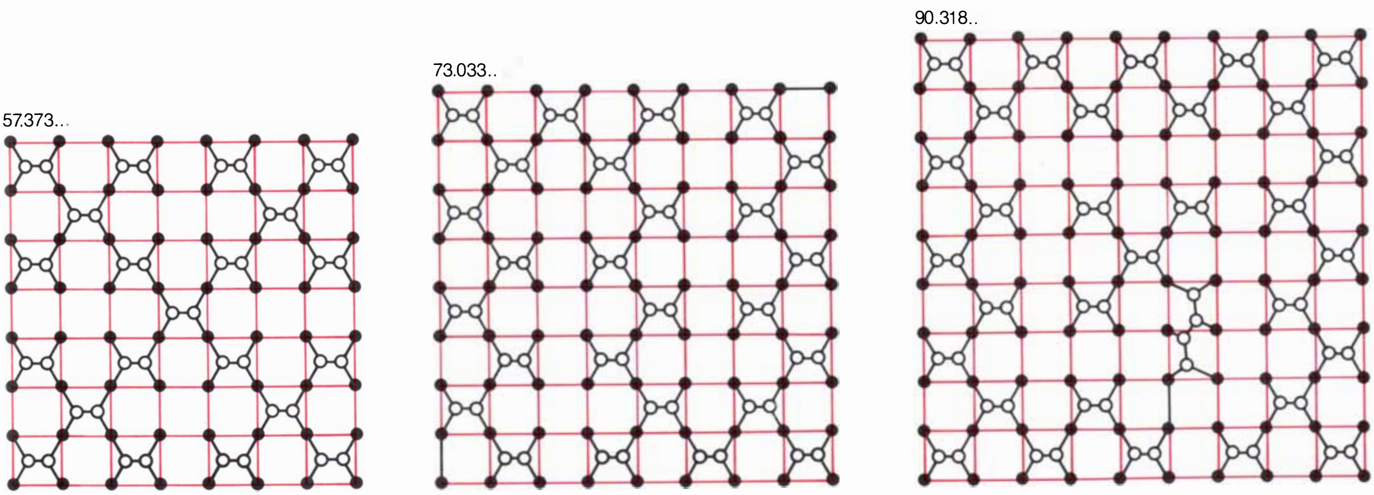
Space does not allow me to provide examples of the best-known patterns for nonsquare rectangular arrays, for which Graham and Chung have many curious results and conjectures. I close by giving the best Steiner tree they have found for the order-22 square [*see illustration above*]. It includes a pattern bounded by six points on two squares, which does not match the familiar *X*. The six-point pattern also appears in the Steiner tree of order 10, and its length is

$$\sqrt{11 + 6\sqrt{3}},$$

or about 4.625. The total length of the tree is approximately 440.021. If any reader can shorten that length, please let me know.



A Steiner tree of length 440.021... for the order-22 square lattice. Is it minimal?



The author's solution to the checkerboard problem is shown on the order-9 square lattice

BOOKS

Pele's hair, fastships, the galactic meek, a visit to Bregma, Nasion and Glabella

by Philip Morrison

VOLCANIC ASH, by Grant Heiken and Kenneth Wohletz. University of California Press ($40). MOUNT ETNA: THE ANATOMY OF A VOLCANO, by D. K. Chester, A. M. Duncan, J. E. Guest and C. R. J. Kilburn. Stanford University Press ($55). That observant surveyor one January day at Walden did not overlook the sample of snowflakes the north wind lodged on his coat: "And they all sing, melting as they sing, of the mysteries of the number six: six, six, six." Atomic order is the jubilant burden of that ancient but transient chorus. The first large book of this pair instead records the grand antiphonal response made by the glassy state to the crystalline. The geometrical mode of the response is not of the snowflake but of the dewdrop, a harmony free of integers; Platonic, not Pythagorean. The volume is by two much traveled geologists from Los Alamos; it is a technical atlas presenting tiny samples of the more dangerous and more enduring snowfalls of the volcano, mainly glassy grains as seen under the scanning electron beam on a scale nearly a hundredfold smaller than Thoreau's myriads of six-spoked wheels of crystalline ice.

The first of the hundreds of carefully categorized ashfall samples displayed is a micrograph of a single sphere of basaltic glass; the next one shows more such spheres accompanied by a perfect cylinder, a bit of a "single strand of Pele's hair," a filament of natural fiberglass.

The quantal properties of the atoms assumed, classical physics gives us a reasonably clear picture of how volcanic glasses form. A typical ashfall begins underground in a yellow-hot upward stream of molten basalt. Low silica content and high temperature guarantee little viscosity and easy flow. The recipe then demands the presence of sufficient dissolved gas; it is enough to have about one part in 1,000 of water or carbon dioxide by weight. As the molten magma leaves its regime of high pressure to approach the low-pressure world aboveground, the dissolved gas nucleates into tiny bubbles. The pleasantest of analogies is precise: the champagne is uncorked. When the bubbles' gas pressure is high enough, they can expand and locally disrupt the liquid surface. Obviously low viscosity favors such a process. If the entire surface is to be disrupted into a continuous fire fountain (the sight so celebrated in the Hawaiian home of Pele, the goddess of volcanoes) or even into the less pyrotechnic but more dangerous staccato of lava shrapnel and larger-caliber projectiles, there has to be enough gas so that most of the volume near the surface is taken up by bubbles.

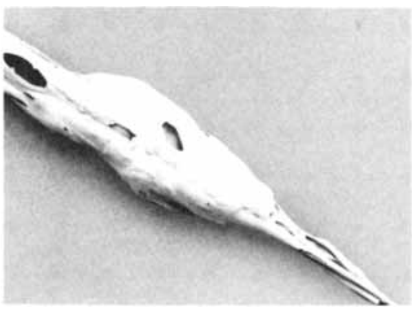
The conditions of fluid flow dominate the process. If the bubbles and the ambient melt rise at about the same speed, the bubbles have little chance to coalesce; they remain rather small. In that low-viscosity, high-temperature extreme we can count on Hawaiian fireworks. If the melt flows slowly enough so that the buoyant bubbles can rise swiftly through it, bigger and faster-rising bubbles will overtake and incorporate many little ones, to grow and rise still faster. Then some can surface as huge bubbles, large as a Volkswagen; the outflung portions of bubble wall material cool in flight. They fall no longer merely as fireworks but as Etna's artillery.

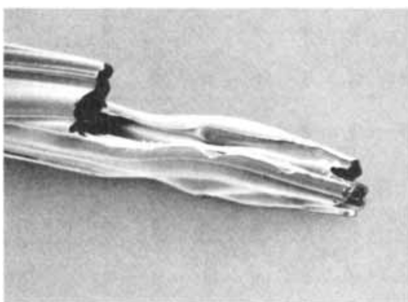
Pele's hair, which is abundant in Hawaii, arises in fluid streams thrown out into long, thin fibers like cotton candy. While the jets from the bubbling surface remain molten, any thin streamer is unstable. It quickly fragments under surface tension into small droplets. If radiation and moving air first cool the material off, the viscosity can increase so much that the stream solidifies before the surface capillary waves can break it up. Thus Pele casts her long, delicate hair to the winds; it is found as tangled bundles, always shattered into shorter lengths by the movement of air and by striking the ground.

Synoptic in scope, the atlas musters and comments on typical samples of famous eruptions of Tambora, Hekla, Krakatau and Mount St. Helens. It treats sequences of ashfalls from prehistoric cases, and it samples tuffs found at long distances from ashfalls ancient and modern. The nomenclature is a rich jargon, but useful glossaries and tables ease the way for nonspecialist browsers. The little orange glass spheres from the Taurus-Littrow valley appear as evidence of an old low-viscosity eruption on the moon, and a sample from the sodium carbonate (trona) outpourings of one unusual volcano in the Rift Valley shows full crystallinity. These grains are not glass at all. Possibly modified by rainwater, they are not dendritic but rather dully tabular crystals, a trona sleet. Something rather like that once faced a few of our forebears crossing a new ashfall at Laetolil a few million years ago.

Pluto and Neptune often meet, at times calmly, at other times explosively. These cases produce distinct grains, often diagnostic of the event. The phenomena and the particulates produced by explosions of thermite and water mixes (a 24-meter high plume is shown) are a specialty of Los Alamos; related experiments have been carried out in reactor-safety studies at Sandia. The coarse steam-burst particle aggregates resemble those in some watery volcanoes.

The Mount Etna book is a summary of work done since World War II on the most accessible and longest-studied volcano in Europe. These authors are British, but they report as well on work done jointly with French and





Pele's hair, a thin strand formed in a lava jet, and enlargement of its broken tip **(right)**

You get perfectly focused pictures. Because the N2020 focuses automatically, whether your subject is moving or still.

What happens when you don't set the exposure? You get perfectly exposed pictures. Because the N2020 offers three programmed modes for auto-exposure, too.

What happens when you don't advance the film? You get perfect pictures anyway. Because the N2020's built-in motor drive advances it for you.

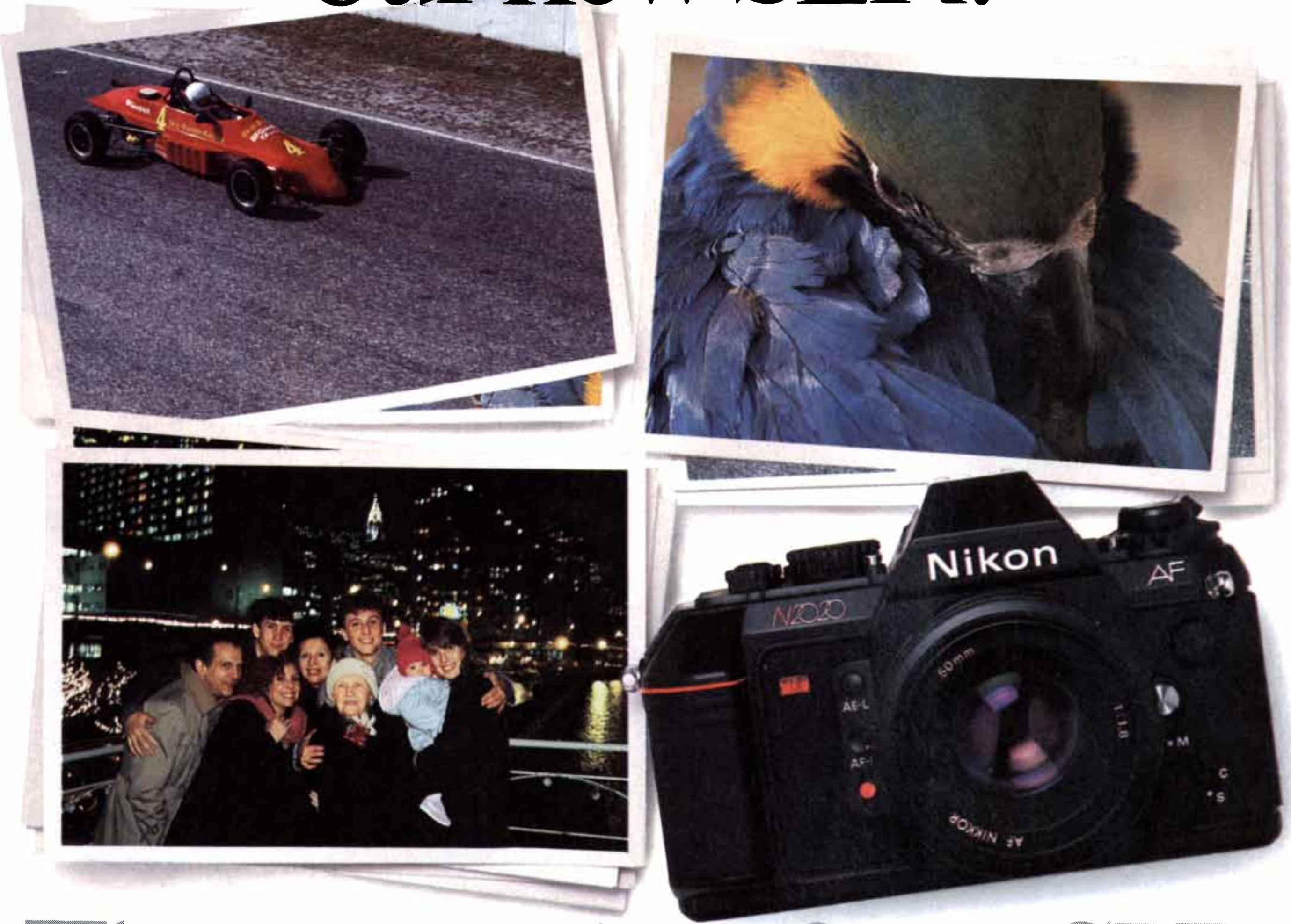
What happens when you don't have much light? You get perfect pictures even in total darkness. Because the new SB-20 Autofocus Speedlight helps the N2020 "see" in the dark.

You also get a choice of unparalleled Nikon lenses including our popular new zooms. And lots of manual options, too.

So what happens when you buy an N2020? Your pictures will be perfect whether you are or not.

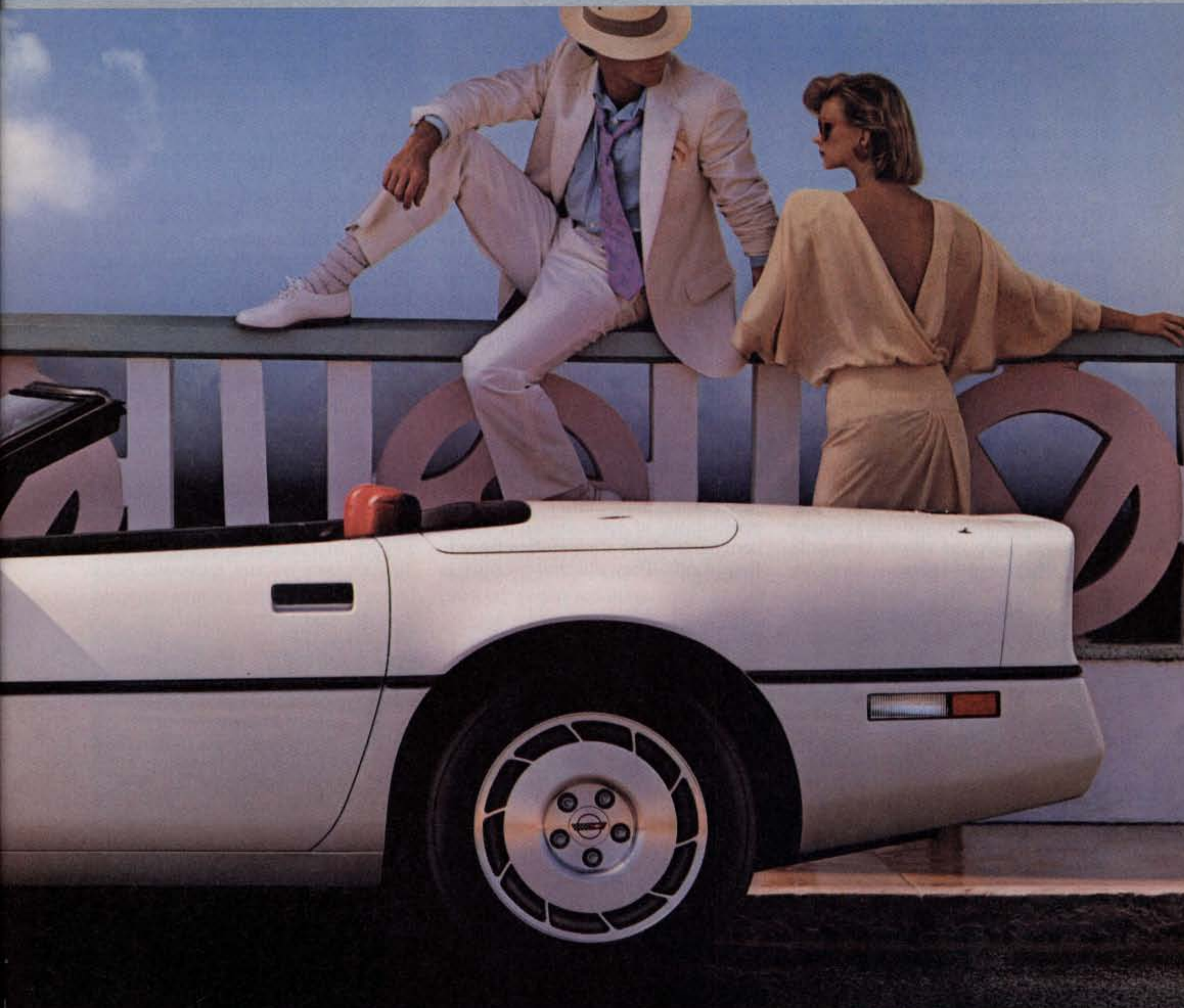
Nikon We take the world's greatest pictures.

What happens when you don't focus our new SLR?



The N2020 Autofocus SLR.

Nikon USA CLUB One year membership in the Nikon USA Club is free with every purchase when you submit the Nikon USA Club application. For further information write: Dept. N9, Nikon Inc., 19601 Hamilton Ave., Torrance, CA 90502-1309. © Nikon Inc., 1986.



particularly Italian volcanologists. British interest in Etna is venerable; next to the caldera there long stood a large cabin called the Casa Inglesi. It was built in 1811 by the British, then occupying Sicily, for visitors to the crater region. (The building was destroyed by an eruption in 1971.) Charles Lyell was a powerful early contributor to our knowledge of Etna, as indeed were such classical authors as Empedocles, Strabo and Lucretius, who explored the domains of myth and science.

The mountain has been admired and feared for a very long time; the poet Pindar called it the "pillar of Heaven." An impressive graph, based on the clear historic record, plots the volume output of lava from Etna since the year 1500. The flow reached a maximum in the 17th century, a period that encompassed the worst recorded toll of Etna, the near-destruction of the town of Catania in 1669 and the concurrent collapse of the Summit Cone.

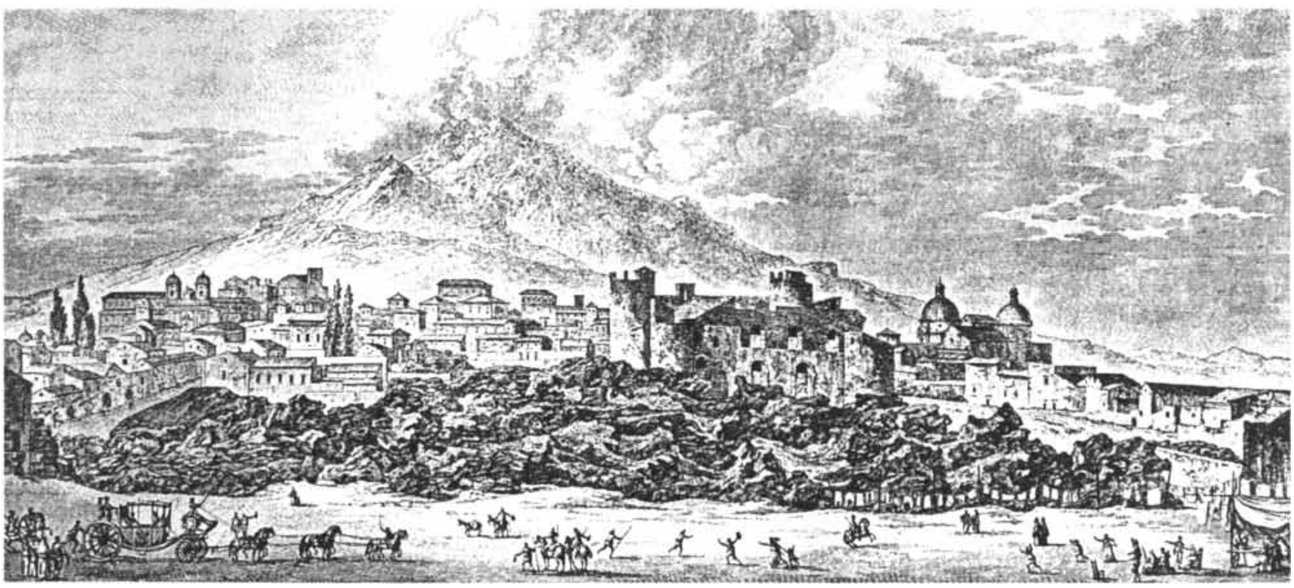
The book's subtitle, which implies volcanological interest alone, does not adequately describe the scope of the volume. Almost half of the chapters are devoted to the human ecology of Etna, to the hazards Etna poses to its population and to how these individuals respond and adjust. There is a fine map of land use around the volcano. The volcano has claimed the central 10 kilometers, perhaps; all around that barren stronghold agriculture prospers. The older soils are good on the lower slopes, and the fertile plains below have been irrigated during the past generation or two to provide intensive grain crops. On the slopes vineyards, citrus orchards and plantations of hazelnuts and pistachio are of high value. The mountain altitudes and even the usual plume have granted local rainfall and mild temperatures. Although use of terraced land on the steeper flanks of the volcano has dwindled during recent decades, new investment in the plains just south of the mountain has more than made up for that loss. Add in the abundant free-spending tourists, curious and daring, and the vacation-home buyers occupying picturesque if infertile land, and a reader is almost persuaded that "over the long term the region gains more from the volcano than it loses through short-term though often catastrophic events."

The technical chapters are not easy; in particular the rheological summary is impressively quantitative. There has been remarkable progress on mathematical models of the complex effusions of bubbly lava. On that basis, using powerful new measures of seismicity, ground tilt and uplift, volcanologists have made a plausible if yet untested map of the internal plumbing of the mountain. The main magma pool lies 20 kilometers deep; a hot column reaches up to within a kilometer of the surface even when there is no visible effusion. It was not always so; three centuries ago, before an extensive slide exposed and cooled the high eastern flank of the mountain, there was a secondary high-lying pool of molten rock. In those days it posed the principal threat of lava outflow and massive collapse.

A closing table outlines the present possibilities of intervention. Can we cool parts of the slow molten front to fend it off? They did that recently in Iceland, but it took a ton of water a second. Only right at the seaside might that be an affordable tactic for controlling Etna. Diversion by breaking through selected old lava walls might work on the high slopes, far from valued property. But down nearer the plains there is nowhere to divert the lava harmlessly. Perhaps dams and barriers might be partly built in advance to protect all the most valued assets, and then be rushed to completion as necessary once danger came. It would cost a great deal of labor. Yet volcanologists, guided by their mathematical models, expect one day to be able to predict and manipulate the course of lava flow.

When all has been proposed, it is the people who live around Etna who will decide. Their past adaptation has been a proud one. A wonderful old full-page engraving shows the town of Catania, hardly 15 miles from the central cone of Etna, viewed in the terrible year 1669. That April, Etna's lava inundated the place; 17,000 of its 20,000 people lost their homes. A horrid pile of lava lies across the scene; its reach fell just short of the castle keep. A nobleman has driven up in his coach-and-six to see the lava wall. Nearer yet to the still-hot lava some "ingenious Catanesi" have set up clotheslines so that their laundry may dry more quickly! Etna is not likely ever to defeat the resilient people who work and wait below it.

INTERSTELLAR MIGRATION AND THE HUMAN EXPERIENCE, edited by Ben R. Finney and Eric M. Jones. University of California Press ($19.95). IN ADVANCE OF THE LANDING: FOLK



Contemporaneous engraving showing laundry hung out to dry **(right foreground)** *in the heat of lava that invaded Catania in 1669*

CONCEPTS OF OUTER SPACE, by Douglas Curran. Abbeville Press, Inc. (paperbound, $16.95). Enrico Fermi was a frequent visitor to early postwar Los Alamos. Once in the summer of 1950 he walked to lunch with three or four friends talking about the flying saucers at that moment ubiquitous in news stories and cartoons. After some speculation on interstellar travel the conversation shifted. A bit later, "apropos of nothing," one witness recalls, "Fermi said, 'Don't you ever wonder where everybody is?' Somehow...we all knew he meant extraterrestrials."

Fermi's point is now a pivot in the speculation about human uniqueness. If you see space rather as Antarctica, a fine place to visit although no agreeable residence, you are inclined to seek electromagnetic messages from distant starlit planets, arguing from some crude symmetry. If instead you see the dark interstellar volumes as so many Atlantics and Americas, you wonder at the absence of perceived fleets of voyagers, and not without a certain hubris you tend to view our own species as first in our galaxy to dream about spacefaring, and historically close to takeoff en masse.

This volume looks outward, in advance not of everyone landing but of many of us taking off. Clearly placed on the serious side of science's border with science fiction, it reports a conference at Los Alamos in the summer of 1983 that imaginatively extends some of Fermi's 1950 implications. The attention of about 20 authors, mostly astronomers, space engineers, anthropologists and historians, focuses on such topics as the means and modes of projected galactic travel, the demographic, evolutionary, genetic, social and moral issues raised by long-term interstellar migration of humans, and the lessons drawn from cross-cultural studies of human migrations of the past. Two final papers are more skeptical of mass space travel. They include a theoretical model of migration rate that points up the uncertainties of such extrapolation, and a summary of the status of the empirical search for microwave signals, much swifter and cheaper than travel.

A fascinating polarity is found in the extrapolations collected here, a duality of view: should we imagine bold, heroic voyagers or quiet, patient ones? We read of fastships and space nomads. Fastships might carry hundreds of expert colonists from the rich, sunlit oasis of home into the dark, the well-fitted ship powered by a fierce microwave beam. (Is anyone listening?) The nomad bands build their slower, lighter ships of cometary matter and power their extended-family lives by the concentration of pale starlight in huge farms of gossamer mirrors, a continental area devoted to each inhabitant.

We read of the real past, of !Kung hunters and gatherers, egalitarians and sharers, who spread by a random hiving-off into the desert emptiness, and of the Vikings, fleeing the populous but narrow arable fields crowded among the fjords, to found predator states whose most ambitious colonies did not survive. If a linear scale factor of tens of billions can be conceded, the Polynesians present the neatest analogue; their voyaging canoes, which over 10,000 years peopled the isolated atolls flung across the ocean wastes, are indeed evocative of wanderers among the stars.

A Malthusian engineering of boundless hope is also represented, its semilog plots rising linearly to the stars. One essay even suggests that the invisible mass that binds the orbiting stars of the spiral galaxies might turn out to be husbanded and converted stars. Gently unwrapped and reassembled into many long-lived little dwarfs, their dim light would be hidden under bushels of radiation-frugal macromachines, the "habitats...of advanced civilizations," albeit preternaturally inconspicuous ones, even shy.

The First Law is employed here scrupulously, so that uses of cosmic sources of matter and energy are optimized in detail. Yet a reader feels that the Second Law, with its insistence on noise, error and waste, its up-front costs for every structure and every choice, has been rather slighted.

The magical key to the future is the self-reproducing and exponentially growing construction system that can unwrap a sun, say by drawing off surface plasma into new-made magnetic channels. Diverted, the gas flies far into space, there to slow and condense onto prepared dust. Eventually the material would be gathered into self-gravitating, tiny substars, cold warehouses for dwarf starry fusion furnaces to light the future. This audacious author, David R. Criswell, remarks that he will leave alternative schemes "to others with more imagination than myself."

There are other heroic forecasts, an evolutionist seeing a diversity of our descendants scattered across the galaxy "within 2 million years," novel hominid taxa that would astound us by their forms and adaptation. Another expert looks confirmingly to a far future Milky Way peopled by our prosperous progeny, by then hundreds of billions of intelligent species, dominated by those few more aggressive races of intrepid voyagers. This future includes rare conflicts. Within such a dilute regime even nations of heroes can support no more than a few million interstellar wars at a time. The editors (Finney an anthropologist at the University of Hawaii, Jones an astronomer at Los Alamos) comment elsewhere that the conquerors who roved in longship and caravel almost never made landfall on any coast to which the meek had not already spread in their slow and silent way.

This volume reasons out the hopes of an enthusiastic moiety of the technologically adept. Hard to test, its arguments by analogy and extrapolation are fine reading, if less than uniformly persuasive. Fermi's question may have a simpler answer. Everyone is trying hard to make a living not far from some long-lived star. They may be "noble in reason...in action how like an angel!" Yet they are finite beings just as we are; therefore they too are confined by logistic curves that having soared past the lively immaturities of exponential growth flatten to diminishing returns along every physical dimension. Such beings may even be freed of the old dream of immortality.

The second book is the work of a Canadian journalist-photographer who spent a few years on the road, seeking from Quebec to the Mojave the outward and visible signs of a new kind of inward grace: the sacrament of saucers from heaven. His remarkable photographs and clear first-hand accounts, always gentle and yet precise, disclose a widespread if small popular subculture that has about it the air of an inauspicious beginning. Once almost 2,000 years ago a traveler might have found something like this among small, serious groups gathered in Rome, Corinth or Ephesus; only 150 years ago noteworthy examples might have been visited in upstate New York. St. Peter's and Salt Lake City were hard to foresee.

The images of this particular salvation take a curious form. About a fourth of some 60 photographs display the structure of a rocket, represented in the round in a matter-of-fact way, at a scale practicable for local carpenters working plywood and sheet metal, an icon that is carefully modeled on the tangible results of the engineer. A somewhat larger number of these images represent no sight ever seen, no real contrivance. Instead they display at some scale in some material (a few are only painted) the prototypical flying saucer, a form familiar only from published artwork in the media, large and small. These are found on gas stations and in gardens, nestling in the woods or next to power pylons. Rather fewer than a third of the photographs document more complex social groupings among which the same roots have fully flowered into costume and ceremony, and a few show assemblages of real technology. Some here are recognizable cargo cults, not on an out-of-the-way Melanesian island but wistful social isolates within the ocean of American life.

There is strong continuity with past phenomena. The history of the late George Adamski, a prolific author and guru who dwelt at Mount Palomar, Calif. (below the observatory site), is revealing. By his own lengthy accounts a familiar of the flying saucers, he repeatedly enjoyed passage on interplanetary voyages, often to the dark side of the moon. Before the 1950's Adamski had been privy to quite distinct mysteries: he had founded a monastery in the 1930's to gather seekers after Cosmic Law (and to legitimate the sale of wine). He once frugally "resurrected a metaphysical tract he had written in the 1930's by substituting 'the Space Brothers' for 'the Royal Order of Tibet' throughout." Adamski is still revered by many, some of whom hold he has returned to earth. His feats went along easily with his status as minor celebrity in image and print, a man received during the unabashed heights of his appeal by members of Congress and by the Queen of the Netherlands. Adamski's photographs did much to establish the canonical form in which we conceive the flying saucer; there is evidence that one of the most celebrated of them presents a close-up of a small aluminum egg incubator.

We see a Hollywood service in which a Spiritual Battery is recharged; it is a little ceremony wherein by turns the gowned communicants gravely place their hands against the device, cradled on its tripod. Such a battery is rated at a capacity of 700 prayer-hours; several are kept at full charge ready to forestall world crises. Their efficacy has often been demonstrated. Some other views of taking part in the celebration of extraterrestrials of power and good will are here; often high planetary dignitaries and their servitors appear costumed in the satins and sequins of remembered theater.

Only two technical scenes occur. One shows a private roomful of relay racks, with which the complex spectral communications claimed are probably only simulated. (That judgment derives from the awry description and from the simple Lissajous figure seen on the big CRT screen central to the display.) Another site, on rolling land near Austin, Tex., was far more ambitious. There, a decade back, an array of lights flashed some coded function of pi; other saucer attractants, including a ruby laser under computer control, were deployed. The expense was borne by a few well-heeled followers.

The latest manifestation reported is not at all a folk phenomenon but a frank example of commercial theater, by Bob Gurr. The closing ceremonies of the 1984 Summer Olympics were ended by a lavishly staged night saucer visit to that packed and darkened stadium, complete with a probing searchlight from the hovering, flashing, responsive craft. "A seven-foot-eight-inch white-suited alien" visitor appeared, to voice a benign message as the fireworks rolled, the entire production carried worldwide on television. A single video image is shown.

In this volume the yearning is clear, the complexity of modern fears and hopes evident, the influence of newly predominant culture images pervasive. Both pathos and pleasure are to be seen in the poignant faces. Yet neither elegance nor wit is often tapped by these visually banal icons; only two beautiful images are found. One is a striking accident of a real effort; the big ring of lights that aspired to lure a passing saucer is magically inviting in the soft Texas night. The other serves as jacket photograph; it is not folk art but an abstracted small saucer form possibly of rock. Painted (it appears) by a genuine artist, Jene Highstein of Park Forest, Ill., it rests in a wide field of sere autumn grass against a twilight sky glowing green.

Both the well-argued forecasts of starship takeoff and the folk celebration of saucer landing may derive from a belief long ago defined by an American writer of insight; the two seem caught alike in the gleam of the green light at the end of Daisy's dock, where Jay Gatsby believed he might grasp "the orgastic future that year by year recedes before us. It eluded us then, but that's no matter—to-morrow we will run faster, stretch out our arms farther.... And one fine morning—"

THE HUMAN SKELETON, by Pat Shipman, Alan Walker and David Bichell. Harvard University Press ($27.50). "The subject is bone" are the opening words of this beautifully illustrated and original text. Its intention is to treat the human variety of the material widely, both as cellular tissue and as developing organ, in structure and in function. The special intent is to provide a strong framework for the interpretation of enduring bony remains so that through forensics or anthropology we can understand more of the past life they framed. The two senior authors are Johns Hopkins anatomists whose current research centers on human paleontology; the third contributor is the medical illustrator who with pen and pencil has created the fastidious figures that mean so much to a reader who needs to be engaged with three-dimensional form.

The wide scope of the text gives it value for a general scientific audience; there may be no other book that takes such a broad view of what began as the most classical of anatomical specialties, significant now in a wide variety of disciplines. That proud history is the source of the greatest difficulty for the general reader as well: a classical jargon saturates the account. All is named in polysyllables of Greek and Latin, little is lettered or numbered. It is easy to understand that a study whose subject matter is an assortment of 206 standard bones needs many specialized nouns. Good enough; we shall try to learn the vomer and the innominate, the more since the diagrams show them well labeled. But even the helpful geometrical map does not make it easy to remember the many terms for relative direction and position of axes and planes or to understand the compound adverbs that the noun forms so freely generate, for example "mesiodistally." A dozen cranial measurements are still useful even though they recall old follies of physical anthropology in its claims for bony destiny. They set out the distances between standardized landmarks on the skull whose osteometric names seem a tour of some strange Mediterranean coast, from Bregma to Nasion with a stopover at Glabella.

Words apart, there is plenty of matter here. The first quarter of the book, after preliminaries, treats the nature of bone, not particular bones. It begins at cellular level and continues on to tissue structure, growth, strength and even the subtleties of joints. Bone is a composite material, the protein fiber collagen providing the matrix for mineral crystals. It performs rather better than oak wood in yield strength and stiffness (here the graph shown is discrepant from the text data). Its microstructure is crack-resistant, and its lack of the simple prestress found in tree trunks seems to fit the complex varying loads bone must handle. The dynamic nature of bone in growth and response to load is strongly emphasized, a response that implies unending remodeling and resorbing. It is the supply of blood to the living cells mediated by complex circulatory design that allows such change. The skeleton is no mineral but a living organ that changes continuously in fabric and in outline. Only so can it remain nearly constant in function.

The lubrication of our most mobile

joints, the authors note, is subtle and still poorly understood. The fluid film between the microscopically rough bearing layers is too thin to lubricate them, although the slippery fluid is known to be essential. The entire joint design speaks against the possibility that such a film can be drawn in by the slow relative motions. A model of a static lubricant film, pumped in by weeping and resoaking, as with a wet sponge pressed on a hard, smooth surface, fails quantitatively. A surface-active lubricant may be present and also, one of the authors has proposed, a novel form of fluid pumping by cartilage distorted under load.

The greater part of the text provides an account of bone function. System by system there are chapters about breathing, the upper limb and its manipulations and the lower limb and walking. The skull both as container and as a mechanism for chewing and speech is also presented. Evidently the muscles that drive the static framework need to be part of the account, and they are, in figure and text alike. A few pages on walking bring out its complexity by beginning with a much simplified model and then extending it over six stages of increasing realism. The account of chewing is also revealing: incisal biting is symmetrical but molar chewing is asymmetrical, the process taking place on one side at a time (some individuals favor one side over the other). The action is a balance between a working side and a foodless side that takes up the reaction from the muscular sling pulling on the mandible. A shear as well as a crushing force that is essential for easy swallowing can thus act on the food.

The most interesting part of the text is the last. It briefly treats the fundamentally statistical diagnosis of age, sex, race and stature from the evidence of the bones. The development and then the attrition of teeth, the fusion of the sutures, the ossification of cartilage layers in elbow, hip, knee and other structures, the degeneration of loaded joints, as well as microscopic studies of teeth and of bone remodeling—all give evidence of age at death. Determinations of sex based on several skeletal measurements, particularly those of skull, pelvis and limbs, can be made that have confidence levels well above 90 percent.

Race or ethnicity demands comparative population statistics; in other words, multiple cranial measurements and their careful analysis. The overall scores are probably a crude measure of genetic relatedness. A few pages of tables list the height of Americans expected from the lengths of various bones; the study, a generation old, included more than 5,000 known and grouped individuals. Height is particularly sensitive to nutritional status in childhood; environmental rather than genetic factors appear to dominate.

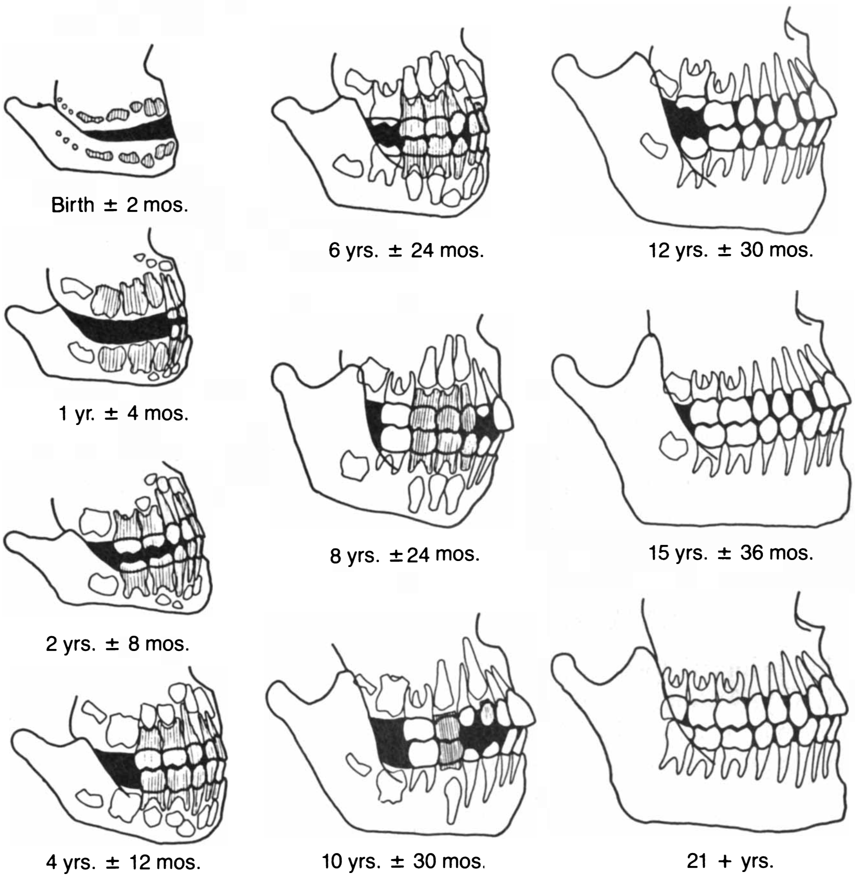
The last two chapters discuss the signs of trauma and disease. It is remarkable that a rare condition, hypervitaminosis A, is diagnosed in one skull of *Homo erectus*. It is known today as a rare congenital enzyme deficiency or as the result of overdose of carnivore liver. Less exotic injuries such as fractures and punctures can often be recognized. That a given injury is the cause of death is less certain from the state of healing; the process can take years.

Reconstruction of living appearance from bone remains has passed careful testing. It is an art pioneered through precise measuring of soft-tissue thickness at many points on the head of cadavers of varying groups, such as males and females. Such reconstruction is clearly mostly a forensic or a historical study.

More recent and more general are techniques that reveal behavior. One is tooth microwear, which can uncover patterns that correlate with broad diet categories (a frugivore's teeth show patterns of wear distinct from those of a carnivore). The length and width correlations of scratches in a molar tooth from several species of primates sharply separate the leaf- and stem-feeding gorilla from the fruit-eating orang. The chimp lies between.

Isotopic analysis of bone collagen is a hopeful new technique. The rise of maize culture results in a change in carbon-isotope abundance. Maize has a photosynthetic pathway different from that of the plants making up the mixed diet that corn replaced. Altered nitrogen-isotope ratios may provide a measure of the length of the food chain, for there is an enrichment of the heavier isotope from plant to secondary carnivore. Marine and land plants also differ isotopically. The original sample is of course the key to success; the organic matrix in bone weathers in burial. The only safe sample is biochemically identifiable collagen.

We are what we eat, what we suffer, what we do; bone will bear the marks longer than any other memory can. The last line of this vivid, rich although nowhere easy book offers the authors' hopes in strong metaphor: the transformation of skeletal remains "back into living organisms."



Dental-eruption diagram from **The Human Skeleton** *(deciduous teeth are hatched)*

For the Technical Professional:

The story behind IBM's new 32-bit RISC PCs.

These aren't your ordinary workstations.

For IBM RT Personal Computers—6150 and 6151—are built around a revolutionary 32-bit, virtual memory, RISC-based microprocessor. Developed and manufactured by IBM.

The significance? Far faster throughput. Because RISC (for Reduced Instruction Set Computer) architecture eliminates many of the unnecessary bells and whistles found in conventional chip design.

What's more, an IBM RT PC can directly address up to four megabytes (Mb) of real memory, over a trillion bytes of virtual memory and up to 210 Mb of DASD, or disk storage.

All of which makes IBM RT PC microprocessors particularly well suited for the total application requirements of technical professionals working in engineering, scientific, industrial and academic environments.

The IBM RT PC 6150 Model 20
Floor-standing unit

CADAM® and CIEDS[1] Software

One of the key programs available for the IBM RT PC—attached to the IBM 5080 Graphics System—is Professional CADAM. And it's compatible with established host-based CADAM.

There's also Computer-Integrated Electrical Design Series (CIEDS), which allows for the schematic capture of integrated circuit designs.

And there is a growing list of applications available from software companies in the following fields: mechanical and electrical CAD/CAM, petroleum production, artificial intelligence, software engineering, project management and graphics.

UNIX[2] System V

The primary operating system is derived from the UNIX System V—with some very important enhancements added by IBM. These enhancements make this operating system—Advanced Interactive Executive or AIX[3]—much easier to use, give it improved performance and reliability, and also add a virtual storage capability.

This operating system gives you multi-user and multi-tasking operations for up to eight concurrent local or remote users working on low-cost ASCII terminals such as the IBM 3161.

There is also a full-function IBM SQL/RT relational data base management facility available.

Understands Many Languages

The IBM RT PCs are highly literate. You'll be able to write and run applications in C, Fortran 77, BASIC, PASCAL and Assembler.

And as for general-purpose applications, you'll be able to run UNIRAS[4] (for engineering and business graphics), RS/1[5] (for data analysis and graphing functions), IMSL[6] (a library of Fortran subroutines), Interleaf's Workstation Publishing Software[7] Solomon III[8] (for accounting), SAMNA+[9] (for spreadsheet and word processing) and Applix IA[10] (for integrated office functions).

PC Compatible

The IBM RT PCs can incorporate as an optional feature an Intel 80286 coprocessor (in effect a PC AT on-a-board) that runs concurrently with the 32-bit RISC processor. That means the RT PCs can support many existing IBM PC applications and the interchange of their related data. It also means the RT PC I/O bus can accept standard PC cards.

Contact your IBM marketing representative. Or call 1 800 IBM-2468, Ext. CZ/428, for participating Authorized Dealers, or for free literature. Software companies interested in writing programs for the IBM RT PC should call 1 203 783-7141.

IBM®

IBM 6-86
DRM, Dept. CZ/428
101 Paragon Drive, Montvale, NJ 07645

☐ Please send me information on the IBM RT PCs.
☐ Please have an IBM marketing representative call me.

Name________ Title________
Company________
Address________
City________ State____ Zip____
Phone________

The IBM RT PC 6151 Model 10
Desk-top unit

Established 1845 SCIENTIFIC AMERICAN June 1986 Volume 254 Number 6

Beyond the Era of Materials

The industrial nations now face a historic change: economic growth is no longer accompanied by increased consumption of basic materials. The economic outlook depends on the capacity to adapt to this shift

by Eric D. Larson, Marc H. Ross and Robert H. Williams

The significance of materials for human society is reflected in the fact that several epochs are named for materials that were exploited intensively during their span: the Stone Age, the Bronze Age and the Iron Age. The Industrial Revolution was based in large measure on radical improvements in the methods for modifying basic materials such as cotton, wool, iron and later steel. Continuing improvements in production techniques have since made a growing number of material products available to more markets. Indeed, since the Industrial Revolution a hallmark of economic growth has been an increase in the consumption of materials.

In recent years there appears to have been a fundamental change in this pattern of growth. In North America, Western Europe and Japan economic expansion continues, but the demand for many basic materials has leveled off. It appears that the industrial countries have reached a turning point. They are now leaving the Era of Materials, which spanned the two centuries following the advent of the Industrial Revolution, and are moving into a new era in which the level of materials use will no longer be an important indicator of economic progress. The new era may turn out to be the Era of Information, but it is probably too soon to name it with confidence.

It is not too soon, however, to examine the causes of the change. Our work suggests four causes. Substitution of one material for another has slowed the growth of demand for particular materials. So have design changes in products that increase the efficiency of materials use. Perhaps more important, the markets that expanded rapidly during the Era of Materials are by and large saturated. And new markets tend to involve products that have a relatively low materials content.

The passage of the Era of Materials has some significant implications for the further economic development of the industrial countries. Those implications constitute a set of lessons that are only now beginning to be learned. The first lesson is that, contrary to much that has been said and written, imports had only a minor role in the weakening of the U.S. basic-materials industries through the 1970's. Instead that weakness was due largely to a leveling off of demand. Since about 1980, however, imports have begun to undermine U.S. strength in materials processing. It is probable that in the decades ahead more of those operations will be carried out overseas, where costs are lower. The affected U.S. industries will fight to retain the later stages of manufacturing and will increase their role in the processing of recycled materials.

Although such trends appear to be disturbing, they by no means imply the death of manufacturing in the U.S. and other industrial countries. There are many opportunities for innovation in high-technology products characterized by a low materials content per dollar of value added, such as high-strength and corrosion-resistant steels and specialty chemicals for the fast-growing pharmaceutical, electronics, pesticide, herbicide and biotechnology markets. Nevertheless, the end of the Era of Materials can be followed by a new era of growth in manufacturing only if the watershed through which industrial societies are passing is widely understood and appropriate adjustments are made.

Since in our view the primary factors that are affecting the changing role of materials involve consumption rather than production, we shall begin by describing the forces that underlie the demand for materials. We measure the consumption of materials in physical quantities (kilograms) rather than economic ones (dollars) to give a sharper picture of the changing role of materials. We represent the physical quantities in two main ways. If the physical quantity is divided by the population, the result is the consumption of a material per capita at a particular time. If the quantity is divided by the Gross National Product (G.N.P.) in constant dollars, the result is an index of the significance of a material in the economy as a whole. By examining these two measures over time it is possible to identify significant trends in the demand for a particular material. When the trends for various materials are compared, it is found that for many materials the historical pattern of demand follows a similar course.

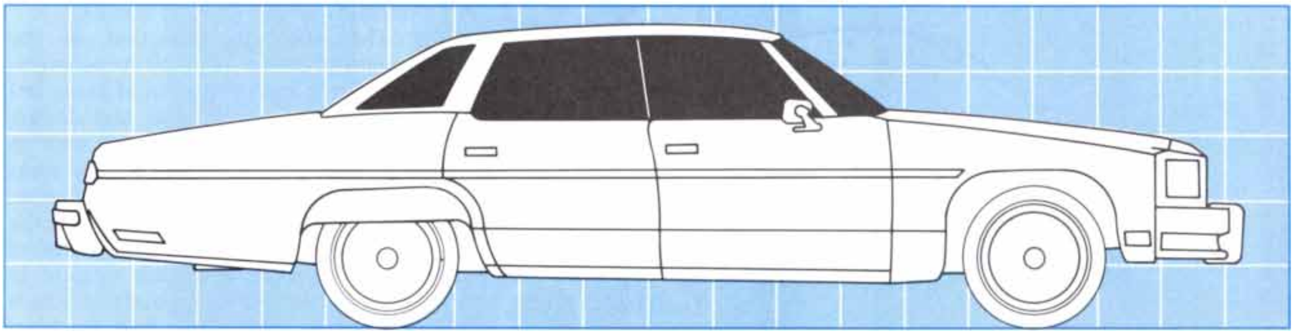
When a material is introduced, consumption rates (measured in kilograms per capita) are typically low, and there are vast potential markets. In the early part of the cycle consumption of the new material generally increases much faster than the economy as a whole. The rapid growth is reflected by a rise in the index of kilograms per dollar of G.N.P. Such rapid growth encourages improvements in processing technology that increase productivity and the quality of the product as they lower its price. These changes further stimulate the growth of demand.

In the second phase of the cycle more sophisticated products are made, the material begins to be a less significant component of the production process and the ratio of value added in manufacturing to the quantity of material increases. Additional innovations make it possible to use the material more efficiently. In this phase the demand for the material measured in kilograms per dollar of G.N.P. peaks and begins to decline, even though consumption in kilograms per capita may still be increasing. Then, in the final phase of the cycle, the markets for the material in bulk become saturated; new markets are largely for specialty products, which have little effect on the total quantity of consumption. The reversal of growth is so complete that even per capita consumption levels off and may begin to decline.

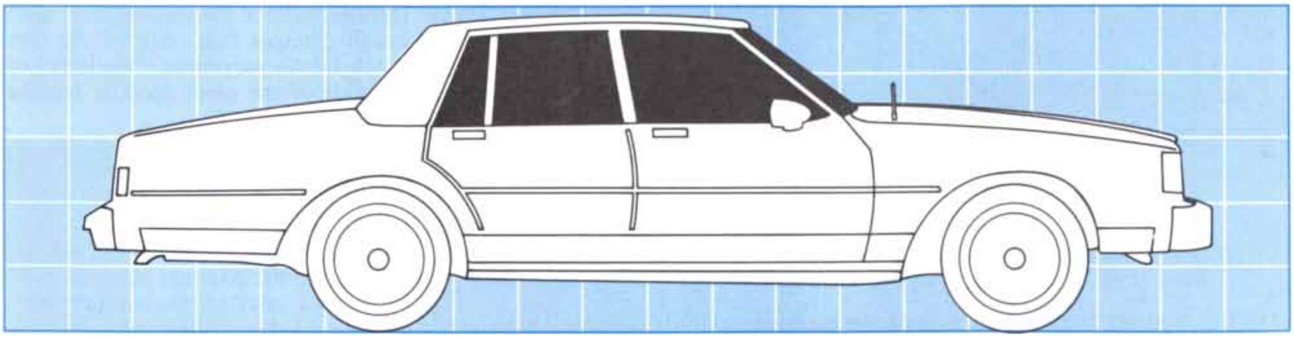
The history of steel consumption in the U.S. provides an excellent example of the demand cycle for a material. During the late 19th century steel made a significant contribution to the construction of cities, heavy industry and railroads. The rapid growth in demand created a favorable climate for innovations such as the Bessemer process, which reduced costs and encouraged continued expansion. As a result of these trends the demand for steel grew faster than the G.N.P. until about 1920. In the 1920's, however, the market for heavy steels of the type employed in building infrastructure stopped growing rapidly. As a result the consumption of steel per unit of G.N.P. began to decline from the historic peak that had been reached in about 1920.

Although the demand for steel had begun to decrease in relation to the economy as a whole, per capita consumption continued to increase: steel was being produced for consumer goods. For example, steel sheets were needed for automobiles and household appliances. Such markets for consumer products were far from saturated in 1920. In the 1950's, however, the growth of per capita demand began to decelerate even as more consumer goods were produced. When consumer markets approached saturation in the 1970's and no new bulk markets developed, per capita demand leveled off. It is currently falling, and the significance of steel consumption in the national economy as measured by kilograms per unit of G.N.P. is roughly what it was 100 years ago, having decreased to 40 percent of the peak level of 1920.

Most other basic materials seem to be following a similar trend. We carried out case studies for six materials in addition to steel. Among them cement and paper are commodities that have been used extensively since the early days of industrialization. In contrast, aluminum, ammonia, chlorine and ethylene are materials that came into widespread use only in the 20th century. In every case we found the cyclical pattern of demand that is



1976 CHEVROLET CAPRICE CLASSIC had a curb weight of 2,007 kilograms (4,424 pounds). During the mid-1970's rising energy prices provided a strong impetus for reducing the size of U.S. automobiles. For the Caprice the major "downsizing" took place between 1976 and 1977: the 1977 model weighed 300 kilograms (661 pounds) less than that of the year before. By means of changes in design and substitution of lighter, stronger materials, weight was saved throughout the car. For example, 54 kilograms was saved in the grille, sheet metal and electrical work, 45 in the suspension and brakes, 30 in the frame and 25 in the wheels, tires and steering gear.

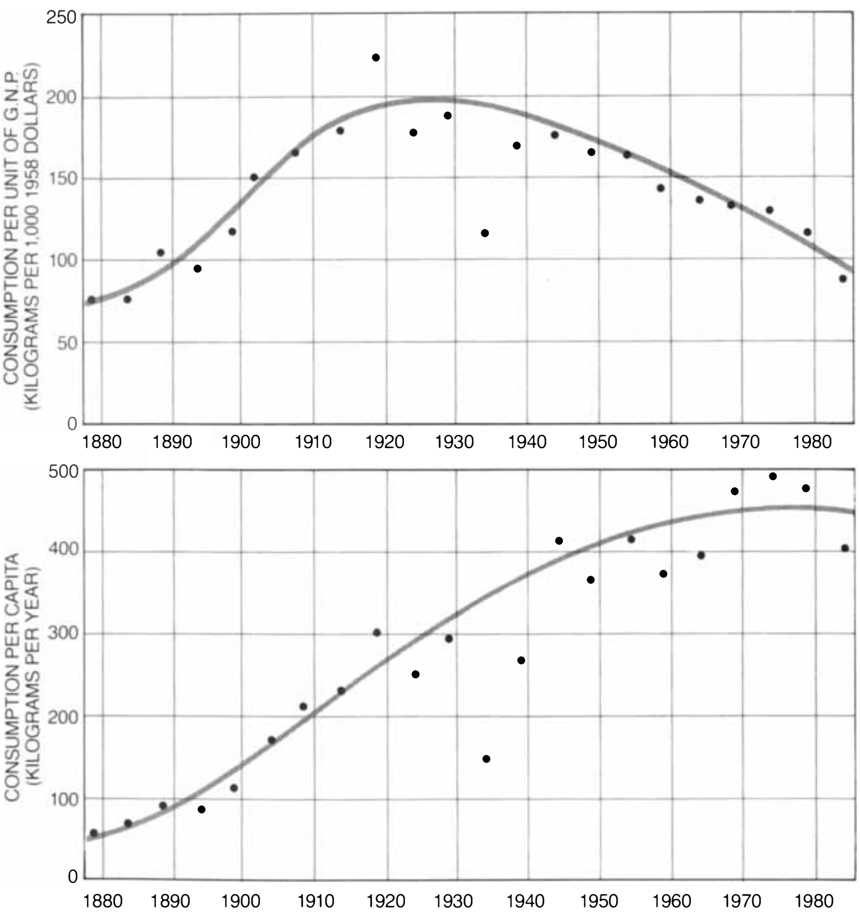


1986 CHEVROLET CAPRICE CLASSIC has a curb weight of 1,617 kilograms (3,564 pounds), or 390 kilograms less than the 1976 model. The disparity in weight between the two Caprices reflects a general trend among U.S. automobiles. Between 1975 and 1985 the average weight of U.S. cars fell from 1,727 to 1,450 kilograms. Part of the reduction was due to decreases in the iron and steel content along with increases in the fraction of aluminum and plastic. More efficient use of materials is among the causes of a leveling off of demand for basic materials in the industrial countries. The leveling off represents a significant turning point: since the beginning of the Industrial Revolution economic growth in industrial countries has been accompanied by increased demand for basic materials.

predicted by classical theories of industrialization. The duration of the cycle was much longer for the traditional materials than it was for the modern ones, but the outcome was the same. For all seven materials consumption per dollar of G.N.P. had begun to decline by the 1970's, and consumption per capita has now essentially leveled off. Similar trends are seen in the industrial countries of western Europe, including France, Germany and the U.K.

Is the leveling off of demand for materials merely a pause in a historical pattern of growth, or does it mark a fundamental structural change in the economy? Because the departures from long-term trends are recent, statistics alone cannot provide the answer. When the statistics are considered along with the underlying causes of the leveling off, however, they suggest that a profound shift is under way. As we noted above, four factors appear to be responsible: substitution of materials, more efficient use of materials, saturation of markets and shifting consumer preferences. Whereas the first two have been operating almost since the beginning of the Industrial Revolution, the second two are relatively recent phenomena that are associated with the final stages of the demand cycle.

The substitution of modern materials has been slowing the growth of demand for traditional materials such as steel and paper. In the automotive market, which accounts for between 15 and 20 percent of the demand for steel in the U.S., lighter materials are replacing steel. The tendency toward substitution was reinforced in the 1970's by the promulgation of minimum standards for fuel economy. To increase fuel economy manufacturers began to build smaller, lighter cars that contained less steel. The average weight of cars made in the U.S. decreased from more than 1,700 kilograms in 1975 to less than 1,500 in 1985. Over the same span the fraction of the average vehicle made up of iron and steel fell from 81 to 69 percent; the fraction made up of aluminum and plastic rose from 6 to 11 percent. Many such substitutions are taking place throughout U.S. industry.

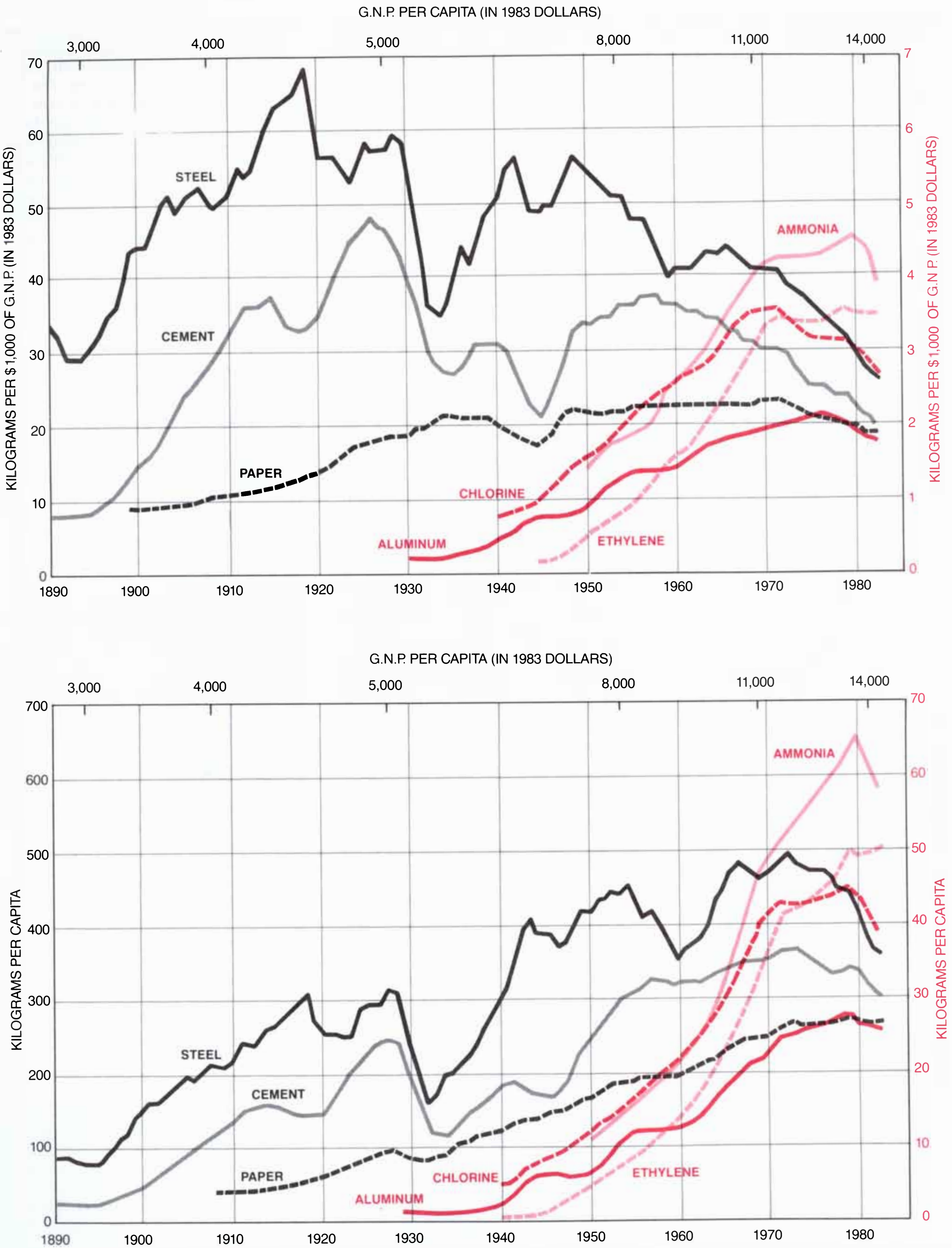


STEEL CONSUMPTION in the U.S. illustrates the classic cycle of changes in demand for a basic material. The upper panel shows the consumption of steel in kilograms per $1,000 of Gross National Product (G.N.P.). The lower panel shows consumption in kilograms per capita. In the early part of the cycle demand increases rapidly according to both measures. At some time consumption per unit of G.N.P., which indicates the relative importance of the material in the economy, reaches a peak and begins to decline. Steel consumption reached that point in about 1920. Per capita consumption continues to grow after that, but in the last stages of the cycle per capita consumption levels off and may even decline.

Since early in the 19th century, improvements in the strength and durability of materials have also made it possible to decrease the amount of material in products without vitiating any of their functions. Indeed, the function of the product may be improved at the same time, as is shown by the history of the locomotive. In 1810, when boilers were made of cast iron or sheet iron, the ratio of weight to power in a typical locomotive was about 1,000 kilograms per horsepower. Some 50 years later steel boilers were introduced, making possible significant reductions in weight; by 1900 the ratio was about one-tenth of what it had been in 1810. As electric locomotives were widely introduced around 1950, the ratio reached 25 kilograms per horsepower. By 1980 it was 14.

The dramatic reduction in the weight-to-power ratio of locomotives was due to a combination of many improvements in materials and design. Similar refinements can be seen in many modern products. Plastic films available today are stronger yet thinner than those sold a decade ago; the new films represent a general trend toward a lower materials content in plastic products. The radial tire, which is much more durable than the bias-ply tire, is an example of an improved design that leads to demand reduction (in this case for rubber). The aluminum can has also undergone an evolution in design: its average weight decreased by about 20 percent between 1970 and 1984, owing largely to improvements in production technology. Thinner walls of the can, together with design changes that reduced the size of the lid (in particular, "necking" of the ends of the can), account for the reductions.

At least two factors drive the increasingly efficient use of materials. One is the rise in the cost of producing the materials themselves, much of it due to higher prices for energy. Since 1972 the average price of electricity and coal (corrected for inflation) doubled for industrial customers in the U.S.; the price of natural gas quadrupled. Many materials-processing industries are quite energy-intensive: they consume large amounts of energy per unit of output. About three-fourths of the cost of producing am-



TRENDS IN CONSUMPTION of seven materials exemplify the overall fate of basic materials in the U.S. Among the seven are two "traditional" materials in addition to steel: cement and paper. The others are "modern" materials: aluminum, ammonia, chlorine and ethylene. Consumption of the traditional materials peaked long before that of the modern ones, but use of all seven is now declining relative to G.N.P. (*upper panel*). For most of these materials per capita consumption is also declining (*lower panel*). The scale at the top of each panel shows G.N.P. per capita. As it suggests, the diminished role of basic materials is related to increasing affluence.

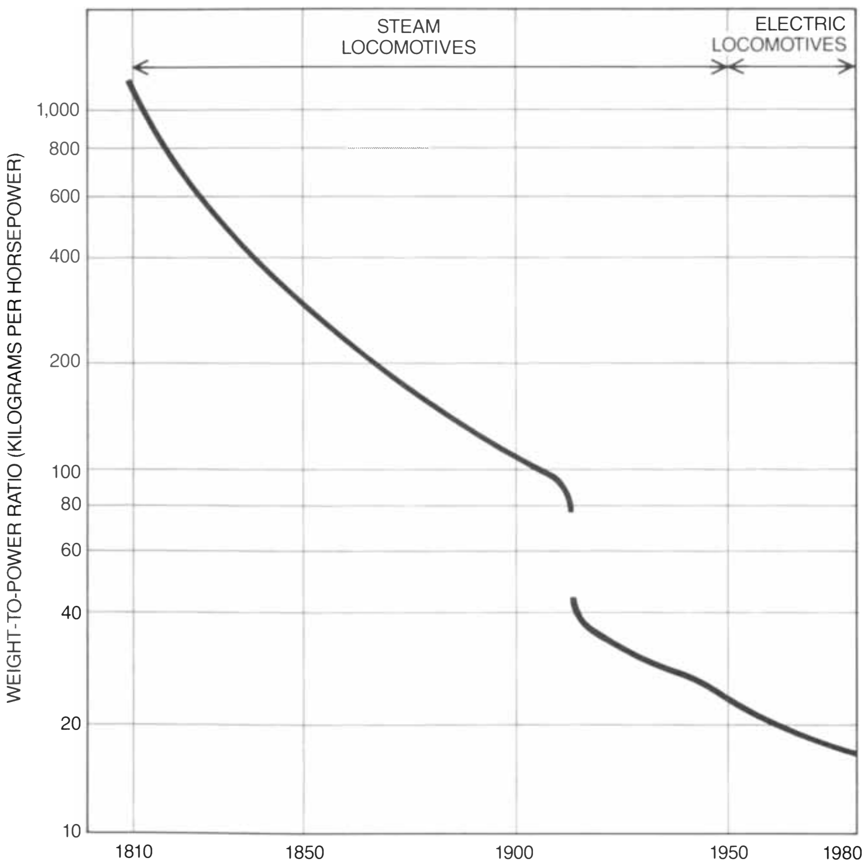
monia is accounted for by natural gas (which for making ammonia is both a feedstock and an energy source). About one-third of the cost of making aluminum is accounted for by electricity. Such industries are naturally quite sensitive to increases in the cost of energy. The increases thus provide a powerful stimulus for making more efficient use not only of energy but also of materials.

The second factor that stimulates efficiency is competition from substitute materials. Such competition particularly affects the manufacturers of traditional materials, whose products must compete against modern materials with more desirable properties. Partly to compete against the makers of aluminum, plastics and fiberglass, the steel industry has put increasing emphasis on the production of specialty steels with properties such as high strength or resistance to corrosion. The automotive industry has responded by utilizing more specialty steels in its products. In 1975 about 5 percent of all the steel in an average car was high-strength or stainless; by 1985 the fraction was 14 percent, and most observers expect it to rise above 20 percent by early in the next decade. Such changes are accompanied by a reduced demand for steel, since each kilogram of high-strength steel typically replaces 1.3 kilograms of ordinary carbon steel or cast iron.

Although materials substitution and improved efficiency of materials use have become more significant in recent years, they are probably not as important in reducing demand as the relatively recent phenomena of the saturation of traditional markets and the shift in consumer markets to new, less materials-intensive products. Consider steel and cement. Much of the underlying demand for steel and cement in bulk in the U.S. in the 19th and early 20th centuries came from the building of infrastructure, including highways, railroads, public transportation, commercial buildings and housing. The era of infrastructure building is now all but over. The rail network has been shrinking for many years, and highway construction slowed in the 1960's with the completion of most of the interstate system. Demand for new housing will undoubtedly decline as the young-adult cohort of the population ceases to expand. The only area where demand for steel and cement continues to be strong is the market for commercial buildings, which is not a large one for materials in bulk.

Other basic-materials industries are facing similar, if less epochal, shifts in the pattern of demand. The most rapidly growing market for paper during the 1950's and 1960's was in corrugated boxes to serve as shipping containers. By about 1970 that market had encountered its limit because all except large commodities and bulk goods were being packaged in corrugated boxes. A parallel saturation of the market for inorganic chemicals in bulk is illustrated by recent trends in the utilization of nitrogen fertilizer, which accounts for about 80 percent of all the ammonia produced in the U.S. From 1955 to 1980 the amount of fertilizer increased more than sixfold. Future growth will be slow, however, both because most cropland is currently receiving high levels of fertilizer and because today's high prices for fertilizer provide a powerful incentive to use fertilizer more effectively. Saturation of the market for organic chemicals is illustrated by the case of synthetic fibers, which have captured most of the apparel, automotive and home-furnishing markets; there are no apparent opportunities for further expansion.

Markets for heavy consumer goods are also reaching the limits of growth, partly owing to the affluence of the industrial countries. Increases in income in the past few decades have made it possible for most households to buy a wide range of basic consumer goods. In the U.S. almost all adults who might drive have cars. Furthermore, people are now keeping their vehicles longer. From the late 1950's until 1975 the average age of passenger cars held fairly steady at between 5.5 and six years. After that it rose quickly, partly as the result of improved durability, reaching 7.4 years by 1983. Most observers of the automobile industry expect new-car sales to remain at the levels of the mid-1970's for the foreseeable future. Similar trends apply to such heavy household appliances as stoves, refrigerators, washers



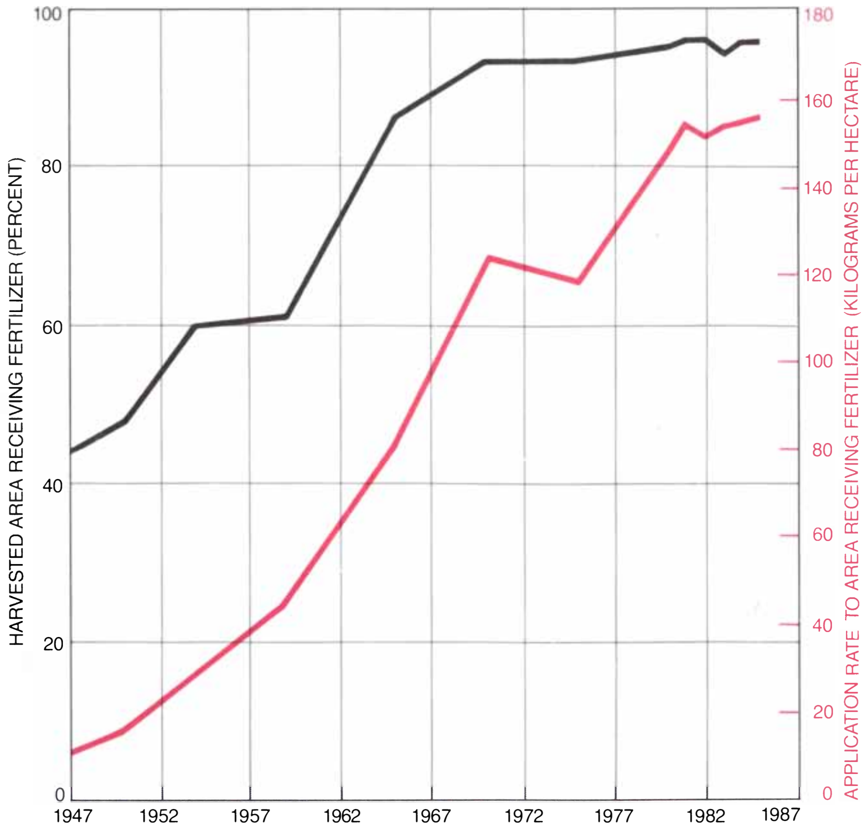
WEIGHT-TO-POWER RATIO of locomotives underwent a decrease of nearly 70-fold between 1810 and 1980; the decrease reflects many improvements in design and materials. In the mid-19th century iron boilers were replaced by boilers made of steel, a change that made possible lighter equipment and higher internal pressures. By 1900 the ratio had decreased by a factor of 10, and it continued falling during two world wars, reaching a level of about 25 kilograms per horsepower as electric locomotives were introduced around 1950. (The gap between 1910 and 1920 results from the disruption of data collection during World War I.) Similar (albeit less dramatic) improvements have been made in many industrial products. Substitution of materials and design changes that lead to more efficient use of materials are two of the factors responsible for the leveling off of demand for basic materials.

and dryers, because most households own a full set. Rather than supplying growing markets, the function of production in these areas is to replace consumer goods as they wear out.

Saturation of traditional markets is reinforced by another and relatively new phenomenon: consumer preferences that change with rising income. The affluent tend not to spend additional income on more of the same—yet another car, for example. Instead marginal income is often spent on items such as a video-cassette recorder, a personal computer and the accompanying software, membership in a health club, better health care or a service that provides stock-market information. Although such goods and services are disparate, they are characterized by a low materials content per consumer dollar. There are now no significant new markets for consumer goods having a high content of materials per dollar.

Our analysis of the four factors responsible for the declining consumption of basic materials suggests that these recent trends in the use of materials will not be reversed: they signal a historic shift in the economies of the industrial countries. It is clear that the transformation has already had a profound effect on the structure of U.S. industry. Indeed, it appears that declining demand rather than the threat of imports was responsible for the stagnation in many basic-materials industries in the U.S. during the past decade. In the 1970's the overall production of basic materials followed a declining trend very similar to that of overall consumption. The fact that the production trend followed the consumption trend so closely implies that although trade was important for certain materials, its overall role did not change much.

Since 1980, however, the pattern has in fact begun to change. For many basic industries the first stages of materials processing have begun to move to countries where production costs are lower. This trend will undoubtedly continue, with the result that production in the U.S. will decelerate even more quickly than demand. The response being made on the part of U.S. manufacturers is to cede primary processing operations to foreign plants while attempting to secure a continuing role in the later fabricating and finishing steps, in part through the development of new products. U.S. manufacturers will also expand production based on recycled materials. Both strategies are promising. The U.S. can draw on its diverse technological resource base to help bring new products to market. Moreover, the fraction of materials needs that can be met by recycling is in principle much greater when demand is not growing than it is in the early phase of a rapid growth of demand.



SATURATION OF MARKETS is among the factors that have contributed to a leveling off of demand for basic materials. The data in the illustration show why the market for nitrogenous fertilizer (which accounts for about 80 percent of all ammonia produced in the U.S.) is largely saturated. The black curve indicates the proportion of U.S. farmland on which corn is being grown that receives nitrogenous fertilizers. The colored curve indicates the amount of fertilizer applied to each hectare of such land. (A hectare is about two and a half acres.) Two factors suggest that demand for fertilizer will not grow much more in the years to come. Almost all corn land already receives fertilizer. Moreover, the benefits gained from applying more fertilizer diminish rapidly above the level reached in the past few years.

A close look at the situation of steel, aluminum, ethylene and paper producers suggests how the outlook can vary from industry to industry under the new market conditions. The steel industry has been affected by a leveling off of demand longer than any of the other three, and the stagnation of demand has had a stifling effect. Since 1950 only two all-new integrated steel mills have been built in the U.S., the last one in the 1960's. (Integrated mills carry out the entire production process, beginning with the processing of iron oxides.) Building new integrated mills is not economic because the capital charges exceed the reduction in operating costs achieved by building the new mill. The introduction of new components into existing mills has also proceeded slowly in the U.S.—far more slowly than in, say, Japan. New steel plants are being built in countries where production costs are not as high as they are in the U.S. The disparity stems from lower labor costs and in some cases from government intervention, which can reduce the cost of capital or even provide direct subsidies. Also many new foreign mills are situated at major ports and so have the advantage of low ore-transport costs.

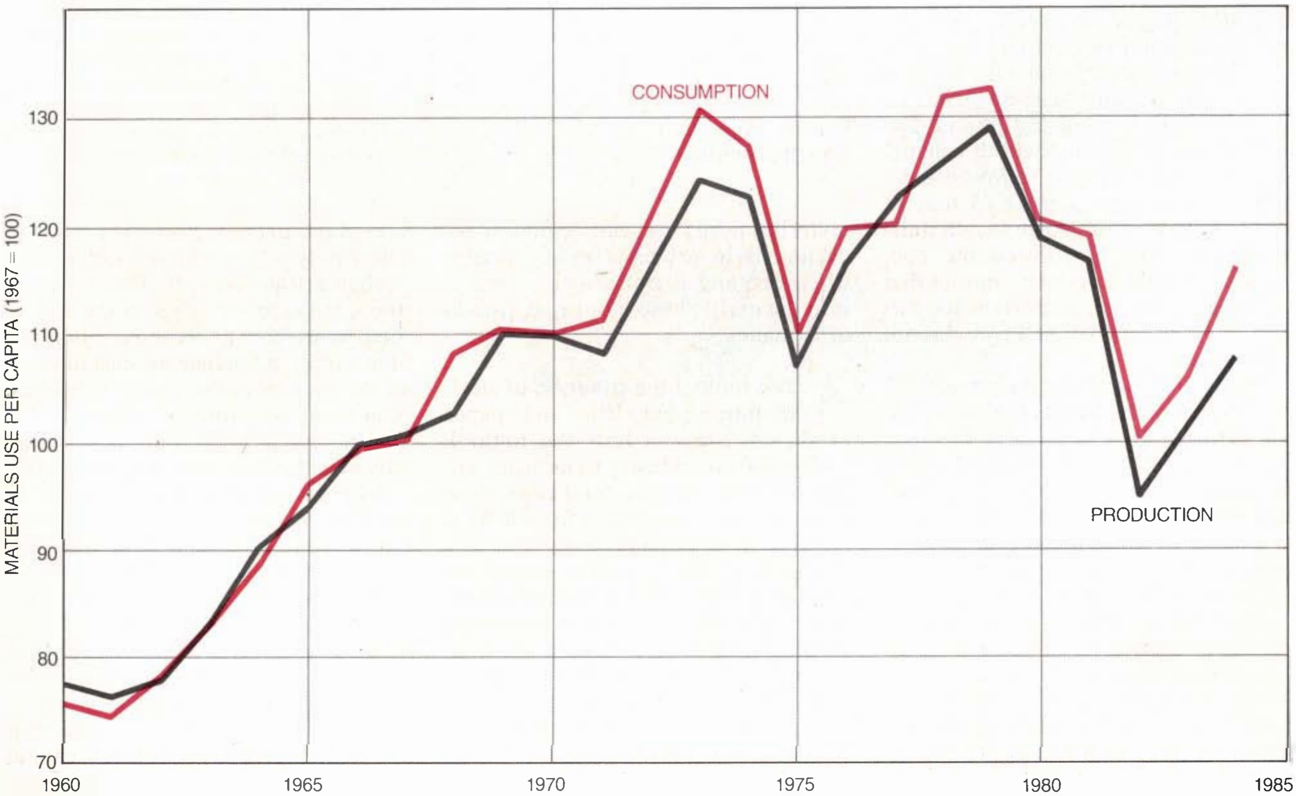
Whereas integrated steelmakers in the U.S. may be affected by a terminal illness, the secondary steel industry, made up largely of minimills (whose raw material is scrap), is in robust health [see "Steel Minimills," by Jack Robert Miller; SCIENTIFIC AMERICAN, May, 1984]. A key part of the successful strategy of the secondary industry has been to choose products that can be manufactured with low capital costs and labor requirements per unit of output. In the late 1970's and early 1980's the minimill sector was growing by about 10 percent per year, and in 1983 it accounted for 18 percent of all steel made in the U.S. Until now the minimills have by and large been limited to products that can be made from steel containing significant impurities (particularly copper). Technological advances may overcome that limitation, either by removing the impurities or by finding ways to improve properties of the steel (such as ductility) in spite of the impurities. Such advances would make it possible, for instance, to produce sheet steel for new cars from automobile scrap. If that were done, secondary producers might take a substantial part of the remaining market from integrated producers.

Aluminum is another industry in which primary capacity is diminishing while the secondary sector flourishes. As we noted above, a large fraction of the cost of making primary aluminum goes to pay for electricity. Indeed, each increase of one cent per kilowatt-hour in the price of electricity leads to about a 10 percent increase in the cost of producing aluminum from ore. The cheapest electric power available in the U.S. is found in the Pacific Northwest, where the price is about 2.7 cents per kilowatt-hour to the aluminum smelters; power from new coal or nuclear plants costs between five and six cents. Unexploited hydropower sites and natural gas and coal at remote sites are natural resources that make possible low-cost electricity generation. In Australia, Brazil and Canada new power plants that utilize such possibilities and that have been built or are under construction can provide electricity at less than two cents per kilowatt-hour. Hence expansion of primary capacity in the U.S. would not be economically competitive.

Making aluminum from scrap, however, consumes only from 5 to 10 percent as much energy as making it from ore. As a result the secondary sector is still capable of rapid growth: between 1970 and 1983 its share of total output increased from 20 to 33 percent. The recent expansion, which is based largely on the recycling of aluminum cans, is likely to continue.

The manufacturers of ethylene are in a somewhat different position than are the manufacturers of steel or aluminum, because the U.S. is still a net exporter of the chemical: in 1984 exports accounted for about 8 percent of total production. Yet the factors that are causing primary metals production to leave the U.S. are also affecting the makers of ethylene. Ethylene, a product of petroleum or of petroleum gas, is an important component of many intermediate chemicals (those that are further modified to yield the final product). From ethylene come polyethylene and other polymers, along with ethylene glycol and ethyl alcohol. Large new plants to make those substances as well as ethylene itself are under construction or have recently been completed abroad in areas having certain specific advantages. In the Far East the advantage is proximity to markets; in the Middle East, Canada and Mexico it is proximity to raw materials, such as ethane, that are hard to transport. Completion of these facilities will substantially reduce the U.S. share of world ethylene production. In the medium term exports of basic ethylene derivatives are likely to decline sharply; in the long term the realization that it is cheaper to make ethylene abroad will inhibit construction of new plants in the U.S. What is true for ethylene and its derivatives also applies to other basic organic chemicals.

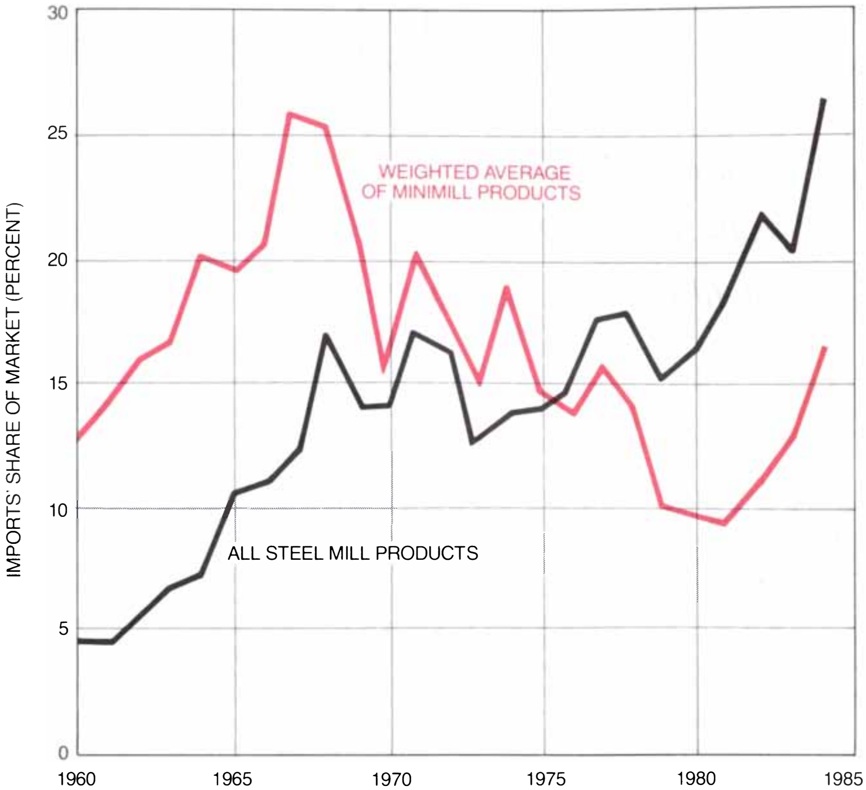
Among the basic materials paper is the one that provides the greatest solace to U.S. industry. For at least the time being the paper industry is assured a niche in the marketplace of the



PRODUCTION OF BASIC MATERIALS followed a trend very similar to the trend for consumption in the U.S. during the 1960's and 1970's. The black curve shows production per capita, the colored curve consumption per capita. The close resemblance between the curves before 1980 suggests that until then imports had little effect on the overall level of production. If imports had been a factor, production would have fallen relative to consumption. Hence it appears that before 1980 the leveling off of demand was the main cause of slow growth in the industries that process basic materials. In the current decade the level of production has fallen with respect to consumption, which suggests that since 1980 both imports and the leveling off of demand are constraining domestic production.

Information Era, owing in large part to rapid growth in the demand for office paper. (The "paperless office" remains a scheme of the future.) Although overall consumption of paper is declining in relation to G.N.P., it is not declining on a per capita basis. Moreover, the likelihood is that the international demand for pulp and paper will increase sharply in the next few years. The U.S. has a unique combination of forest resources, production facilities and technical skills, which may make it possible to export more paper goods to supply the needs of other countries. In most cases those needs are for two commodities: linerboard, the high-strength facing from which corrugated shipping containers are made, and high-quality pulp, which goes into a variety of products. Many countries that have their own paper mills still must import these materials, and so aggressive marketing could lead to considerable growth in U.S. exports.

The combination of mature domestic markets and increased foreign competition facing the basic-materials processing industries of the U.S. and other industrial countries has far-reaching implications for economic planning. Those who make economic policy in the industrial countries must recognize that reforms such as subsidizing the steel industry would not restore rapid growth, because they would have no effect on the underlying stagnation in demand. The materials industries cannot be sustained by protecting antiquated technology. Substantial innovation will be necessary to bring these industries into conformity with the present reality. In the past the rapid growth of demand was a spur to technological innovation, but that stimulus is gone. Perhaps external stimuli, such as tax and research-and-development policies that encourage innovation, will be needed if the materials industries are to adapt to changed circumstances.



RECYCLED MATERIALS may provide a way for manufacturers in industrial countries to lessen the effect of imported materials. The black curve represents the fraction of all steel products used in the U.S. that were made elsewhere. The colored curve shows the corresponding fraction of minimill products. (Minimills are steel mills that employ scrap as raw material.) Although imports are increasing in the steel industry as a whole, they are being driven out in areas where minimill products are concentrated. The authors contend that manufacturers in the industrial countries will have to cede much of the processing of primary materials to countries with advantages in labor or raw-materials costs. Such losses may be partially offset by growth in industries based on recycled materials, along with expansion of industries that entail extensive fabrication and finishing of diversified products.

The recent changes certainly call for increased effort on the part of economic planners. Yet in some respects they make planning easier. Specifically, the shift away from primary materials processing could stabilize or even reduce overall energy requirements for industry. Since the processing of basic materials consumes much more energy per dollar of output than fabrication and finishing activities do, even a small shift away from processing can have a profound effect on the energy consumption of industry (which in 1984 accounted for about two-fifths of all energy consumed in the U.S.). Our analysis suggests that aggregate materials production will remain roughly constant in the U.S. between 1984 and 2000 (when measured in kilograms of output weighted by the energy consumed in manufacturing each product). Since we expect industry to improve the efficiency of its energy use at a rate of from 1 to 2 percent per year during that period, the result may well be a decrease in industrial energy consumption, perhaps of as much as 20 percent, so that considerably less capital would be needed to provide energy for industry than is suggested by conventional estimates.

A decrease in energy requirements would also slow the upward trend in the world oil price that will inevitably occur in the decades ahead as the market economies of the world again become heavily dependent on oil from the Persian Gulf region, where the bulk of the earth's remaining oil resources lie. The economic benefit hence would be complemented by a reduction in the world tensions that arise from competition for increasingly scarce resources. In addition, declining demand for materials might bring with it better prospects for resolving environmental problems that tend to grow worse as the level of materials consumption increases.

The end of the Era of Materials is therefore by no means to be regarded only with distress. Like any other profound historical transformation, it brings with it benefits as well as heavy costs for those with an investment in the passing era. What is of crucial importance is to recognize that a fundamental and perhaps irreversible change is taking place. The industrial countries are witnessing the emergence of an information-centered society in which economic growth is dominated by high-technology products that have a relatively low materials content. In this society basic materials will continue to be used, and at very high rates compared with the rates in other societies. The critical economic fact is that their use will no longer be growing. In the years to come economic success and failure will be determined by the ability to adapt to this reality.

Fibronectins

These adhesive proteins act as biological organizers by holding cells in position and guiding their migration. Detailed analyses of fibronectins are revealing molecular bases for their functions

by Richard O. Hynes

Within the complex architecture of a multicellular organism most normal cells keep their places. They are anchored to basement membranes and connective tissue made up largely of a fibrous mesh of proteins and other substances. In the adult body a few cell types routinely move through this extracellular matrix, and during embryonic development and wound healing certain cells migrate extensively. The movement is highly organized, and most of the cells reach their destinations unerringly. How is this organization of cells, both fixed and dynamic, maintained? Part of the answer lies in a variety of large glycoproteins (proteins with attached sugars) that bind cells to the extracellular matrix.

The best-understood of these anchoring and organizing molecules are the versatile proteins known as fibronectins or, collectively, as fibronectin. The name, from the Latin roots *fibra* (fiber) and *nectere* (to bind or connect), reflects the most obvious roles of these proteins. They serve as cables and as connectors: fibronectin molecules can assemble into fibrils, bind to cells and link cells to other kinds of fibrils in the extracellular matrix. Fibronectin also bears an important but still poorly understood relation to the internal organization of cells, and its adhesive character makes it a crucial component of blood clots and of pathways followed by migrating cells.

Over the past decade the study of fibronectin has advanced from functional description to detailed structural understanding. The molecule was dissected first into discrete protein domains, and they in turn were found to be built of smaller protein modules. At each level, structure could be correlated with properties, leading to a molecular map of function. The molecular studies have clarified the biological roles of this group of proteins.

Paradoxically, the first clues to the importance of fibronectin in organized cell behavior came from studies of cancer, which is characterized by disorder. In 1973, while I was working at the Imperial Cancer Research Fund in London, I compared the surfaces of tumor cells with normal cells by attaching a radioactive marker to the cell-surface proteins and separating them by electrophoresis, which arrays proteins in a gel according to their molecular weight. The radioactive marker made it possible to distinguish individual surface proteins; a comparison of the gels made from normal cells with those from cancer cells showed that one protein with a high molecular weight that was abundant on the surface of normal cells was absent from the surface of cancerous ones. At almost the same time Carl G. Gahmberg and Sen-itiroh Hakomori made the same observation at the University of Washington, and it has been confirmed by many other investigators since then. The missing protein is now known as cell-surface fibronectin.

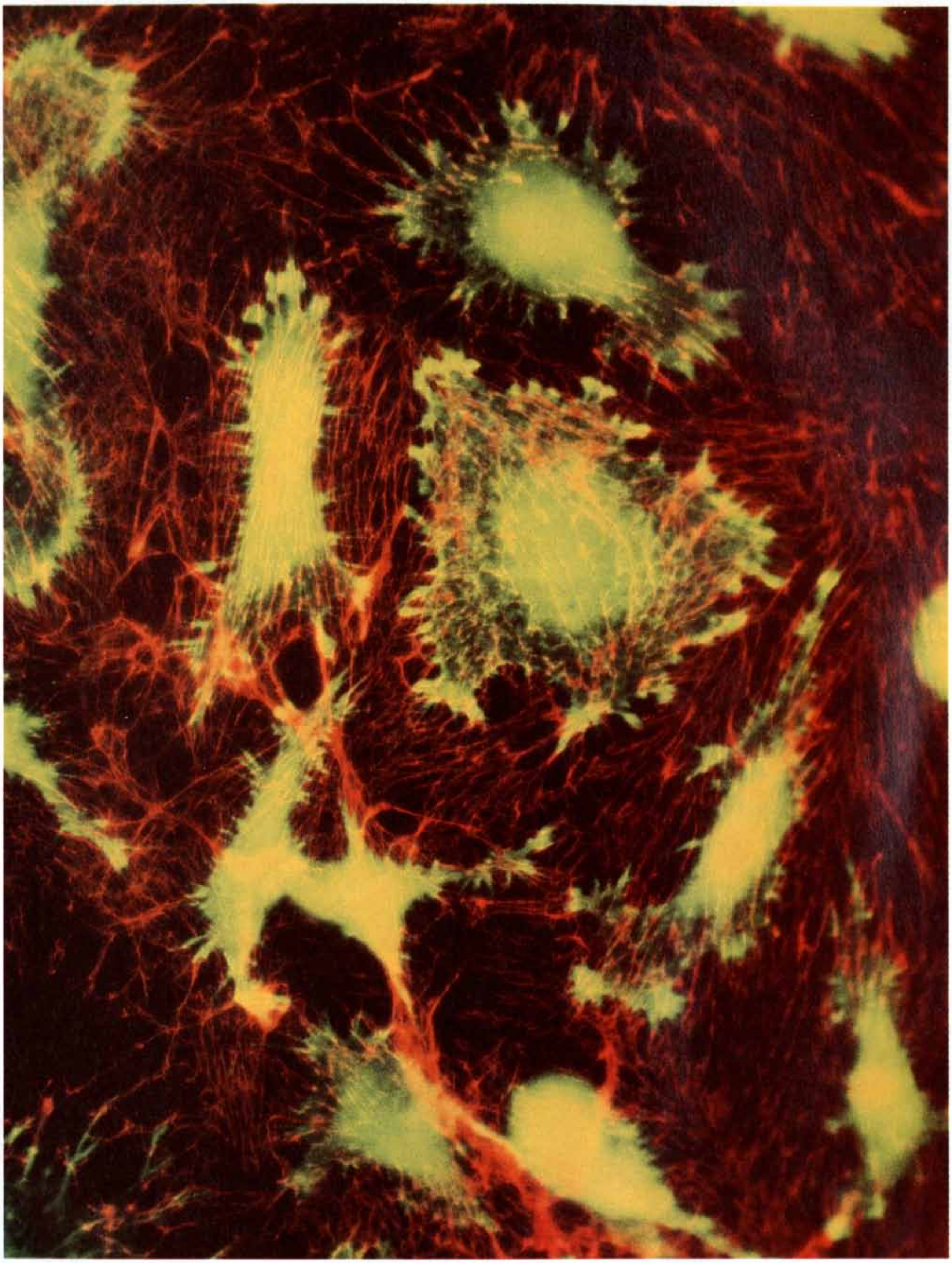
Another line of research added significance to the finding. Cells that are isolated from the body and cultured in a petri dish generally require a supplement of serum derived from blood plasma. The serum contains growth factors that stimulate cell division, and it also helps the cells to attach themselves to the culture dish. Frederick L. Grinnell of the University of Texas Health Science Center at Dallas and Robert J. Klebe of the University of Texas Health Science Center at San Antonio traced the adhesion-promoting effect of the serum to specific proteins. Erkki Ruoslahti and Antti Vaheri of the University of Helsinki showed that one of the proteins, now known as plasma fibronectin, is biochemically similar to cell-surface fibronectin. It seemed likely that the cell-surface protein also plays a role in normal cell adhesion, and its absence might account for the fact that cancer cells generally adhere very poorly to their substrate.

Kenneth M. Yamada and Ira Pastan and their colleagues at the National Cancer Institute and Iqbal Ali, working with me at the Massachusetts Institute of Technology, confirmed the importance of fibronectin in cell adhesion. We added fibronectin purified from normal cells to cultures of tumor cells and observed a dramatic improvement in cell adhesion. We noted a concomitant change in cell shape: in attaching themselves to the glass or plastic substrate the tumor cells, which ordinarily are rounded, became flattened against the surface.

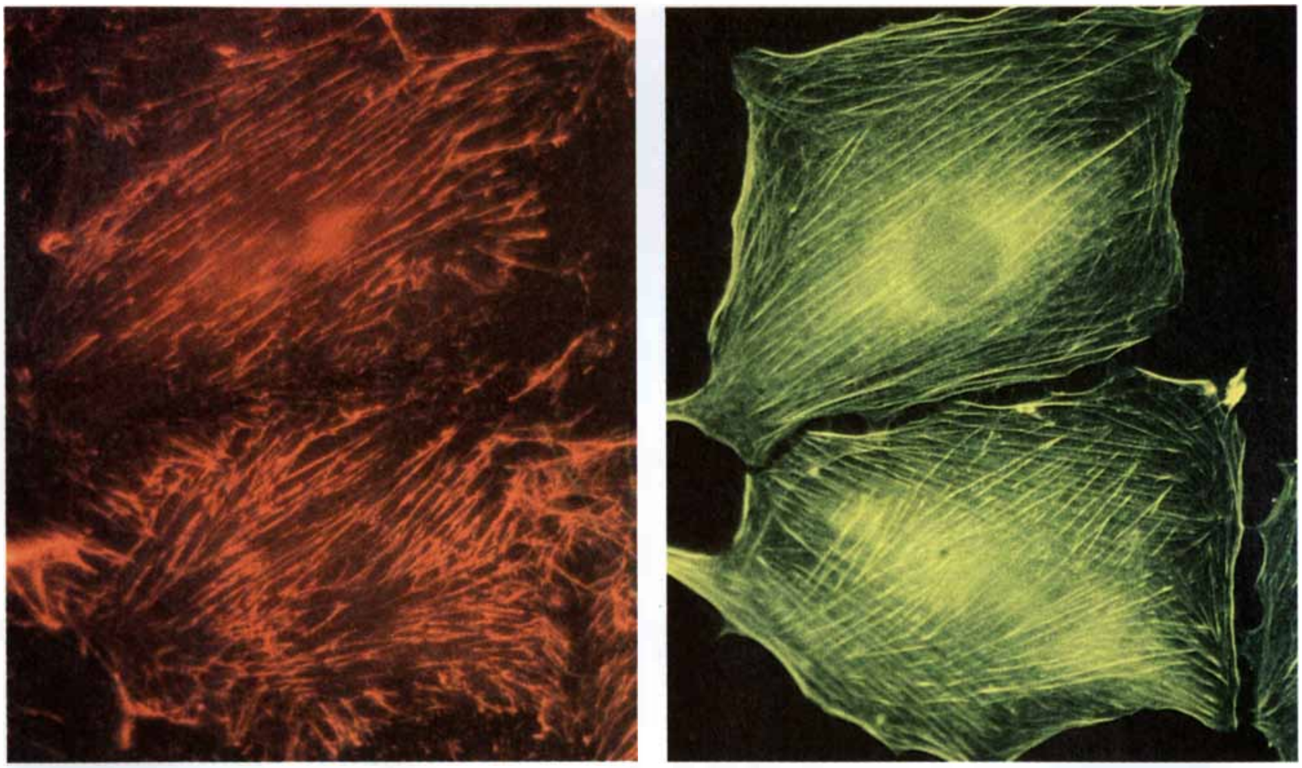
We also found that the tumor cells treated with fibronectin underwent a change in internal structure. In normal tissue the cytoskeleton, an internal lattice of structural and contractile proteins, gives form to the cells. In tumor cells the cytoskeleton is disorganized. The addition of fibronectin to the tumor cells restored a major constituent of the cytoskeleton, the web of filaments made of the protein actin, to its normal highly organized configuration. (It is important to note that the fibronectin did not restore tumor cells to normal in every respect; in particular, the treated cells went on multiplying too much.)

Prompted by that finding, Antonia Destree, Denisa Wagner and I investigated the structural relation between extracellular fibronectin and intracellular actin in normal cells. Soon after the discovery of fibronectin it had been observed that the protein polymerizes around cells, forming long fibrils a few nanometers (billionths of a meter) in diameter that make contact with the cell membrane; they can be stained with an antibody that binds specifically to fibronectin and is labeled with a fluorescent marker.

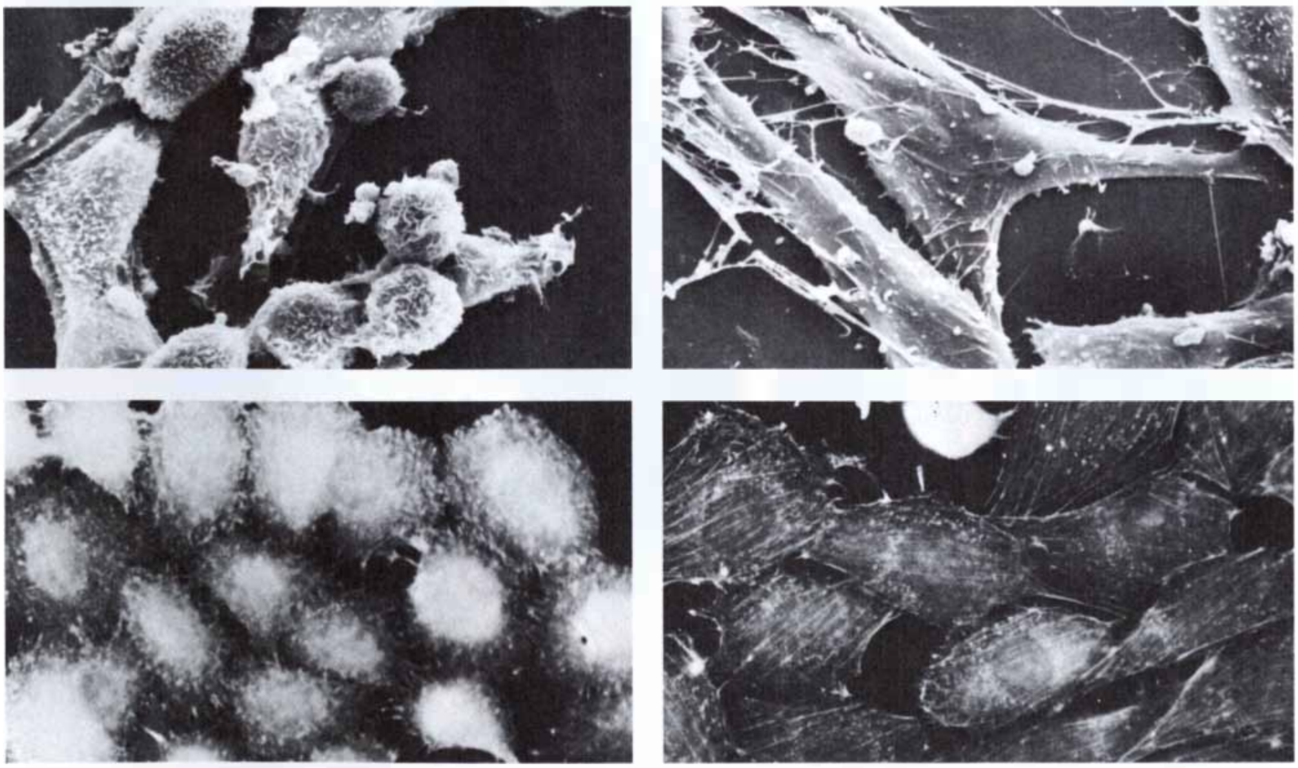
By staining the actin inside the cell and the fibronectin on the cell exterior



ADHESIVE WEB OF FIBRONECTIN (*red*) **underlies fibroblasts, or connective-tissue cells (*green*), in culture. The fibroblasts themselves laid down the web and adhered to it. Not long before the photograph was made an agent that blocks adhesion was added to the medium. The cells remain attached through needlelike processes but will eventually be released. The fibroblasts were stained with an antibody that placed a fluorescent marker on actin, a protein in the cells; another antibody attached a different marker to fibronectin.**



ALIGNED MATRICES of fibronectin (*left*) and actin (*right*) are evident in cells stained with fluorescently labeled antibodies specific for each protein. The photographs were made under different light to excite the stains separately. They distinguish the fibronectin fibrils, which bind the cells to their substrate, from the intracellular web of actin. The alignment of the external and internal matrices reflects the presence of other proteins that have been found to extend through the cell membrane, linking the two sets of fibrils.

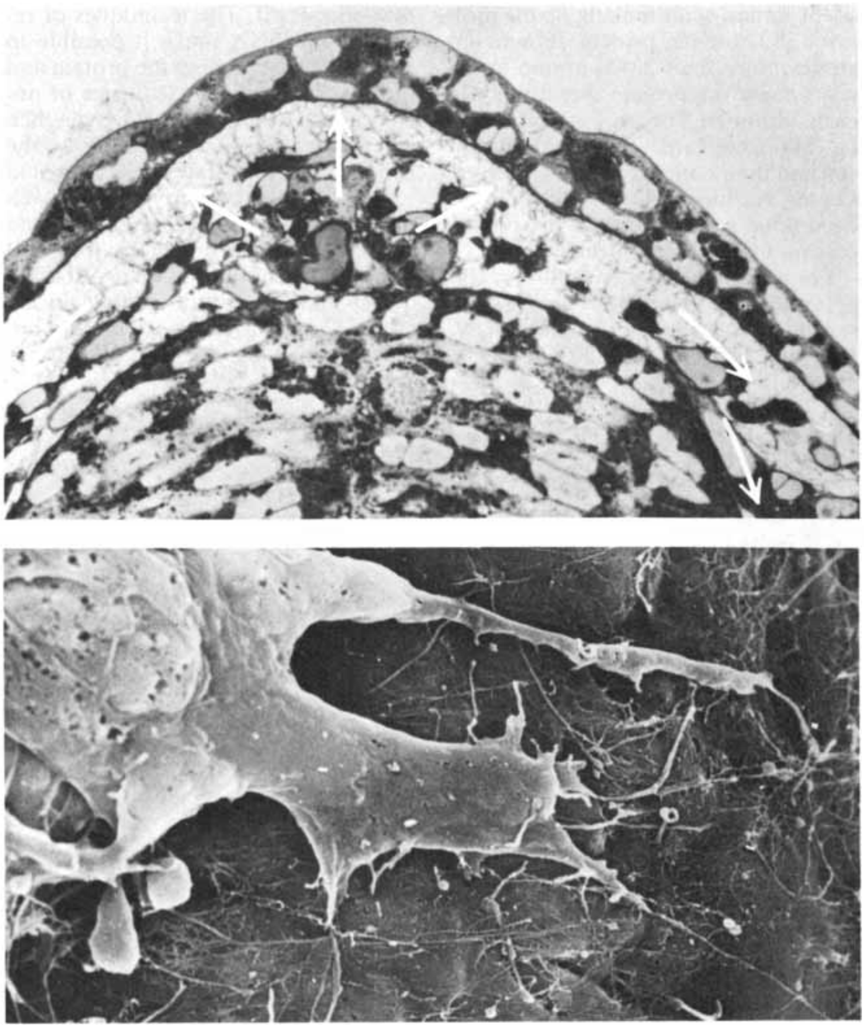


CELL ADHESION AND ORGANIZATION are improved when fibronectin is added to tumor cells, which generally lack the protein. Untreated tumor cells adhere poorly to their substrate and are rounded in shape (*top left*). A fluorescently labeled antibody stains the disorganized internal actin matrix typical of such cells (*bottom left*). The addition of fibronectin from cultures of normal cells causes tumor cells to adhere and flatten against the substrate (*top right*). The actin network in treated tumor cells regains its normal organization (*bottom right*). Kenneth Yamada of the National Cancer Institute provided the scanning electron micrographs at the top.

with different fluorescent markers and examining the cells under the microscope, my colleagues and I were able to show that individual fibrils running inside and outside the cell membrane coincide. Irwin I. Singer at the Institute for Medical Research in Bennington confirmed the finding at higher magnifications. It was clear that a physical connection must span the cell membrane between the cytoskeletal actin filaments and the extracellular fibronectin, linking the matrices.

The tumor-cell model continued to yield clues to fibronectin's role in normal cells. Ali found that the addition of fibronectin to tumor cells in culture stimulated them to move across the surface of the culture dish. In an effort to determine whether fibronectin also plays a role in cell migration in vivo, David R. Critchley and Marjorie A. England of the University of Leicester in England, Byrne W. Mayer and Elizabeth D. Hay of the Harvard Medical School and I studied early chicken embryos. By applying fluorescently labeled antibodies to the tissue we detected fibronectin in areas of extensive cell migration. Janet E. M. Heasman and Christopher C. Wylie of St. George's Hospital Medical School in London and I found the same relation between fibronectin and cell migration in amphibian embryos.

In an extensive series of investigations Jean Paul Thiery and his colleagues at the Institute of Embryology in France demonstrated that neural-crest cells, which are the precursors of the peripheral nervous system and other tissues, migrate through the embryo in spaces that are filled with fibronectin-rich fibrils. Thiery's group studied the same cells in vitro and found they migrate more extensively when they are exposed to fibronectin. Those results and others suggest that fibronectin-rich pathways guide and promote the migration of many kinds of cells during embryonic development.



FIBRONECTIN LINES MIGRATION ROUTES of neural-crest cells in the embryo of an axolotl, a kind of amphibian. The section (*top*) cuts across the neural tube, a hollow structure that is the precursor of the brain and spinal cord; above it are the cells of the neural crest, which migrate outward, around the neural tube, to form the peripheral nervous system and certain other tissues. The cells travel (*arrows*) through a fibronectin-rich protein matrix, light in color in this micrograph. A scanning electron micrograph (*bottom*) shows a neural-crest cell moving across the neural tube; its processes adhere to fibronectin fibrils, presumably to gain traction. Jan Löfberg of the University of Uppsala provided the images.

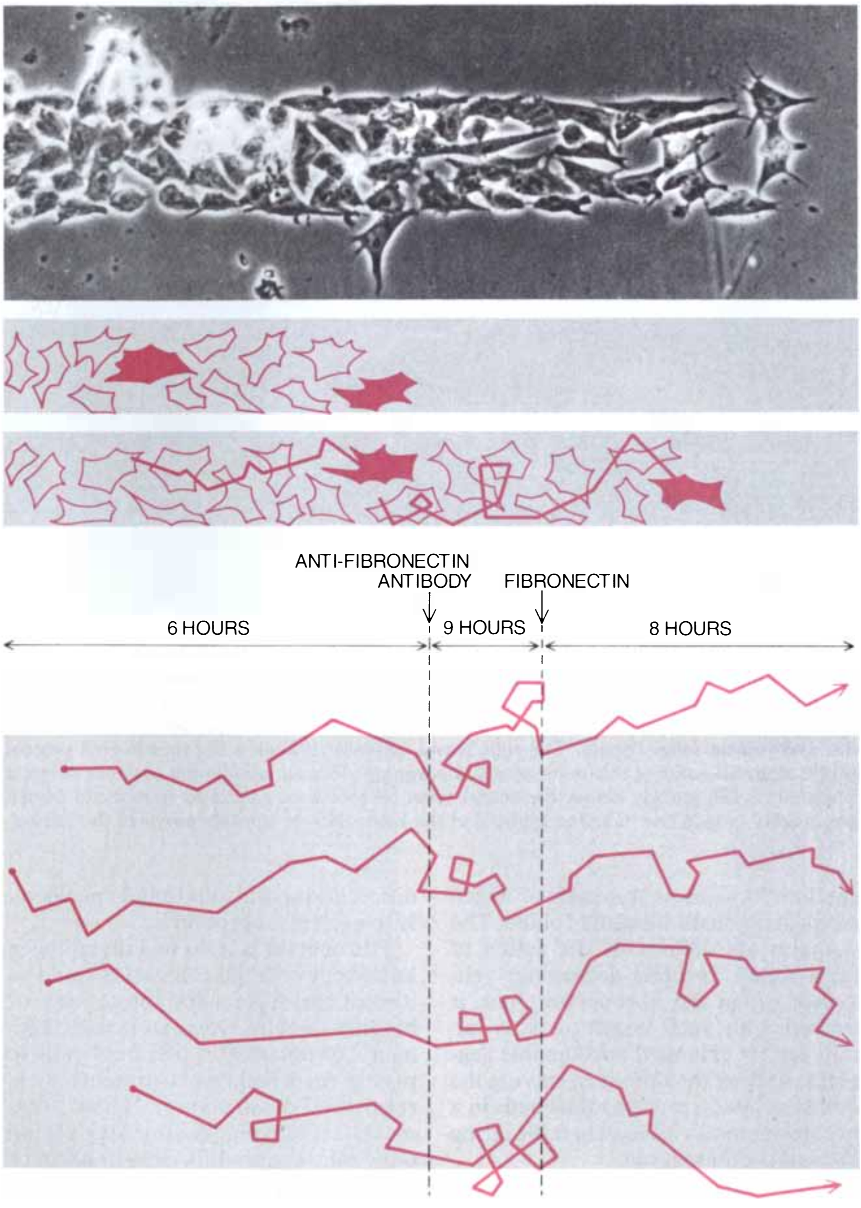
What is the molecular basis for fibronectin's varied roles? A great deal of biochemical work done in laboratories throughout the world has led to a model of the fibronectin molecule in which the protein's binding functions and its structure are clearly correlated. The molecule is a dimer: it consists of two similar subunits. Each subunit has a molecular weight of about 250,000 daltons (one dalton is the mass of a hydrogen atom), and they are joined at one end by disulfide bonds. The protein chain of each subunit forms an elongated structure 60 to 70 nanometers long and two to three nanometers thick; that structure in turn is subdivided into a series of smaller domains, within each of which the protein chain is tightly folded. The domains are defined by the action of proteolytic (protein-degrading) enzymes: when the fibronectin chain is treated with such enzymes, it is cut only in the extended and flexible segments joining the domains, leaving the domains intact, much as the beads in a necklace remain intact when the string connecting them is cut.

Each domain, it appears, is responsible for one of fibronectin's binding functions. For example, earlier work had shown that fibronectin binds to collagen and fibrin, two other glycoproteins that are important constituents of extracellular matrices. By degrading fibronectin with proteolytic enzymes and exposing it to matrices of collagen and fibrin, one can identify different domains that bind specifically to each glycoprotein.

Fibronectin is able to link collagen and fibrin with the cell surface, and so it must also have a domain capable of binding to cells. One can isolate fragments of fibronectin that bind cells to plastic (to which most proteins stick, regardless of structure). These fragments contain the cell-binding region only: they cannot link cells to fibrin or collagen. Several other binding functions have also been mapped onto the fibronectin molecule, yielding a portrait of the protein as a versatile connector, able to link cells and extracellular proteins in complex structures through its various binding domains.

To examine the structure of fibronectin at a higher resolution it is necessary to determine the linear sequence of the amino acids making up the molecule. Since each protein subunit includes more than 2,000 amino acids, sequencing the protein directly is not easy, although Torben Petersen, Karna Skorstengaard, Staffan Magnusson and their colleagues at the University of Aarhus in Denmark have indeed done so for much of the bovine plasma-fibronectin molecule.

For a protein as large as fibronectin a faster way to find the amino acid sequence is to decipher it indirectly, from the DNA sequence of the gene that encodes it. The techniques of recombinant DNA make it possible to isolate DNA encoding the protein and quickly determine its sequence of nucleotides. The DNA sequence can then be reverse-translated according to the genetic code to deduce the amino acid sequence of the protein. The approach has enabled Jean Schwarzbauer and John Tamkun in my laboratory and Alberto Kornblihtt, Karen Vibe-Pedersen and Tito Baralle of the University of Oxford to sequence rat and human fibronectins. The data for bovine, human and rat fibronectins are in good agreement and lead to a picture of the molecule in which structure and function can be correlated on a finer scale.



CELL MIGRATION IN CULTURE is guided and promoted by fibronectin. A stripe of the protein on the surface of a culture dish defines a route for migrating neural-crest cells (*top*). Diagrams (*middle*) show the locations of specific cells early in their migration and after an interval; their paths in the interim are plotted. In another experiment (*bottom*) neural-crest cells were first allowed to travel along stripes of fibronectin for six hours. Antibody to fibronectin was then added to the culture medium, inactivating the protein. The movement of the cells was slowed and became less directed. After nine hours the antibody was washed off and the fibronectin was replenished. The cells then resumed their progress. Jean Paul Thiery and his colleagues at the Institute of Embryology in France did the experiments.

The amino acid data show that the fibronectin molecule is made up mostly of short amino acid sequences repeated many times. The sequences are of three general types; the members of each set are similar but not identical. Each of the structural and functional domains of fibronectin contains one or more of the repeats. For example, there are two fibrin-binding domains on each subunit of the protein; each consists of the sequence known as type I, repeated three times in one domain and five times in the other. The repeats are themselves functional: individual type I sequences are probably capable of binding to fibrin, for instance, and a single type III repeat serves as the functional core of the cell-binding domain.

Tamkun and Erich Odermatt have shown that each of these small, functional modules is encoded in the fibronectin gene as a unit, usually a single exon: a block of DNA that is expressed as protein and is set off from other exons in the same gene by noncoding stretches of DNA known as introns. For example, the three-unit fibrin-binding domain is encoded in three exons, one for each type I repeat, and the same relation probably holds for the five-unit domain. The one-to-one correspondence of exons to protein modules probably also characterizes the type II repeats, but most of the type III repeats are encoded by a pair of adjacent exons, which seem to specify distinct substructures within the protein module.

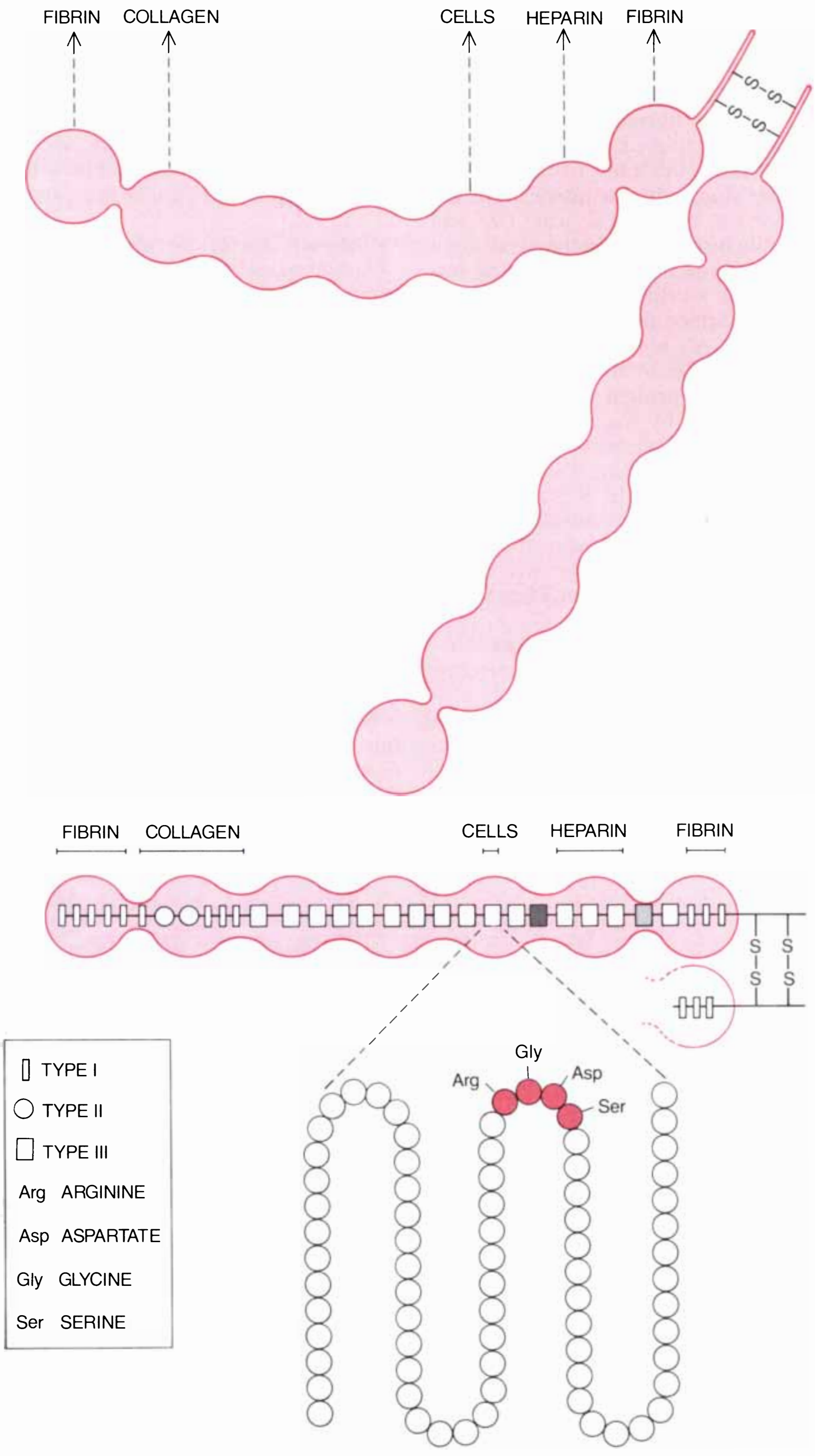
The structure of the gene for fibronectin thus supports the general notion, first proposed by Walter Gilbert of Harvard, that exons correspond to functional modules within proteins. It also suggests that the gene evolved from three primordial minigenes, each encoding a protein resembling one of the three kinds of protein modules seen in fibronectin. In the course of evolution each minigene became duplicated once or many times; the individual copies were assembled in series to make the present-day gene, which contains about 50 exons, as Hideyasu Hirano, Yamada and their colleagues at the National Cancer Institute first showed. At the same time the copies diverged from one another, although they retain common features that testify to their relatedness.

The degree of divergence between protein repeats encoded by these genetic units yields clues to the age of the fibronectin gene. A given repeat shows many more similarities to the corresponding repeat in the fibronectin of a different mammal than it does to other repeats of the same type in the same fibronectin molecule. The 90-amino-acid type III repeat forming the core of the cell-binding domain, for example, is identical in more than 90 percent of its sequence in bovine, human and rat fibronectin, whereas in the fibronectin from a single species the similarity among the various type III repeats is only 30 percent.

It appears that the minigenes were duplicated and began diverging long before the ancestors of the mammals started down their separate evolutionary paths. Indeed, it is probable that the fibronectin gene arose from the primordial minigenes early in the evolution of multicellular organisms; one can argue that the advent of such organisms required the evolution of adhesive proteins such as fibronectin, capable of holding cells together. Preliminary evidence suggests fibronectin is present in even the simplest present-day multicellular organisms: sponges, sea urchins and other invertebrates.

If the fibronectin gene was assembled from a set of functionally independent building blocks, might the individual building blocks have been incorporated into other genes as well? Gilbert has proposed that exons can be shuffled around in the course of evolution, with the result that they can appear in otherwise unrelated genes. Comparisons of fibronectin with the amino acid sequence of another protein yield good evidence of exon shuffling. As Laszlo Patthy of the Enzymology Institute in Budapest first noted, tissue plasminogen activator, an enzyme that binds to fibrin and breaks it down, contains a segment that is homologous with the type I repeats of fibronectin. Like the type I repeats, the homologous protein segment is responsible for its host molecule's affinity for fibrin, and it too is encoded by a single exon.

Although the exons making up the fibronectin gene correspond directly to functional units in protein, not all of them are expressed in every fibronectin molecule. The gene itself does not vary, but studies of the messenger RNA's that direct the assembly of fibronectin molecules, done in my laboratory and at Oxford, have shown it can give rise to at least 12 versions of the protein. Different cell types synthesize different sets of variants. Liver cells, for example, make only plasma fibronectin. It has become clear that the variation reflects variable patterns of RNA splicing: the process by which introns are removed from the primary



FIBRONECTIN MOLECULE consists of two similar subunits (*top*) joined by a pair of disulfide bonds (*S–S*). The action of proteolytic (protein-degrading) enzymes reveals that each subunit is subdivided into domains within which the protein chain is tightly folded, and hence resistant to degradation. Individual domains (shown schematically as bulges) account for fibronectin's ability to bind to various proteins and surface membranes. The amino acid sequence of fibronectin reveals structure within the domains: small, repeated protein modules whose similarities in sequence allow them to be classified into three types (*bottom*). The modular construction of one of the fibronectin subunits is shown, with the three types of repeats represented as rectangles, circles and squares. An apparently unrelated module is indicated in light gray; a type III repeat that is missing in some variants of the fibronectin molecule is shown in dark gray. The repeats vary somewhat; the amino acid sequence arginine-glycine-aspartate-serine distinguishes the type III repeat of the cell-binding domain from other type III repeats. The sequence is crucial to fibronectin's ability to bind to cells.

RNA transcript, a direct copy of the gene, and exons are brought together to form messenger RNA, which is then translated into protein.

When the fibronectin gene's RNA transcript is spliced, most of the exons are retained, but a few of them can be cut out along with the introns. The deletions can occur in at least two and possibly more places in the fibronectin gene. Schwarzbauer, Tamkun and Jeremy Paul, working in my laboratory, have confirmed that the resulting messenger RNA's are translated into proteins that differ by the presence or absence of the protein structural modules encoded by the variably spliced exons. Depending on the outcome of the alternative splicing, the fibronectin subunits (half of the dimeric protein) can vary from 2,145 to 2,445 amino acids in length; between those extremes various combinations of included and omitted protein modules are possible.

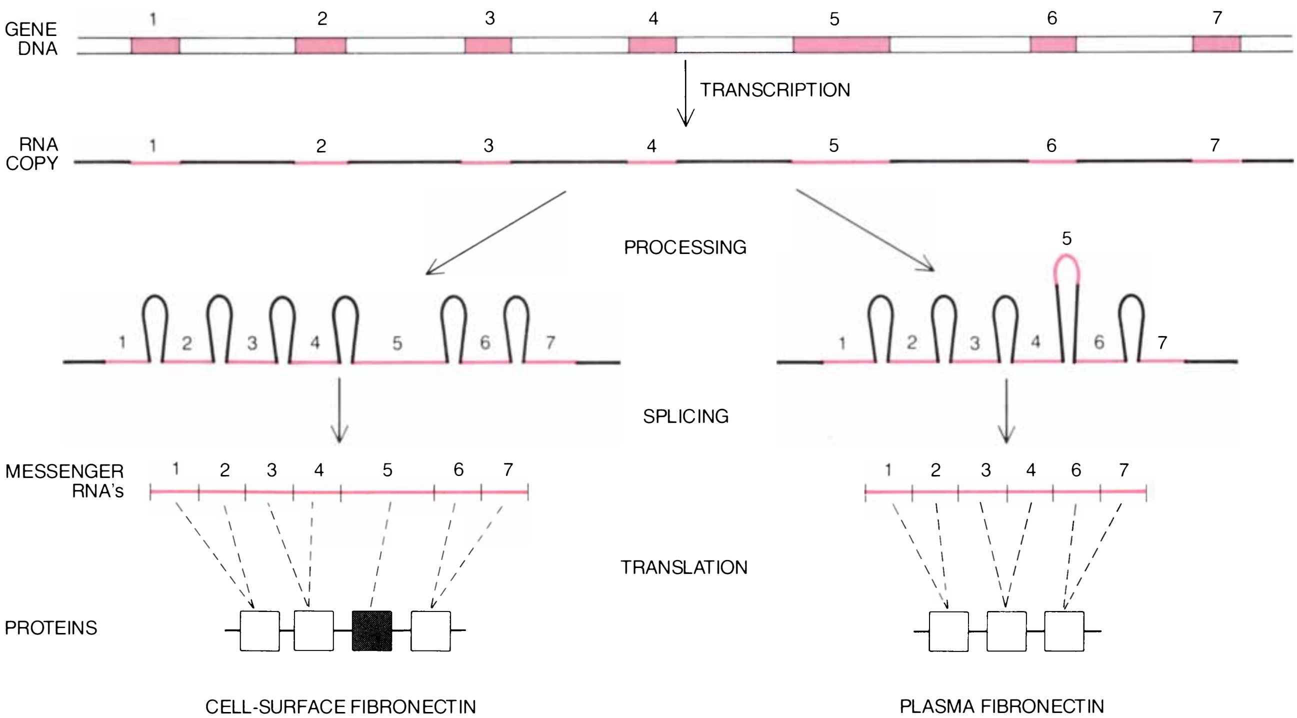
Alternative splicing may have functional consequences. The fibronectins synthesized by liver cells and by fibroblasts (connective-tissue cells) are a case in point. The plasma fibronectin made in the liver always lacks a specific type III repeat, and it circulates freely in the blood; that repeat is included in the fibronectin made by fibroblasts—a form of the protein that readily polymerizes into insoluble fibrillar matrices. The precise relation between alternative splicing and such functional variations is not understood, however, nor is it clear how the pattern of splicing might be regulated.

The new molecular understanding of fibronectin has enabled investigators to probe the details of its interactions with other proteins. So far fibronectin's interaction with the proteins of the cell surface is the best-understood. The smallest fragment of fibronectin yet isolated that binds to cells, thereby enabling them to stick to glass or plastic, is 108 amino acids long. Michael D. Pierschbacher and Ruoslahti and their colleagues at the La Jolla Cancer Research Foundation, who identified the fragment, synthesized a series of peptides (sequences of several amino acids) corresponding to parts of it. They then sought the shortest synthetic peptide that would enable cells to adhere.

Remarkably, they found that a peptide four amino acids long could promote cell adhesion under some conditions. Made up of the amino acids arginine, glycine, aspartate and serine, in that order, the peptide did not stick to plastic or glass on its own; the surface had to be chemically modified or covered with another protein to which the peptide could be coupled. Under those circumstances the synthetic peptide, or longer peptides containing the same sequence, caused cell adhesion.

When the peptide was added in soluble form to a dish containing cells and a fibronectin-coated surface, it tended to prevent cells from binding to the surface, indicating that it competes with intact fibronectin for binding sites on the cells. The fact that the short peptide must be added at a fairly high concentration to compete effectively shows it has a lower affinity for the cell surface than has the complete fibronectin molecule. Clearly the sequence of four amino acids is only a part of the cell-binding domain, but it is an important one, defining an active site for interaction with cells.

Pierschbacher and Ruoslahti, and also Yamada and his colleagues, determined that the initial sequence of three amino acids, arginine-glycine-aspartate, cannot be altered even in minor ways without inactivating the synthetic peptide. The peptide retains its activity, however, when one of several other amino acids is substituted for



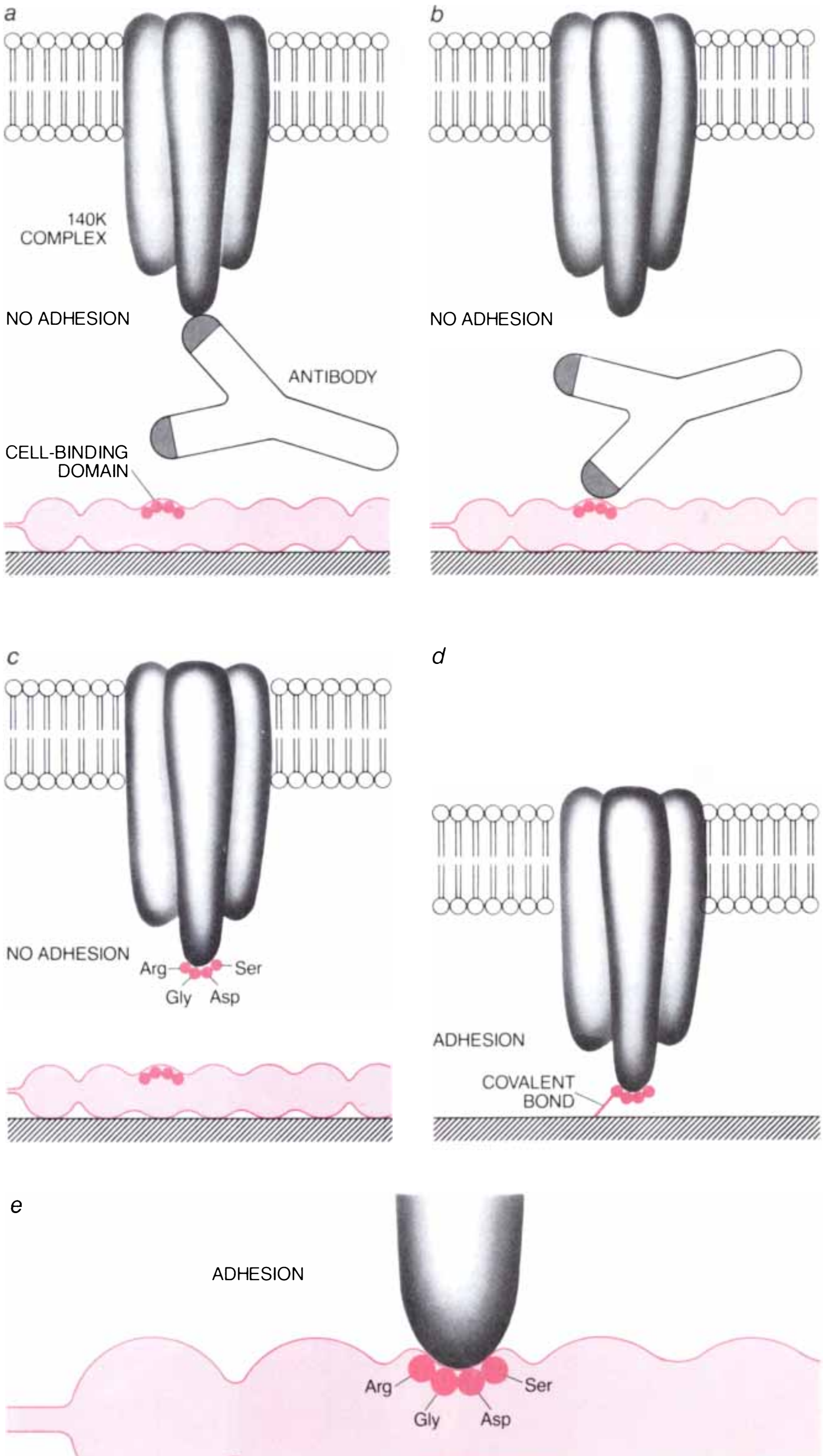
ALTERNATIVE SPLICING enables the single fibronectin gene to be expressed as multiple forms of the protein. Two examples are cell-surface fibronectin, which readily polymerizes into fibrils, and plasma fibronectin, which circulates in the blood as a soluble molecule and lacks a type III protein repeat that is present in the other form. The repeat is encoded in a single large exon (*5*); it is flanked by other type III repeats, each of which is specified by two smaller exons. All the exons are transcribed in the initial RNA copy of the gene (a part of which is shown at the top). In fibroblasts and other cells only the introns, the noncoding segments between the exons, are excised from the RNA as it is processed; the exons are spliced to make messenger RNA, which is translated into cell-surface fibronectin (*left*). In the cells of the liver the large exon is cut out of the copy with the introns. The result is plasma fibronectin (*right*).

serine in the fourth position. It may be significant that one of the segments of fibronectin that vary because of alternative splicing contains the sequence arginine-glycine-aspartate-valine. The segment may therefore serve as a second cell-binding site, and its presence or absence may bear some relation to functional variations in fibronectin.

Fibronectin not only binds to the cell surface but also interacts somehow with the actin filaments in the cell's interior. The interaction must be mediated by a structure in the cell membrane, probably a protein or proteins, that spans the membrane between the fibronectin fibrils outside the cell and the internal actin cytoskeleton. How can one identify the molecules in question? Presumably the same protein complex that makes the transmembrane connection also interacts with fibronectin during cell adhesion. Hence molecules that interfere with the adhesion of cells to a substrate should serve as probes for identifying the crucial membrane proteins.

Clayton A. Buck, Caroline Damsky, Karen Knudsen and Rick Horwitz of the Wistar Institute of Anatomy and Biology and the University of Pennsylvania and Jeffrey Greve and David I. Gottlieb of Washington University in St. Louis developed such probes by preparing antibodies to a range of cell-surface proteins. They tested individual antibodies for the ability to block the adhesion of cells to fibronectin-covered surfaces. Both groups found adhesion is blocked by antibodies that recognize and bind to a complex of three cell-surface glycoproteins, each of which has a molecular weight of about 140,000 daltons; together they are known as the 140K complex. The Pennsylvania group and Wen-Tien Chen of Georgetown University, who worked with Yamada, made use of fluorescent antibodies to show that the 140K complex occurs at places where both the fibronectin fibrils and the actin filaments meet the cell membrane.

It therefore seemed probable that the 140K complex participates in the transmembrane connection. With the candidate proteins identified, the next step was to seek more direct evidence of binding between the complex and fibronectin. Robert Pytela, Pierschbacher and Ruoslahti exposed heterogeneous mixtures of proteins, which had been extracted from cells, to a matrix of fibronectin and showed that the fibronectin in effect picked out the 140K complex from the mixtures. Horwitz, Yamada and their colleagues proceeded differently, purifying the 140,000-dalton proteins recognized by their adhesion-blocking antibodies



EXPERIMENTAL STRATEGIES for manipulating cell adhesion reveal the underlying molecular interactions. Antibody to the so-called 140K complex, a cluster of three glycoproteins in the cell membrane, prevents the cell from sticking to a fibronectin-coated surface (*a*), which suggests the complex is crucial to cell adhesion. Antibody to the cell-binding domain of fibronectin also blocks adhesion (*b*), confirming that the 140K complex interacts with that domain. A synthetic peptide four amino acids long that simulates a peptide sequence in the cell-binding domain also inhibits adhesion, probably by competing with the full domain for the 140K binding site on the cell membrane (*c*). The sequence mimicked by the synthetic peptide must define an active site in the cell-binding domain; this supposition is confirmed by the fact that the peptide alone can promote adhesion when it is coupled to a surface (*d*). Hence the picture has emerged of an adhesive interaction between the 140K complex and the sequence of four amino acids in the cell-binding domain of fibronectin (*e*).

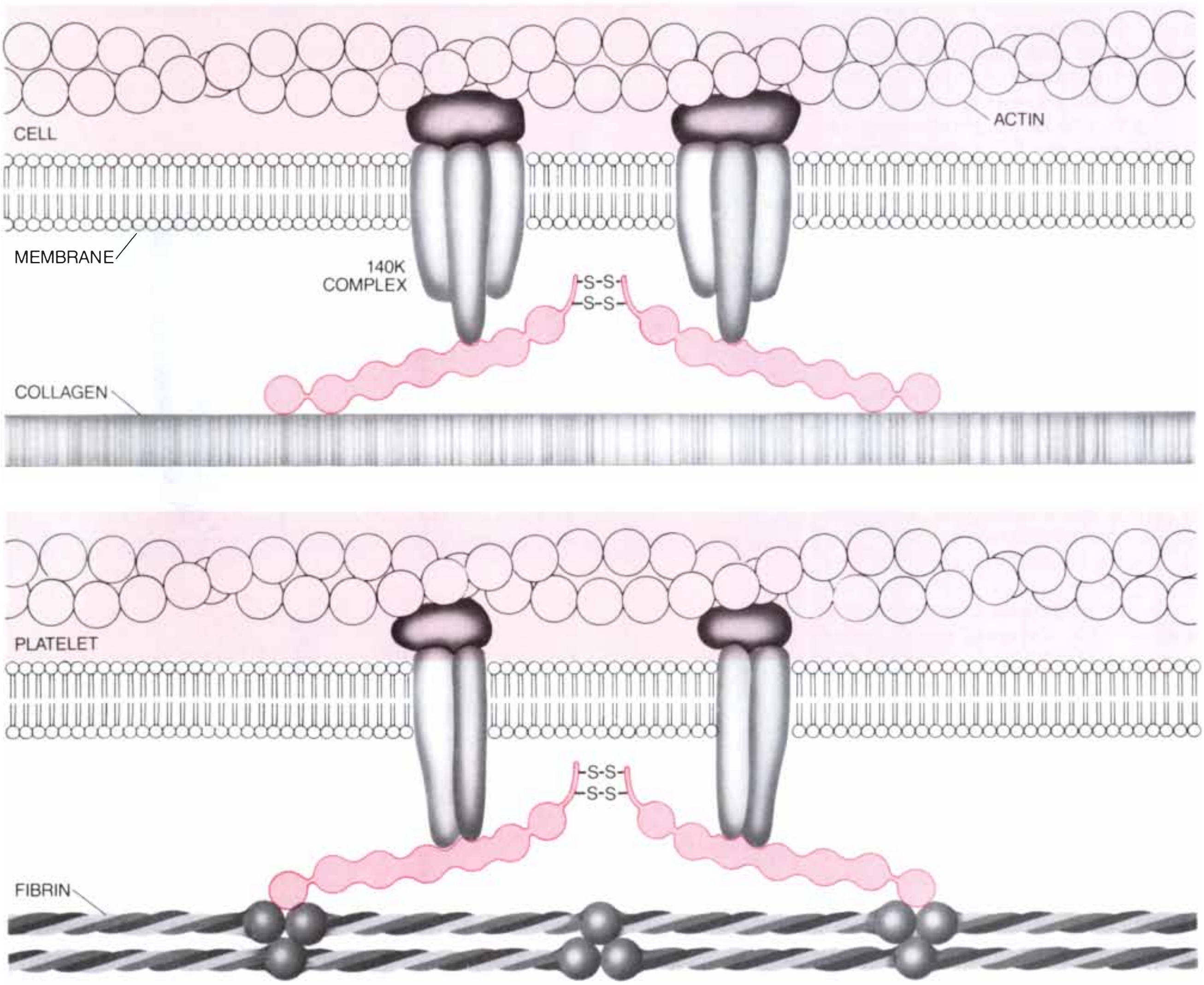
and then demonstrating that the proteins bound to fibronectin.

In both sets of experiments, peptides containing the arginine-glycine-aspartate sequence inhibited fibronectin binding, which indicates that the 140K complex not only binds to fibronectin but also does so specifically at the cell-binding domain. Taken together, the evidence is strong for the involvement of the 140K complex in the transmembrane connection between fibronectin and actin. A more precise understanding of how the complex functions awaits a detailed structural analysis; in my laboratory we are now determining the amino acid sequence of the 140K proteins by recombinant-DNA methods.

What light does this detailed molecular information shed on fibronectin's part in the dynamic aspect of multicellular organization—on cell migration? To confirm that fibronectin interacts with migrating cells through its cell-binding domain, Thiery and his colleagues Jean-Claude Boucaut and Thierry Darribère have injected antibody to the domain into early amphibian and chicken embryos; in separate tests they also examined the effect of the four-amino-acid active-site peptide, which competes with the domain. Both of the reagents inhibited the movement of cells through the developing embryos.

Presumably the same complex on the cell surface that binds to fibronectin during cell adhesion is also critical in migration. Marianne Bronner-Fraser of the University of California at Irvine has injected antibody to the 140K complex into early chicken embryos and found that it inhibits cell migration, whereas irrelevant antibodies do not. Hence the molecular interactions underlying cell adhesion also seem to be operating in embryonic cell migration.

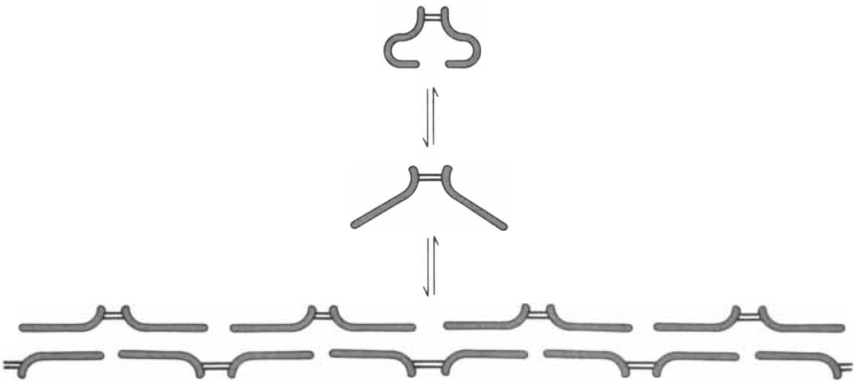
Cells move around the adult body less extensively than they do in the embryo, but during wound healing a large number of cells migrate into the affected area. They travel across the blood clot, a structure composed largely of fibrin but also of much fibronectin, which binds to the fibrin as the clot forms. Deane F. Mosher of the University of Wisconsin Medical



FIBRONECTIN MEDIATES ADHESION of cells and platelets to fibrils of other proteins. In linking fibroblasts and other cells to the extracellular protein collagen (*top*) each fibronectin subunit binds to the 140K complex; another protein complex (*dark gray*) inside the membrane completes the link between fibronectin and the actin filaments within the cell. At their far end the subunits interact with collagen through their collagen-binding domain. Fibronectin's role in the adhesion of platelets to fibrin, a phenomenon that is important in the stoppage of bleeding and the formation of blood clots, differs only in details (*bottom*). The fibronectin-binding complex in the platelet membrane consists of only two glycoproteins, which differ from those of the 140K complex but interact with the same domains on the fibronectin molecule. The fibronectin in turn is linked to the strand of fibrin through its fibrin-binding domains.

School showed that the blood-clotting enzyme called factor XIII also cross-links the two proteins by causing them to form a covalent bond. In separate work Grinnell found that fibroblasts and other cells that heal the wound adhere to the clot and migrate over it only when fibronectin is cross-linked to the fibrin.

In injuries to the cornea of the eye a thin film of fibrin and fibronectin develops at the site. The corneal epithelial cells that heal the wound migrate across the protein film. Teruo Nishida and his colleagues at Osaka University have shown that additional fibronectin speeds the repair process, and they have had some success in treating corneal ulcers with eyedrops containing fibronectin. If supplementary fibronectin also turns out to promote the healing of skin wounds, perhaps a fibronectin Band-Aid will one day appear on the scene.



MODEL OF FIBRIL FORMATION has been proposed by Helmut Hormann of the Max Planck Institute for Biochemistry in Munich. Under the normal conditions of salinity and *p*H the fibronectin subunits are folded (*top*). Under certain circumstances the molecule opens, its two subunits splaying at the disulfide bonds joining them (*middle*). Hormann suggested that splayed molecules can assemble into a fibril, with intermolecular bonds forming between antiparallel subunits (*bottom*). Details of this model, including the nature of the bonds between the subunits, are uncertain, and other models have been suggested.

Even before a wound can start to heal, bleeding must stop and the clot must form. Principal actors in these processes are not cells at all but the sticky cell fragments known as blood platelets. The platelets cling to the edges of the broken vessel and clump together to form a plug. Later they mediate clot retraction (the shrinking of the clot that helps to close the wound) by contracting, each of them tugging on its neighbors and on the fibrin matrix. Those functions obviously require platelets to adhere tightly to the matrix. Karen Smith and I showed in 1977 that the addition of fibronectin to platelets in vitro results in firmer adhesion to the substrate, causing platelet spreading: the extension and thinning of platelets that accompanies full adhesion. It is likely that fibronectin is also necessary in vivo for full platelet adhesion.

Mark H. Ginsberg and Edward F. Plow of the Scripps Clinic and Research Foundation have analyzed in detail the interaction of platelets with fibronectin. They and others have presented evidence that platelets, once they have been activated by factors released by other platelets at the clotting site, develop specific binding sites for fibronectin. By preparing antibodies to the surface membrane of activated platelets and testing the antibodies for their ability to inhibit the adhesion of platelets to a layer of fibronectin—a method similar to the one used to characterize the 140K complex—Ginsberg and Plow identified a complex of two membrane glycoproteins as the surface binding site.

John Gardner in my laboratory and Pytela in Ruoslahti's have isolated the platelet glycoproteins by allowing them to bind to fragments of fibronectin or to the active-site peptide. The glycoproteins seem to interact with the same active site on fibronectin as the 140K complex does, but they are probably distinct from their counterparts on fibroblasts and other cells. The difference between the surface glycoproteins of fibroblasts and those of platelets may reflect their differing styles of adhesion. For example, platelets must bind tightly to the clot matrix to stanch bleeding, and they must remain attached during the mechanical stresses of clot retraction. Cells such as fibroblasts and neural-crest cells, in contrast, do not remain anchored to a fibronectin-rich matrix but must be able to migrate across it. Their adhesion must be great enough to give them traction but not great enough to arrest their movement.

Can what has been learned so far about the structure of fibronectin and its role in normal cells help to explain the observation that touched off the research 13 years ago, the loss of the protein from the surface of tumor cells? The simplest explanation for the loss is that malignant cells stop making fibronectin, or make much less of it than normal cells. That hypothesis has been confirmed in some tumor cells by exposing them to a radioactive amino acid, which the cells then incorporated into proteins; the application of an antibody able to bind to radioactively labeled fibronectin revealed that abnormally small amounts of the protein had been synthesized.

In other cases the cancer cells continue to synthesize the protein but fail to bind it and assemble it into fibrils. Such cells could be making a defective form of fibronectin, perhaps because of an abnormal pattern of alternative splicing. Alternatively the cells might somehow break down fibronectin after synthesizing it. Several groups have found that cells transformed in culture by tumor viruses make enzymes that can degrade fibronectin close to the cell surface.

The cause of the tumor cells' loss of fibronectin might, on the other hand, involve not the extracellular fibrils themselves but rather the transmembrane complex: the large multimolecular structure that includes the cell-binding domain on fibronectin, the 140K complex on the cell surface, its binding site on the actin filaments, and several other proteins. Within a nexus so crucial to organization and cell behavior, defects at very different points could all lead to the disarray characteristic of tumor cells, that is, the disappearance of fibronectin from the cell exterior, the loss of cell adhesion and the disordered cytoskeleton.

That disarray, observable in vitro, would seem to bear a relation to a central feature of cancer in the body: the lifting of normal controls on cell position and migration. The result is metastasis, the migration of tumor cells to unrelated tissues elsewhere in the body. In normal tissue fibronectin probably acts in concert with other adhesive glycoproteins and their cell-surface binding sites to keep cells in place and regulate migration. Hence alterations in any of a number of proteins could underlie the vagrant behavior of cancer cells. The fibronectins are the best-understood of these glycoproteins, however, and the others probably function similarly. The study of fibronectins, initiated by work with tumor cells, may eventually come full circle to shed new light on the bases of malignant behavior.

Is Nature Supersymmetric?

Supersymmetry could represent the next step in the quest for a few simple laws that explain the nature of matter. Physicists have been seeking evidence to test the theory

by Howard E. Haber and Gordon L. Kane

About 25 centuries ago the Ionian Greeks argued that the apparent complexity of the universe could be understood in terms of a few simple underlying laws. Remarkable progress has been made toward realizing that goal. It appears that the basic constituents of matter have been identified. A few forces can account for the behavior of any form of matter, ranging from subatomic particles to galaxies. To complete the description of the laws of nature, however, further insight is still needed. For the past decade a large number of theoretical physicists have extensively explored an approach called supersymmetry. A supersymmetric theory incorporates and extends the successful discoveries of past years in an attempt to construct a new and more comprehensive theory. It also makes testable predictions.

Perhaps the most convenient point of entry into the concept of supersymmetry is the standard picture of the fundamental constituents of matter. All matter consists of molecules, which in turn consist of atoms. An atom consists of a number of protons and neutrons bound together in a nucleus, which is surrounded by a "cloud" of electrons. Individual elements are distinguished by their number of protons.

Until recently protons and neutrons were thought to be fundamental particles. Experiments carried out with high-energy particle accelerators during the past two decades have revealed that they are not; protons and neutrons appear to be composed of elementary particles known as quarks. Quarks are observed to carry a fraction (+2/3 or −1/3) of the electric charge of the proton. There are six "flavors," or types, of quarks. They are called up, down, charm, strange, top and bottom.

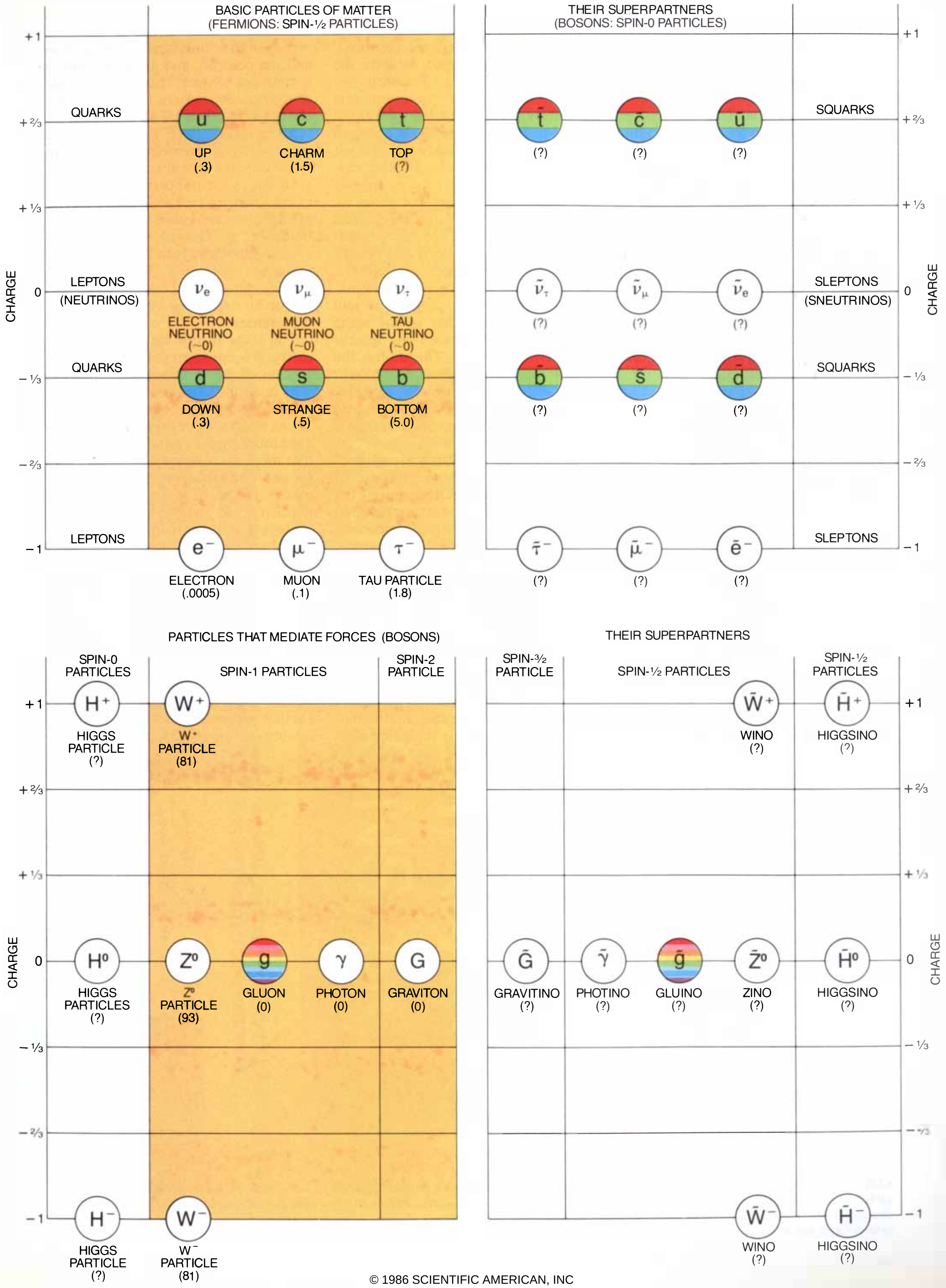
An individual quark is not expected to be isolated, or observed alone; quarks are always part of composite particles known as hadrons. Hundreds of hadrons have been identified and catalogued. They include the proton and neutron as well as the more exotic pion and kaon. A proton, for instance, is composed of two up quarks and one down quark, and a neutron is composed of one up quark and two down quarks. For an appropriate analogy one can liken quarks to the ends of a string and hadrons to the entire string, including the quarks. Suppose one tries to isolate a quark by colliding two hadrons. If a quark tries to get out after the collision, it stretches the string, which breaks. The result is more strings, that is, more hadrons (mainly pions, since they are the lightest hadrons).

Unlike protons and neutrons, electrons do seem to be fundamental particles. In fact, they are part of another family of so-called elementary particles known as leptons. There are six flavors of leptons too: the electron, the muon, the tau particle, the electron neutrino, the muon neutrino and the tau neutrino.

All interactions between leptons and quarks can be accounted for by four kinds of force: gravitation, electromagnetism, the strong force and the weak force. The electromagnetic force binds electrons and nuclei to make atoms. The atoms, although they are electrically neutral, interact through a residual electromagnetic force to form molecules. The strong force binds quarks to make protons, neutrons and all other hadrons, and the residual strong force between protons and neutrons is the so-called nuclear force that binds them into nuclei. The weak force is responsible for such phenomena as some nuclear decays and aspects of the fusion process that releases energy from the sun. In reality there are only three fundamental forces: a great accomplishment of the past two decades has been the demonstration that the electromagnetic and weak forces are manifestations of the same force, known as the electroweak force. The strengths of the forces vary widely. The strength of the electromagnetic force between two protons, for instance, is roughly 10^{36} times greater than the strength of the corresponding gravitational force.

The forces are transmitted by the exchange of a number of particles. The photon, the quantum of electromagnetic radiation, is the carrier of the electromagnetic force. Eight particles known as gluons mediate the strong force. The photon and the gluons can be interpreted as particles having zero

SUPERSYMMETRY posits that for every ordinary particle there exists a "superpartner" having similar properties—except for a quantity known as spin. According to existing theory, there are two kinds of ordinary particles: those that are the basic constituents of matter (*yellow section of top table*) and those that mediate the forces, or interactions, among those constituents (*yellow section of bottom table*). The basic constituents of matter are known as leptons and quarks. More generally they can be categorized as fermions. A fermion is a particle that behaves as though it carries a spin, or an intrinsic angular momentum, equal to half-integer units (1/2, 3/2 and so on) of Planck's constant, which is itself the fundamental unit of angular momentum in quantum theory. The basic particles that mediate forces, on the other hand, are bosons: particles whose spins are integer units (0, 1, 2 and so on) of Planck's constant. According to supersymmetry, each ordinary fermion should have a superpartner that is a boson (*right half of top table*) and each ordinary boson should have a superpartner that is a fermion (*right half of bottom table*). The bracketed number below each particle is its mass, when known, in billions of electron volts. (The graviton has not yet been found, but it is expected to be massless.) Not shown are the antiparticles of the particles in the top table. (An antiparticle is identical with another particle in mass but opposite to it in electric charge; the W^- boson, for instance, is the antiparticle of the W^+ boson.)


BASIC PARTICLES OF MATTER
(FERMIONS: SPIN-½ PARTICLES)
THEIR SUPERPARTNERS
(BOSONS: SPIN-0 PARTICLES)
CHARGE
+1
+⅔
+⅓
0
−⅓
−⅔
−1
QUARKS
UP (.3)
CHARM (1.5)
TOP (?)
LEPTONS (NEUTRINOS)
ELECTRON NEUTRINO (~0)
MUON NEUTRINO (~0)
TAU NEUTRINO (~0)
QUARKS
DOWN (.3)
STRANGE (.5)
BOTTOM (5.0)
LEPTONS
ELECTRON (.0005)
MUON (.1)
TAU PARTICLE (1.8)
SQUARKS
SLEPTONS (SNEUTRINOS)
SQUARKS
SLEPTONS
(?)
PARTICLES THAT MEDIATE FORCES (BOSONS)
SPIN-0 PARTICLES
SPIN-1 PARTICLES
SPIN-2 PARTICLE
HIGGS PARTICLE (?)
W+ PARTICLE (81)
HIGGS PARTICLES (?)
Z0 PARTICLE (93)
GLUON (0)
PHOTON (0)
GRAVITON (0)
HIGGS PARTICLE (?)
W− PARTICLE (81)
THEIR SUPERPARTNERS
SPIN-3/2 PARTICLE
SPIN-½ PARTICLES
SPIN-½ PARTICLES
WINO (?)
HIGGSINO (?)
GRAVITINO (?)
PHOTINO (?)
GLUINO (?)
ZINO (?)
HIGGSINO (?)
WINO (?)
HIGGSINO (?)

mass. The weak force is mediated by three particles: the positively charged W^+, the negatively charged W^- and the neutral Z^0. Unlike the photon and the gluons, these particles are heavy: they have masses nearly 100 times the mass of the proton. All these carriers have been experimentally observed. The mediating agent of the gravitational force, as yet only conjectured, is the graviton.

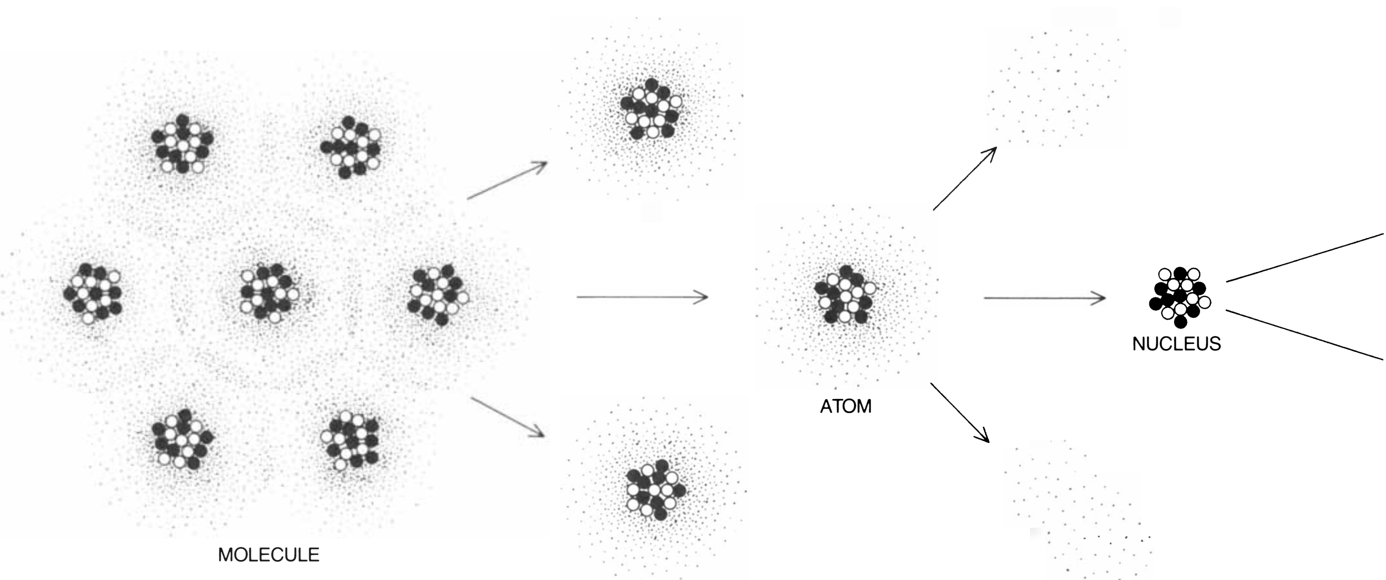
The theory that describes the quarks and the leptons and their interactions has come to be called the standard model. For the standard model to be mathematically consistent a so-called Higgs particle must exist. (The simplest version of the model contains an electrically neutral Higgs particle; more general models allow for electrically charged Higgs particles as well.) It is thought the masses of the W^+, W^- and Z^0 particles and of the quarks and the leptons are generated through interactions with the Higgs particle. The standard model predicts how the Higgs particle should interact with the other particles, but the mass of the Higgs particle itself is not predicted. The expected properties are such that so far no experiment could have found a Higgs particle, and since the mass is not known it is hard to plan experiments to search for it.

How have physicists made sense of the panoply of particles described in the standard model? First, the particles can be divided into two fundamental classes: fermions and bosons. Leptons and quarks, the basic constituents of matter, are fermions. The basic particles that mediate the four forces are bosons. Fermions behave as though they carry an intrinsic angular momentum, called spin, equal to half-integer units (1/2, 3/2 and so on) of Planck's constant, which is itself the fundamental unit of angular momentum in quantum theory. Bosons have spins that are integer units (0, 1, 2 and so on) of Planck's constant. The effects of the half-integer spin difference between fermions and bosons are profound. Fermions are "antisocial" and tend to occupy different energy states; bosonsare "gregarious" and tend to clump together in the same energy states. All leptons and quarks are spin-1/2 fermions. The photon, the W^+, W^- and Z^0 particles and the eight gluons are spin-1 bosons. The graviton is expected to be a spin-2 boson and the Higgs particle is expected to be a spin-0, or spinless, boson.

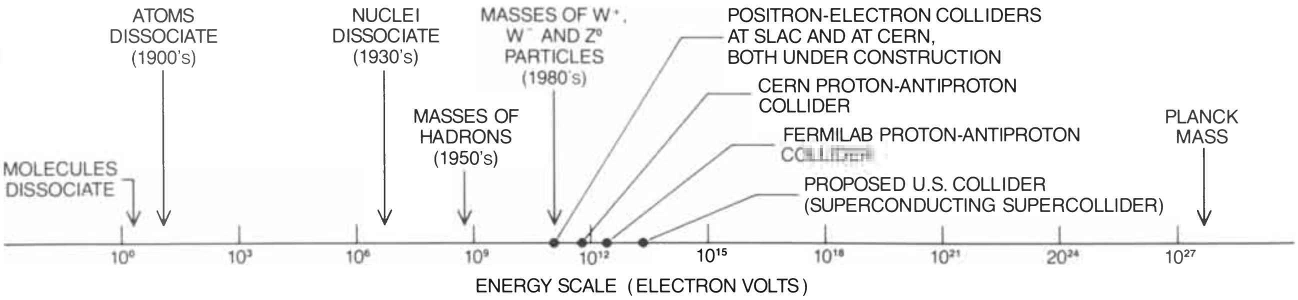
An important unifying element of the standard model is the concept of symmetry. The interactions among the various particles are symmetric (that is, invariant, or unchanged) in the face of a number of subtle interchanges. Suppose, for example, several protons are arranged in close proximity to one another (as in a nucleus), so that the strong force among them is much greater than the repulsive electromagnetic force. Imagine that the strong forces acting among the protons are then measured. If one now replaces each proton with a neutron, the forces remain unchanged. In fact, mathematically one can imagine replacing each proton with a "mixture" of the proton and the neutron, and again the forces remain unchanged. This is an example of a symmetry where the same interchange is made at all points in space.

More generalized symmetries are those in which interchanges vary from point to point in space and time. Such symmetries are important elements of gauge theories [see "Gauge Theories of the Forces between Elementary Particles," by Gerard 't Hooft; SCIENTIFIC AMERICAN, June, 1980]. All interactions described by the standard model can be successfully accounted for by exploiting such generalized symmetries.

The stage is now set for supersymmetry. It is apparent that in spite of the success of the standard model, physicists must look beyond it if they hope to understand completely the properties of matter; some aspects of the standard model are mysterious and suggest that more discoveries will come. First, no one can explain why the standard model takes the form it does. The mathematical structure of the theory is elegant and surprisingly simple, and the observed interactions show many symmetries. Yet a number of other forms (different choices of symmetries) would theoretically have been equally plausible and elegant. Second, there is no understanding of the physical origin of the masses of the fundamental particles and the strength of the forces acting between them. Why do they have the values they do? Most particle physicists hope that eventually such parameters can be calculated



STRUCTURE OF MATTER as it is currently understood is depicted schematically. All matter consists of molecules. Molecules consist of atoms. An atom consists of a number of protons (*black spheres*) and neutrons (*white spheres*) bound together in a nucleus, which is surrounded by a "cloud" of electrons (*dots*). Protons and neutrons appear to be composed of quarks, particles that carry fractional charge. A proton, for instance, is composed of two "up" quarks and one "down" quark, and a neutron is composed of one up



HIGHER ENERGIES are required to probe ever deeper into the fundamental structure of matter. At energies of about one electron volt molecules dissociate into atoms. At about 10 electron volts atoms dissociate: they are stripped of their electron cloud. Atomic nuclei fragment when they are made to collide at energies higher than a million electron volts. Many hadrons, or entities composed of quarks, have masses of about a billion electron volts. (Einstein's famous equation $E = mc^2$ states that mass and energy are equivalent and can be referred to interchangeably.) The carriers of the weak force, the W and Z^0 bosons, have masses close to 100 billion electron volts. The energy of the Planck mass (about 10^{28} electron volts) was attained only in the earliest epoch of the universe.

rather than just measured. Although there are no direct clues at present as to how to extend the standard model, supersymmetry seems to many physicists to be a likely direction in which to look. In the past few years efforts have been under way to search for evidence of supersymmetry in nature.

The search for experimental evidence of supersymmetry centers on the discovery of new particles. The reason is that the theory requires that for every ordinary particle there exist a "superpartner" with identical properties—except its spin differs by half a unit. In other words, supersymmetry differs from all previous theories in that it relates the two fundamental classes of particles, fermions and bosons, to each other. In addition the strengths of the interaction forces among the proposed superpartners are identical with the strengths of the interaction forces among the various normal particles.

The spin-0 superpartners of the fermions are named by adding the prefix *s-* to the normal particle name; for example, the spin-1/2 electron and the quark have the spin-0 partners selectron and squark respectively. The spin-1/2 superpartners of the bosons are named by adding the suffix *-ino* to the root of the normal particle name; for example, the partner of the spin-1 photon is the spin-1/2 photino and the partner of the spin-1 gluon is the spin-1/2 gluino. (For the Higgs particle there is a minor technical complication. The supersymmetric theory requires both electrically charged and neutral Higgs particles; a positively charged H^+, a negatively charged H^- and three neutrals, collectively denoted as H^0, are needed.)

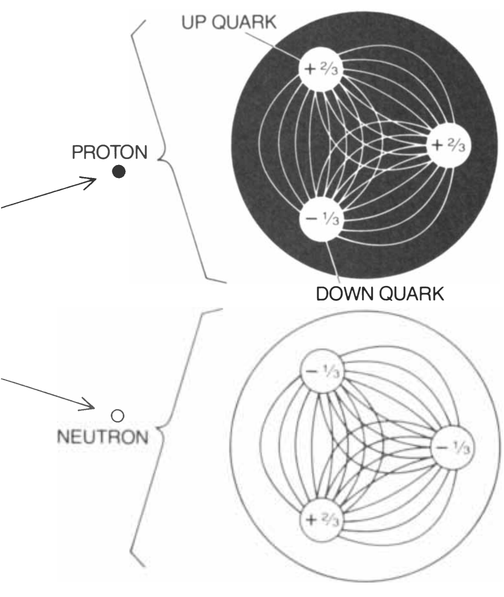
What advantages are gained by multiplying the number of elementary particles by 2? First, supersymmetry can solve a fundamental problem: it provides a mechanism by which a single theory can account for two important energies, or masses, that differ by many orders of magnitude. (As expressed by Einstein's famous equation $E = mc^2$, mass and energy are equivalent and can be referred to interchangeably.) Those energies are the masses of the W^+, W^- and Z^0 particles, roughly 10^{11} electron-volt units of energy, and the so-called Planck mass, about 10^{28} electron volts [*see illustration above*]. The Planck mass is an important quantity to include in any theory that tries to account for gravity in a unification of the four fundamental forces; if elementary particles with a mass of 10^{28} electron volts existed, the strength of the gravitational force among them would be greater than the strength of each of the other fundamental forces.

Ordinarily one would expect that if the masses of the W and Z^0 particles were computable in a fundamental theory that contained the Planck mass, they would turn out to be roughly the same order of magnitude as the Planck mass, rather than 10^{17} times smaller. In the supersymmetric theory, however, delicate cancellations occur that allow the W and Z^0 masses to be many orders of magnitude smaller than the Planck mass, just as they are observed to be. These delicate cancellations are not contrived but are guaranteed by the mathematical structure of the supersymmetric theory.

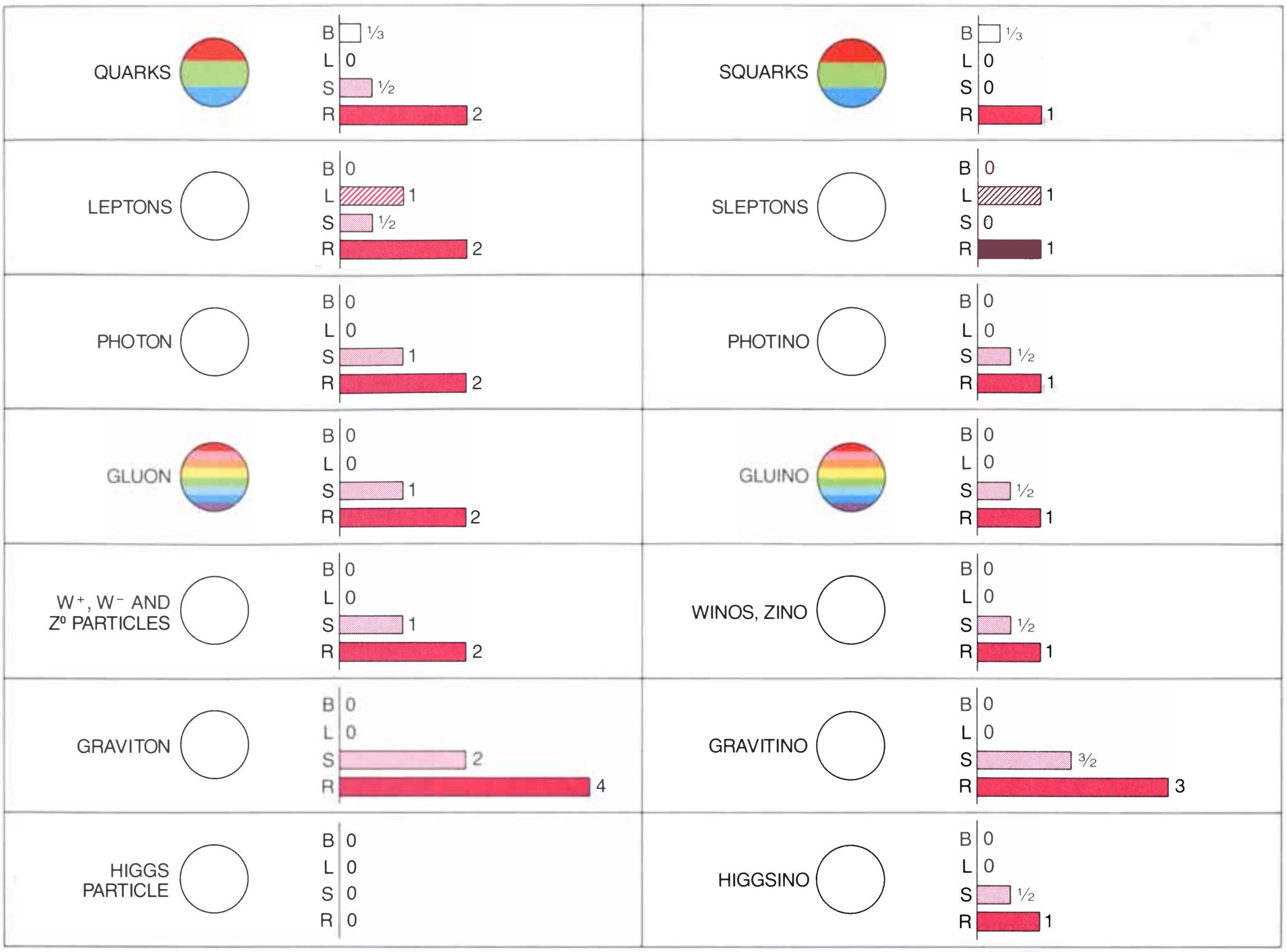
We should point out, however, that not all widely separated energies are problematical. For example, existing theories can easily account for the energy levels observed in both hadrons (such as protons and neutrons) and atoms, even though the energy levels of the hadrons are roughly 10^9 electron volts and those of the atoms are roughly 10 electron volts. The reason is found in the hierarchy of structure: atoms are made up of nuclei and electrons, nuclei are made up of protons and neutrons, and protons and neutrons are made up of quarks. If, however, one wants to develop a theory of *fundamental* particles and interactions, one no longer has recourse to such a hierarchy.

A second feature of supersymmetry is its close relation to Einstein's theory of gravity. Ever since Einstein introduced general relativity physicists have attempted, without much success, to unify gravity and quantum mechanics. It is now widely believed by theorists that a quantum-mechanical theory of gravity will be a supersymmetric one [see "Supergravity and the Unification of the Laws of Physics," by Daniel Z. Freedman and Peter van Nieuwenhuizen; SCIENTIFIC AMERICAN, February, 1978].

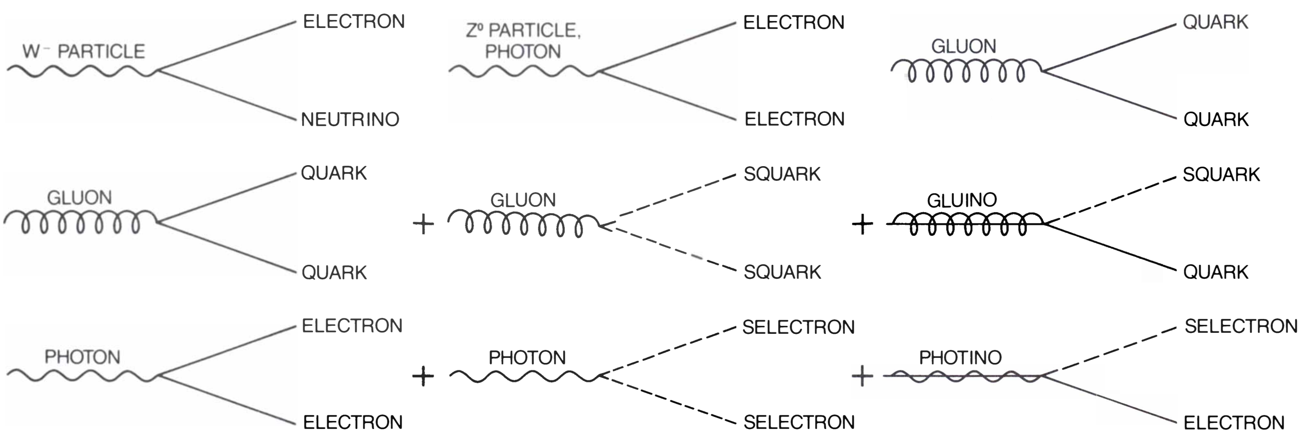
Although many physicists have contributed to the development of supersymmetry, a mathematically consis-



quark and two down quarks. An individual quark is not expected to be isolated, or observed alone; quarks are bound into composite structures by gluon "lines of force."



FUNDAMENTAL PARTICLES can be categorized by such quantities as baryon number (*B*), lepton number (*L*), spin (*S*) and *R* number (*R*). (The *R* number is given by the formula $R = 3B + L + 2S$.) Lepton number, baryon number and spin vary among the different particle types, but the *R* number is even for all ordinary particles (*left side of chart*) and odd for all predicted superpartners (*right side*). The pattern has important consequences. Specifically, one property of most supersymmetric theories is that the *R* number cannot change from even to odd (or from odd to even) during reactions among particles. If, for instance, protons are collided, supersymmetric particles must be produced in pairs or the *R* number would change from even to odd. Assuming that a supersymmetric particle is indeed produced, its decay products must contain an odd number of supersymmetric particles. As a consequence the least massive of all the supersymmetric particles must be stable, since there are no lighter supersymmetric particles into which it can decay.



FEYNMAN DIAGRAMS offer a convenient method of illustrating reactions among elementary particles. The various lines (straight, wavy, curly, broken and combinations thereof) represent particular particles. The lines are joined to represent interactions, called scattering, that can take place among particles. Shown here are vertexes: building blocks of more complex reactions. The vertexes in the top row illustrate typical processes described in the prevailing theory, which is known as the standard model. The interactions of a supersymmetric theory are easily obtained by replacing any two lines of the vertexes of the standard model with the corresponding supersymmetric partners. Such replacements for the gluon-quark-quark vertex yield the vertexes in the middle row; replacements for the photon-electron-electron vertex yield the vertexes in the bottom row. Feynman diagrams are more than pictures; they are a shorthand for a well-defined mathematical procedure for determining the probability that any given scattering process will take place.

tent formulation of the theory was first made in the early 1970's by a number of different groups working independently: André Neveu and John H. Schwarz, then at Princeton University; Pierre M. Ramond of the University of Florida; Yu. A. Gol'fond and E. P. Likhtman of the Physics Institute of the U.S.S.R. Academy of Sciences; V. P. Akulov and D. V. Volkov of the Ukrainian Physical-Technical Institute, and Julius Wess of the University of Karlsruhe in West Germany and Bruno Zumino of the University of California at Berkeley.

If indeed nature is supersymmetric, the supersymmetry must be a "broken symmetry," that is, a symmetry that holds true approximately, or for parts of the theory only. Imagine what would happen if nature were exactly supersymmetric. Selectrons would have the same mass as electrons and would bind to protons by the electromagnetic force. The properties of atoms formed in such a way would be very different from those of normal atoms. As fermions, electrons must occupy different energy levels in atoms; as bosons, selectrons would occupy the same energy levels. If atoms contained selectrons instead of electrons, the structure of the periodic table of elements would be completely altered. Since such atoms have not been observed, the mass of the selectron (if the selectron exists) must be larger than the mass of the electron, and so the symmetry is broken.

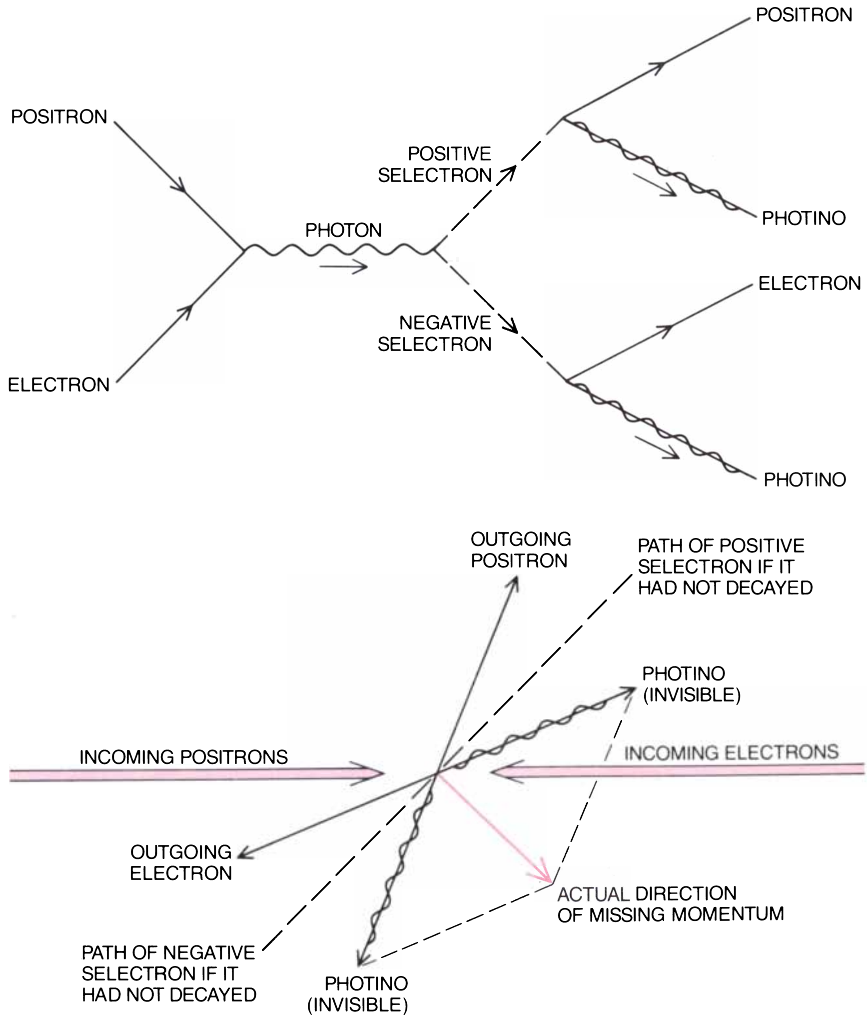
The appeal to a broken symmetry may at first seem to be an ad hoc solution. In reality, however, the concept of broken symmetry is a familiar and powerful tool of physics. One of its great triumphs is seen in the theory of electroweak interactions. In that theory a symmetry naively implying that the photon, W^+, W^- and Z^0 particles are all massless is broken. As a result the W^+, W^- and Z^0 particles are expected to have mass. Moreover, their masses are correctly predicted by the theory. It is easy to construct a theory that maintains all the desirable aspects of a perfect supersymmetry but in which the supersymmetry is broken in such a way that the masses of the superpartners are much larger than the masses of the corresponding standard-model particles.

Can the masses of the superpartners be predicted? A possible answer is suggested by the arguments regarding the wide difference in mass between the W and Z^0 bosons and the Planck mass. The delicate cancellations required to explain the disparity remain operative in a slightly broken supersymmetric theory, but only if the superpartner masses are not much larger than the W and Z^0 boson masses.

As work on the mathematical structure and properties of supersymmetric theories has progressed, physicists have increasingly wondered how experimental evidence for supersymmetry might be detected. If nature is supersymmetric, how would investigators recognize it? In 1979 Pierre Fayet, now at the École Normale Supérieure in Paris, first studied these questions, partly in collaboration with Glennys R. Farrar, now at Rutgers University. By the early 1980's the mathematical motivations for the study of supersymmetry had become clearer and a better understanding of how to build realistic models incorporating supersymmetry was developed. At that point we and others began to study in detail the question of how to recognize evidence for supersymmetry.

Today that question is essentially answered. Physicists know how to calculate the predictions of supersymmetric theories and how to decide whether a particular experimental signal, or class of events, does or does not substantiate a supersymmetric theory. Furthermore, it is known how to set quantitative limits on a supersymmetric theory if no signal is observed in a given set of experiments.

Suppose two normal particles, such as two electrons or two protons, are accelerated to high energies and supersymmetric particles are produced in any collisions that ensue. What pre-



SUPERSYMMETRIC PARTICLES might be produced by particle accelerators called electron-positron colliders. (A positron is identical in mass with an electron, but its charge is positive.) Collisions between electrons and positrons could yield positively and negatively charged selectrons, as is shown in the Feynman diagram at the top and in the drawing at the bottom depicting the collision and its products. The positively charged selectron would decay into a positron and a photino and the negatively charged selectron into an electron and a photino. Both photinos would escape detection, so that an event of this type, if it occurred, would be characterized by the fact that on the average half of the energy would be missing. Such an event would also exhibit missing momentum; the direction would be given by the "vector sum" of the photinos' momentum. No candidates have been found to date.

dictions does supersymmetry make about such particles?

A conservation law of supersymmetry leads to two results. First, supersymmetric particles cannot be produced alone; they must be produced in pairs. Second, if a supersymmetric particle is produced, its decay products must contain an odd number of supersymmetric particles. As a consequence the least massive of all the supersymmetric particles must be stable, since there are no lighter supersymmetric particles into which it can decay. There are several candidates for the least massive supersymmetric particle. For now we shall assume it is the photino.

The existence of a least massive supersymmetric particle has important consequences for the detection of supersymmetry in the laboratory. (In addition the least massive supersymmetric particle could have cosmological consequences: it might account for the so-called dark matter, or missing mass, of the universe.) All supersymmetric particles except the least massive one decay into particles that must include the least massive one at the end of their decay chain. The decays occur so fast that the decaying particle does not have time to travel a detectable distance. If a supersymmetric particle were produced, only its stable decay products, which would be standard-model particles and the least massive supersymmetric particle, would be observed in the laboratory. To detect supersymmetric particles, therefore, the least massive supersymmetric particle must be detected.

Unfortunately the probability that the least massive supersymmetric particle could be directly detected is so low as to be negligible: it interacts very weakly with ordinary matter and would escape any detector. As a result the total energy, including the rest mass, of the detected product particles would be less than the total energy of the colliding particles. In other words, the signature of supersymmetry is missing energy: the energy is carried away by the least massive supersymmetric particle.

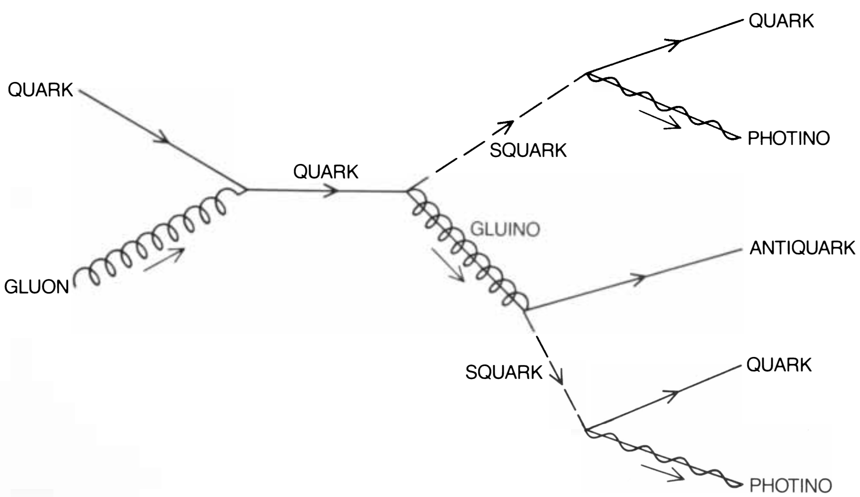
Can a particle be discovered indirectly by observing that a substantial amount of energy is missing? The answer is yes. That is how the existence of the neutrino was deduced about 50 years ago. Much more recently (in 1983) the existence of the *W* boson was confirmed by a similar method. In experiments done at CERN, the European laboratory for particle physics in Geneva, *W* bosons were produced by the collision of protons and antiprotons. (An antiproton is identical in mass with a proton, but its charge is negative.) Some of the *W* bosons subsequently decayed into an electron and a neutrino. The neutrino escaped the detector, leaving a high-energy electron. By measuring the energy of the electron and determining what the energy of the neutrino must have been for energy conservation to hold, the existence of the *W* boson was inferred and its mass was determined.

Similar experiments for detecting supersymmetric particles are now being done with electron-positron colliders and proton-antiproton colliders. (A positron is identical in mass with an electron, but its charge is positive.) Collisions between pairs of electrons and positrons could produce pairs of positively and negatively charged selectrons. The positively charged selectron would rapidly decay into a positron and a photino and the negatively charged selectron would decay into an electron and a photino. Both photinos would escape, so that an event of this type, if it occurred, would be characterized by the fact that on the average half of the energy would be missing.

Since the outgoing positron and electron arise from different decays, their directions would be uncorrelated: they would not point in opposite directions, as would be expected in a normal collision. Moreover, if a plane were defined by the incoming particle beams and the outgoing positron, the outgoing electron would not in general lie in that plane, again contrary to the result of a normal collision. Searches for events with these characteristics have been made at electron-positron colliders such as the CESR at Cornell University, PEP at Stanford University and PETRA in Hamburg. So far no candidates have been found.

The negative results do not rule out supersymmetry. They only mean that selectrons lighter than the highest energy available at electron-positron colliders do not exist; the mass of a selectron must be at least 23 times greater than the mass of a proton (roughly a billion electron volts), or about 28 percent of the mass of a *W* boson. (Certain indirect methods allow somewhat larger masses to be ruled out.) Such quantitative limits can be obtained because supersymmetry is a well-defined theory in which one can calculate whether a detectable number of events due to supersymmetric particles of a given mass would be expected in a given experimental situation. Without such calculations a positive result could still be interpreted as evidence for supersymmetry, but a negative result would simply imply that too few superpartners were produced, with no information on the possible allowed masses. Because it is reasonable to expect the masses of supersymmetric particles to be as large as a few *W*-boson masses, colliders able to explore regions of higher energy are needed.

The available energies at proton-antiproton colliders are somewhat higher, but there the comparison between theory and experiment is subtler. It is subtler because of the composite nature of the proton: calculations must take into account the fact that a proton consists of three quarks and a number of gluons that bind the quarks together. As might be expected, in a moving proton about half of the momentum is carried by the gluons and the other half by the quarks. Then each of the three quarks must carry about a sixth of the total momentum. Existing machines, limited essentially by economic considerations, have produced 23-billion-electron-volt (BeV) collisions between electrons and posi-



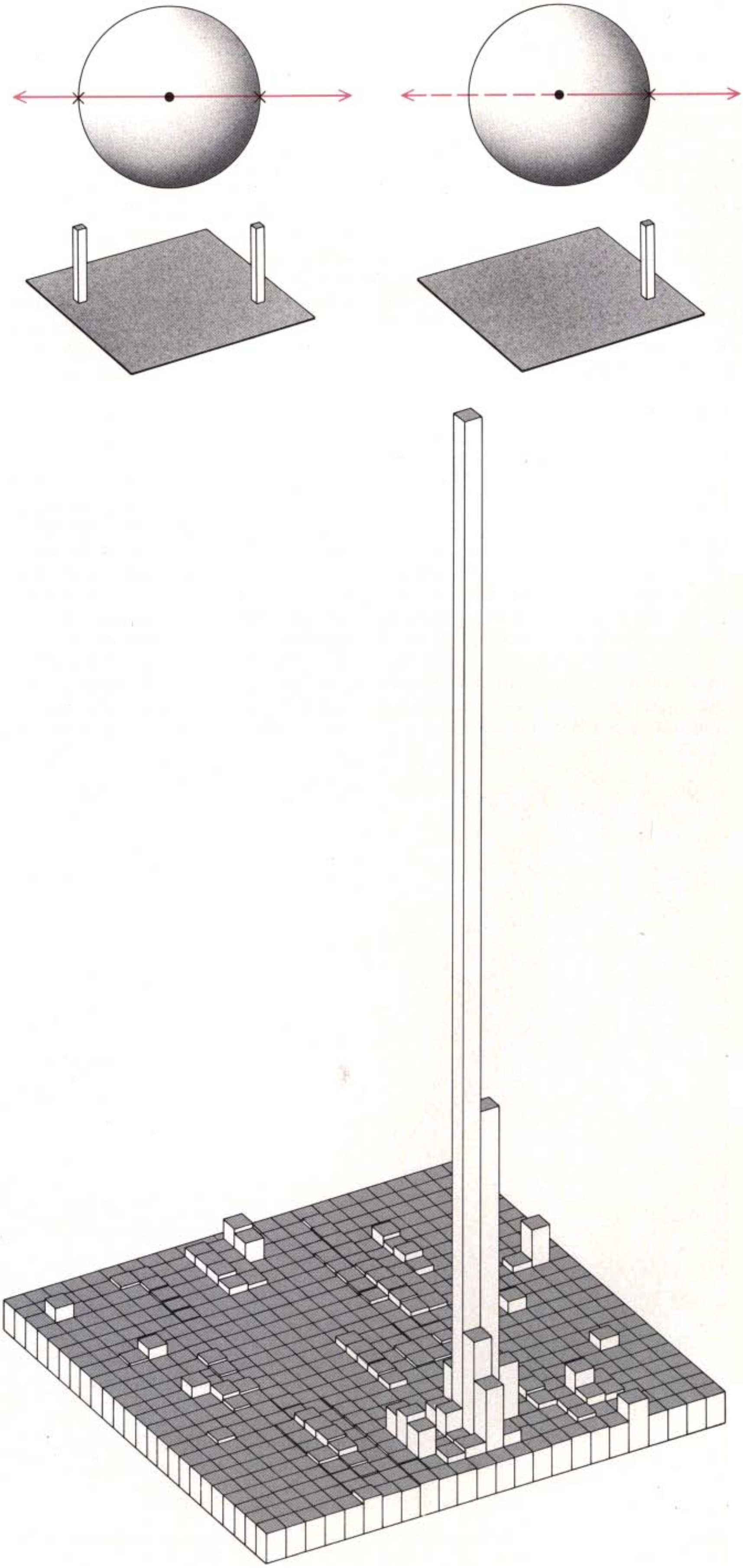
PROTON-ANTIPROTON collisions might also produce supersymmetric particles if, for instance, a quark from one proton or antiproton collides with a gluon from another one.

trons and 315-BeV collisions between protons and antiprotons. As a result the quarks and gluons in a proton, even though they carry only a fraction of the proton's momentum, can still be exploited to probe mass scales higher than those accessible to present-day electron-positron colliders.

A number of different collisions can take place at proton-proton or proton-antiproton colliders. Suppose a quark from one proton strikes a gluon from an antiproton and produces a squark and a gluino. The squark would rapidly decay into a quark and a photino, and the gluino would decay into a quark, an antiquark and a photino. An isolated quark or antiquark cannot emerge from the collision; instead it must somehow turn itself into hadrons. Because the emitted particle is energetic in this case, it turns itself into a group of hadrons all moving in roughly the same direction as the original quark. In the parlance of physics such a group of hadrons is commonly referred to as a "jet" of hadrons.

In the example, three jets of hadrons (one for each of the two quarks and one for the antiquark) and two photinos would emerge into the region around the interaction. When the gluino decays, its energy is shared among the quark, the antiquark and the photino. The energy need not be shared equally, however; sometimes one or another of the three decay products gets most of the available energy. Although the energy imbalance and other related effects make comparison of theory and experiment more difficult at proton colliders than at electron colliders, the comparison can be made.

As an example, in 1982 Jacques P. Leveille, then at the University of Michigan, and one of us (Kane) proposed that evidence for supersymmetry would be obtained if certain observations (such as one or more jets of



MISSING ENERGY is the signature of supersymmetry. Suppose a particle decays into two particles, both of which are registered by a spherical detector. Then if the detector is "unrolled" on a flat surface, two peaks would appear, one for each particle (*top left*). If the detector registers just one of the particles, only one peak would appear (*top right*) and energy would seem to be missing. Since supersymmetric particles would escape all known detectors, their production would be characterized by one-peak events. Such events, one of which is shown here, have been found at the UA1 (for Underground Area 1) detector at CERN, the European laboratory for particle physics. Events predicted by the standard model can appear in a similar configuration, however, and so it is still unclear whether or not supersymmetric particles have been produced.

particles in the presence of a large amount of missing momentum) were seen at a high-energy hadron collider. Much excitement was generated the following year when events of this type were reportedly seen at the UA1 (for Underground Area 1) and UA2 detectors at CERN. A number of groups of theorists have examined the possibility that supersymmetric particles were being produced. We have done detailed analyses in collaboration with R. Michael Barnett of the Lawrence Berkeley Laboratory. Other analyses have been made by John Ellis of CERN and Henry Kowalski of DESY (the electron accelerator in Hamburg), by Vernon D. Barger of the University of Wisconsin at Madison and his collaborators, and by Ewald Reya of the University of Dortmund in West Germany and D. P. Roy of the Tata Institute of Fundamental Research in India.

The number of missing energy events observed in the original, albeit small, data sample initially suggested the possibility of new particles. Events predicted by the standard model, however, can appear in a configuration similar to the configurations observed at CERN. Statistical analyses are therefore needed to prove that inherently new phenomena are being seen. Now that more data have been accumulated the situation remains unclear, although the possibility that the discovery of supersymmetric particles has occurred is less likely than it seemed initially. More data will be collected at the CERN collider this year and may make it clearer whether any of the few candidate events corresponds to the production of supersymmetric particles.

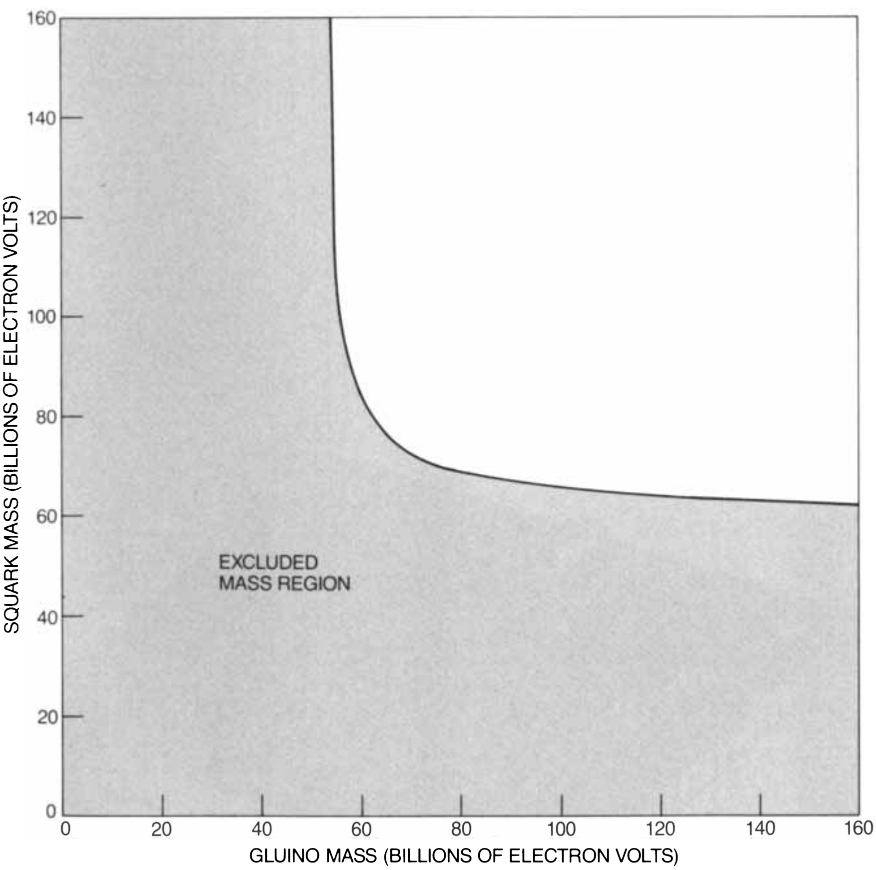
Together with Barnett we have done a comprehensive analysis of the data from CERN. First, we assumed the photino is the least massive supersymmetric particle. Working from the additional assumption that the missing energy events can be accounted for by the standard model, we concluded that gluinos and squarks must be heavier than about 75 proton masses, or 70 BeV [*see illustration below*]. Although this is a large mass, it is still not as large as the *W*-boson mass (81 BeV).

Yet if it turns out that some of the events do correspond to the production of supersymmetric particles, our analysis implies the photino is not the least massive supersymmetric particle. This conclusion is remarkable, given that the masses of the supersymmetric particles are not known. We were able to arrive at it because the supersymmetric theory is strongly constrained, thereby yielding detailed predictions that can easily be tested by experiment.

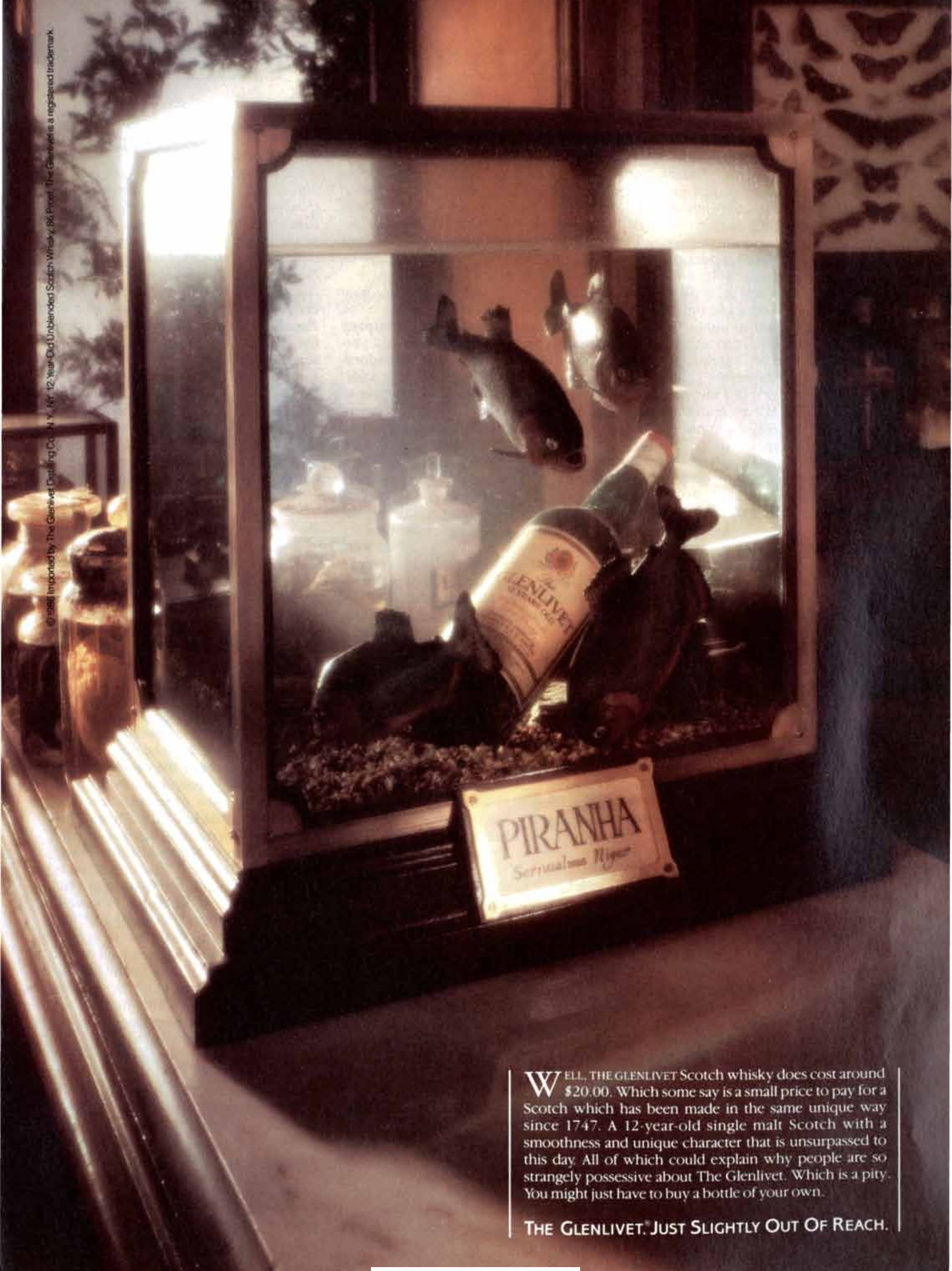
In collaboration with Mariano Quiros of the Institute for the Structure of Matter of the Council for Scientific Research (CSIC) in Spain we have argued that the least massive supersymmetric particle might be the higgsino (the supersymmetric partner of the Higgs boson). If this turns out to be the case, the photino would be unstable and would decay into a photon and a higgsino. In this scenario the limits that can be placed on squark and gluino masses are somewhat weaker.

As machines with larger energies and intensities become available in the future, they could produce and detect supersymmetric particles of greater mass. The electron-positron colliders that will start up in the next few years (TRISTAN in Japan in 1986, SLC at the Stanford Linear Accelerator Center in 1987 and LEP at CERN in 1989) will be able to detect sleptons up to about 50 BeV in mass. The proton-antiproton collider at Fermilab, which should begin to yield data at the end of this year, will be able to detect squarks and gluinos of mass between 100 and 150 BeV, depending on its intensity. Before 1990, then, sleptons of 50 BeV and squarks of about 150 BeV will have been either found or excluded.

To go beyond those masses will require machines that are being planned but have not yet been approved. The U.S. particle-physics community has committed itself to proposing a proton-proton collider called the Superconducting Supercollider (SSC), having energies of 20,000 BeV per beam and an intensity about 1,000 times as great as that of the CERN or Fermilab proton-antiproton colliders [see "The Superconducting Supercollider," by J. David Jackson, Maury Tigner and Stanley Wojcicki; SCIENTIFIC AMERICAN, March]. At the SSC squarks and gluinos of masses up to more than 20 times as great as the *W*-boson mass could be found. When such a machine is in operation, if not before, physicists expect to find experimental clues pointing toward a theory that goes beyond the standard model. In particular, data from the SSC could definitively decide whether nature is supersymmetric on the scale of the electroweak force and could thus help in understanding the laws of nature on that scale. The alternative is that supersymmetry could at best be a mathematical property of quantum field theories, relevant to energies far greater than those that investigators can ever hope to probe directly.



ANALYSIS OF EXPERIMENTAL DATA from CERN shows that if squarks and gluinos do exist, their mass is probably greater than about 70 billion electron volts (*white region*).


PIRANHA
Serrasalmus Niger
WELL, THE GLENLIVET Scotch whisky does cost around $20.00. Which some say is a small price to pay for a Scotch which has been made in the same unique way since 1747. A 12-year-old single malt Scotch with a smoothness and unique character that is unsurpassed to this day. All of which could explain why people are so strangely possessive about The Glenlivet. Which is a pity. You might just have to buy a bottle of your own.
THE GLENLIVET.® JUST SLIGHTLY OUT OF REACH.
© 1986 Imported by The Glenlivet Distilling Co., N.Y., NY. 12-Year-Old Unblended Scotch Whisky. 86 Proof. The Glenlivet is a registered trademark.

SCIENCE AND THE CITIZEN

Bugs in the System

As genetically altered bacteria and other life forms produced by biotechnology companies enter the marketplace and the environment, they are provoking concern. The release of such organisms, members of the public and some scientists fear, could have unpredictable ecological consequences. Certain critics even raise the specter of some unforeseen, catastrophic impact on public health.

For anyone who is concerned about the possible hazard, bureaucratic implementation of the review procedures required by Federal law and the interpretation of the law by manufacturers have been decidedly discomfiting. The suspension this spring of two licenses for the use of genetically engineered organisms dramatized the issue sharply. The Environmental Protection Agency halted what was said to be the first release of such organisms into the environment when the agency learned that the maker had already experimented outdoors with its engineered bacteria, designed to slow the formation of frost on crops. Soon afterward it was revealed that the bacteria might not have been the first modified organisms to enter the environment after all. Under licenses issued with little publicity by the U.S. Department of Agriculture a livestock vaccine consisting of genetically altered virus was field-tested last year and put on the market in January. The department temporarily stopped the sale of the vaccine after the disclosure.

The attempt by Advanced Genetic Sciences (AGS) of Oakland, Calif., to commercialize its "ice-minus" bacteria had already run into trouble. For two years the company had sought permission for a field test of the bacteria, two species of *Pseudomonas,* a common genus, from which the DNA encoding a protein that nucleates the formation of ice crystals had been deleted. The EPA issued the permit in November, 1985, but apprehensive residents of Monterey County, Calif., the site of the small strawberry patch that was to have been sprayed with the bacteria, objected. Local groups voiced safety concerns that had previously been raised by Jeremy Rifkin, a determined critic of biotechnology. In January, AGS postponed the test, which was meant to determine the ability of the modified bacteria to displace natural frost-nucleating strains.

The EPA then discovered that an earlier experiment, in which AGS workers injected the modified bacteria into the bark of fruit and nut trees to see whether the organisms cause disease in plants, had been done not in a greenhouse but in the company's rooftop grove. After investigating the incident the agency fined AGS $20,000 and rescinded the testing permit. The EPA views the rooftop test as an unauthorized environmental release.

The Biologics Corporation of Omaha, the maker of the modified virus, was the next biotechnology company to come to grief. Its product is a version of the virus that causes pseudorabies, a herpeslike disease that is widespread among swine. The deletion of part of the gene for the enzyme thymidine kinase renders the engineered virus unable to cause disease, making it a safe vaccine. More than a year ago, without formally submitting the application to its own Agricultural Recombinant DNA Research Committee, the Agriculture Department approved the testing of the swine vaccine. Pigs were inoculated experimentally last year in Illinois, and in January the department authorized the sale of the vaccine.

Regulators and scientists disagree about whether the use of the vaccine constitutes a release into the environment. Nonetheless, the outcry that arose in April when Rifkin drew attention to the department's decision led the agency to suspend the marketing license for two weeks while it documented more fully the procedures by which it had approved the vaccine.

The way still seems to be clear for the testing of several other genetically engineered organisms. Agracetus, a company in Middleton, Wis., may proceed with an experiment for which it received Federal approval last November: a field test of tobacco plants genetically altered to resist crown gall disease. A scientific advisory panel to the EPA has recommended that the agency allow the Monsanto Company to test a microbial pesticide consisting of a common soil bacterium with an added gene. Taken from another bacterium, the gene encodes a toxin that is lethal to cutworm, a pest of corn roots. Rifkin has criticized the scientific review of Monsanto's application and may go to court to block the test.

Strong criticism of Federal biotechnology review procedures, specifically those of the Agriculture Department, has also been aired by the General Accounting Office. In March it issued a report (prepared before the revelations about the pseudorabies vaccine) concluding that the process by which the Agriculture Department reviews the safety of genetically altered organisms is poorly defined.

How great is the danger such procedures and reviews are designed to eliminate? According to Patrick W. Flanagan, head of the ecology program at the National Science Foundation, the pseudorabies vaccine and the ice-minus bacteria "are the lowest-risk organisms we'll see." Other workers point out that ice-minus strains of *Pseudomonas* occur in nature, and that the change in the pseudorabies virus serves only to weaken it, not to give it some trait not found in the wild strain.

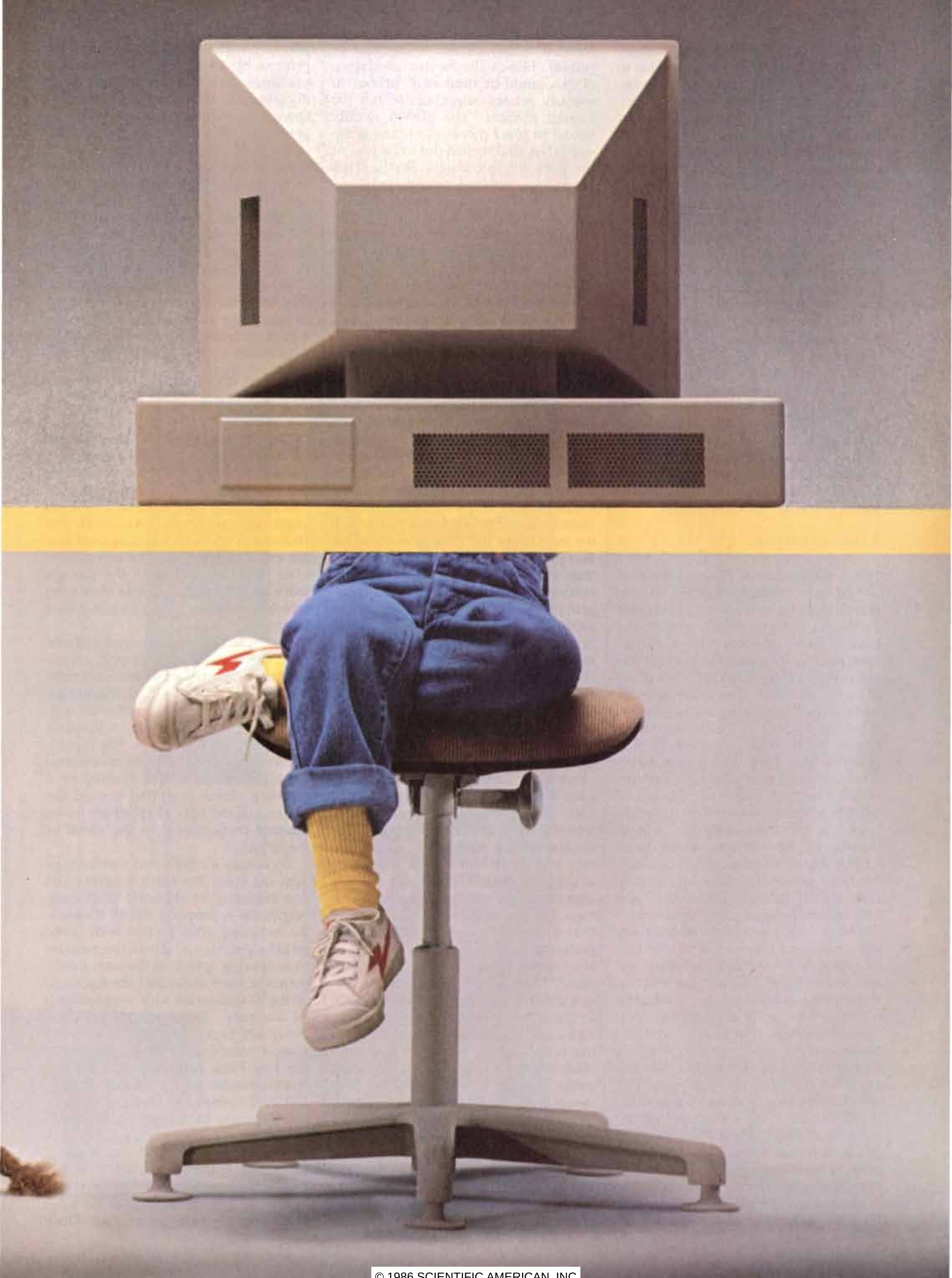
Yet instances that do raise legitimate concerns are easy to imagine: a large-scale release of ice-minus *Pseudomonas* is one of them. Ice-positive bacteria in the atmosphere are thought to nucleate the formation of some cloud droplets and precipitation particles. It is conceivable, Flanagan says, that spraying large agricultural areas with an ice-minus strain could change the normal ratio of ice-positive to ice-minus bacteria in the overlying atmosphere, thereby affecting precipitation patterns. "We don't know how to evaluate the risk," says Flanagan, who nonetheless believes tests yielding a "ballpark estimate" of the hazard are possible. He adds that until ways to gauge such environmental risks are developed, large-scale releases of genetically engineered microbes into the soil, air or water should not be done.

A concern about the unpredictable and potentially troublesome effects of introducing alien organisms into the environment has led Rifkin to call for the development of a "science of predictive ecology" before releases of any kind are authorized. That condition is tantamount to calling a halt to the entire enterprise, if Robert K. Colwell, an ecologist at the University of California at Berkeley, is right. "In principle," says Colwell, who served as an outside reviewer of data on the ice-minus bacteria, "there can't be a fully predictive ecology." The number of unpredictable events bearing on an organism's fate in the environment is simply too great, he explains. Colwell adds: "People who say we have to prove safety don't understand the nature of scientific proof." He believes tests can be devised for particular organisms that give a reasonable estimate of ecological safety.

Molecular biologists tend to play down the possible hazards of environmental release. According to Walter Gilbert of Harvard University, recom-


Head into the future, in this Century.
Buicks are more popular every year, and Century is the most popular of all. Perhaps this is because Century embodies all of Buick's virtues: roominess, comfort, luxury and convenience.
Perhaps also because these virtues are engineered into a contemporary, stylish, front-wheel-drive package, with responsive performance and road manners. Century has power-assisted rack-and-pinion steering and MacPherson strut front suspension. Engine choices range from an electronically fuel-injected 2.5-litre to a sequential-port fuel-injected 3.8-litre.
Then again, perhaps Century's popularity is due to refinements, which will keep it contemporary for years to come. This year, Century sports a rakish new front-end treatment that enhances its appearance and aerodynamics.
Or, perhaps Century's popularity is a result of all these attractions, in a satisfying combination.
Buckle up and visit your Buick dealer. Then, head into the future, in this Century. For more information, a test drive or a brochure, call:
1-800-86-BUICK (1-800-862-8425).
Wouldn't you really rather have a Buick?
BUICK
GM
Buick Century.


Electronics.
Still growing. And the future belongs to those who are committed to innovation— to Eaton.
EATON
Growing into the future.
Automotive • Electronics • Defense • Capital Goods



binant-DNA work in the laboratory has taught that "it is much harder to create hazards, by accident or intentionally, than people might think." "The lesson of the science of the last decade," Gilbert notes, "is that accidental or deliberate changes essentially weaken the organism." Other molecular biologists point out that organisms with a novel genetic makeup arise continually in nature through processes of genetic recombination, mutation and the horizontal transfer of genetic material. Thus, they argue, what is going on in biotechnology laboratories is in principle nothing new.

Many investigators call for a scientific consensus about the nature and magnitude of the risks posed by the release of engineered organisms, one that can serve as the basis for a coherent regulatory process. In the meantime the intense public scrutiny and the need to set sound regulatory precedents would seem to require that companies and regulators tread more carefully than they have, even when the organisms at issue are clearly safe.

Visionary Genes

The techniques of molecular biology have made it possible for research workers to resolve, quickly and cleanly, a centuries-old problem: the genetic basis of color vision. It has long been known that the normal human eye has three different types of receptors for color. The receptors are cone-shaped cells lining the retina; enzymes embedded in their outer membranes make them sensitive to red, green or blue light. Using new methods for cloning genes, Jeremy Nathans of the Stanford University School of Medicine and his co-workers have isolated the genes encoding the protein portion of the enzyme, which gives each receptor its particular sensitivity. Nathans' group has gone on to show how defects in those genes cause the commonest form of color blindness.

The starting point of the recent work was an enzyme responsible not for color vision but for black-and-white vision in dim light. That enzyme is called rhodopsin, and, found in rod-shaped retinal cells, it is the best-studied of all visual pigments. From cows' eyes Nathans and his colleagues isolated a considerable quantity of the messenger RNA (mRNA) for rhodopsin. The mRNA was used to make complementary DNA (cDNA) sequences that encode the pigment. The cDNA was then put in solution with DNA fragments making up a human genome (the full complement of genetic material).

In solution two pieces of DNA will bind to each other if the genetic information they encode is identical or similar. Hence the bovine rhodopsin cDNA could be used as a "probe" to identify related sequences within the human genome. The cDNA readily bound to some pieces of human genomic DNA that turned out to be part of the gene for human rhodopsin. That result was expected, since human and bovine rhodopsin were assumed to be closely related. Under slightly different conditions the probe will also bind to DNA sequences to which it is less closely related. When the conditions were altered, the cDNA bound to three other sets of DNA fragments.

Were those other fragments genes, or were they merely "noise," or artifacts of the binding procedure? Working with human retinal tissues collected during autopsies, Nathans and his co-workers found that the three other DNA sequences are transcribed into mRNA in the human eye; transcription is the first step in the production of a protein from the gene. Thus it appeared that the sequences isolated by the rhodopsin probe were indeed functional genes. The next step was to find the location of the three genes, and for that purpose a set of hybrid mouse-human cells were employed, each containing a known portion of the human genome. Using the three new genes as probes to bind to the DNA of the hybrid cells, Nathans and his collaborators found that one gene was on chromosome 7 and the other two are close together on the *X* chromosome. (Females have two X chromosomes, males one *X* and one *Y*.)

Final confirmation that the three genes are involved in color vision came from studies of color-blind subjects. Roughly 8 percent of men (along with a much smaller proportion of women) have difficulty distinguishing red from green. Red-green color blindness may result either from alterations in the green-sensitive pathway or from alterations in the red-sensitive pathway. The group led by Nathans found that those in whom the red-sensitive pathway is altered have a change in one of the genes on the *X* chromosome. Those in whom the green-sensitive pathway is altered have a change in the other gene on the *X* chromosome (that gene may be present in multiple copies). Hence it was concluded that the pair of *X*-chromosome genes encode the red and green pigments and, by a process of elimination, that the gene on chromosome 7 encodes the blue vision pigment.

The recent results, published in two articles in *Science,* enabled Nathans and his co-workers to propose a striking hypothesis concerning the origin of red-green visual disorders. The hypothesis rests on recombination, the process by which two chromosomes exchange genetic material during cell division. In some instances the exchange is unequal: one chromosome gets more than the other. If that were to happen within the region of the *X* chromosome where the red and green genes lie, the result might be the deletion of, say, a green-pigment gene. On the other hand, the outcome might be the addition of a "green" sequence to the gene coding for the red pigment. If such unequal recombination were to take place during the cell division that leads to the production of eggs or sperm the aberrant *X* chromosome would ultimately be passed to a man who would have a defect in either red or green vision.

Big-Bang Bias

The big-bang theory of the universe holds that all galaxies are rushing uniformly away from one another as the result of a primordial explosion that took place some 15 to 20 billion years ago. A group of American and British astronomers has suggested that the expansion may not be so uniform after all. The results of the group's work were first presented in Hawaii at a North Atlantic Treaty Organization workshop.

The discovery of the nonuniform motion grew out of a six-year survey of the recessional velocities and relative distances of 400 elliptical galaxies, all of which lie within a spherical volume of the universe having a diameter of 350 million light-years. The investigators could determine the recessional velocities because light emitted by a receding object is shifted toward the red end of the optical spectrum by an amount proportional to the speed of the object.

To obtain a rough measurement of how far from the earth a galaxy lies one measures its apparent brightness; brightness is proportional to distance. By including other factors such as the total brightness and size of the galaxies as well as the speeds of the stars within them the team measured the distances of the 400 galaxies with unprecedented accuracy. The participants in the group are David Burstein of Arizona State University, Roger L. Davies of the Kitt Peak National Observatory, Alan Dressler of the Mount Wilson and Las Campanas Observatories of the Carnegie Institution of Washington, Sandra M. Faber of the University of California at Santa Cruz, Donald Lynden-Bell of the Institute of Astronomy of the University of Cambridge, Roberto Terlevich of the Royal Greenwich Observatory and Gary

A. Wegner of Dartmouth College.

They found that galaxies in one hemisphere of the 350-million-light-year spherical volume are receding more slowly than galaxies in the other hemisphere. Based on the big-bang model one would expect, for instance, that all galaxies at a distance of 200 million light-years from the earth should be receding at a speed of about 3,000 kilometers per second, or about 2,000 miles per second. Instead the team determined that galaxies in one hemisphere recede at a speed of about 2,300 kilometers per second whereas those in the other recede at about 3,700 kilometers per second. In other words, the uniform expansion caused by the big bang is skewed by a net drift velocity of 700 kilometers per second.

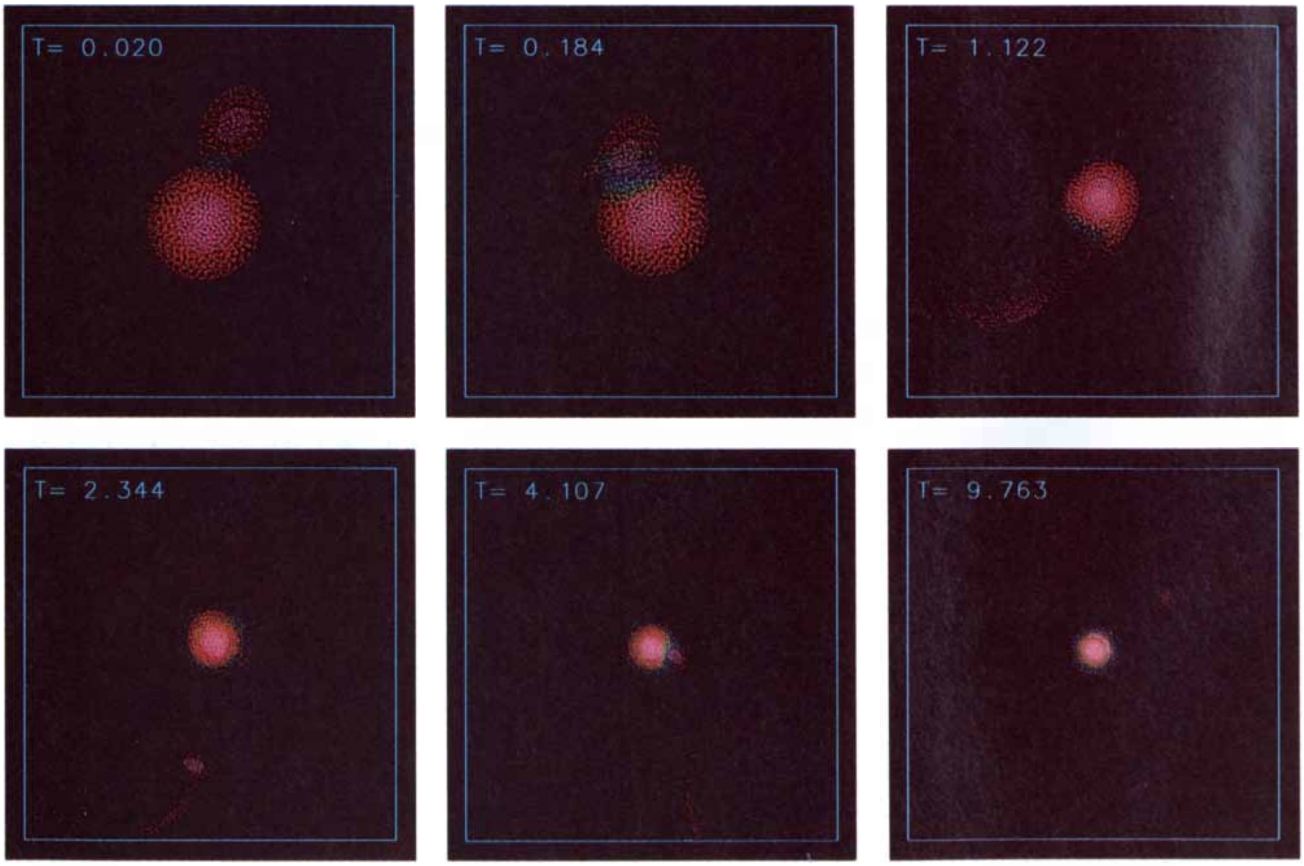
What is the cause of the net drift? A simple but speculative interpretation is that a distant, immense but as yet undiscovered mass is pulling the entire local universe toward it by the force of gravity. Alternatively, the drift could be the remnant of massive explosions that took place after the big bang. It is interesting that similar explosions (although smaller in scale) have also been postulated to account for the observation that the universe is made up of gigantic bubbles: spherical or slightly elliptical regions of space apparently devoid of matter, whose outer surfaces are defined by galaxies.

Whence the Moon?

At a time when closeup images of the moons of Uranus are shown on television one might think fundamental questions about the earth's own moon had been settled long ago. Far from it: it is not even clear how the moon formed. According to one theory, it is a piece of the earth's mantle that was spun off early in the planet's history; according to another, it formed in an independent orbit but was captured by the earth's gravity; still another holds that the moon and the earth accreted close together and simultaneously. Now a set of computer simulations lends support to a fourth mechanism: the impact on the earth of a planetary body a little larger than Mars.

Contradictions have always plagued the first three theories. One type of problem is dynamical. For example, to spin off a large part of its mantle the earth would have had to rotate about once every two hours, much faster than is compatible with the current angular momentum of the earth-moon system. Then there is the question of the moon's anomalous composition. Compared with the earth it is severely depleted in volatile chemical elements, a fact that none of the three theories accounts for well. The second and third theories also do not explain adequately why the earth has a large iron core but the moon does not, even though both bodies probably accreted in the same part of the solar system.

The giant-impact hypothesis avoids these problems. First proposed in 1976 by A. G. W. Cameron of Harvard University and by other workers, it holds that the moon formed when a planet-size object grazed the earth and part of the object settled into orbit. Until recently the physical feasibility of the theory had never been tested through detailed calculations. To that end Cameron joined forces with Willy Benz and Wayne L. Slattery of the Los Alamos National Laboratory, who



MOON MAY HAVE BEEN FORMED by a collision between the earth and an object whose mass was about a seventh that of the earth. A computer simulation of the hypothesized collision is shown here; the time since the impact is given in hours. Both the earth and the impacting object are assumed to be molten and to consist of an iron core (*purple*) and a granite mantle (*red*). The heat of the impact vaporizes rock from both mantles (*blue*); volatile elements escape into space. The impacting object is shattered, and its iron core is separated from its mantle. The mantle debris then reaccretes to form the moon, while the iron core crashes back into the earth.

had developed a computer model capable of simulating a giant impact.

The workers assumed that both the impacting object and the earth were molten, and that in each case one-third of the mass was in an iron core and the rest in a rocky mantle. (Since the moon is roughly as old as the earth, the collision must have happened before the earth cooled and acquired a crust.) By varying the initial conditions, notably the mass of the impacting object, the workers calculated what type of collision, if any, would produce a moon.

When the mass of the object was between .12 and .15 times that of the earth—the ratio for Mars is about .11—the collision yielded not just a moon but a moon of the observed size, angular momentum and gross composition. In the computer simulation the collision shatters the impactor, and the rocky debris of the mantle subsequently becomes separated from the iron core. The gravitational fields of the three bodies—earth, impactor core and impactor mantle—conspire to keep the mantle debris in orbit. Within hours the gravitation of the debris itself draws it together into a roughly spherical moon.

Hence according to this scenario, the moon consists primarily of material from the impactor's mantle minus the volatile elements, which are vaporized by the tremendous heat of the impact and lost to space. The impactor core is sacrificed: it crashes back into the earth, and because it is heavy it soon settles to the earth's own iron core. If the scenario is correct, the primeval cataclysm that produced the moon also left its traces underground.

Time Deposits

How does a continent come to have inland deposits of oil and gas? Writing in *Geology,* Jack Oliver of Cornell University speculates that the thrust sheet formed by the edge of one continent when it collides with another may act as "a great squeegee," driving seawater into the crust. The seawater would carry organic matter, the precursor of oil and gas.

Oliver's hypothesis stems from an investigation he and his colleagues at Cornell made last year of the effects of a collision between Africa and North America. The collision, which took place about 300 million years ago, obliterated an ocean, known to geologists as Iapetus. It also deformed the North American crust, helping to produce the Appalachians. Roughly 110 million years later, when the continents rifted apart (creating the Atlantic), a piece of the African crust remained attached to North America. Today the piece is known as Florida.

The Cornell investigators found the suture that welds this African remnant to the rest of North America; it lies under southern Georgia. They also found something unexpected: the thumps produced by their seismic surveying apparatus indicated that the suturing had trapped pockets of fluids, perhaps from Iapetus, at depths as great as 18 kilometers.

Oliver proposes that great volumes of such fluids may have entered the North American crust at the time of the African collision. He notes that the collision caused a sheet or wedge at the forward margin of Africa to thrust westward over sediment at the edge of North America. The sediment must have been porous; indeed, as much as half of the sediment's volume could have been seawater from Iapetus. Moreover, the sediment must have been rich in hydrated minerals. The pressure of the overlying wedge could have forced fluid from the pores and the minerals into permeable North American strata. In turn, Oliver observes, the fluid might act as a lubricant and facilitate continued overthrusting. Some of the fluid might produce "veins and dikes and other features that characterize the metamorphic interior of mountain belts," in this case the Appalachians. Finally, some of the fluid might penetrate "well into the midcontinent," carrying "heat, minerals, and petroleum, or the ingredients for petroleum."

Evidence for the last of these possibilities is embodied, Oliver writes, in the patterns of distribution of U.S. oil and gas deposits. Immediately west of the Appalachians gas is commoner than oil; at greater distances oil is commoner than gas. Since gas forms at higher temperatures than oil, the pattern may have resulted from the temperature distribution of fluids squeegeed into the crust as the Appalachians were created.

A similar pattern can be discerned southwest of the Appalachians. There, at a time more recent than the collision of North America and Africa, a crustal body collided with the U.S. from the south. The identity of the body is uncertain. Nevertheless, the collision raised the Ouachita mountain belt, which winds through Texas, Oklahoma, Arkansas and Mississippi. Moreover, the arriving landmass must have encountered the equivalent of what the Mississippi River's drainage basin is today: a great delta of sediment pervaded by seawater and organic matter. North of the Ouachita belt lie the oil and gas fields of states such as Texas and Oklahoma. The boundary of this province nearest the line of the Ouachita belt "is rather abrupt and well defined"; the boundary farthest from the belt is an irregular "featheredge." It is the distribution one would expect for a fluid pushed by a squeegee.

Squeezed Light

One of the most profound yet counterintuitive laws of quantum theory is the uncertainty principle. It states that pairs of "complementary" variables, such as position and velocity, cannot be measured simultaneously in atomic systems. If, for instance, one determines the precise location of an electron, all information about the electron's velocity is lost. Conversely, knowledge of the electron's velocity can be obtained only at the expense of knowledge of its location.

Although a similar uncertainty relation is expected between the phase and the amplitude of a light wave, corroborating tests are more difficult. One such test has been done in recent months by Richard E. Slusher and his colleagues at the AT&T Bell Laboratories. They have successfully "squeezed" noise, or unwanted fluctuations, in phase at the cost of increasing it in amplitude. (Conversely, they have also squeezed noise in amplitude at the cost of increasing it in phase.) They describe their work in *Physical Review Letters.*

Slusher and his colleagues Leo W. Hollberg, Bernard Yurke, Jerome C. Mertz and John F. Valley generated the squeezed light by employing a technique known as four-wave mixing. They directed two laser beams at each other from opposite directions. The laser beams met in a material known as a nonlinear medium: a substance whose properties are affected by the characteristics of light impinging on it. In this case the nonlinear medium was a beam of sodium atoms oriented at right angles to the laser beams.

The laser beams interacted with the sodium atoms in such a way that two output beams emerged in opposite directions along an axis tilted at a slight angle with respect to the axis of the two input laser beams. Because of the special properties of the nonlinear medium, when the output beams were combined by mirrors, the result was a single beam of squeezed light. The noise in either the phase or the amplitude of the squeezed light can be reduced by varying the position of the mirrors that direct the beams emerging from the nonlinear medium.

Although so far the degree of success is small (the investigators report a 7 percent drop in noise), some physicists hope to reduce noise eventually by a factor of 10. If such a reduction is achieved, squeezed light would have a number of applications. It might be employed in ring-laser gyroscopes (detectors of rotational motion now found on board many aircraft), high-precision interferometers such as the ones designed to detect gravity waves and optical communication systems.

Tipping the Scale

Obesity tends to run in families. A study that may be the most convincing word to date has found that genes are clearly a culprit; the childhood family environment alone seems to have little effect.

The investigators, Albert J. Stunkard of the University of Pennsylvania and his collaborators in Copenhagen and Houston, report their work in the *New England Journal of Medicine.* They calculated fatness, as measured by a "body-mass index," in 540 Danish adults who were raised by adoptive parents. (The body-mass index is a person's weight in kilograms divided by the square of the person's height in meters.) The workers also calculated the fatness of the biological and adoptive parents of those 540 subjects and grouped the adoptees into one of four categories: thin, median weight, overweight or obese.

Stunkard and his colleagues found a strong relation between obesity in the adoptees and obesity in their natural parents. In contrast, they found no such relation between adoptees and their adoptive parents. In fact, the investigators found a strong relation between each of the four adoptee weight classes and the body-mass indexes of the natural, but not of the adoptive, parents. That is, just as obese biological parents tended to produce offspring who eventually grew obese, so did thin natural parents tend to have offspring who became thin adults.

While the new findings appear to underscore the strong influence of genetic inheritance on adult weight, some investigators question the study conclusions. They note that the body-mass index may not actually measure fatness; very muscular people who have little fat could have a high body-mass index.

Even if genetic endowment does strongly predispose people to obesity, many investigators (including Stunkard) believe it would be a mistake for obese individuals to forsake weight-control efforts. As one expert points out, obesity is not a single entity, and it may have a stronger genetic component in some people than in others. Stunkard and his colleagues themselves say their data do not indicate that the environment is without effect; the interaction of hereditary and environmental influences is not clear. Furthermore, Theodore B. Van Itallie, who wrote the *New England Journal's* editorial discussing Stunkard's report, notes that many people have been able to control obesity and that such control can aid in preventing or managing hypertension, hypercholesterolemia and non-insulin-dependent diabetes, among other diseases.

Van Itallie's prescription for obesity-prone individuals is to learn to maintain a comparatively high level of physical activity and to "eat defensively." He admits making major lifestyle changes is difficult. Nevertheless, he adds, "in the absence of safe and effective antiobesity medications, there are few, if any, satisfactory alternatives."

Brand New

Much as ranchers have traditionally branded their animals, a Houston-based biotechnology company has begun to label its live-virus vaccines. In the labeling procedure, which the company calls DNA branding, workers insert the name of the company—NovaGene—into the genetic material of the weakened virus that constitutes a given vaccine. The first product to undergo such branding is directed, appropriately enough, against a disease of animals that have long been subjected to branding themselves: cattle. This disease, known as infectious bovine rhinotracheitis (IBR), is a respiratory ailment caused by a herpes virus.

Genetic engineers construct the IBR brand by assembling bits of DNA that code for amino acids whose standard single-letter symbols approximate the spelling of "NovaGene." The letter *N* represents the amino acid asparagine; therefore the DNA code for asparagine starts the brand. This code is followed, in order, by the codes for glutamine (symbolized by a *Q;* no amino acid is represented by an *O*), valine (*V*), alanine (*A*), glycine (*G*), glutamic acid (*E*), asparagine (*N*) and glutamic acid (*E*). Once the brand is complete, workers splice it into the DNA of the vaccine virus, where the brand becomes part of the virus's genetic material and replicates whenever the virus replicates.

According to Malon Kit of NovaGene, who invented DNA branding, the vaccine maker applies two strategies to ensure that the brand is inactive and therefore will not affect the performance or the safety of the vaccine. The brand is positioned away from start codes, the DNA segments that initiate gene expression. Without such codes nearby the brand goes unexpressed; that is, the encoded amino acids are not produced. In addition the brand is kept small, and so if it were somehow expressed, the resulting protein (amino acid chain) would in all likelihood be too small to function.

If the brand is not expressed, how is it detected in an animal that has been given the vaccine? When detection is necessary, workers isolate virus from tissue and expose it to a DNA probe that binds to the brand.

Kit expects DNA branding to be a boon to manufacturers of pharmaceuticals because it will unambiguously indicate whether a given manufacturer produced a vaccine. For instance, imagine that a cattle owner sues a manufacturer, contending that the company's vaccine causes postvaccination complications. If the brand is absent from virus isolated from the affected animals, the court will know that the illness was caused by an unauthorized duplication of the vaccine or by some other infectious agent. Branding may also provide regulators with an "audit trail" by which they can monitor the usage and effectiveness of a vaccine in a given herd. The branded IBR vaccine is not yet available commercially. It awaits U.S. authorization for field-testing.

Viewing Vessels

Images of a patient's circulatory system, called angiograms, are often necessary in the detection and treatment of occlusive arterial diseases such as atherosclerosis. Radiologists can obtain such images using conventional X-ray techniques by injecting a liquid containing iodine, which absorbs X rays, into the circulation through a catheter. Such a procedure entails some risk for the patient, particularly if it is the coronary arteries that are to be made visible. The procedure can precipitate arrythmias, hemorrhage and toxic reactions.

Two new ways to visualize blood vessels now promise to reduce, if not entirely eliminate, the drawbacks associated with conventional X-ray angiography. One method, developed by E. Barrie Hughes, Edward Rubenstein, Robert Hofstadter and their colleagues at Stanford University, is an extension of what is called digital-subtraction angiography. In this type of angiography the patient is X-rayed twice after the iodine-based liquid has been injected into the circulation. The two exposures, taken a few thousandths of a second apart in time, differ in the energy of the illuminating X-radiation: one is made at an energy slightly above and one at an energy slightly below the X-ray energy maximally absorbed by iodine. The intensity of the X rays passing through the patient are digitally recorded for each exposure and then subtracted from each other. In this way virtually all the obscuring nonvascular tissue can be "subtracted out" and thereby rendered invisible in the digitally reconstructed image.

The investigators found that X-radiation of the appropriate energies and intensities for their technique could best be generated at the Stanford Synchrotron Radiation Laboratory. Synchrotron radiation is intense electromagnetic radiation emitted by charged particles (electrons in this case) that are magnetically forced into curved trajectories while traveling at a speed close to that of light. Two thin, sheetlike beams of "tuned" X rays suitable for scanning a patient can be readily filtered out from such radiation.

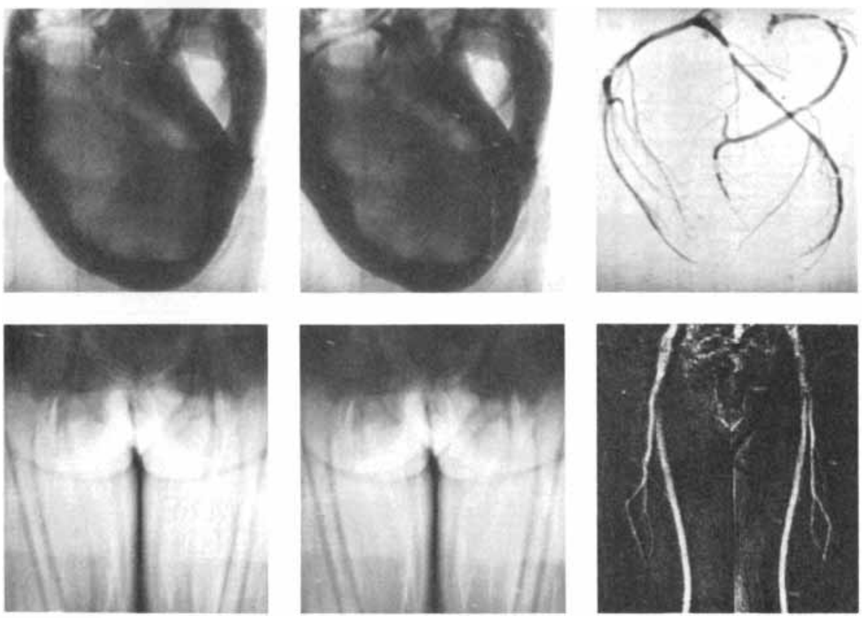
The investigators have achieved extraordinary sensitivity with their technique. Angiograms can be produced with an iodine concentration in the blood far less than that required by conventional methods. Indeed, an intravenous injection of the contrast agent, in place of the usual arterial catheterization, is sufficient to ensure that enough iodine is in the blood perfusing the arteries to be imaged.

Another way to visualize blood vessels, developed by Van J. Wedeen of the Massachusetts General Hospital and his colleagues, is currently more widely applicable. The method makes use of a commercially available magnetic-resonance imaging system and dispenses completely with the need for an injected contrast agent.

When nuclei of hydrogen atoms are subjected to a burst of radio waves while they are under the influence of a strong magnetic field, they emit a characteristic signal known as the magnetic-resonance signal. The amplitude, frequency and phase of the signal are affected by, among other things, the gross motion of the material incorporating the resonating nuclei. Hence magnetic-resonance imaging systems can distinguish vessels in which blood is flowing.

Because conventional magnetic-resonance images reveal the vessels as sectioned structures within a thin slice of the body, Wedeen and his colleagues had to reprogram a magnetic-resonance imaging system so that the three-dimensional structure of the body could be projected onto an image plane. In such an image, however, the various resonance signals from the tissues in the projected volume obscured the signals that are characteristic of flowing blood.

The investigators therefore resorted to much the same tactic as the one employed in digital-subtraction angiography to remedy the problem. They recorded the velocity-sensitive magnetic-resonance response signal emitted as the ventricles of the heart are contracting, so that arterial flow velocity is high, and subtracted them from the signal recorded when the heart cavities are filling with blood and arterial flow velocity is low. The stationary foreground and background tissues are thereby canceled, leaving an image only of those structures in which a change in the velocity of blood flow is detectable in the course of a cardiac cycle: the arteries.



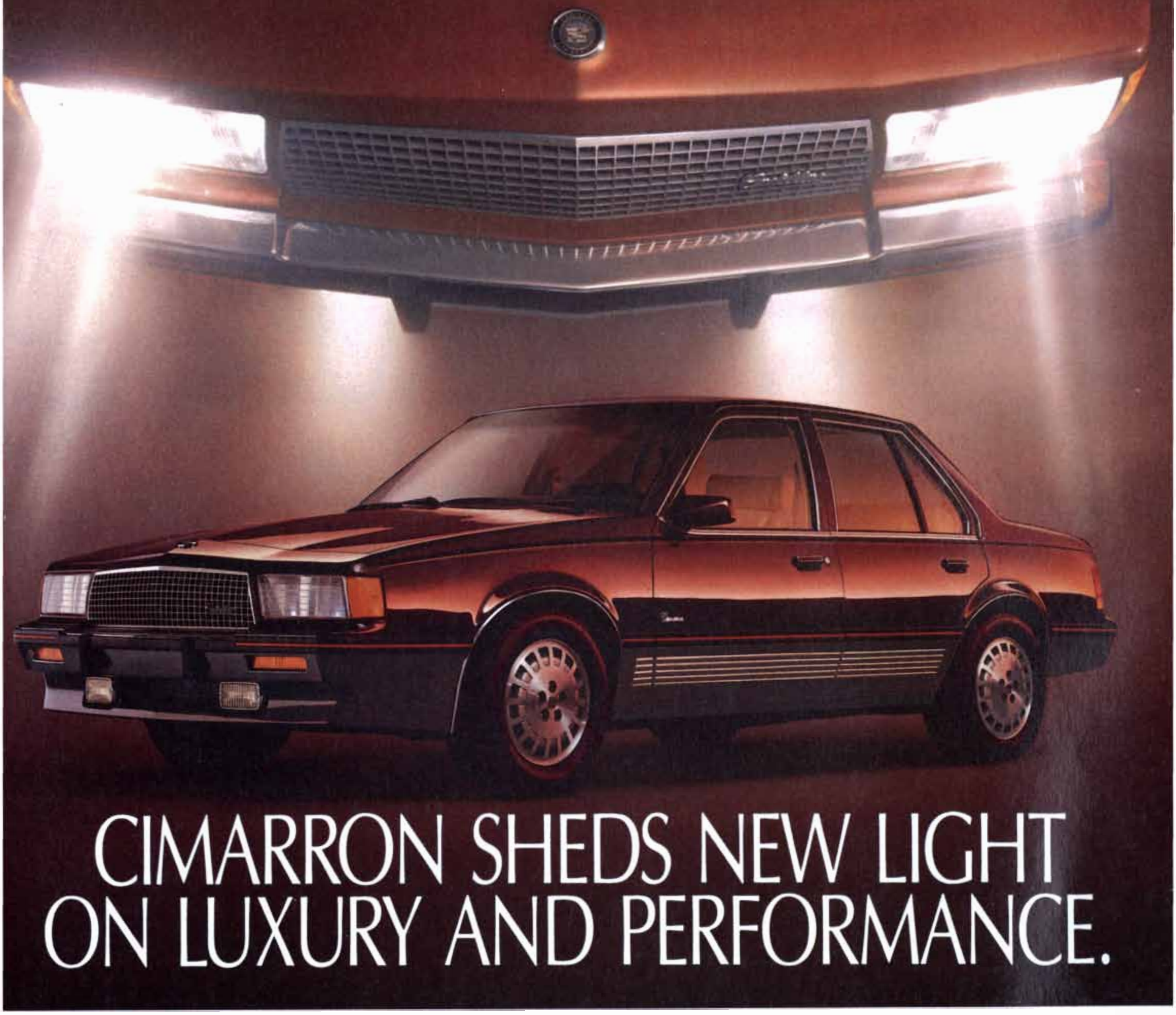
BLOOD VESSELS can be rendered clearly visible by two imaging methods that "subtract out" obscuring nonvascular tissue. In one technique (*top row*) a pair of X rays (*left and middle*) that differ slightly in the energy of the illuminating radiation are subtracted from each other to reveal (*right*) the coronary arteries of an in vitro pig heart. In the other imaging method (*bottom row*) magnetic-resonance signals from human thighs are recorded at systole, when blood-flow velocity is high (*left*), and at diastole, when it is low (*middle*). The difference between the two images (*right*) reveals the femoral arteries as bright tracks.

Guilty Looks

Can a judge's tone of voice or behavior influence the jury in a criminal trial? It can, according to a law student, a professor of psychology and a municipal-court judge who looked into the question. Even when a judge does not say anything prejudicial, his or her behavior can signal a point of view to jurors, the investigators found. They recently reported their findings in the *Stanford Law Review*.

The authors of the study are Peter David Blanck, president of the law review, Robert Rosenthal of the Department of Psychology and Social Relations at Harvard University and LaDoris Hazzard Cordell, the presiding judge of the Municipal Court in Santa Clara County, Calif. They assembled data by giving questionnaires to five judges, 331 jurors and 61 prosecution and defense lawyers who took part in 34 criminal-misdemeanor trials in California municipal courts. In addition they made videotape recordings of each judge as he or she gave instructions to the jury in the course of several trials. Eighty Stanford students were employed to rate the taped data according to several verbal and nonverbal criteria.

One finding was that a judge was more likely to expect a guilty verdict when the defendant already had a serious criminal record or was of low socioeconomic status. The judge may reveal his expectation through his tone of voice, facial expressions, attitude toward prosecuting or defending attorneys or his involvement in details of the trial. Older judges were rated as being more likely than younger ones and females as being more likely than males to reveal their attitudes. Raters thought judges' nonverbal behavior in the process of delivering instructions to the jury tended to be more professional with defendants of a higher socioeconomic status or with older and better-educated jurors. "Future researchers," the authors say, "should address the question of what kind of judges, in terms of communicative skills and perhaps other personality variables, are prone to influence prejudicially what kind of jurors and other trial participants."



EXCEPTIONAL VALUE.

The first thing you'll notice about the Cimarron is its distinctive new front end, redesigned with composite headlamps and wraparound turn indicators and parking lamps.

But this new look is a mere reflection of the Cimarron. Look closer and you'll discover a luxury car. And a performance car. All designed with an eye toward exceptional value.

CADILLAC LUXURY.

Cimarron's interior is complete with dozens of unexpected comfort and convenience features ... impressive even by Cadillac standards. Like genuine Sierra Grain leather seating areas. Power six-way driver's seat. Cruise control. Air conditioning. And just about every imaginable power assist. All standard.

RESPONSIVE PERFORMANCE.

Cimarron is designed to be agile and fun to drive. For even greater performance, you can specify an optional robust 2.8 liter V6 engine with multi-port fuel injection and a fully gas-charged suspension.

IT ALL ADDS UP.

Distinctive styling. Cadillac luxury. Eye-opening performance. What other car in its class has so much value to offer?

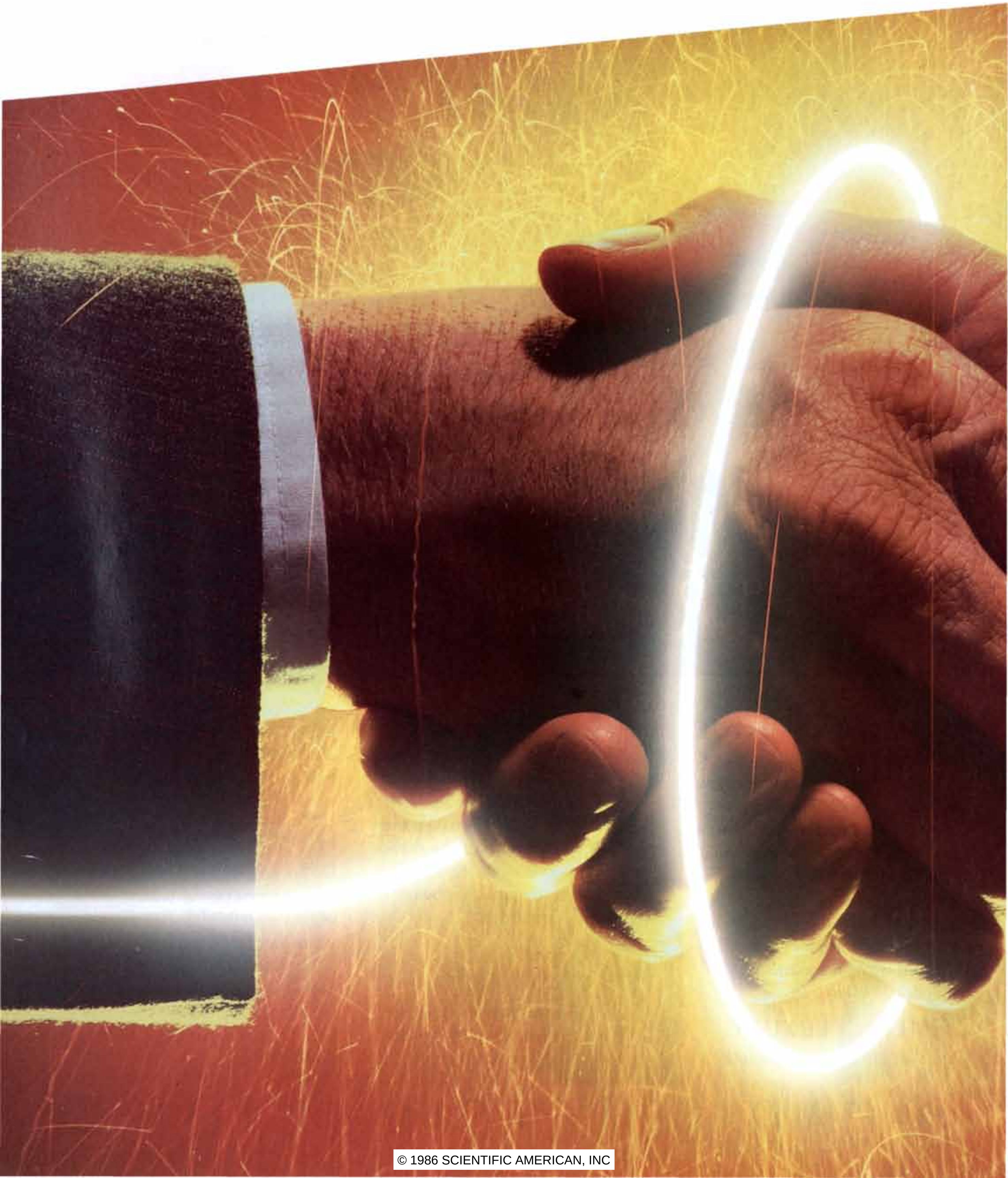
But don't take our word for it. Test drive a Cimarron at your Cadillac dealer today. We think you'll find the experience very illuminating.

CADILLAC • CIMARRON

GM

Let's Get It Together...Buckle Up.

Now business



can feel positive again about communications.

We've made the complex world of communications easier to deal with. We're the companies of Ameritech.

Nobody understands communications better. We send more messages to more people in more ways than any other company in our region.

Nobody offers a broader scope of services. There's virtually nothing we can't provide for the movement of information. Everything from fiber optics and digital technology to Yellow Pages directories that reach millions of readers each day.

You can count on us to meet your future needs as well. Nobody is moving faster than Ameritech to bring more and newer communications services to our region. Nobody is making a bigger investment to accomplish that purpose. Some $3.7 billion during the last two years.

We're the Bell companies of Illinois, Indiana, Michigan, Ohio and Wisconsin. We also serve you through leading companies in: Cellular mobile phone service. Voice and data products and systems. Computer software. Lease financing. Directory publishing. And technology development.

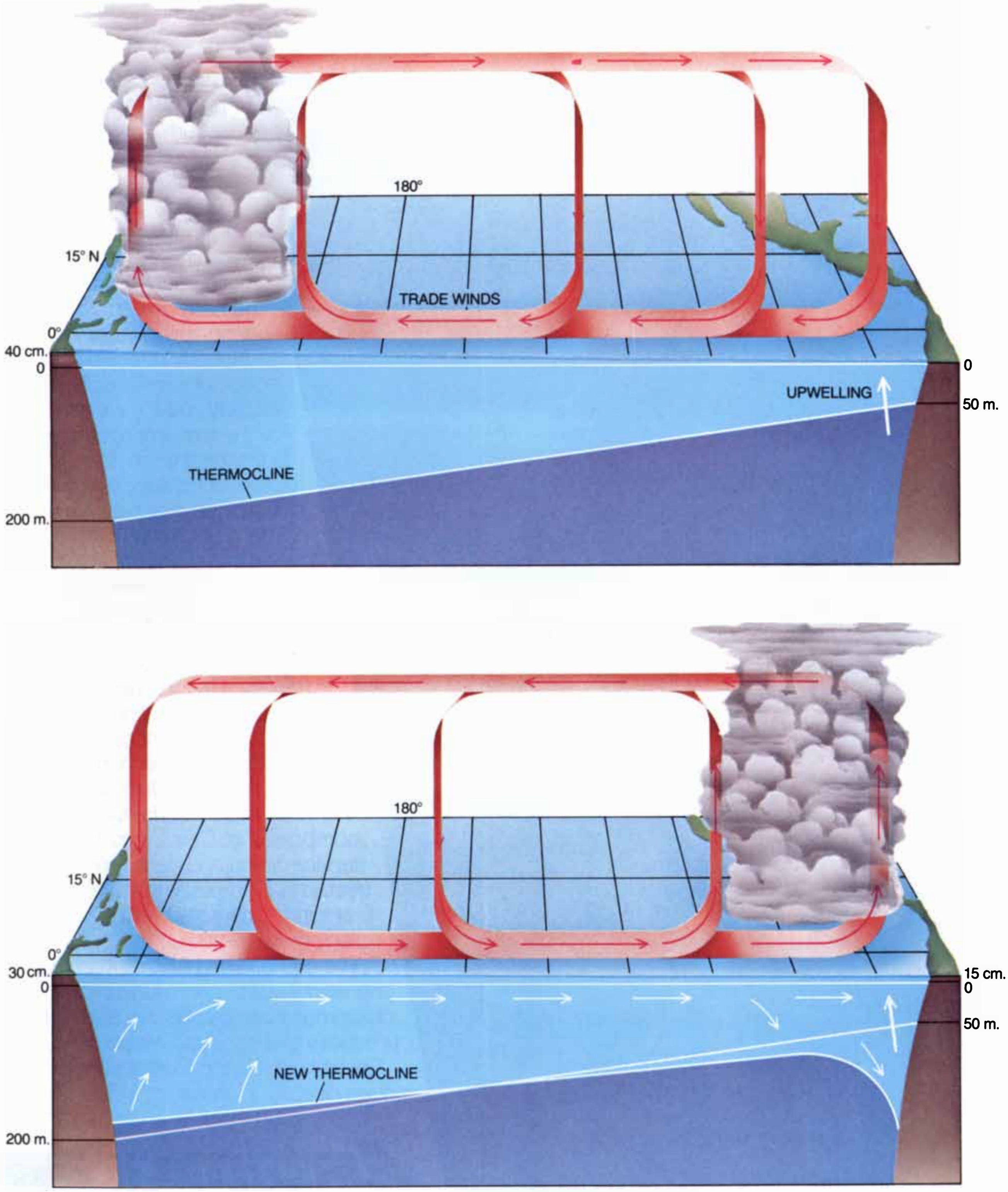
The companies of Ameritech are taking the worry out of communications. Our customers can be positive about the products and services we provide.

Let us help you, too. Contact any of our companies directly or phone us: 1 800 562-2444.

AMERITECH

AMERICAN INFORMATION TECHNOLOGIES

Helping you communicate℠



EL NIÑO is a manifestation of the Southern Oscillation, an atmospheric-pressure seesaw between a high-pressure center in the southeastern Pacific and a low-pressure center over Indonesia and northern Australia. Under ordinary conditions (*top*) the pressure difference between these centers drives easterly trade winds along the Equator. The winds pile up warm water and raise the sea level in the western Pacific by about 40 centimeters. They also depress the thermocline, the boundary between warm surface water and the underlying cool layers, to a depth of about 200 meters. Off South America, where the trade winds drive surface water offshore, the thermocline is shallow, and cool water wells up to the surface. Near Indonesia the trade winds converge with westerly winds, leading to rising air and heavy rain. The air flows eastward at high altitudes and sinks in the central and eastern Pacific, where the weather is dry. During El Niño (*bottom*) the east-west pressure difference becomes so low that the trade winds collapse in the western Pacific. The warm water piled there flows back toward the east; at the same time subsurface waves called Kelvin waves travel across the Pacific and depress the thermocline off South America, where the upwelling water becomes warm. Both effects warm the sea surface. During the most recent Niño (1982–83), the severest in a century, the wind directions and hence the weather pattern were completely reversed.

El Niño

This anomalous warming of the sea surface in the equatorial Pacific is associated with a vast fluctuation in atmospheric pressure. Global changes in weather patterns are linked to the equatorial anomalies

by Colin S. Ramage

Of all the environmental sciences meteorology has the most immediate impact on society. Conversely, the social demand for weather forecasts exercises a guiding influence on meteorological research. An example is the field of seasonal forecasting, in which the public has an intense interest but in which the results so far have been no better than what one would expect from educated guesswork. The failure is not surprising. Given that weather events such as the arrival of a storm can be predicted only three to five days in advance, why should one expect to predict a seasonal weather pattern, which is just the product of individual events, months in advance? Yet the demand for such prophecies is great, and so forecasters persevere.

In doing so they look for ways of bypassing their inability to make long-range forecasts of individual weather events. In particular they look for large-scale, even global, changes in the atmospheric circulation that influence the weather on a regional scale. The assumption is that if the precursors and effects of large-scale changes were understood, they could serve as rough guides to the meteorological future.

By far the most prominent (and promising) guide yet observed is the phenomenon known as El Niño, an anomalous warming of surface water in the equatorial Pacific. El Niño occurs at irregular intervals in conjunction with the Southern Oscillation, a massive seesawing of atmospheric pressure between the southeastern and the western tropical Pacific. Both phenomena have been known to oceanographers and meteorologists for decades. Since they occur in one of the world's most sparsely inhabited regions, however, they were for long of interest to only a few investigators.

The phenomena would have remained obscure had it not been discovered that they are linked to weather in other regions, including the U.S. During the Niño–Southern Oscillation event of 1982–83, for example, California was beset by floods, while the drought in Africa intensified. The discovery of this global connection implied that the oceanic and atmospheric anomalies of the equatorial Pacific might be the key to accurate seasonal weather forecasts elsewhere. Although the promise has not yet been fulfilled, the pressure to attempt forecasts has already led to progress in understanding the anomalies.

Anomalies Defined

For more than a century the name El Niño, the Spanish term for the Christ child, has been applied by fishermen to the annual appearance at Christmastime of warm water off the coast of Ecuador and northern Peru. Ordinarily the ocean surface there is rather cool compared with typical equatorial water; it is kept that way by the northward-moving Peru Current, which drives surface water offshore, so that cool water wells up from the depths. The cool water is rich in nutrients, mainly phosphates and nitrates, that nourish photosynthesizing plankton and that ultimately support the Peruvian anchovy fishery, the largest fishery in the world. When at Christmastime a warm southward current displaces the cool water, reducing the upwelling of nutrients, fishing is disrupted, but only slightly. The warming extends no farther south than northernmost Peru, and it usually ends by March or April.

Occasionally, however, El Niño is much more intense, extensive and prolonged. Rather than dropping back to normal in March or April, sea-surface temperatures rise all along the coast of Peru and in the eastern and central equatorial Pacific. The temperatures may stay high for more than a year. Such relatively intense Niños were observed in 1953, 1957–58, 1965, 1972–73, 1976–77 and, most recently, in 1982–83, when the sea surface off Peru warmed by more than seven degrees Celsius. The cumulative impact on the anchovy fishery of the events since 1972 has been catastrophic: the annual catch fell from a peak of more than 12 million metric tons in 1970 to less than half a million tons in 1983. In scientific circles the term El Niño is now generally applied only to these intense events rather than to the annual, relatively mild warming (between one and two degrees C.) of the sea surface.

The first major step toward understanding El Niño was taken in 1966 by Jacob Bjerknes of the University of California at Los Angeles, who noted that the anomalous warming of the sea is associated with the Southern Oscillation. The Southern Oscillation, first observed in 1924 by Sir Gilbert Walker, is a transpacific linkage of atmospheric pressure systems. When pressure rises in the high-pressure system centered on Easter Island, it falls in the low-pressure system over Indonesia and northern Australia, and vice versa. To quantify the phenomenon Walker defined the Southern Oscillation index, which is calculated by subtracting pressure in the western Pacific from pressure in the eastern Pacific. The index is positive when the difference between east and west is higher than normal and negative when the difference is lower than normal.

The causes of the Southern Oscillation are not known. What makes it interesting from a seasonal-forecasting viewpoint are the marked fluctuations it undergoes over periods of years, fluctuations that are associated with widespread departures from normal conditions in temperature and precipitation. For example, Walker himself had noted that the summer monsoon rains in India tend to fail when the Southern Oscillation index is low and to be plentiful when it is high. Four decades later Bjerknes realized that El Niño too is associated with a low index, beginning as the index falls from high values and reaching a climax when the index is lowest. Thus the 1972–73 Niño that devastated the Peruvian anchovy fishery came at a time when the index had fallen to one of its lowest values ever, and it was accompanied by a severe drought in India. Moreover, drought also struck the Soviet Union, New Guinea and Hawaii, while Peru, the Philippines and California were hit by heavy floods.

Clearly the effects of a Niño–Southern Oscillation event (which for the sake of brevity I shall simply call a Niño) are felt well beyond the equatorial Pacific. Because the 1972–73 Niño was widely recognized soon after its onset, investigators were encouraged to think future Niños and their global weather effects might prove to be predictable some months in advance.

The Canonical Niño

By the mid-1970's a generally accepted view of the sequence of events preceding and accompanying El Niño had emerged. The sequence, which came to be known as the canonical Niño, was based on Bjerknes' description of the Southern Oscillation and on a model of the oceanic circulation developed by Klaus Wyrtki of the University of Hawaii at Honolulu. The models themselves were derived from monthly average values of variables such as atmospheric pressure, wind speed and direction and sea-surface temperature, measured at various places in the Pacific during the Niños of 1957–58, 1965 and 1972–73.

The canonical description of El Niño focuses on the behavior of the trade winds over the tropical Pacific. In the central and eastern Pacific the trade winds form part of the circulation around two persistent high-pressure systems, the South Pacific high centered near Easter Island and the North Pacific high centered off California. Around high-pressure centers the circulation is anticyclonic: clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere. Accordingly the trade winds blow toward the Equator along the coasts of North and South America. The northeast and southeast trade winds meet in the so-called intertropical convergence zone, which generally moves between the latitudes of four degrees north in April and eight degrees north in September. Because El Niño begins near the Equator (for reasons that are not entirely clear), the southeast trade winds are particularly important in the Niño cycle.

Along the coast of South America the southeast trades drive the Peru Current, pushing surface water offshore and allowing cold, nutrient-rich water to well up to the surface. The westward push of the trade winds continues across the eastern and central Pacific. According to Wyrtki's model, the resulting stress on the sea surface raises the sea level in the western Pacific. In effect water piles up in the west. The piled-up water deepens the warm surface layer of the ocean; put another way, it depresses the thermocline, the boundary between the well-mixed surface layer and the colder, deeper layers. (The thermocline is actually not a sharp boundary but a stable zone in which the vertical temperature gradient is large.) Whereas off the South American coast the thermocline begins at a depth of only 50 meters—which explains why the upwelling water is cold—in the western Pacific it begins at a depth of about 200 meters.

The southeast trade winds themselves are ultimately driven by the pressure gradient between the South Pacific high-pressure system and the low-pressure system over Indonesia and Australia. Hence the Southern Oscillation index, which charts the pressure difference between those systems, is also a measure of the strength of the winds. When the index is high, the pressure gradient is large and the winds are intense.

El Niño is heralded by a precipitous drop in the index and a corresponding collapse of the trade winds in the western Pacific. The collapse generally begins in October or thereabouts. No longer supported by the winds, the warm water piled in the western Pacific flows back toward the east, and the sea level rises east of the International Dateline (which near the Equator follows the 180-degree longitude line). The flow takes the form of subsurface waves known as Kelvin waves that travel along the Equator, reaching the coast of South America in two to three months. The Kelvin waves have two effects: they generate anomalous eastward currents and they depress the thermocline.

Both effects tend to warm the sea surface, the first by physically bringing in warm water from the west and the second by preventing the upwelling of cool water from within or below the thermocline. The latter effect is the more important one, and it is particularly significant off South America, where the thermocline is ordinarily shallow. The sea surface there begins to warm in December or January, when the first Kelvin waves reach the coast. At this point it is usually not yet clear whether the warming is simply part of the normal annual cycle or

HISTORICAL RECORDS of the Southern Oscillation index (*top*) and of the sea-surface temperature off the Peruvian coast (*bottom*) show the two variables are linked, although the association is not perfect. The Southern Oscillation index is calculated by subtracting the atmospheric surface pressure at Darwin, Australia, from the pressure at Easter Island. When the index is positive, the east-west pressure gradient is greater than normal and the trade winds are stronger; conversely, a negative index indicates weak trade winds. The graph shown here is based on monthly averages of the surface pressure and has been statis-

whether a true Niño is getting started.

As a Niño develops, the trade winds continue to relax near Indonesia and are eventually replaced by surface westerlies. The westerlies trigger intensified Kelvin waves that further depress the thermocline off South America. Although the southeasterly trade winds along the South American coast do not collapse and thus continue to drive upwelling, the upwelling water is now warm (and nutrient-poor) rather than cold. Consequently the westward current off equatorial South America not only is weakened by the eastward push of the Kelvin waves but also is much warmer than before. The warming of the sea surface therefore begins to spread to the west along the Equator. In the eastern and central Pacific, where the surface is ordinarily also cooled by wind-driven upwelling, the warming is intensified by an actual diminution of upwelling, because the trade winds there do become significantly weaker.

Indeed, the normal wind pattern along the Equator can reverse itself completely during El Niño. Normally when the easterly trades reach the Equator, they blow along it, carrying progressively warmer and moister air toward the low-pressure system over Indonesia. There they converge with surface westerlies. The warm air rises, the moisture condenses and heavy rains fall. The rain-depleted air flows eastward in the upper troposphere (at altitudes of between nine and 12 kilometers), cools and then sinks over the central and eastern equatorial Pacific, where the weather is generally sunny and dry.

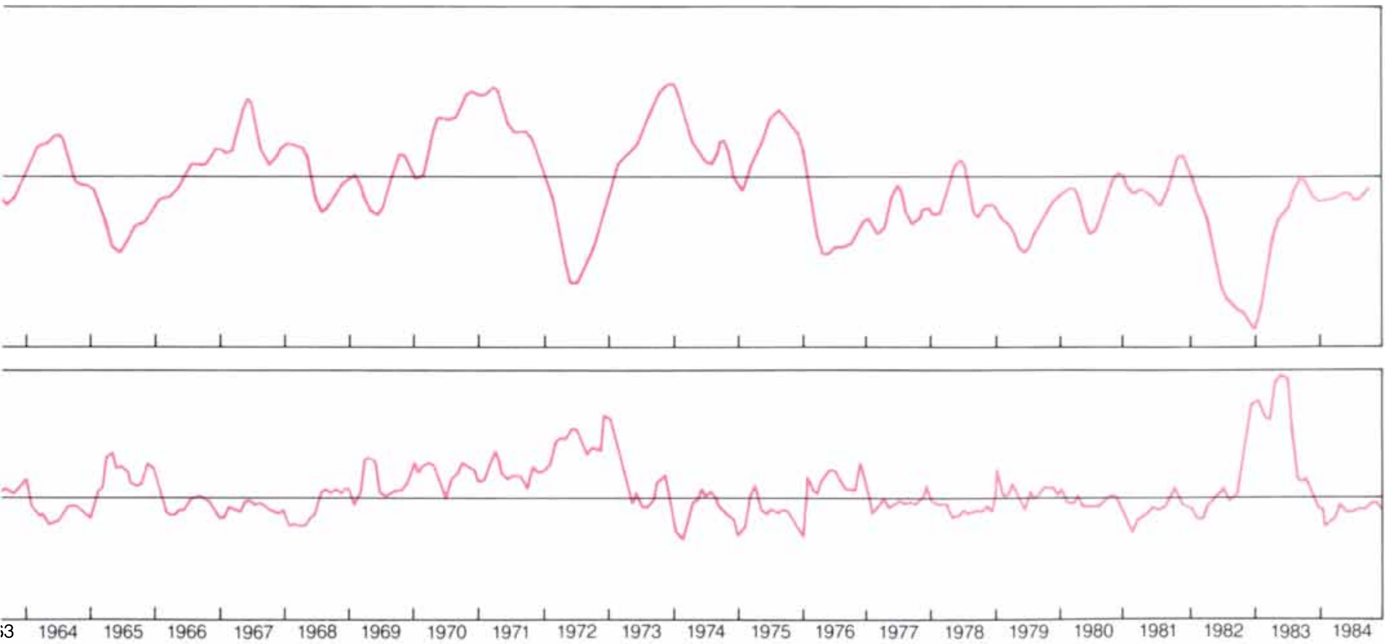
According to the Wyrtki model, the flow direction in this circulation cell is reversed during El Niño by a positive-feedback interaction between the atmosphere and the sea surface. The surface westerlies that develop east of Indonesia trigger Kelvin waves that warm the central Pacific; because air rises over warm water, the anomalously warm sea surface pulls the rising branch of the circulation cell eastward. The westerlies follow behind, intensifying and triggering more Kelvin waves. Ultimately the rising branch shifts to the central and eastern Pacific, producing heavy rainfall in that normally dry region. (In the extreme Niño of 1982–83 the region of rising air extended all the way to the South American coast.) Air flows to the west rather than to the east in the upper troposphere and sinks over Indonesia; as a result the weather there becomes unusually dry.

From June to August, El Niño temporarily relaxes, and the sea surface off South America cools two or three months later. Then, toward the end of the year, yet another warming occurs. Soon after that the Southern Oscillation index rises and the easterly trade winds freshen. By March or April, about 15 months after El Niño began, conditions in the equatorial Pacific have largely returned to normal.

Predictions

Although every Niño has differed from the canonical sequence, those studied most closely seemed at first to share enough similarities and precursory signs to render future Niños predictable. In particular, Wyrtki suggested that the Southern Oscillation index had to rise and the trade winds had to intensify, piling up more water in the western Pacific, before a collapse of the winds could trigger El Niño. Wyrtki's model was based largely on the 1972–73 Niño, which did indeed meet the preconditions he specified.

The idea that stronger than normal trade winds were a reliable predictor of El Niño gained further support from the work of Eugene M. Rasmusson and T. H. Carpenter of the National Oceanic and Atmospheric Administration (NOAA) Climate Analysis Center, who prepared composite anomaly charts for all the Niños between 1949 and 1973. The workers assumed that by averaging together observations for comparable periods of different Niños, the common precursory signals could be made to stand out more sharply against the background noise of random differences between the

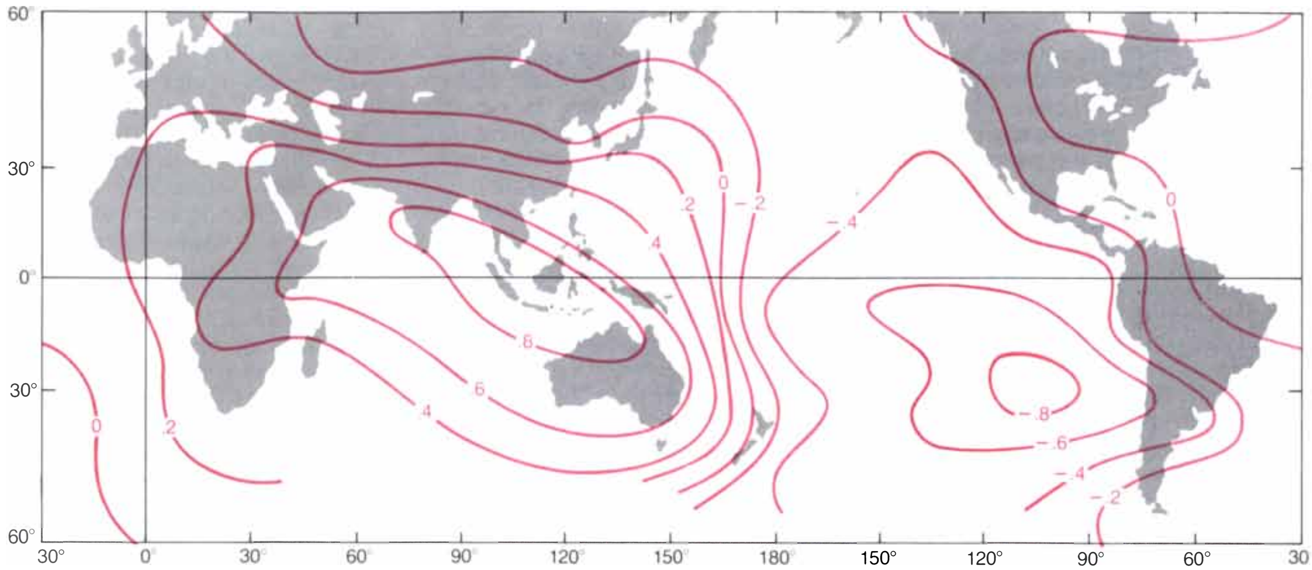


tically "smoothed." The graph of the sea-surface temperature at Puerto Chicama, Peru (7.7 degrees south), shows the departure from normal monthly average temperatures. According to an analysis of these two and other variables by William H. Quinn of Oregon State University, strong Niños occurred in 1957–58, 1972–73 and 1982–83. Moderate Niños occurred in 1953, 1965 and 1976–77. Except for the late-starting 1982 event, the anomalous warming of the sea surface off Peru has tended to begin around March, after the Southern Oscillation index has fallen from positive values and the trade winds have slackened. Late in the year the sea surface cools; then the warming resumes, climaxing early the following year. Some Niños, such as that of 1972–73, were preceded by an intensification of the trade winds. A wind buildup in 1974, however, was not followed by a Niño, and the 1982 event began without a buildup.

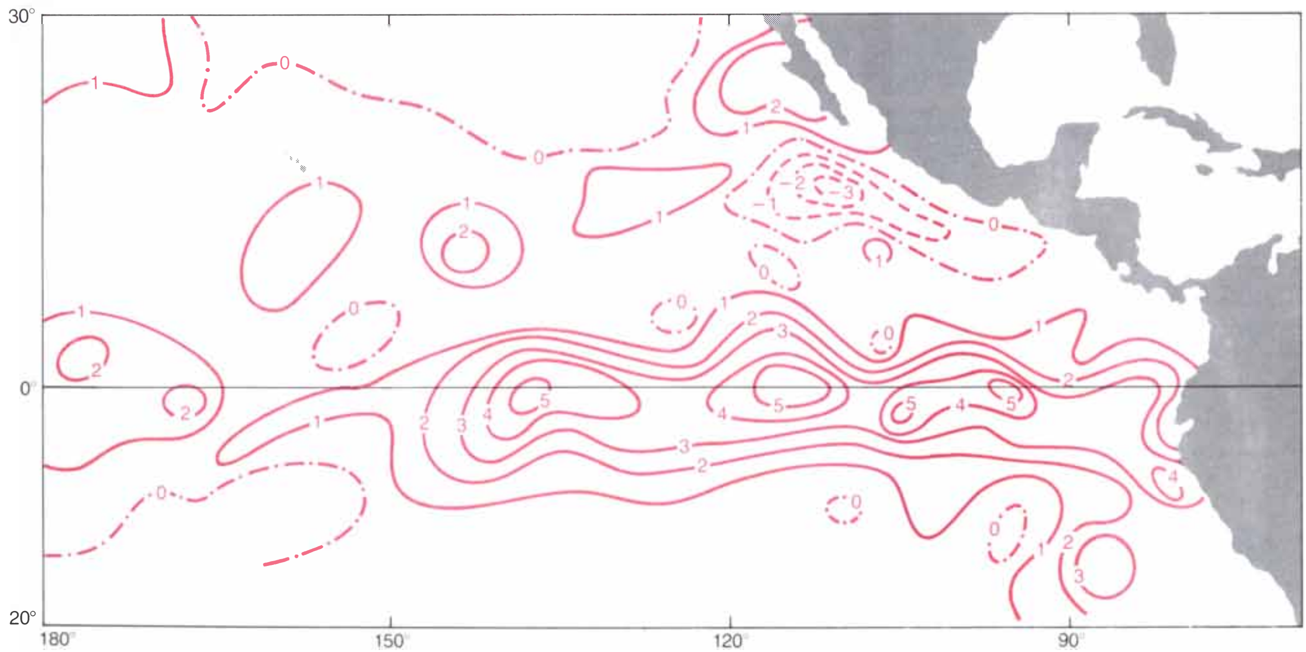
events. Strong trade winds in the western Pacific in the few months before a Niño began emerged as one of the most significant precursors.

Yet even before Rasmusson and Carpenter had developed their composite charts, the Niño of 1976–77—or rather, the absence of a Niño in 1975—had downgraded trade-wind intensification from a sufficient condition to a necessary one for El Niño. In 1974 the Southern Oscillation index rose and the trade winds strengthened. Accordingly a Niño was expected in 1975, but no anomaly developed. Instead, after an initial dip, the index rose again. Finally it fell in 1976 and there was a moderate Niño.

The next Niño, in 1982–83, was even more damaging to the reputation of strong trade winds as a predictor. After 1976 the thermocline remained depressed off the South American coast; as a result nutrients remained scanty near the sea surface and the anchovy population failed to rebound. At the same time the Southern Oscillation index hovered around normal. In other words, the trade winds did not intensify, which implied, according to the Wyrtki model, that no Niño was in the offing. Consequently all observers were surprised when the index began



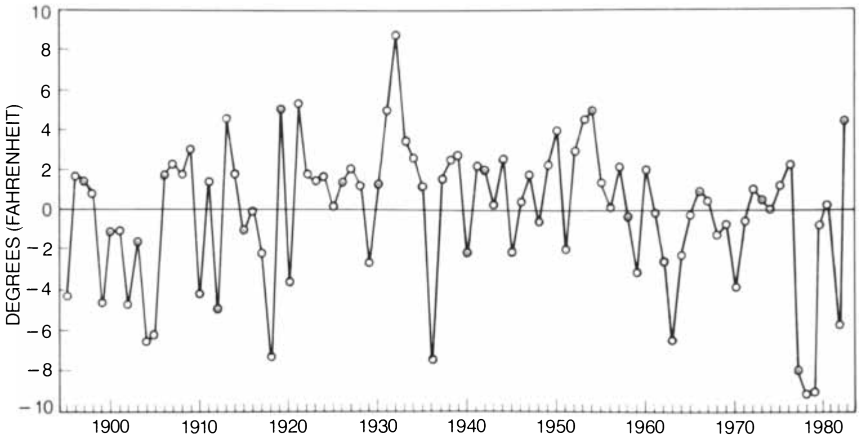
GLOBAL ATMOSPHERIC INFLUENCE of the Southern Oscillation is evident on this map showing the correlation between the mean annual surface pressure at Jakarta, Indonesia, and the pressure elsewhere. A value of 1 or –1 in a given region would be a perfect correlation: the pressure in that region would always change by the same relative amount as the pressure at Jakarta. The pressure near Indonesia shows a strong negative correlation with the pressure near Easter Island, which is at the other end of the seesaw.



SEA-SURFACE WARMING during El Niño is concentrated in the central and eastern equatorial Pacific. The map shows the difference in sea-surface temperature (in degrees Celsius) between August, 1972 (a Niño year), and August, 1979 (a normal year). This particular comparison is instructive because it illustrates the difficulty in basing weather forecasts on El Niño. In the areas where the largest temperature differences occurred rain was scanty in both the Niño and the non-Niño month, even though forecasting models predict heavy rainfall over an anomalously warm sea surface. Global weather patterns were also largely the same in the two months.

to plummet in March, 1982. By June it had reached record low levels, and the greatest Niño in at least a century was under way. Disastrous floods struck the Pacific coast of the Americas, a drought settled over Australia and the drought in the Sahel in Africa grew even worse. Meanwhile the islands of French Polynesia in the central South Pacific suffered record-breaking rains. Whereas Polynesia is usually hit by a tropical cyclone only about once every three years, in 1983 no fewer than six cyclones ravaged the islands between February and mid-April.

The disaster of 1982–83 showed that in addition to being an insufficient condition for El Niño, strong trade winds are not even a necessary condition. The attempts to predict El Niño on the basis of precursory signals failed, it is now clear, because they relied on a set of data that did not represent all possible Niños. The extraordinarily intense 1982–83 event may not have been entirely without precedent; George Kiladis and Henry Diaz of the NOAA/University of Colorado Cooperative Institute for Research in Environmental Sciences have found evidence that the Niño of 1877–78 closely resembled the one of 1982–83. Unfortunately the data from that long ago are too meager to have been incorporated into composite anomaly charts or other detailed statistics.



WINTER TEMPERATURES IN ILLINOIS reveal no sign of a link with El Niño. The graph shows the departure of the average temperature in a given winter from the long-term mean. El Niño years (*colored dots*) have been marked by both cold (1977) and warm (1983) winters. Most of the significantly anomalous winters have occurred in non-Niño years.

Uncanonical Nature

The failure to predict El Niño underlines the current lack of understanding of how the anomaly develops. What was once thought to be a rather straightforward cycle, consisting in each case of the same sequence of events, turns out to be a highly variable phenomenon. The Wyrtki model assumed that the Pacific ocean-atmosphere system generally occupied one of only two stable states, the first represented by El Niño and a low Southern Oscillation index, the second represented by an "anti-Niño" and a high index. Once El Niño began, it was thought, a positive feedback between ocean and atmosphere reinforced and prolonged the cycle; once the cycle was over the same type of interaction accelerated the return to anti-Niño conditions. In recent years the distinction between the two states has become blurred as a spectrum of states has been discovered between them. For example, summer drought in India and above-normal rainfall in the central Pacific are regarded as hallmarks of El Niño, but the severe Indian drought of 1979 and the record monthly rainfall at Canton Island (2.8 degrees south, 171.7 degrees west) in December, 1977, both took place in non-Niño years.

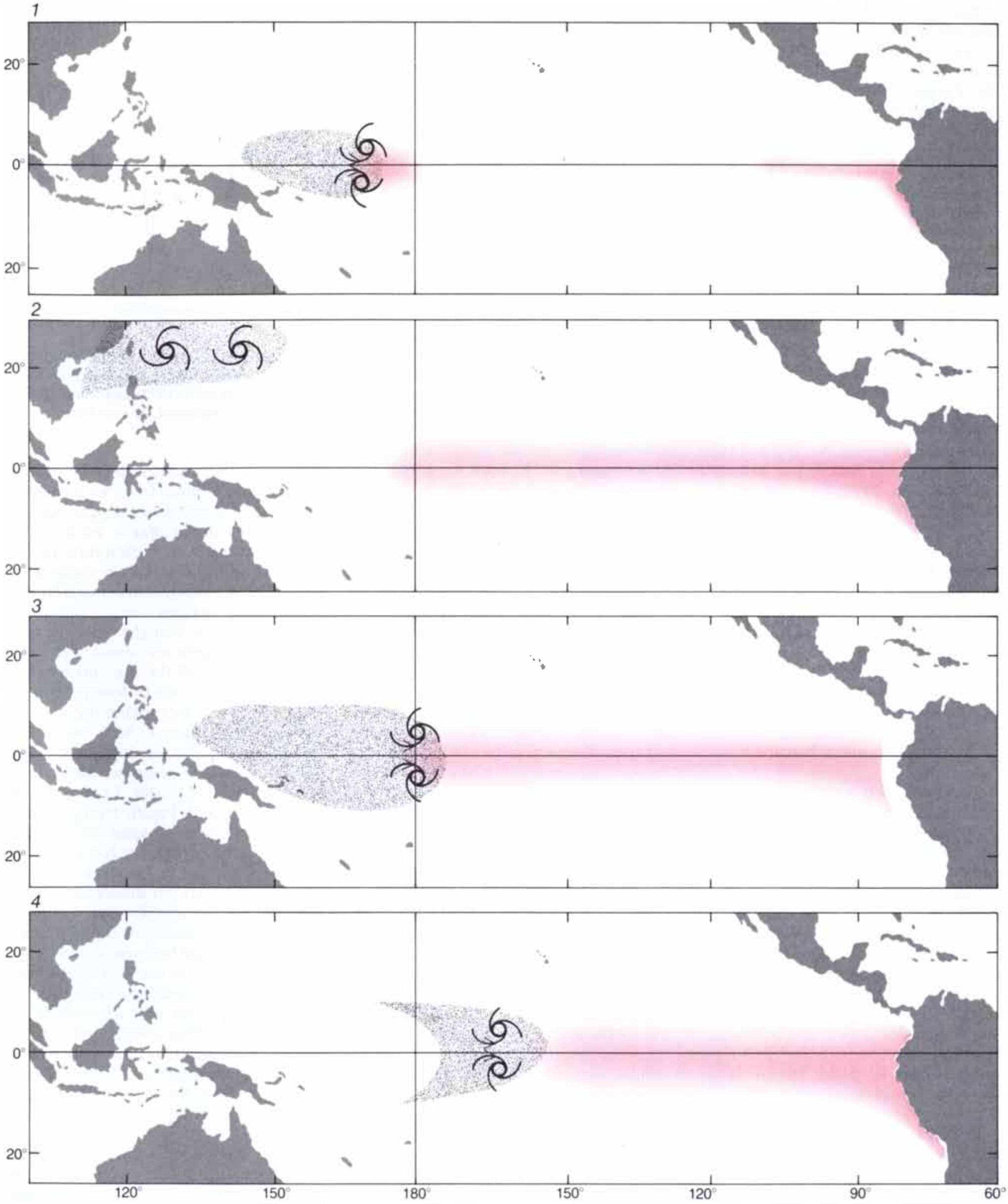
Somehow the variability of the Niño cycle seems to have been obscured by the methods employed in studying it. One serious flaw in virtually all the research has been the emphasis on the Niños themselves, whether individual or composite events. The much longer intervals between Niños have been largely neglected. And yet a theoretical model that links an anomalously warm sea surface in the equatorial Pacific to, say, the failure of the monsoon in India ought also to explain why the rains sometimes fail when the sea-surface temperature is at or below the normal level. Current models do not.

A second flaw in the models is their reliance on monthly and seasonally averaged values for atmospheric and oceanic variables. The problem with this approach is that it masks short-term deviations from normal conditions: the weather events that go into the averages. According to Roger B. Lukas and his colleagues at the University of Hawaii at Manoa, the five-to-seven-day weather cycle, and not just the monthly pattern, may have an important influence on the development of El Niño.

Lucas and his group have suggested a scenario in which El Niño is triggered by bursts of westerlies lasting for about a week over a stretch of a few hundred kilometers in the western Pacific. I believe the anomalous westerly winds are often generated by tropical cyclones. Early in most Niño years cyclones form in two surface pressure troughs, one on each side of the Equator, that belong to the large Indonesian low-pressure system. Since the winds of a cyclone run counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere, the winds on the equatorial side of the storms always blow from west to east. These strong westerlies trigger Kelvin waves that move warm water eastward, causing it to accumulate in the central Pacific. The Kelvin waves also warm the South American coast by depressing the thermocline.

The pressure troughs and the cyclones they spawn are found above the warmest areas of the sea surface, because the warm water lowers atmospheric pressure by heating the air and causing it to expand. As a result the troughs follow the sun, lagging slightly behind it on its seasonal excursions away from the Equator. From June to August the South Pacific trough disappears; the cyclones spawned by the North Pacific trough are too far north of the Equator to generate Kelvin waves, which are an inherently equatorial phenomenon. (The amplitude of a Kelvin wave decreases rapidly away from the Equator because it is inversely proportional to the Coriolis force, an effect of the earth's rotation that increases in intensity with distance from the Equator.) The absence of Kelvin waves emanating from the western Pacific in the summer is reflected a few months later in the dip in sea-surface temperatures along the South American coast.

From September on, the North Pacific trough follows the sun south. In doing so it moves over the warm pool in the central Pacific left from the previous spring's burst of Kelvin waves. The warm pool intensifies the trough. Once again tropical cyclones develop and trigger Kelvin waves that move the warm water farther to the east. The warm water tends to pull the trough and the cyclones with it; during El Niño cyclones form well to the east of where they ordinarily do. In Novem-



TROPICAL CYCLONES may trigger and sustain El Niño by generating week-long bursts of westerly winds along the Equator. From January to May in Niño years (*1*) cyclones typically form on both sides of the Equator in pressure troughs (*shaded area*) belonging to the large Indonesian low-pressure system. The cyclone westerlies set off subsurface Kelvin waves that traverse the Pacific in two to three months. The Kelvin waves push warm water toward the central Pacific and depress the thermocline off South America, thereby warming the sea surface in both regions (*color*). From June to August (*2*), as the warming spreads, the South Pacific trough disappears; the North Pacific trough and its cyclones, which follow the sun, are too far north to generate Kelvin waves at the Equator. Hence the sea surface off South America cools from September to November (*3*). By then the northern trough has moved back near the Equator, and it may have moved over and been intensified by the pool of warm water left over from the spring. If it nonetheless spawns no cyclones, El Niño ends. If cyclones form, new bursts of Kelvin waves push the warm water farther to the east, and it in turn pulls the trough behind it (*4*). As Kelvin waves strike South America from December to February the sea surface there warms again.

ber of 1957 and 1982, for instance, cyclones developed in a region near the Line Islands (at a longitude of roughly 160 degrees west) that is normally completely inactive.

According to this scenario, tropical cyclones are crucial both in initiating and in prolonging El Niño. The anomalous westerlies they generate in the western Pacific trigger the oceanic anomaly; a prior buildup of the trade winds is not required. Once it has started, El Niño can easily peter out if it is not sustained by further surges of westerly winds. A key juncture in Niño evolution comes at the end of northern summer, when the surface trough moves back toward the Equator. The warm water there favors the formation of cyclones, but if for some reason none develop, El Niño ends.

If the scenario is correct, the problem of predicting the time of onset, the duration and the intensity of El Niño becomes one of predicting the occurrence of tropical cyclones—in other words, it becomes about as hard as predicting the weather. A surface trough over the ocean is almost always a necessary condition for tropical cyclogenesis, but it is by no means a sufficient condition. If it were, some 5,000 tropical cyclones would form around the world every year rather than the observed average of about 80. The circulation interactions needed to trigger a tropical cyclone are rare and not well understood, and the interactions that trigger cyclones at the right place and time to launch or sustain a Niño are rarer still. They are independent of El Niño itself. They may, for example, include interactions of tropical winds with mid-latitude weather systems, the forecasting of which, it will be remembered, was one of the main purposes of studying El Niño in the first place.

Mid-Latitude Effects

Whether or not it is possible to forecast the onset of El Niño, one might still hope to predict its effects outside the Tropics once it has begun. One way of predicting such effects is to apply the laws of physics to the atmosphere, using high-speed computers to solve numerically the equations that govern its motions. Many investigators have designed numerical models to simulate the impact on the atmosphere of anomalous westerly winds and a prolonged warming of the sea surface in the equatorial Pacific.

All models predict that heavy rainfall in the central and eastern Pacific should heat the upper troposphere. (When water vapor condenses to form rain, it releases the latent heat it acquired in evaporating from the sea surface.) The energy added to the upper troposphere at the Equator is transferred to higher latitudes through an intensification of the so-called Hadley circulation, in which air flows from the Equator to the poles at high altitudes and returns to the Equator at the surface. The Hadley circulation in turn transfers energy to the westerly jet streams that drive the large storm systems in the mid-latitudes.

John M. Wallace and John D. Horel of the University of Washington have suggested that the effects on North American weather should be most pronounced in the winter following the onset of El Niño. Unfortunately no consistent pattern has yet been detected. According to Thomas R. Karl of the National Climatic Data Center, the moderate Niño of 1976–77 coincided with an abnormally cold winter in North America; in contrast, the strong Niño of 1982–83 coincided with an abnormally warm winter. Four other abnormal winters during the period from 1975 to 1982 coincided with no Niño at all.

A look at Illinois air-temperature records for the winters between 1894 and 1983 is even more discouraging. There were 17 moderate or strong Niños during that period; on 10 occasions the winter in Illinois was somewhat warmer than normal and on seven occasions it was colder. Seven of the 90 winters deviated substantially from the norm, but only one of those (1976–77) coincided with a Niño. A statistical analysis done by Tim P. Barnett of the Scripps Institution of Oceanography has confirmed that Pacific sea-surface temperature is a poor predictor of winter air temperature over the central U.S. The correlation with air temperature over the coasts of southern California and of the Gulf of Mexico is more significant, according to Barnett, but even in those regions the effects of a warm Pacific interact with local conditions in ways that are as yet unpredictable. The Niño winter of 1972–73 was exceptionally wet in southern California; forecasters who expected the same in 1976–77 were stunned by a near-record drought.

Prospects

I do not mean to suggest there is no hope of ever using El Niño as a basis for seasonal weather forecasts. The search continues for its precursors and for a pattern of effects. The former endeavor is being greatly aided by a new ocean-atmosphere data set, compiled at the initiative of Joseph O. Fletcher of the NOAA, that incorporates more than 70 million weather reports made by ships at sea between 1854 and 1979.

Some investigators are trying to predict El Niño not by looking for precursory signals in statistical data but by devising numerical models that attempt to reproduce the physical evolution of the tropical ocean-atmosphere system. On the basis of such a coupled model Mark A. Cane, Stephen E. Zebiak and Sean C. Dolan of Columbia University's Lamont-Doherty Geological Observatory have predicted that a moderate Niño will begin this spring and peak late in the year. The central idea underlying their model is that during a Niño warm water moves not only from west to east along the Equator but also to the north and south, away from the Equator. A Niño can therefore recur only when the equatorial heat reservoir has been refilled by an influx of warm water from higher latitudes. According to this model, the heat content of the upper equatorial ocean, as measured by the depth of the thermocline, is a key indicator of the likelihood of a Niño.

Since data on thermocline depths in the equatorial Pacific are sparse, the Lamont-Doherty model derives those values from wind data, which are its sole input. It then forecasts the evolution of winds, thermocline depth, ocean currents and sea-surface temperatures by taking into account the feedbacks among all these variables. Cane and his colleagues have tested the model's ability by making retrospective forecasts for 12 of the past 15 years. (The three years in which Niños were under way were excluded.) They report that in nine of the 12 years the forecasts of whether a Niño would occur were unambiguously correct; in the other three years the forecasts were ambiguous. By this fall the model's first real-time forecast will have been put to the test. It is important to note, however, that the model makes no attempt to predict the effects of a Niño on the weather in the U.S. or in other mid-latitude regions.

The question of whether El Niño will become a useful guide for seasonal weather forecasters remains open. Yet its impact on meteorology, if not on the weather, has already been decidedly favorable. Because of public demand for forecasts, research meteorologists have been put in the unusual but salutary position of having to test their hypotheses almost immediately. When the forecasts fail, as they invariably have, the hypotheses are modified sooner than they would be otherwise, and so work progresses at a faster pace. If routine daily weather forecasting were incorporated into meteorological research in this way both disciplines would be enhanced.

The Heart as a Suction Pump

A new model suggests some energy from each contraction is stored within the muscle and provides the power for a suction that aids filling. The effect is amplified by the motion of the heart as a whole

by Thomas F. Robinson, Stephen M. Factor and Edmund H. Sonnenblick

In the earliest machine guns, after each bullet was fired a new round had to be cranked into place manually before it could be set off by the gun's mechanism. Similarly, according to the currently accepted model of the human heart, after each contraction the heart remains passive as it is refilled by blood driven entirely by pressure in the veins. We propose a new model of the heart's function. Our model is analogous to modern automatic weapons, in which the force generated by the firing of one cartridge provides the energy needed to load the next round. In our model some of the energy of each systole, or contraction of the heart muscle, is stored in the heart and then powers the following diastole, or expansion.

The human heart has four chambers, the left and right atria and the left and right ventricles [*see illustration on page 86*]. The atria collect blood as it returns from the body (in the case of the right atrium) or from the lungs (in the case of the left atrium). During diastole one-way valves open between the atria and the ventricles and blood pours into the ventricles, filling them. Then, during systole, the ventricles contract, closing the valves to the atria and sending blood through other one-way valves to the lungs (in the case of the right ventricle) or to the body (in the case of the left ventricle).

The currently accepted model of the heart's function derives largely from work done near the end of the 19th century by Otto Frank in Germany and later by Ernest H. Starling in England. According to the Frank-Starling law, the energy imparted to the blood by the contraction of a ventricle, independent of any control by nerves or hormones, is proportional to the length of the ventricular muscle fibers at the end of the preceding diastole. Once systolic contraction is complete the subsequent diastolic filling is a passive function of venous pressure, which stretches the relaxed muscle of the ventricle wall. According to the Frank-Starling law, the energy expended in contraction has no essential role in the diastolic filling of the ventricles.

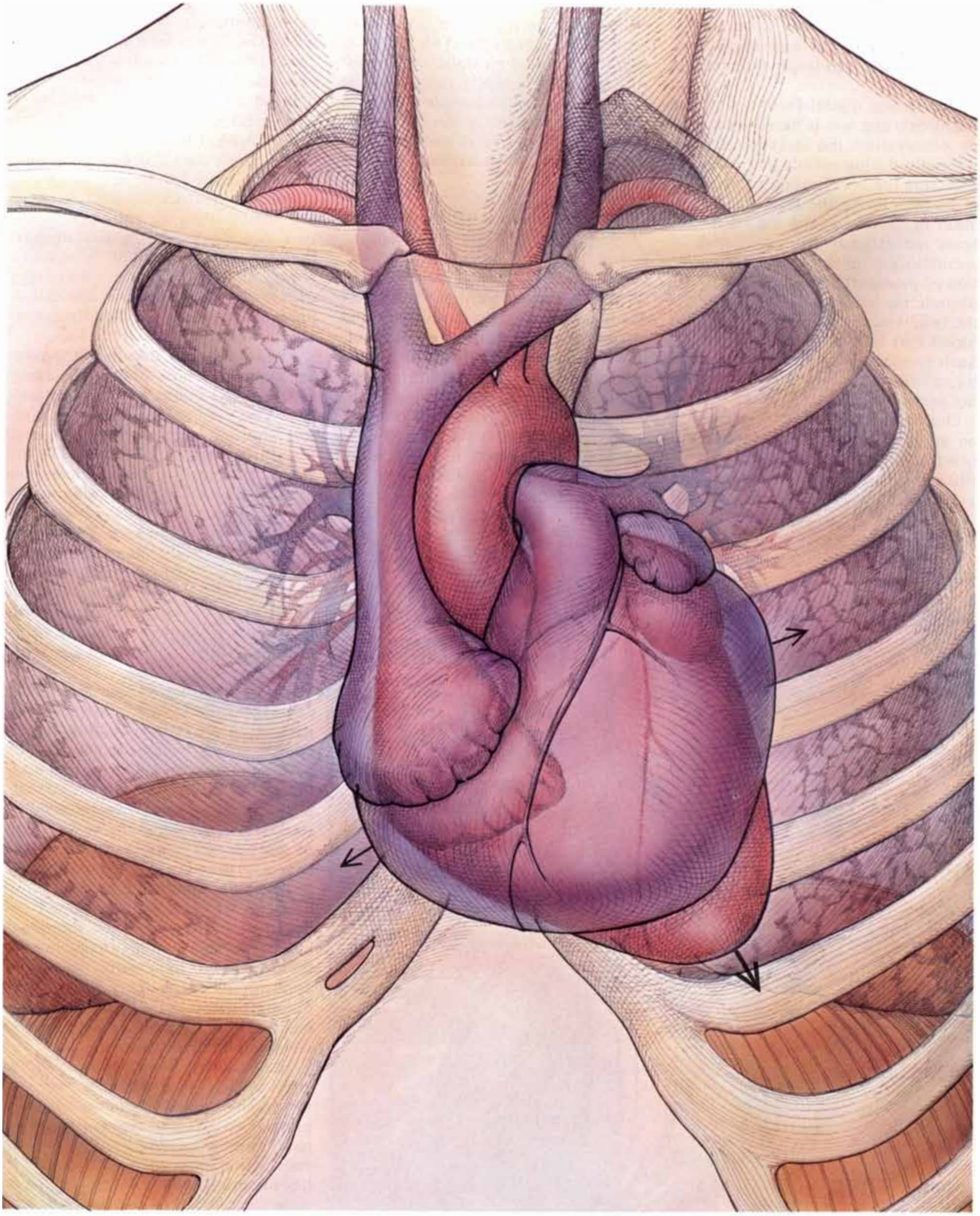
The Frank-Starling law has dominated the thinking of most cardiologists and cardiovascular investigators. It has a number of limitations, however, some of which stem at least in part from the experimental techniques by which the law was derived. For example, the Frank-Starling law is most adequate as a description of the depressed heart, in which cardiac output is quite low and the venous filling pressures are high. The reason is partly that the original formulations of the law were based on data obtained with hearts excised from animals; in excised hearts cardiac function is abnormally depressed.

Another limitation is that the mechanism by which the heart is filled in the Frank-Starling model is relatively static; it does not reflect the dynamic interplay between systole and diastole. Here again the original experiments may be partly responsible: the technology of the time allowed measurement only of static pressures and average blood flows. In addition the hearts that were studied had been excised and fixed in position, and so they did not display the movement within the body that characterizes the normal heart. In recent years it has become possible to make accurate real-time dynamic measurements of the flow of blood and of the heart's motion in living subjects.

In our model the dynamic relation between systole and diastole is critical to the proper function of the heart. The systolic contraction provides much of the energy that drives the process of diastolic expansion. This energy is stored and recovered in two ways. One is by the gross motion of the heart itself. When the heart contracts, it propels blood upward and thereby, in accordance with Newton's law of action and reaction, propels itself downward within the body. This recoil stretches the great elastic vessels and connective tissue that hold the heart in place. As the heart subsequently relaxes it springs upward, meeting the inflow of blood head on. This raises the velocity of the blood with respect to the heart and helps to power the filling process.

The other way the energy of systole is stored is in the deformation of the heart itself. We propose that systolic contraction compresses the elastic elements of the heart and its muscle fibers in such a way that even without any external filling the natural tendency of the ventricles is to expand. This expansion creates a negative pressure, or suction, that pulls blood into the ventricles from the atria. In the Frank-Starling model the heart is assumed to act as a static pressure pump. In this new model the heart, which moves within the thorax and expands naturally after systole, acts as a dynamic suction pump.

Our proposal is primarily a response to three unsatisfactory aspects of the Frank-Starling law. First, the law implies that cardiac output should be uniquely determined by the venous filling pressure on the right side of the heart. The reason is that any blood arriving at the left side of the heart must first have been pumped through the lungs by the right side, and the final volume (and hence the length of the muscle fibers) of the right ventricle is (in the Frank-Starling model) entirely determined by the filling pressure. The fact is, however, that in healthy individuals the right atrial pressure is very low—on the order of a few millimeters of mercury. It is difficult to see how small changes in this already low pressure could control the filling of the



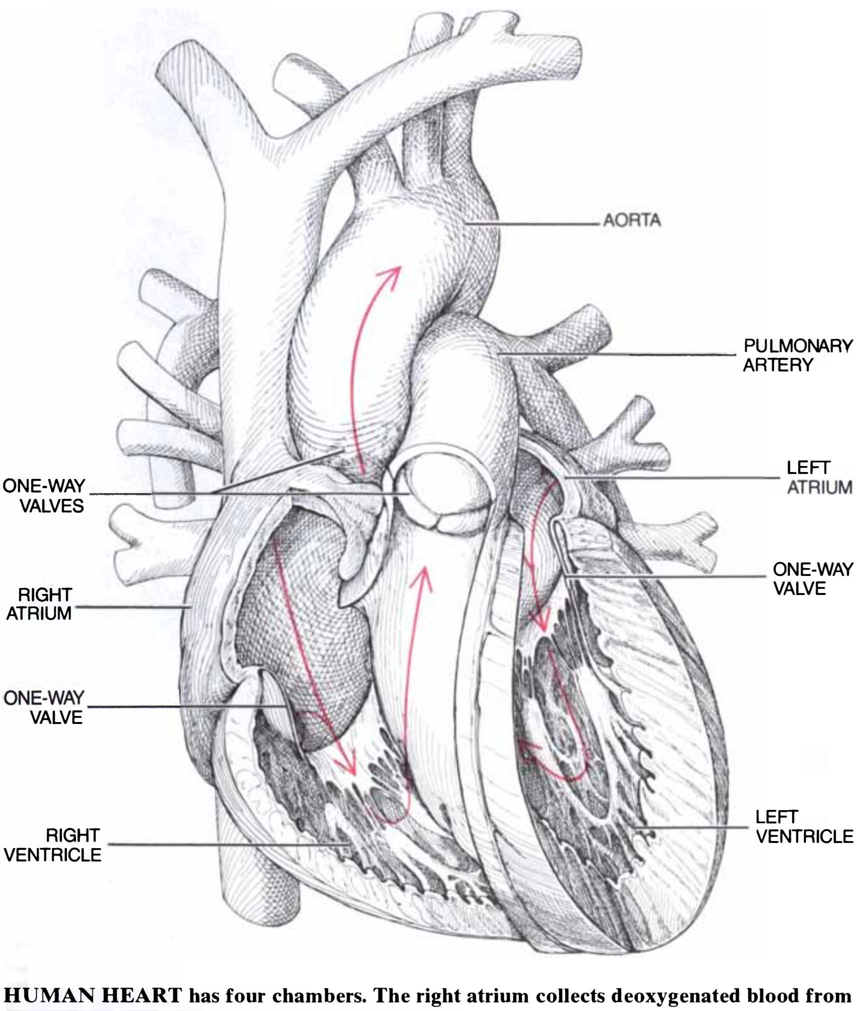
PUMPING ACTION of the heart takes place in two stages: systole (*redder shade*), in which the heart contracts, sending blood to the body and lungs, and diastole (*bluer shade*), in which the heart relaxes and the ventricles (the heart's main pumping chambers) fill with blood. In the traditional model of the heart's action the heart in diastole is a passive organ: the filling of the heart depends entirely on the pressure in the veins that feed it. In the model proposed by the authors the heart actively expands during diastole (*small arrows*), creating a suction and thereby drawing blood into the ventricles. The heart's gross motion also contributes to diastolic filling. When the heart contracts, it jets downward (*large arrow*): it recoils because the blood pumped from the heart moves upward, toward the head. The heart's downward motion stretches the elastic tissues that hold it in place, and so during diastole it springs back upward, meeting head on the inflow of blood. The resulting high relative velocity of blood and heart accelerates the heart's filling process.

entire heart, particularly when changes in body position and respiration should cause much larger fluctuations in pressure.

Our second dissatisfaction with the Frank-Starling law is based on a simple observation: the heart moves substantially during systole and diastole, reflecting the acceleration of volumes of blood. During the systolic acceleration of blood toward the head, the heart moves away from the head, in accordance with the law of conservation of momentum. In the subsequent diastole the heart moves back toward the head and the ventricles engulf the blood that has moved into the atria. Such events must be the result of inertia and elastic recoils, not of static filling pressures.

Our third dissatisfaction is based on another observation. Mammalian hearts continue to empty and refill when they are excised and placed in a dish of buffered solution—in which there can be no changes in external diastolic filling pressure. This observation suggests that diastolic filling pressure does not uniquely determine the length of the heart's muscle fibers and so cannot entirely explain the output of the heart. Some other means of ventricular filling, such as suction, must also play a role.

This conclusion is bolstered by two further experiments, one on the left ventricle and one on isolated segments of heart muscle. Edward L. Yellin of the Albert Einstein College of Medicine of Yeshiva University and other workers have done experiments in which the mitral valve, which separates the left atrium from the left ventricle, is occluded so that the ventricle cannot fill. In these cases the pressure in the ventricle falls below zero after the systolic contraction and a negative filling pressure is developed. The strength of the negative pressure depends on the degree to which the ventricle was emptied during the systolic contraction. The reason is that the left ventricle has a certain equilibrium volume, at which the filling pressure is exactly zero: a positive external pressure is needed to increase the ventricle's volume and a negative external pressure is needed to decrease it. In systole the ventricle is contracted until it is smaller than the equilibrium volume. Any suction during diastole must therefore be dependent on the strength of preceding systolic events: the energy of elastic recoil is somehow stored during systolic contraction and then, during diastole, it creates a suction.



HUMAN HEART has four chambers. The right atrium collects deoxygenated blood from the body and passes it through a one-way valve into the right ventricle, which pumps it through another one-way valve to the lungs by way of the pulmonary artery. Oxygenated blood returning from the lungs is collected in the left atrium before filling the left ventricle. The left ventricle then pumps the oxygenated blood back to the body by way of the aorta.

The equilibrium volume of the average large dog heart, for example, ranges between 15 and 30 cubic centimeters. The volume at the end of diastole normally ranges between 40 and 50 cubic centimeters. If about 60 percent of the blood in the ventricle were ejected, the ventricle's volume at the end of systole would be between 16 and 20 cubic centimeters. The ventricle would therefore tend to expand, generating a negative pressure, or suction, during the early phase of diastole. For any given heart the suction is greatest when the size of the heart at the end of systole is least [*see illustration on opposite page*].

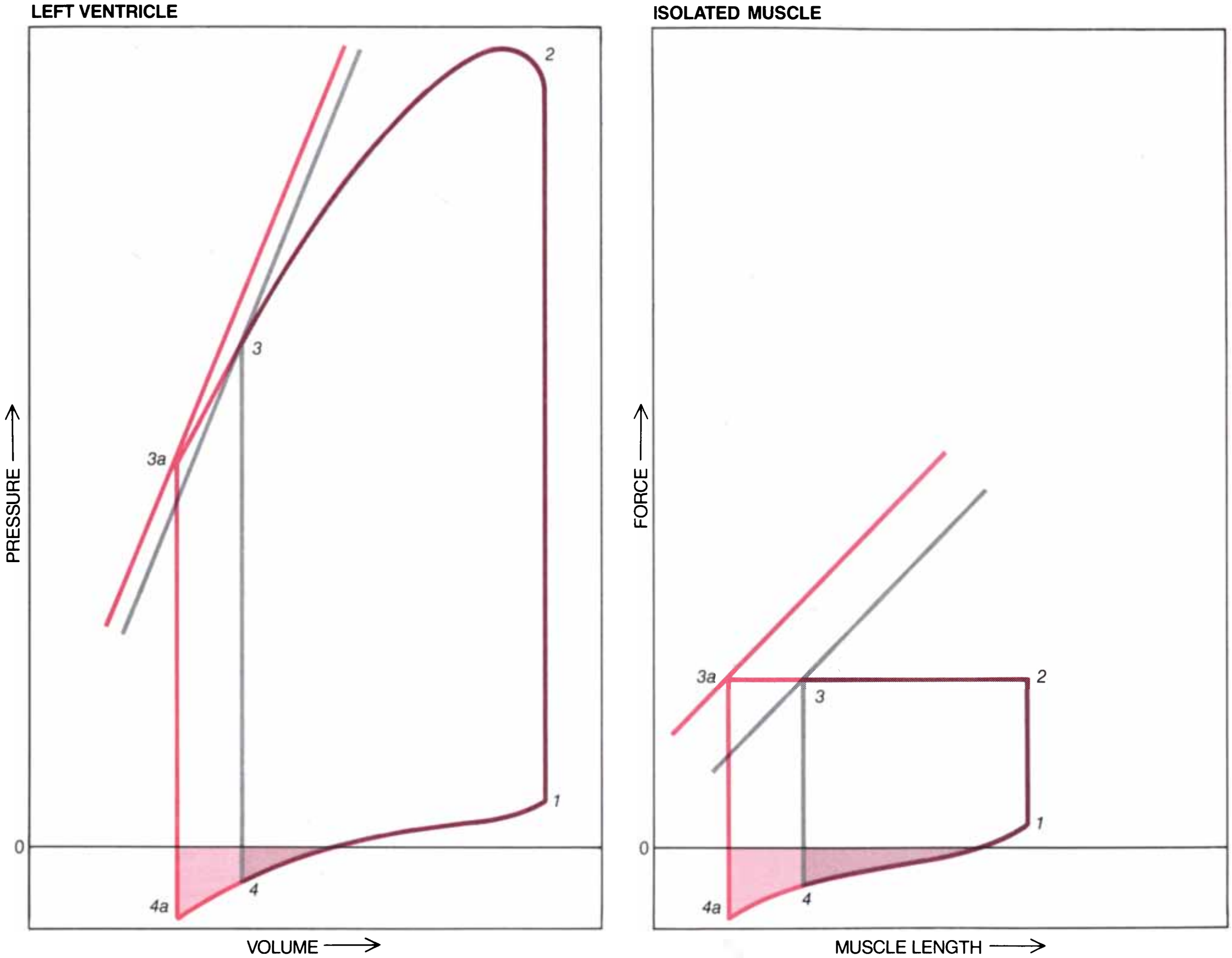
(It is interesting to note that in the condition called heart failure the ventricle does not contract completely in systole, and the volume to which it fills in diastole increases. The normal elastic energy is therefore not stored during systole and cannot be released as recoil during diastole, and so the filling of the ventricle does depend on higher venous pressure. The Frank-Starling law is, then, applicable to the failing heart, in which elastic recoils are reduced or eliminated, even though it has serious shortcomings in explaining the function of the normal heart. The applicability of the Frank-Starling law to failing hearts helps to explain why it has not been extensively modified by clinicians or investigators.)

The restoring force observed in the left ventricle as a whole is also observed in isolated segments of heart muscle, which spontaneously reextend to a "normal" resting length after any contraction. Just as the left ventricle has an equilibrium volume, so linear segments of heart muscle have an equilibrium length. When a strip of the muscle is placed in a bath of buffered solution and stretched between fixed points like an elastic band, there is a length at which the resting tension of the muscle segment is roughly zero; when the segment is stretched, the resting tension rises exponentially. When a muscle stretched to about 20 percent beyond its equilibrium length is excited, it generates more than twice as much force as an excited muscle fiber at equilibrium length. At lengths shorter than the equilibrium length the muscle remains slack when it is not excited, and when it is excited, it must shorten before it can develop any force. When it relaxes, it once again becomes slack: recoil reextends the muscle.

What occurs within the heart and the cardiac muscle to create such elastic recoils and suctions? We have found factors contributing to these forces at every level of structure in the heart, from the gross anatomy of the heart itself to the microscopic structure of the interior of muscle cells.

A cell of cardiac muscle, like a skeletal-muscle cell, consists of a periodic array of quasi-cylindrical structures called sarcomeres, which lie end to end along the length of the muscle cell [*see illustration on next page*]. Each sarcomere is bounded by two protein structures called *Z* disks, which are perpendicular to the long axis of the cell. From each *Z* disk many thin filaments of actin, a protein, extend toward both ends of the muscle cell. Half of these filaments extend into each of the sarcomeres of which the *Z* disk is a boundary. The thin filaments extending from one of the *Z* disks that bound a sarcomere do not normally extend far enough to touch or overlap the filaments extending from the *Z* disk at the sarcomere's other end.

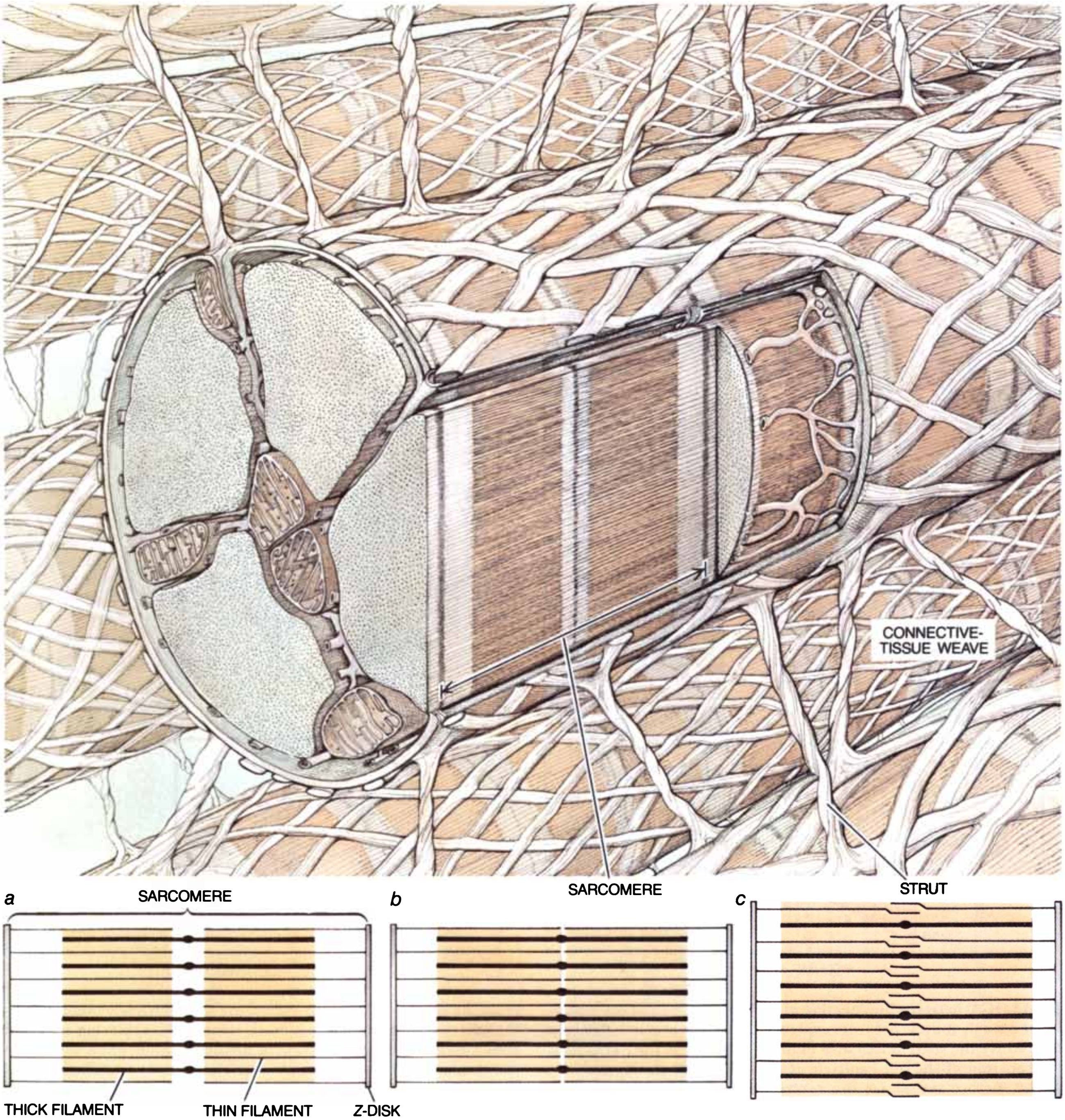
Interdigitated with the thin filaments are a number of thick filaments made



ANALOGY between the function of the whole heart and the behavior of a single muscle provides evidence that individual muscles help to produce the suction that aids filling. The graph at the left depicts the function of the left ventricle at rest (*gray*) and during exercise (*color*). Before systole (*1*) the ventricle, which has just expanded, has a low pressure and a high volume. The ventricle then contracts, driving the pressure up until the valve to the aorta opens (*2*); during this period the volume remains constant because the blood in the ventricle is an incompressible fluid. As blood rushes into the aorta both pressure and volume in the ventricle decrease until the ventricle has been emptied of as much as two-thirds of the blood it contained (*3*). Then the valve to the aorta closes and the muscle walls expand. Since valves to the atrium and the aorta are both closed, pressure in the ventricle drops rapidly (*4*) until there is a negative pressure (*shaded region*), which causes a suction, drawing in blood from the atrium and thereby increasing the ventricle's volume. During exercise the heart empties more completely, achieving a smaller final volume (*3a*). This leads to a greater suction when the valves are closed and the walls of the ventricle expand (*4a*). The graph at the right traces the behavior of a single muscle supporting a weight; the muscle's length and the force it produces are analogous to the volume and pressure of the ventricle. At first the muscle is relaxed and extended (*1*). When the muscle is excited, it begins to generate force, but its length does not change until the force is equal to the weight of the load being lifted (*2*). Then the cell shortens while producing a constant force. If the weight is released, the force decreases rapidly and overshoots (*4*) the point of zero force. This represents the muscle's tendency to reextend after a contraction. When the muscle is "potentiated" (as it is during exercise), it contracts further (*3a*) and so it has an extremer tendency to reextend (*4a*).

primarily of the protein myosin. The thick filaments occupy the region of the sarcomere that is not normally reached by thin filaments, and each thick filament overlaps thin filaments from both *Z* disks bounding the sarcomere. When the muscle is stimulated, thin filaments from both *Z* disks are pulled along the thicker filaments toward the center of the sarcomere. This action pulls the *Z* disks closer together, shortening the sarcomere. It is the nearly simultaneous contraction of many sarcomeres placed end to end within the cell that accounts for the remarkable ability of muscle cells to shorten quickly.

During the cardiac cycle the sarcomeres in the heart's muscle cells shorten from about 2.05 micrometers (millionths of a meter) to 1.85 micrometers



HEART MUSCLE CELL (*top*) consists largely of an array of fibers, each made up of many units called sarcomeres placed end to end. A sarcomere (*bottom,* a) is bounded at both ends by so-called *Z* disks. Thin filaments made of the protein actin extend from each *Z* disk toward the center of the sarcomere. The thin filaments are interdigitated with thick filaments, which are made of the protein myosin and do not extend to the *Z* disks. Contraction of the muscle (*b*) results when thin filaments are drawn along the thick filaments toward the center of the sarcomere, pulling the *Z* disks closer together. When the muscle is fully contracted (*c*), thin filaments from opposite ends of the sarcomere overlap one another within the lattice of thick filaments. This increases the cell's girth and may stretch the *Z* disks and other lateral elements (not shown) of the cell skeleton; the stretched cytoskeletal elements may provide a restoring force that elongates the contracted cell by decreasing its girth. Muscle cells are wrapped in a crisscross weave of connective tissue; in the human heart other strands of connective tissue, called struts, which are twisted like ropes, tether muscle cells to one another.

and elongate again to 2.05. How is this elongation achieved? Some fraction of the energy of contraction must be stored in the muscle cell, but where? Physiological evidence suggests that the cause of the elongation may lie both in the structure of the sarcomeres themselves and in the components of the muscle cell that form the cytoskeleton, the set of interconnected fibrous elements that helps to maintain a cell's shape.

The thin filaments within the sarcomere are each about one micrometer long, and so as the sarcomeres contract, thin filaments from the opposite ends of a sarcomere (that is, filaments extending from different *Z* disks) meet at the center of the sarcomere when it is about two micrometers long. As the sarcomere contracts further, thin filaments from different *Z* disks begin to overlap within the lattice of thick filaments. The latticework of filaments becomes crowded in such a "double overlap" state and the sarcomere expands laterally: its cylindrical shape increases in girth. This lateral expansion stretches the *Z* disks and other lateral structural elements of the cell, providing a mechanism whereby some of the energy of contraction is stored to be reused as energy of expansion.

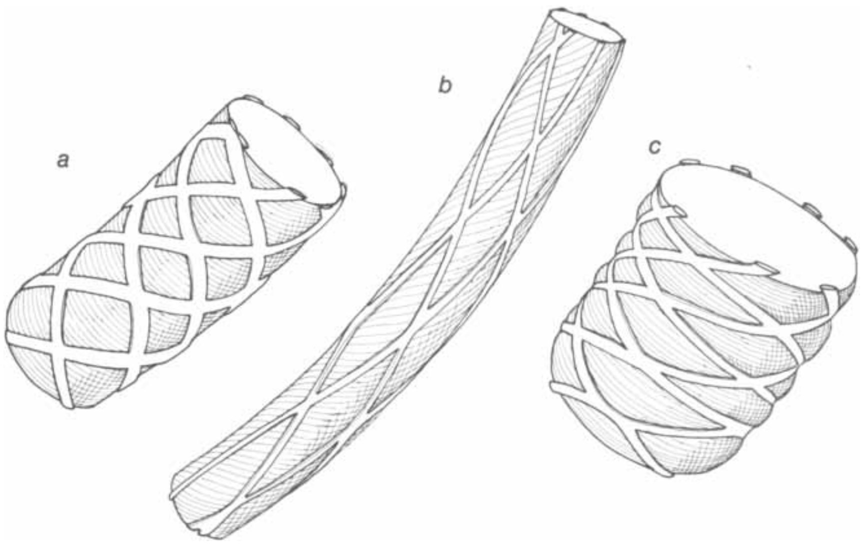
It is not just the muscle cells' microstructure that stores the energy of systolic contraction. There is now increasing evidence that the connective tissue surrounding individual muscle cells, as well as the tissue linking the cells, also plays a role.

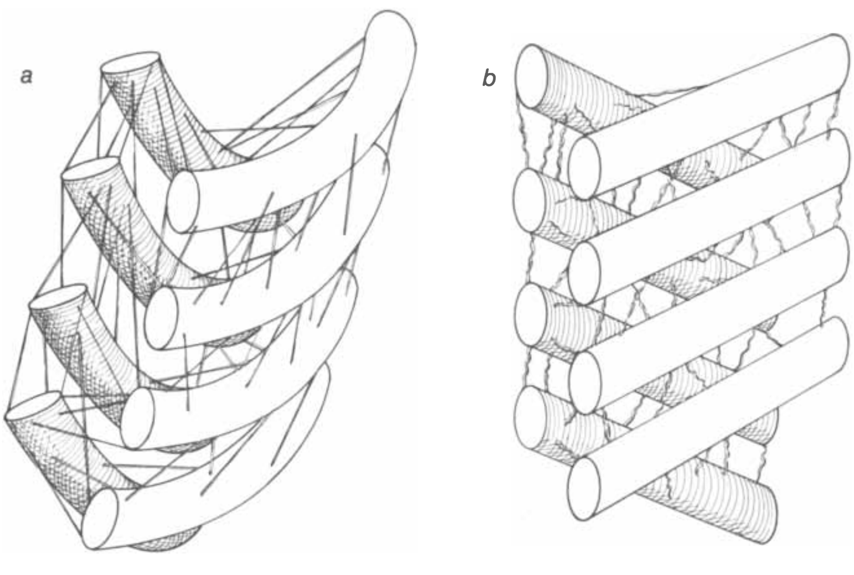
For example, strands of connective-tissue fibers are wrapped helically around the outside surface of each cylindrical muscle cell. The fibers are wound both counterclockwise and clockwise, and so they form a crisscross pattern not unlike that of a net hammock. The fibers consist mostly of collagen, a tensile element that resists stretching, and they also contain varying amounts of elastin. Elastin is an extracellular protein that has stress-strain properties remarkably similar to those of rubber: it can be stretched up to 170 percent of its slack length without being irreversibly deformed.

The netlike structure of connective tissue may provide some of the restoring force that causes individual muscle cells to return to their original length after a contraction. As the cell contracts and grows fatter, it strains against the enclosing weave of fibers. After the contraction the fibers might then squeeze the cell into a smaller diameter, lengthening it.

A similar weave of collagen fibers surrounds entire muscles (which are groupings of muscle cells). Again, like a net hammock, this structure is made primarily of inextensible fibers, but it can nonetheless be stretched considerably without damage because of the pattern in which it is woven: the strands are crisscrossed to form parallelograms, which can be stretched along their main diagonal until all four sides are nearly parallel.

We have found that when strips of isolated rat heart muscle have been stretched until the wrapping collagen fibers reach this limit, the resistance





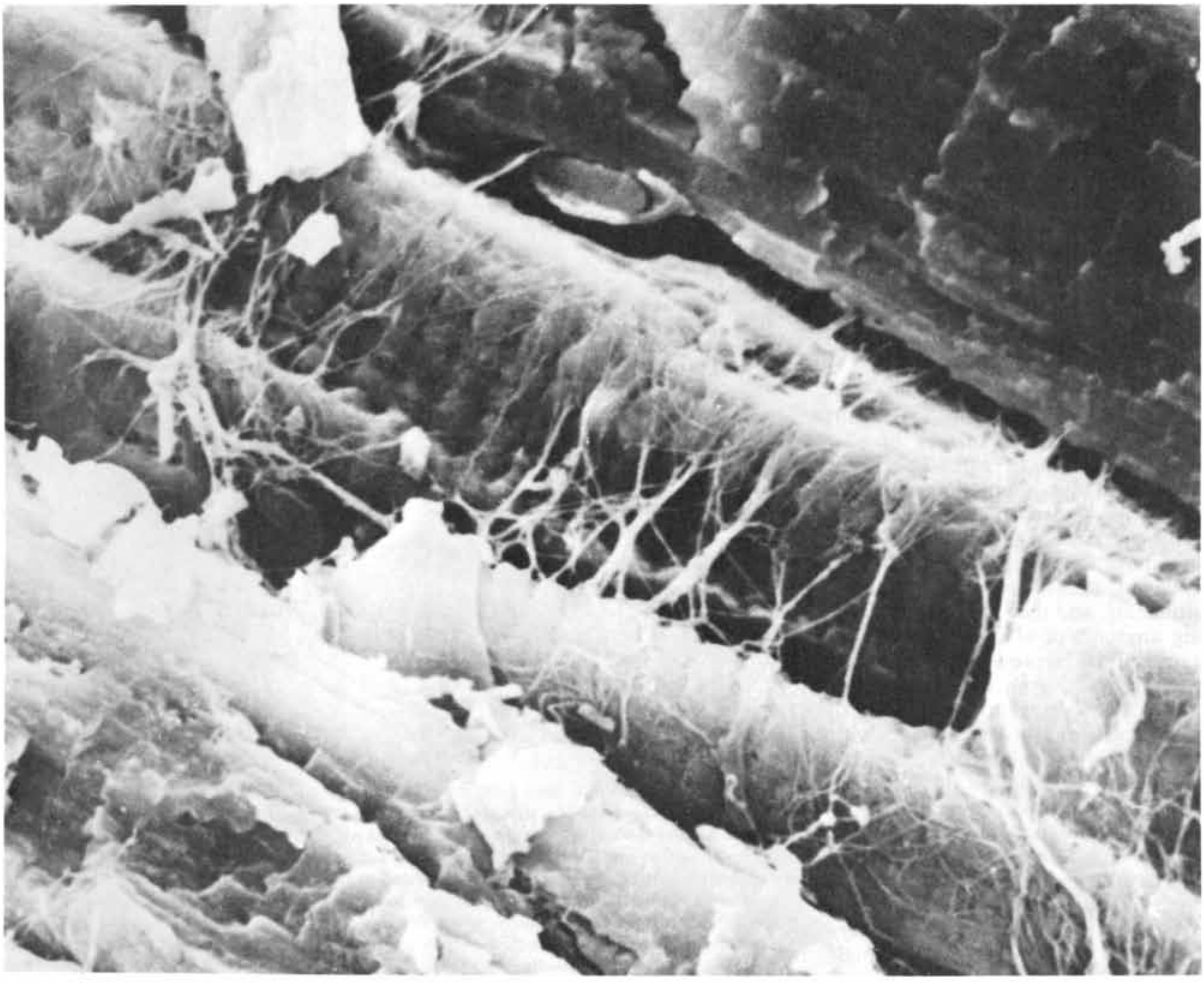
CONNECTIVE TISSUE may contribute in two ways to the heart's tendency to expand actively during diastole. The strands of connective tissue wrapped around individual cells (*top*) are wound both clockwise and counterclockwise around the cells, forming a network with a woven appearance not unlike that of a net hammock (*a*). Several experiments indicate that this netlike structure of parallelograms protects the muscle fiber from overstretching: when the fiber is stretched (*b*), the parallelograms are elongated along a diagonal and resist further stretching. Similarly, when the muscle is fully contracted (*c*), the parallelograms of connective tissue would be stretched along the other diagonal, perhaps resisting further contraction and tending to reextend the muscle. Struts of connective tissue that tether individual cells to one another (*bottom*) may be stressed and stretched (*a*) when the heart is in any configuration other than its natural resting one (*b*). Such tension may help to expand the heart after systolic contraction has compressed it beyond its equilibrium volume. The illustration shows muscle cells from two adjacent layers of the heart muscle.

to further stretching of the muscle increases dramatically. This is also the greatest degree to which the muscle can be stretched and still produce maximum force. The simplest and most appealing interpretation of these findings is that the cross-woven structure of the collagen fibers prevents the sarcomeres from being overstretched. It seems reasonable to assume, then, that the network of collagen also provides a restoring force when the muscle is contracted rather than stretched, so that the parallelograms of collagen fibers are fully extended along their other diagonal. There is no definitive proof of the hypothesis yet, but it is supported by preliminary findings from several laboratories.

There is another way connective-tissue fibers might contribute to the heart's tendency to create a suction by expanding actively during diastole. In the mammalian heart individual muscle cells are linked to one another laterally by ropelike elements known as struts, which are composed mainly of collagen. The struts appear to be twisted helically, much as the cables that support suspension bridges are twisted to provide greater tensile strength. Is it possible that these struts act as tethers, resisting deformation of the heart and storing some energy of deformation to be applied later in diastolic suction?

One way to determine whether the struts contribute to diastolic suction is to compare the function of mammalian heart muscles, in which muscle cells are interconnected by struts, with the function of amphibian hearts, in which the muscle cells are not interconnected (although the individual cells are indeed wrapped in fibers of collagen).

If a beating rat heart is placed in a saline bath, the heart continues to beat, although more slowly than normal, and it jets rapidly through the fluid. Fluid is forcefully expelled through the great vessels during systole and is sucked into the ventricles during diastole. This simple experiment shows there is indeed significant diastolic suction in the rat heart, since there is no external filling pressure to drive fluid into the heart. If a similar experiment is done with a frog heart, the heart beats very weakly and moves through the liquid only a little if at all. A possible explanation is that the diastolic suction in the frog heart is not sufficient to draw much fluid into the chambers and the corresponding ejection of fluid is not forceful enough to move the heart in the opposite direction. It is appealing to suggest that the rat heart has a more vigorous pumping action and a higher degree of suction



THIN STRUTS link the lateral surfaces of muscle cells in the mammalian heart in this scanning electron micrograph. Struts consist primarily of collagen, a protein that resists stretching, but they also contain some elastin, a highly elastic protein. They coordinate the movement of separate parts of the heart muscle, and they may provide some of the force that causes the contracted heart to expand.

in part because of the connecting struts that link individual muscle cells and coordinate the contraction of groups of cells.

The suggestion is given further credence by comparable studies we have made of two invertebrates. One of the invertebrates, the squid, sucks water into a central muscular pump chamber and then expels it forcefully as the pump chamber contracts, thereby jetting through the water [see "Jet-propelled Swimming in Squids," by John M. Gosline and M. Edwin DeMont; SCIENTIFIC AMERICAN, January, 1985]. The squid's hollow, muscular pump chamber is covered by a collagenous mantle, which is tethered to individual muscle fibers by strutlike elements. These are reminiscent of the struts that connect individual muscle fibers to one another in the rat heart.

The muscle structure of the octopus, on the other hand, is similar to that of the frog heart. In the octopus each muscle cell is surrounded by weaves of connective-tissue fibers, but there is no evidence of intercellular connections. The octopus, unlike the squid, normally does not jet through the water but moves sinuously among rocks along the sea floor. (It does occasionally contract sections of its internal cavity to create a small jet effect.)

Based on preliminary studies of the squid, it appears that intercellular connective tissue in the mammalian heart and in the squid enhances the efficiency of a contraction and the subsequent suction, which fills a central cavity after the ejection of fluid. The rat heart and the central cavity of the squid both have such interconnections, and both are able to generate suction and to jet through a liquid medium; the frog heart and the octopus have no such interconnections and do not show significant suction or jetting.

Perhaps the struts contribute to suction by coordinating the action of many individual muscle cells; by tethering the cells together, they may resist deformation of the heart when it is contracted beyond its equilibrium volume, and so in the fully contracted heart tension in the struts might help to store some of the energy of systolic contraction.

The proposal that connective tissue has a role in the diastolic expansion of the heart is supported by experiments we have done with Joseph M. Capasso of the Albert Einstein College of Medicine. We have found that the ratio of the velocity of contraction to the velocity of reextension is much smaller in a whole papillary muscle (the muscle that controls motion of the mitral valve) than it is in isolated muscle cells. That is, the cell regains its length faster when it is part of a tissue than it does when it is isolated from other cells and the connective tissue that binds them together. We attribute the difference in the ratios to the effects of connective tissue on the dynamics of the contraction and elongation of muscle cells.

The creation of suction is not the only way the energy of a systolic contraction may be applied to the following diastolic expansion. In addition to changing the heart's shape (and the shape of its muscle fibers) contraction changes the heart's position in the body. The major blood vessels supplying the human heart and those leading away from it are connected to the wide area known as the base (which is actually at the top of the heart, the part nearest the head). Like a squid in water or a rat heart in a dish of saline, when the heart contracts, driving blood upward, it jets downward, in the direction opposite to that of the expelled fluid. We propose that some of the kinetic energy the heart thus acquires during systole is later applied to the filling of the ventricles during diastole.

Various techniques make it possible to study the gross movement of the heart in the chest. In an early technique, known as ballistic cardiography, the subject lay on a wheeled table and the minute motions of the table in response to changes in the momentum of the subject's heart were recorded. In the more modern method known as two-dimensional echocardiography, a real-time motion-picture image of the heart is generated by sound waves reflected at ultrasonic frequencies off the surface of the heart. It is also possible to measure the speed and direction with which blood flows in certain regions of the body by focusing sound waves on areas where blood is flowing and then recording the Doppler shift of the reflected waves.

Such techniques have shown that the heart moves dramatically within the chest. During systole the base of the heart moves away from the head, stretching the compliant connective tissue and blood vessels that hold it in place. These connecting members convert a fraction of the heart's kinetic energy into potential energy of stretching and exert an upward force on the base, which is drawn back toward the head during diastole. As the rapidly expanding heart moves toward the incoming blood under the influence of this elastic recoil force, it swallows up the returning blood and fills rapidly.

The recoil contributes greatly to the accelerated early filling that is necessary for efficient operation at rapid heart rates; it also allows filling and emptying at low pressures and increases the overall efficiency of the pump mechanism by applying the energy of systole to power diastole.

The heart's gross motion aids filling in another and subtler way as well. The relatively thin-walled right ventricle is shaped like a pocket that partially encloses the left ventricle, as a hand might wrap around a closed fist. The forceful recoil of the relaxing left ventricle can actually lift the right ventricle up and over the blood in the right atrium. The blood, which flowed into the right atrium during the prior systole, need only remain relatively motionless in order to be enveloped by the moving right ventricle. By this mechanism enhanced contractility of the left side of the heart can increase the output of the right side: when the contractility of the left ventricle is enhanced, the left ventricle will have greater systolic emptying and a greater recoil, and it will push the right ventricle over the blood in the right atrium with greater force.

In the model we propose, the early filling of the ventricle is controlled by relaxation and expansion of the heart muscle and by the heart's ballistic recoil. The final volume of the ventricle is then adjusted by the pressure of the incoming blood. In certain pathological states the heart loses some of its recoil and suction and fills incompletely. In these situations the conventional Frank-Starling law becomes accurate for describing cardiovascular function. The diastolic filling pressures increase in order to enhance the volume of the subsequent stroke, but higher pressures can also cause venous congestion, which may ultimately result in congestive heart failure.

In our model the heart is viewed as a dynamic suction pump that depends for its function on the coordinated and integrated properties of a hierarchy of structures at all levels of detail. The levels range from reexpansion of the muscle cell, caused by cytoskeletal elements and double overlaps within the sarcomeres, through tension in the connective tissue wrapped around individual muscle cells and linking one cell to another, to the movement of the whole heart in the chest. The motion of the heart is seen as a unified system, each phase of which contributes to the heart's efficient operation during the next phase. With further refinement this model may provide a complete explanation of cardiac function, an explanation that includes the Frank-Starling model as a limiting case.

Sexual Selection in Bowerbirds

The bower, or mating site, of these extraordinary birds of Australia and New Guinea is the center of intense competition among males. The female's mating choice is based on its architectural adornment

by Gerald Borgia

Naturalists have long been captivated by the complex and highly elaborate bowers of bowerbirds. Charles Darwin approvingly quoted the contemporaneous assessment of the ornithologist John Gould: "These highly decorated halls of [bowerbird] assembly must be regarded as the most wonderful instances of bird-architecture yet discovered." The interest of the two men is not hard to understand. The bowers of bowerbirds are found from dense forests to open grasslands in New Guinea and Australia, and they have no parallels in the animal world.

There are 18 bowerbird species, and the males of 14 of them decorate clearings or build bowers. There are several kinds of bower. The avenue bowers are formed of two vertical walls of sticks built on a broad platform base. In most species the end of the avenue opens onto a display area where decorations are exhibited. The maypole bowers are made of sticks woven around a sapling or a fern and surrounded by a circular raised court. Two species build a massive hutlike structure about 1.5 meters high around the maypole. The structure encloses a domed runway that opens onto a cleared exhibition area. The golden bowerbird places sticks on adjacent saplings joined by a cross branch, which is used as a display perch. Two other species clear and decorate display courts, but they do not build bowers. The males of one such species build a mat of ferns decorated with shells, and they drape lichen over nearby trees. The males of the other species clear a court on the forest floor decorated only with large leaves.

The decorations associated with the bowers vary greatly among species and include naturally found objects: snail shells, pebbles, feathers, insect parts and bits of bone. Near human settlements the decorations include manmade objects that match the color of natural objects: coins, clothespins, plastic bottle tops, pieces of glass, jewelry, paper, teaspoons, nails, screws, thimbles and the like. The stealing habits of the great gray bowerbird are well known to the aboriginals of northern Australia. Fathers-in-law in that society are notorious for taking whatever they want, and the aboriginal word for father-in-law, *juwara,* is also applied to the bowerbird.

Darwin's fascination with bowerbirds went far deeper than mere appreciation of their coloring, playfulness and architectural skill. When bowers were first noticed by Western observers, they were thought to be nests. As early as 1865, however, Gould carefully observed that bowers were sites for sexual display and mating. Darwin, who was familiar with Gould's work, discussed bower building in his book *The Descent of Man and Selection in Relation to Sex.* Apart from its use as a stage, however, the function of the bower had remained obscure until the recent resurgence of interest in sexual selection. Bowerbirds and their bowers afford a unique opportunity to evaluate competing theories of sexual selection. With the new theoretical tools biologists are now beginning to understand the evolution of one of the most extraordinary behaviors in animals, and it is possible to give a much fuller answer to the question: Why do bowerbirds build bowers?

Darwin viewed sexual selection as a process separate from natural selection because the main selective forces in sexual selection are social rather than environmental. Indeed, exaggerated male display characteristics, such as the ornate plumes, calls and dances seen in a variety of avian groups, cannot be reasonably explained without considering the effects of social interactions. Darwin described two elements in sexual selection: male competition for females and female choice of males. Male competition is evident in many species and its importance is generally accepted. The role of female choice, however, remains a hotly debated issue.

The codiscoverer of the process of natural selection, Alfred Russel Wallace, disputed the importance of female choice in animal courtship. He doubted that females of nonhuman species had the mental capacity to differentiate among males. The capacity of females to choose is no longer at issue, but the criteria on which the choice might be based and its relation to male competition must still be resolved. In a wide variety of animals as different as scorpionflies and mockingbirds male contributions of food or other materials appear to be the basis of female choice. The males of most bird species take an active role in parental care, and the likelihood of male assistance as perceived by the female seems to be critical in determining the pairing patterns. Male contributions are generally labor-intensive, and in such species the females benefit by choosing unpaired males as mates. The result is that among birds most pairings are monogamous.

For a few avian species, however, such as the prairie chicken, the cock-of-the-rock, the peafowl and the bowerbirds, the males offer no direct assistance to the females and provide them only with sperm. In such species the males have evolved extreme characteristics of sexual display, and their effectiveness in attracting mates varies widely among individuals. Yet in spite of the lack of tangible inducement, the females of such species show a strong preference for particular males. There is much discussion about why such preferences exist and how they might be related to the evolution of elaborate displays in males.

There are several divergent views about how sexual selection functions when males contribute only sperm

SATIN BOWERBIRD (*Ptilonorhynchus violaceus*) of eastern Australia is shown standing on the platform at the north end of his bower. The deep, iridescent satiny blue coloration identifies the bird as a mature adult male; females and younger males of the species are dark green on the back and spotted yellow-white on the underside. The bower is built by the male on a court cleared on the ground. It is made of sticks woven into two vertical walls that enclose an avenue on the surface of the broad platform base. Bowerbirds also build nests in which the young are born, but the bower and the nest are distinct in shape, building location and function. The decorated platform of the bower serves as a stage for the display of the male and as a mating site. The structural quality of the bower and its decorative embellishments are major factors in the female's choice of a mate, and so the bower is the focus of the competition among male bowerbirds for mates. The satin bowerbird, intensively studied by the author and his colleagues, is the only bowerbird species for which patterns of female choice are known in detail. The photograph was made by Hans and Judy Beste.

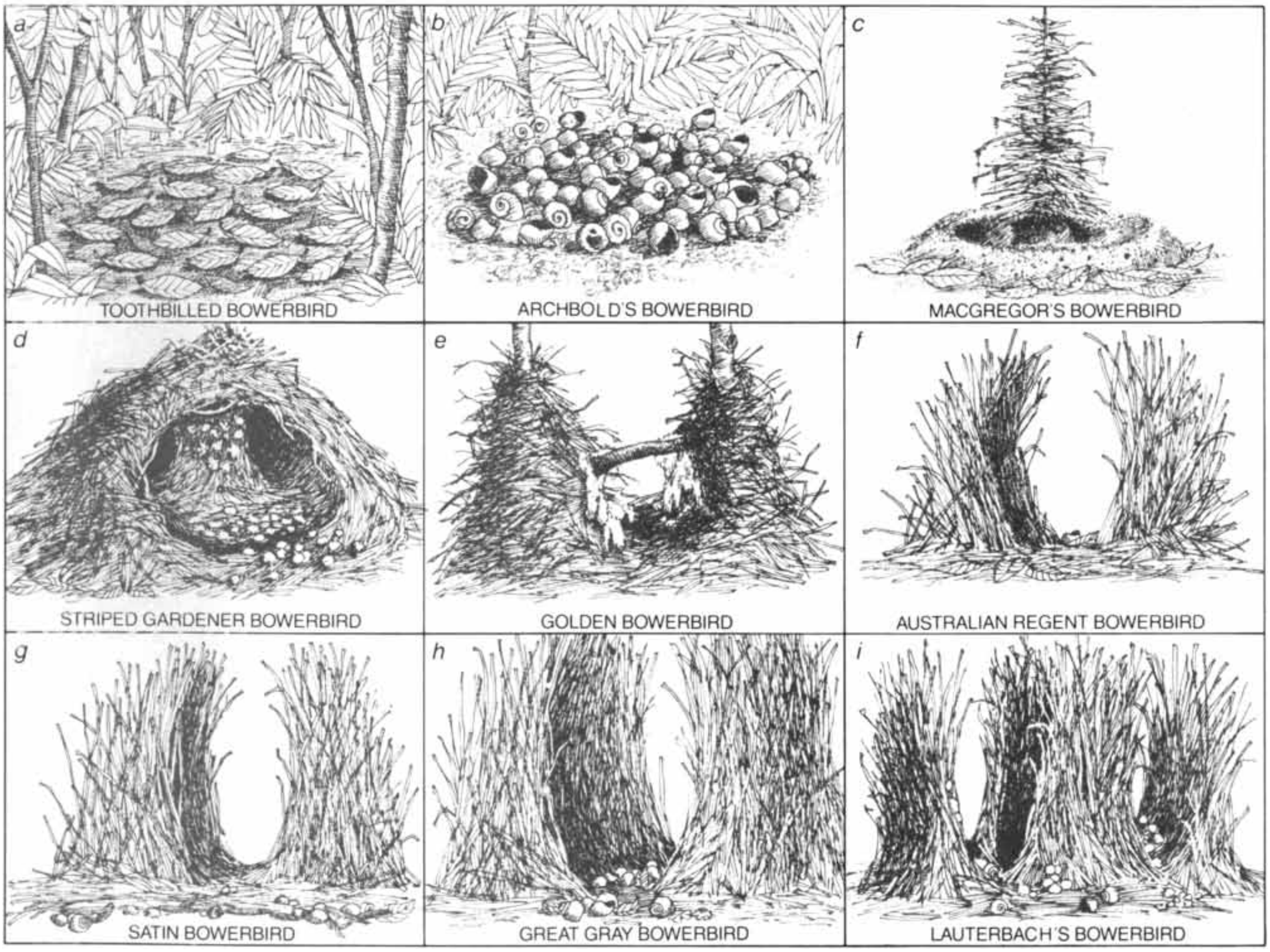
to females. The two most widely discussed general hypotheses are known as the good-genes models and the runaway models. In both kinds of models the female actively chooses her mate, and both models give a plausible explanation for the evolution of bowers. Each of them, however, depends on unproved assumptions, and neither has been shown to have operated in a natural population. Two other models, the proximate-benefit model and the passive-choice model, have been much less discussed, but they may turn out to be important in understanding the evolution of male display.

The hypotheses collectively known as good-genes models stress that a female that discriminates among her potential mates can enhance the overall health and vigor of her offspring. Thus according to the model, male displays evolve because they provide females with information about the relative quality of a male as a sire. But to what aspects of the male should a discriminating female attend?

Richard D. Howard of Purdue University, Tim Halliday of the Open University and several other biologists have suggested females favor older males over younger ones. The older males have presumably demonstrated their hardiness simply by having lived to an advanced age. For example, suppose older males tend to carry heritable traits promoting survival, such as disease resistance or predator avoidance. Then all else being equal, if the good-genes hypotheses are correct, the female should prefer the older males as mates. How is she to make that choice? One plausible strategy might be to attend to the more elaborate display characteristics among her suitors on the assumption that the more practiced and elaborate the male's display, the older the male. Thus the female's search for an older mate might explain the elaboration of male display characteristics.

Females might also seek the genetic enhancement of their offspring by favoring active males able to court vigorously or by favoring males with bright plumage, which could indicate male health and disease resistance. Hence the sexual-display patterns of the males, including their exaggerated plumage and decorated bowers, may have evolved to provide information to females about heritable, fitness-enhancing traits. Complex traits, such as the male's overall vigor, may summarize the effects of genes throughout the genome. Such traits could be quite useful to the female in choosing a genetically superior partner.

In many animal species males compete with one another for access to females. Displays that show the dominance of a male in aggressive encoun-



MALE BOWERBIRD CONSTRUCTIONS vary from simply ornamented clearings on the ground (*a*) to elaborately decorated bowers. The bowers are classified into three or four major kinds. The mat bower (*b*) is a mat of lichens on the forest floor. The mats are well decorated with piles of snail shells. The maypole bower (*c–e*) is built of sticks woven around a central pole, often a sapling or a fern, and the structure is surrounded by a circular raised court. In two species males build a large hut over the maypole and pile decorations on a court near the entrance (*d*). One species (*e*) piles sticks on adjacent saplings joined with a cross branch that functions as a display perch. Some biologists classify this bower as a separate kind. The avenue bower (*f–i*) is built of walls of sticks that enclose the avenue, which opens onto a platform. The decorations may be exhibited on the platform, on the avenue or on both, depending on the species. Lauterbach's bowerbird (*i*) builds an avenue bower in which a second set of walls is perpendicular to the main avenue.

ters may reliably indicate his superior fitness relative to other nearby males. I have therefore suggested females may prefer to choose among males that give ostentatious displays because the females are seeking dominant males. Males giving such displays without harassment must be dominant because subordinate males attempting to give the same display would be challenged by higher-ranking males.

If mates are selected according to display patterns alone, one would expect that all mature males would have the plumage needed for a display but that only some of them would earn the opportunity to show it off. There is an analogue to this prediction for the decoration of bowers. Male bowerbirds continually attempt to destroy other bowers and steal their decorations, and that behavior is analogous to the competition for the opportunity to display plumage. Together with Stephen G. and Melinda A. Pruett-Jones of the University of California at San Diego, I have proposed that a male's ability to maintain a decorated bower of high quality may serve to indicate to the female his relative quality as a sire.

Amos Zahavi of the University of Tel Aviv has suggested the female might gain genetic benefits for her offspring if she favors a so-called handicapped male: a male displaying debilitating physical characteristics or behaviors. According to Zahavi, the elaborate sexual displays are genetically inherited handicaps. The female allegedly gains genetic benefits by choosing a male with a handicap because such a male has survived in spite of a highly disadvantageous trait.

The model has met with much criticism. The most significant flaw is its reliance on the environment to select only the fittest of the handicapped males. If the females are to find the best sires in the population, there must be a very high rate of male mortality attributable to the handicap. Otherwise the overall genetic superiority of the handicapped survivors in the population could not be guaranteed. The model's requirement of high mortality among the sons of handicapped males makes it unlikely that female preference for superior males could evolve according to such a scheme.

AUTHOR'S STUDY AREA is in a valley formed by Wallaby Creek in the Beaury State Forest of New South Wales. The inset map shows the location of the site as well as the range of two subspecies of satin bowerbird: *P. violaceus violaceus* (*color*) and *P. violaceus minor* (*gray*). Eucalyptus is a dominant canopy tree over much of the area; the understory varies and includes grassland and thick thorn scrub. Rain forest predominates in the low areas, along creeks and on the eastern side of the ridges west of Wallaby Creek. Sites of established male bowers of the study area are shown as colored circles; the diameter of the circles indicates the number of matings per season for the bower's owner. Sites of temporary bowers, built by younger males and destroyed soon after, are shown as black circles.

The runaway model for sexual selection was proposed by Ronald Fisher in 1930 to explain the evolution of exaggerated characteristics found only in males. He was the first to note that the pattern of female choice could be self-reinforcing: it could cause its own spread among females and the spread of the male-display characteristics on which it is based.

To illustrate the runaway process suppose there is a population in which there are two kinds of male and two kinds of female. The males differ in the presence or absence of a display characteristic, such as a red tail feather. The females differ in that some of them (the "choosers") mate only with males having a red tail feather, whereas others (the "nonchoosers") do not distinguish among males on the basis of tail-feather color.

In such a population the males having a red tail feather can mate with both chooser and nonchooser females. In contrast, males without a red tail feather can mate only with nonchooser females. Thus red-tailed males have more opportunities to mate and produce a greater proportion of progeny than males without a red tail feather; the proportion of red-tailed males in the population thereby increases. The sons of red-tailed males and chooser females carry a greater-than-random proportion of the genes that lead to choosing behavior in females. As red-tailed males mate more often, the proportion of the traits specifying a female preference for red-tailed males also increases.

Once a pattern of female choice is established, there can be continued selection for more exaggerated male traits. The outcome of the runaway process depends strongly on the choice pattern. If females consistently prefer males having extreme characteristics, such male characteristics are expected to evolve. If, however, females prefer males with a less extreme trait, such as a single red tail feather, males in the population with one red tail feather will tend to predominate over males with an all-red tail. The survival costs of sexual display can affect the process. For example, as extreme characteristics develop in males, new female-choice patterns may arise that cause a reverse runaway toward less extreme male display characteristics.

Very little is known about how new

female-choice patterns arise. Moreover, when they do arise, it is not clear whether they favor only a slight enhancement of male characteristics or an extreme development. Several recent models of runaway selection conclude that female choice can be completely arbitrary and so can give rise to the evolution of arbitrary male characteristics.

The good-genes models and the runaway models thus make differing predictions about how extreme display traits are established. According to the good-genes model, natural selection should favor female preferences for traits that indicate differences in male qualities as sires. In the runaway models such an outcome is not necessarily expected. A runaway selection may not enhance the vigor of the offspring, and it might even promote traits that reduce their fitness.

The extent to which traits that reduce fitness might evolve is still a matter of debate. It would seem more likely that if different kinds of female preference were expressed in a population, the kinds tending to enhance fitness would have an evolutionary advantage over the ones that did not. The extent to which traits enhancing fitness win out over less advantageous ones must depend on at least two factors: the ease with which established traits can be replaced and the frequency

PATTERNS OF BREEDING BEHAVIOR of the satin bowerbird are depicted in the six-part diagram. The male builds his bower out of sticks (*top left*). When a female visits the bower, the male moves down to it from a perch in a nearby tree and begins his courtship (*top right*). After watching the display the female may crouch and tilt forward, and the male immediately mounts her (*middle left*). Other males may compete with the owner of the bower for females by destroying the bower in the absence of the owner (*middle right*) or by stealing its decorations (not shown). A competing male may even try to interrupt or displace a male during copulation (*bottom left*). After successful copulation the female lays her eggs and rears her young in a nest that is separate from the bower (*bottom right*).

with which competing patterns of choice are found in a population.

Comparing the good-genes models with the runaway models is further complicated because the models need not be mutually exclusive. For example, if the initial female choice depended on the effects of good genes, it could lead to the selection of female-preference traits that are self-reinforcing. It is therefore unrealistic to expect the predictions of these models to be easily distinguishable in natural populations. Nevertheless, if the controversy is to be resolved, it will only be through studies of natural populations. One approach is to determine whether or not the patterns of male display are really arbitrary. Alternatively, one could show that they do indeed indicate the relative quality of the males as sires.

I noted above that there are at least two more possible explanations for the evolution of bower-building behavior. According to the proximate-benefit model, there may be immediate benefits for the female that chooses a male giving an extravagant display. For example, bright male plumage may better enable females to detect parasites and so avoid contact with males likely to transmit the parasites. Females that discriminate against infested males gain an obvious, immediate benefit over less choosy females. Other kinds of immediate gain might also arise for the discriminating female. A male giving a display without interference from other males might be able to offer superior protection from predators, and males with elaborate plumage are more likely to be mature and so carry viable sperm.

Finally, some biologists have suggested female choice has had little to do with the evolution of exaggerated male display. According to the passive-display model, put forward by Malte Andersson of the University of Göteborg and Geoffrey A. Parker of the University of Liverpool, elaborate male displays may have evolved as advertising devices. Males with the most ostentatious displays are more readily visible to females and so such males have more chances to mate. Gains from the additional matings repay the males for the extra cost of surviving while giving an extravagant display. The model does not require that females exhibit any active preference for males with larger displays. If a female simultaneously encounters two males that differ in the extent to which their displays are elaborated, the model predicts the size of the displays should not influence her mating decision. For bowerbirds this model implies that the bower is a device for advertising to females the presence of a courting male and that males with larger, better decorated bowers are sexually successful because they are more often found by females.

Although substantial efforts in constructing theoretical models have sharpened the questions one would like to answer about how elaborate displays have evolved, there have been few attempts to test the models in natural populations. In part the lack of testing is a result of the difficulty of finding observable subjects with appropriate characteristics. The special characteristics of the bowerbird's display make it possible to study female preferences in detail. I have already mentioned that male bowerbirds give no material assistance to females or to their young, and the females are free to choose among males from widely separated display sites. One can therefore assume that biases observed in female preferences for males are related to characteristics of the male's display, including his bower, plumage and behavior in the presence of the female.

Furthermore, there may be a functional equivalence between the decorations on the bower and brightly colored plumage that may make it possible to manipulate the general display patterns experimentally. More than two decades ago E. Thomas Gilliard [see "The Evolution of Bowerbirds," by E. Thomas Gilliard; SCIENTIFIC AMERICAN, August, 1963] suggested there is an inverse correlation between the degree of plumage elaboration in males and the size and degree of the decoration of the bowers. Thus, Gilliard noted, the decorated bowers may play the same role in courtship as showy displays of plumage; in fact, he suggested, bowers are a kind of displaced plumage that allow the animals building them to dispense with bright coloration. The development is called the transfer effect. If it is real, it suggests the same forces shaped both the evolution of bower building and decorating behavior and the displays of showy plumage.

Unlike plumage, the bower and its decorations can be easily manipulated and quantified with no direct effect on the bird. Matings take place at the bower, and so cameras monitoring the bowers can record the choice of a mate by the females and the mating success of the males. The observations can be compared with the quality of the bower and the elaborateness of its decoration. Finally, male behavior that can be observed near the bowers is an important indicator of how competition can distinguish among potential suitors. Males often steal bower decorations and destroy the bowers of other males, and the patterns of such aggressive behavior can be compared with the quality of a male's display and his success in finding a mate.

In 1980 I began an intensive study of the satin bowerbird (*Ptilonorhynchus violaceus*) in eastern Australia. J. M. Marshall, Reta Vellenga and Richard Donaghey, who was then at Monash University, had done important early work, and their efforts made the life history of satins the best documented of any bowerbird species. Donaghey allowed me to take over a study population with which he had worked, so that the histories and identities of some males had already been known for four years. At that time the details of female choice were not known for any bowerbird species.

The satin bowerbird ranges along the southern and central east coast of Australia. My research site is in a valley formed by Wallaby Creek in Beaury State Forest of New South Wales [*see illustration on page 95*]. Vellenga had shown that the male is slightly larger than the female and gets his satiny blue plumage in his sixth year. The coloration of adult males differs markedly from the dark green back and spotted yellow-white underside of females and juveniles. The male builds bowers made of sticks on courts cleared on the ground.

The north end of the bower faces the sun at midday and opens onto a display platform, which the mature male covers with bright yellow straw and yellow leaves. A variety of decorations are laid on the platform, including blue parrot feathers, blue and yellow blossoms, insect parts, in particular the outer coverings of cicadas, and other natural objects. Large objects including the shells of land snails are arranged at the outer edge of the platform, and feathers are distributed evenly over the yellow platform between its outer edge and the avenue of the bower. Small objects held by the male in his mouth during courtship are found in a small pile near the bower.

The carpet of yellow straw and leaves on the bower platform creates a bright glow that is particularly noticeable at bowers in forests. The male prunes the leaves above the platform, apparently to allow sunlight to illuminate the platform. The display of shiny blue objects, which are relatively uncommon, and their placement on a yellow background suggests an attempt to give an unambiguous and highly visible signal.

In order to study the behavior of individuals we capture bowerbirds flocking in open pastures. Each bird is fitted with a unique color-band combination, measured and assessed for plumage, the color of its beak and legs, scratches or other evidence of fighting and external parasites. We assign the bird to an age category according to the colors of its plumage and its beak.

My volunteer assistants, graduate students and I record the behavior of males at feeding sites, noting the number of times a male is attacked and how often he attacks others. At the bowers we observe the birds from hides, and we also continuously monitor activity at bowers with remote-control super-8 motion-picture cameras throughout the mating season. The cameras expose a frame every two seconds as long as there is a break in an invisible beam of infrared light passing through the bower. This record enabled us to identify individual visitors to the bowers and to note their activities, including the destruction of the bower, the stealing of its decorations, courtship and copulation. We also made daily records of the bower quality and the movement of marked decorations at more than 33 bowers in the course of the study.

At Wallaby Creek mating begins in early November and continues until late December. In mid-October males become active around the bowers: rebuilding of the bowers at permanent sites is completed, and intensive decorating is begun. Young males visit the bowers and the bower owners often display to them. By early November, however, the bower owners become less tolerant of male visitors; they spend more time near the bower and engage more actively in destroying nearby bowers and in stealing their decorations. Females overwintering at the southern end of the valley start moving north toward the highest concentration of male bowers. In mid-November the matings begin. The males then spend most of their time perched in trees near the bower, calling frequently and moving down to the bower to display to females, to protect it from marauders, to build it or to "paint" it with the saliva generated in chewing bits of vegetation. Matings peak at the end of November and are mostly finished by mid-December.

When a female visits the bower, the male begins his display, often holding a decoration in his beak. He faces the female while he stands on the platform. He gives a whirring call while prancing, fluffing up his feathers and flapping his wings to the beat of the call. Calls are punctuated with periods of silence, quiet chortling, buzzing or mimicry of other birds. The female's initial response is to enter the bower and "taste," or nip, at a few sticks. Then she intently watches the courtship. If she is ready to copulate, she crouches and tilts forward. The male immediately mounts her. At any stage she may leave, thereby ending the courtship. Typically a female mates only once. She later lays two eggs in a nest that is usually outside the area defended by her mate. The hatching of the eggs coincides with the emergence of large numbers of cicadas.

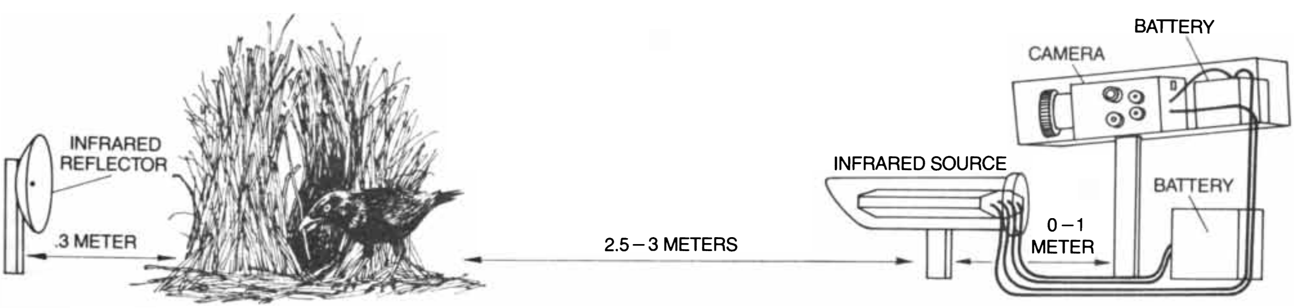
Established males destroy bowers other males may try to build nearby, but during the mating season young males can establish temporary bowers at sites removed from the permanent ones. There is intense courtship at the temporary sites among males that appear to be practicing and learning the display. Such males also visit permanent bowers. If the owner is not present, the visitor may paint the bower, attempt a display or court a visiting female. During courtship by a bower owner other males may hide in the surrounding vegetation and then try to interrupt or displace a copulating male.

Most hypotheses about the evolution of elaborate displays assume the quality of the display affects the female's willingness to mate. For satin bowerbirds mating success varies widely among males: we observed one male that mated with 33 females during the season, whereas many other males did not mate at all. When we ranked bowers for quality, we found a strong positive correlation with male mating success. Neat and well-built bowers with symmetrical walls, fine, densely packed sticks and a highly sculptured appearance were owned by particularly successful males.

To test the importance of bower decorations in the female's choice of a mate we removed the decorations from the bowers of a randomly selected group of males. We then compared their success in mating with the success of a control group we did not disturb. As we had predicted, the decorations do indeed influence mating: we found a significantly higher rate of mating in the control group than we did in the experimental group. Moreover, the number of decorations on the platforms of undisturbed bowers—particularly the number of feathers—was correlated with male mating success in each year of the study.

The discovery that decorations are important in male display led me to a study of decoration stealing. We found that blue feathers are stolen at a much higher rate than other decorations in proportion to their frequency on the bower platforms, and that (because they are rare in the habitat) stealing is the principal means of obtaining them. By monitoring the number of feathers on bower platforms throughout the mating season we discovered that the number on the platforms of successful males peaks at the height of the mating season. At the same time the number of feathers on the platforms of unsuccessful males is reduced.

In another experiment we introduced individually marked blue feathers to the bowers of a group of males, and we added no feathers to the bowers of a control group. We then reversed the treatments of the experimental and the control groups, and we recorded the movements of the feath-



EXPERIMENTAL MONITORING SYSTEM employed by the author in his study of the satin bowerbird is diagrammed schematically. An infrared beam, invisible to the bowerbird, is projected through the avenue of the bower to a reflector. When the beam is interrupted, a super-8 motion-picture camera exposes one frame every two seconds. Birds were also observed from hides. The system enabled the author to monitor the behavior and identity of bower owners and visitors at 33 bowers for the 50-day mating season.

ers in each instance. Yet whatever the initial placement of the feathers, the same males, namely the most active thieves, tended to accumulate the feathers on their bower platforms. The result is strong evidence that stealing is the most important factor in the final number of feathers displayed on the platforms. Because the quality of the bower decoration affects the success of the male in mating, feather stealing appears to affect mating success. More dominant males tend to be more successful at feather stealing.

We also found that the average quality of a male's display depends on the frequency with which it is destroyed by marauding visitors. The more frequent the destruction, the lower the overall quality of the bower. The pattern of destruction can indicate male dominance to the female because females generally limit their bower visits to a small area. If a bower is maintained in relatively good condition, it can serve as a signal to the female that the owner can defend it from attack and destroy the bowers of his neighbors as well. The behavior patterns we found for bower decorating are consistent with the patterns of bower destruction: the bowers of older, dominant males are destroyed less often than those of younger, subordinate males.

How do such observations conform with the models for the evolution of selection patterns I have described above? There are several lines of evidence to suggest females favor dominant males. They choose mates that are able to keep bowers in good repair and well decorated. Such males tend to be the ones that are dominant at feeding sites; their decorative preferences are for objects scarce in the habitat and prized by other males. The age of the male also seems important. Older males maintain bowers of better quality and decorate them more elaborately, and they are more successful in protecting their bowers from destruction. Moreover, older males give more refined courtship calls.

It may be that both male dominance and age are important to the female's choice. Bower building appears to take some practice, but only the most aggressive young males are able to practice building and decorating bowers in the face of repeated destruction by other males. A female that chooses an older, established male with a well-built, well-decorated bower and a refined courtship call has evidence that her prospective mate not only has been able to survive to a relatively old age but also has been able to do it while learning to build and maintain a high-quality bower under the rigors of male competition.

The remaining good-genes model, the handicap, is not supported by the behavior of bowerbirds. The model predicts that large differences in male mortality are associated with differences in the quality of male displays. That is not what one finds. The mortality of displaying males is low, and it appears to be independent of the quality of male displays.

The runaway model can lead to a large number of possible outcomes. Versions of the model that suggest arbitrary outcomes yield no prediction about the kinds of traits that should evolve; hence they cannot easily be falsified. In some cases an arbitrary choice resulting from runaway selection might give rise to the same behavior as a choice made on the basis of male dominance. For example, if females had evolved a tendency to favor males with well-built, well-decorated bowers, males able to steal decorations and destroy bowers, or in other words the dominant males, would thereby be selected.

The existence of patterns consistent with other models weakens the case for the runaway models, particularly if the patterns can be shown to recur in various species. In satin bowerbirds we found that males favored scarce decorations. If a similar pattern in other species were found, it would support the suggestion that the female assessment of male dominance is important in the choice of a mate.

Can passive choice explain the patterns of mate selection we observed? Probably not. We found that sexually successful males tend to receive more female visitors to their bowers than less successful males, and they mate with a greater proportion of the females. Generally females visit several bowers before mating, and their choice is correlated with the overall mating success of the males. In contrast, passive choice seems unlikely. Bowers are on the ground and often under cover; the call of the bower owner from his nearby perch is much more noticeable to a naive female than the bower itself. Furthermore, females are long-lived and therefore probably familiar with all the bower sites in the area they search; it seems unlikely that mere prominence would affect the choice of such females.

There may be some proximate benefit from the bower for the female as a protection from intruding males. We often saw males trying to interrupt matings by bower owners. Usually the owner chased the intruder away and the female remained in his bower, although occasionally copulation was interrupted. When females were on the ground outside a bower, however, they seemed to be uneasy and commonly flew away if intruders came near.

To summarize, the evidence available suggests female satin bowerbirds actively differentiate among males according to the quality of their displays. The females may also choose sires according to the decorations associated with the bowers. The protection hypothesis offers an alternative explanation for the evolution of bower building, but it does not explain why the bowers are decorated. The runaway model cannot be excluded as a possible alternative explanation.

The work on satin bowerbirds is a first step in understanding the evolution of exaggerated characteristics, and we have established the plausibility of several models of that evolution. Nevertheless, studies of mating choice in other bowerbirds will be necessary if one is to explain why bowerbirds build bowers. To what extent do other species show a preference for rare decorations? Do they steal the decorations of other bowerbirds and destroy their bowers? How do males learn to give their displays? Will the findings suggest the same causal relations between male behavior and female choice? Why are bowers built by some species but not by others? Such questions will surely be resolved by further patient observation.

TRANSFER EFFECT, first proposed by E. Thomas Gilliard, is illustrated by the bowerbirds and their bowers in the photographs. The male Australian regent bowerbird (*Sericulus chrysocephalus*), which displays bright plumage, builds a bower of indifferent structural quality and makes little attempt to decorate it (*left*). In contrast, the male great gray bowerbird (*Chlamydera nuchalis*), a bird of dull color, builds an elaborate bower that is richly decorated with shells, flower petals and a variety of manmade objects (*right*). Gilliard suggested the size and degree of decoration of the bower are inversely correlated with the elaborateness of the plumage color. Decorated bowers may therefore have the same function for the male as colorful plumage displays, and any hypotheses that explain the evolution of the bower may also explain the evolution of colorful plumage. The photographs were made by Hans and Judy Beste.

Create new places.

For some, following the road is fine. Others, however, prefer to follow their imagination. And for them, there's the Mitsubishi Montero.

Montero's more than a car or truck. It's a system of unique technologies designed to deliver remarkable two-wheel and four-wheel drive performance. From its rugged six cross-member frame and smooth riding independent torsion bar front suspension, to its 2.6 liter MCA-Jet™ engine with patented Dual Engine Stabilizers,™ Montero creates a new driving experience wherever it goes. But unique technology is only part of the Montero legacy. At $11,464,* this Montero Sport includes AM/FM/cassette stereo, plush carpeting, soft velour cloth upholstery and full instrumentation including an on-dash inclinometer. The driver's seat even has its own adjustable suspension system. Other Monteros start as low as $9,369.*

The Mitsubishi Montero. Sophisticated performance. Discriminating comfort. And engineered, above all else, to let your imagination take you where you want to be.

Mitsubishi Montero

Mitsubishi takes you where you want to be.™

Winner '86 Paris-Dakar Rally, Marathon Class 4th Straight Year.

Call 1-800-447-4700 for your nearest Mitsubishi Motors Dealer.
*Mfr's. Suggested Retail Price. Actual price set by dlrs. Tax, license, freight, options, other dlr. charges extra.

The Perception of Apparent Motion

When the motion of an intermittently seen object is ambiguous, the visual system resolves confusion by applying some tricks that reflect a built-in knowledge of properties of the physical world

by Vilayanur S. Ramachandran and Stuart M. Anstis

Producers of motion pictures, television programs and even neon signs have long banked on the fact that human beings have a quirk in their visual system. When it is confronted with a rapid series of still images, the mind can "fill in" the gaps between "frames" and imagine that it sees an object in continuous motion. For instance, a series of neon arrows lighted up in succession are perceived as being a single arrow moving through space. The illusion of continuous motion is called apparent motion to distinguish it from "real" motion, which is perceived when an object moves continuously across a viewer's visual field. When Sir Laurence Olivier appears to be fencing in a film, he is in apparent motion, whereas a person walking across the theater in front of the screen is in real motion.

In the century or so since the motion picture was invented, filmmakers and television workers have learned to create many compelling illusions of motion, but their progress has been furthered mainly by rule-of-thumb empiricism. Psychological research is only now beginning to describe the mechanisms by which the visual system—the retina and the brain—perceives apparent motion.

The starting point of our own investigations was the premise, set forth by Bela Julesz of the AT&T Bell Laboratories and Oliver J. Braddick of the University of Cambridge, that to perceive an intermittently visible object as being in continuous motion the visual system must above all detect what is called correspondence. That is, it must determine which parts of successive images reflect a single object in motion. If each picture differs only slightly from the one before it, the visual system can perceive motion; if successive pictures differ greatly, the illusion of motion will be destroyed.

Our main question, then, was: How does the visual system go about detecting correspondence? One popular view holds that the brain does so by acting like a computer. When an image stimulates the retina, the eye transmits the image to the brain as an array of tiny points of varying brightness. The brain then compares each point to every point in succeeding frames. By means of complex computations the brain finally discerns the one set of matched points composing a single object that has changed its position—has moved. Attempts to build machines that "see" are generally based on this principle.

The scheme seems logical enough when a simple, unambiguous display is presented. For instance, if a small dot is shown in one frame and is followed by an identical dot placed slightly to the right, the visual system will readily identify the dot in the first frame as an object and find it again—displaced—in the second frame [*see top illustration on page 104*].

The scheme becomes problematical, however, when correspondence is to be detected in more intricate displays. For example, suppose two identical dots are shown in vertical alignment on a computer or television screen and are then replaced by congruent dots shifted to the right. In theory the visual system is now confronted with two possible correspondences: the dots in the first frame could be seen to jump horizontally along parallel paths to the right, or they could be seen to jump diagonally, in which case they would have to cross paths. In practice viewers always see the dots moving in parallel, never crossing.

In another display a computer-generated random-dot pattern forms the first image; then a square region is cut out of the middle and shifted horizontally to create the second image [*see bottom illustration on page 104*]. To the unaided eye the second image appears to be identical with the first and to have no separate central square. Now the images are superposed and then alternated rapidly so that the outer dots are in perfect register, or correlate, and so appear to be immobile. The middle region, where the dots are out of register, appears to move: a well-delineated square is perceived to be oscillating from side to side.

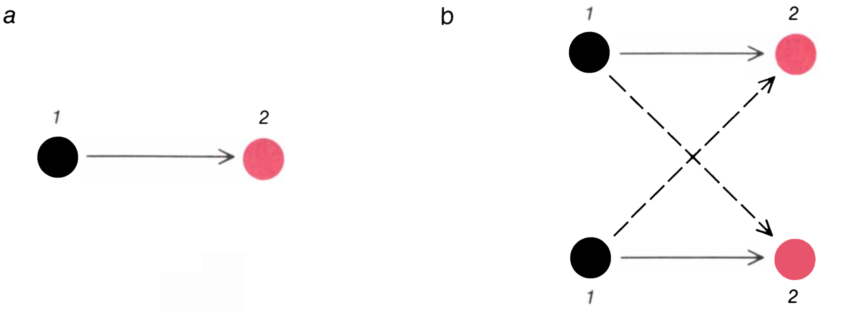
To produce these two illusions by means of point-to-point matchings the brain would somehow need to invalidate hundreds of potential matches, deeming them to be false. While it is possible that the brain laboriously matches all the points and then subjects the matches to a series of elimination tests, our investigation suggests an entirely different approach to detecting correspondence: the visual system applies strategies that limit the number of matches the brain needs to consider and thereby avoids the need for complex point-to-point comparisons.

We believe perception of apparent motion is controlled in the early stage of visual processing by what is in effect a bag of tricks, one the human visual system has acquired through natural selection during millions of years of evolution. Natural selection is inherently opportunistic. It is likely that the visual system adopted the proposed visual short cuts not for their mathematical elegance or aesthetic appeal, as some would suggest, but simply because they worked. (We call this idea the utilitarian theory of perception.) In the real world anything that moves is a potential predator or prey. Hence being able to quickly detect motion and determine what moved, and in what way, is crucial to survival. For example, the ability to see apparent motion between widely separated images may be particularly important when detecting the motion of animals that are seen intermittently, as when they move behind a screen of foliage or a tree trunk.

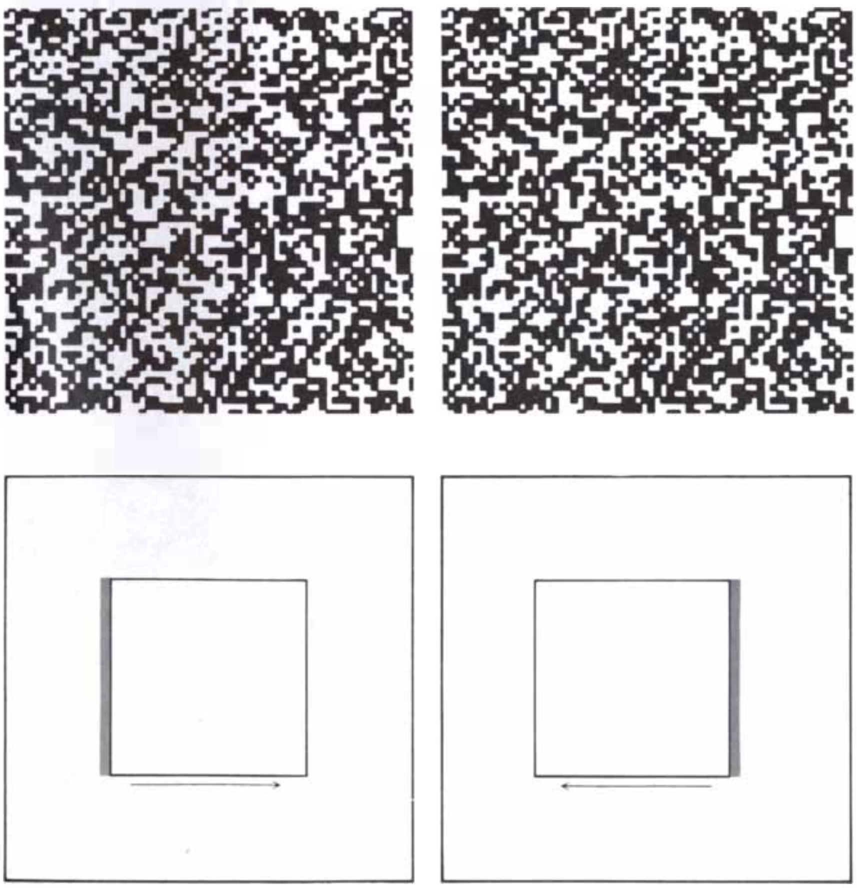
One trick of the visual system is to extract salient features, such as clusters of dots rather than individual dots), from a complex display and then search for just those features in successive images. This significantly reduces the number of potential matches and thus speeds the perceptual process; after all, the probability that two chunks of a visual scene will be similar is much smaller than the probability that two points of brightness will be similar.

Among the features the visual system might attempt to extract from images are sharp outlines and edges or blotches of brightness and darkness; the latter are technically called areas of low spatial frequency. We have evaluated each of these and found that

SUCCESSION OF FRAMES (*top to bottom, left to right*) **capture a sneeze. They are from an early motion picture made in Thomas A. Edison's laboratory in about 1890. In order to perceive continuous motion when still images such as these are flashed, the visual system must above all detect correspondence; that is, it must identify elements in successive frames as being a single object in motion.**



DOT DISPLAYS that produce the illusion of motion are illustrated. In the simplest possible display (*a*) a single spot of light (*black*) is presented briefly on a computer screen and then is replaced by an identical spot displaced to the right (*color*). Numbers indicate the order of presentation. Rather than seeing two separate dots, the viewer perceives the first dot as moving horizontally (*arrow*). A slightly more complex display (*b*) is ambiguous and can be interpreted in two ways. Two vertically aligned dots (*black*) are flashed and then replaced by an identical pair displaced to the right (*color*). In theory the first dots can appear to move horizontally in parallel (*solid arrows*) or to move diagonally (*broken arrows*). In practice viewers always see the horizontal motion, a finding that raises the question: How does the visual system detect correspondence when it is faced with ambiguity? Evidence indicates that it does so by extracting salient features from images and also limiting "legal" motions to those consistent with certain universal laws of matter and motion.



IMAGES COMPOSED OF RANDOM DOTS are shown (*top*); they produce apparent motion when they are superposed and then flashed alternately. The two computer-generated images are identical except that dots in a square central region of the second image (*right*) are shifted to the left with respect to their position in the first image (*left*), as is schematically shown at the bottom. No central square is visible in either image alone, but when the images are alternated, a central square is seen oscillating horizontally against a stationary background. Computer-generated dot patterns were first introduced by Bela Julesz of AT&T Bell Laboratories and by Donald M. MacKay of the University of Keele in England.

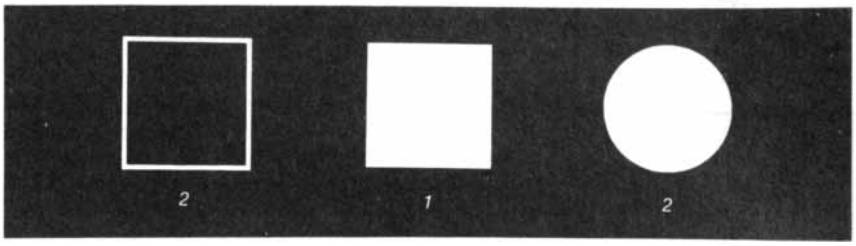
the visual system is likely to detect correspondence between regions of similar low spatial frequencies before it detects more detailed outlines or sharp edges. In other words, the visual system is likely to notice a dark blur moving in a forest long before it identifies the outline of an individual tree swaying in the breeze.

To demonstrate this principle we initially presented a white square on a black background for a tenth of a second and then replaced it with a congruent outline square to the left and a white circle to the right. (All the experiments described in this article presented images to viewers at speeds too fast for thinking; the objective was to eliminate the influence of high-order cognition and focus on the processes responsible for early perception.) Would the viewer see the white square move toward the outline square (which had the same sharp corners as the first square) or toward the circle (which had the same shading as the original square)? Subjects almost always saw the latter effect, providing evidence that the visual system tends to match areas of similar brightness in preference to matching sharp outlines.

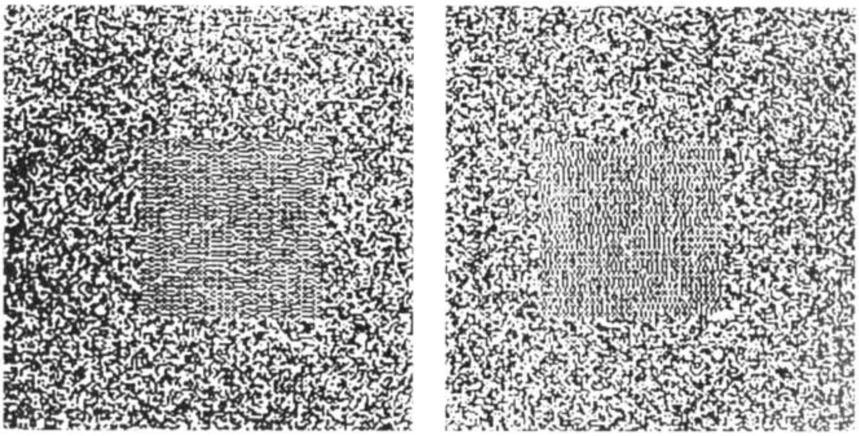
Texture is another feature that appears salient to the visual system. We and our colleagues at Stanley Medical College in Madras, India, presented to subjects two images of random-dot patterns; each image had an inner square with a visual texture different from that of the outer region [*see lower illustration on opposite page*]. The inner square of the second image was the same size and texture as the inner square of the first image, but it was rotated 90 degrees and was shifted horizontally.

We eliminated the possibility that correspondence could be detected on the basis of nontextural cues by ensuring that the dots in the two images would lack point-to-point correlation when the images were superposed and that the average brightness was the same in the inner and outer textures. We could therefore predict that if a shift of texture (such as between the inner and outer regions of the images) is a feature that enables the visual system to detect correspondence, viewers would see the inner square oscillating whenever the two images were alternated rapidly. If, on the other hand, texture is of no help in detecting correspondence, viewers would simply see visual "noise" and no coherent motion. Observers did see the oscillating square, indicating that texture is indeed an important cue for the detection of correspondence by a viewer.

Clearly the mechanism for perceiving apparent motion can accept various inputs for detecting correspondence. We have found a preference for seeing low spatial frequencies and textures; other investigators, such as Shimon Ullman of the Massachusetts Institute of Technology, have found that under certain circumstances line terminations and sharp edges also serve as cues. Perhaps the visual system perceives motion cues hierarchically, first scanning for coarse features before homing in on finer features, rather like an anatomist who first looks through a microscope set at low power before switching to higher magnification. One bit of evidence supporting this view is that subjects do indeed sometimes see the white square in the experiment cited above move toward the outline square, but only when the images are presented slowly and there is time to scrutinize the image.



FEATURES OF OBJECTS that might be extracted to detect correspondence are compared in this experiment. A solid square (*center*) is shown against a dark background and is then replaced with an outline square on the left and a solid circle on the right. The viewer who is confronted with these images usually sees the square move toward the circle rather than toward the outlined square, suggesting that regions of shadow or brightness (low spatial frequencies) are more likely to be detected initially than sharp edges or fine outlines.



TEXTURED DISPLAYS shown here are generated by computer. When they are superposed and alternated, they demonstrate that visual texture can serve as a cue for detecting correspondence. The inner squares, which are shifted horizontally with respect to each other, differ from the outer regions in texture, or distribution of dots, but not in brightness, eliminating the possibility of detecting correspondence on the basis of brightness. In addition the dots in the right-hand image do not correlate with those in the other image, eliminating the possibility of detecting correspondence by point-to-point matching. Therefore the fact that viewers see an inner square oscillate horizontally when the images are alternated can only be explained by the ability of the visual system to detect changes in texture.

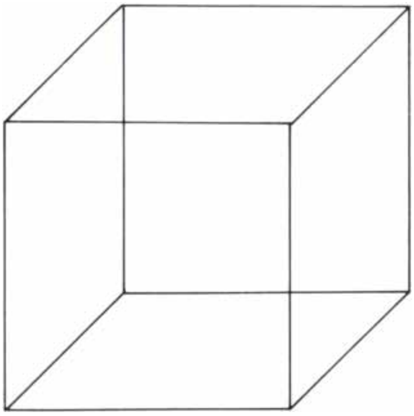
In addition to extracting salient features a second trick of the visual system is to limit the matches it will consider to those yielding perceptions of motion that are sensible, or could occur in the real, three-dimensional world. In other words, as David Marr of M.I.T. first suggested, the visual system assumes the physical world is not a chaotic and amorphous mess, and it capitalizes on the world's predictable physical properties. For instance, if the pairs of jumping dots described above were actually rocks, they would collide if they moved diagonally in the same depth plane and so would fail to reach opposite corners; the only logical perception of the dots' motion is therefore that the two dots in the first frame move in parallel to their positions in the second frame. Sure enough, when these dots are viewed through a stereoscope (a double-lens viewer) and seem to be in separate planes, observers do see them cross; in the real world, objects in different planes—such as airplanes at different altitudes—can indeed cross each other without colliding.

In order to examine the notion that the visual system assumes the world has order, we presented subjects with various motion displays that could be interpreted in more than one way and observed how subjects resolved the ambiguity. We found that one rule applied by the visual system is reminiscent of Isaac Newton's first law of motion, namely that objects in motion tend to continue their motion along a straight path. The visual system perceives linear motion in preference to perceiving abrupt changes of direction.

We demonstrated the power of this rule with an illusion that incorporated a "bistable (dual state) quartet": two dots briefly presented at diagonal corners of a square and then replaced by identical dots at the other two corners. A bistable quartet can be perceived in two ways, somewhat like the familiar Necker cube, which viewers see oscillating between two perspectives. With approximately equal frequency observers of a bistable quartet see two dots oscillating horizontally or two dots oscillating vertically.

The bistable quartet was embedded in the center part of two horizontal rows of dots that appeared to be streaming in opposite directions [*see bottom illustration on next page*]. Only one dot in each row was visible at a time. When the streaming dots reached the center of the screen, the bistable quartet became visible. At that point viewers could in theory see the dots continue in a horizontal path or could see them make a 90-degree turn followed by a second 90-degree turn, to produce two U-shaped trajectories. In practice observers invariably saw horizontal streaming, indicating that the tendency to see linear motion overcame the ability to see the dots in the quartet move vertically. The U-shaped motion was seen only when the parallel rows were brought very close to each other; then Newton's law came in conflict with a competing tendency to see motion between the closest identical points. The proximity principle gains increasing power as objects are moved closer to each other.

A second rule that limits the possibilities for correspondence is that objects are assumed to be rigid; that is,



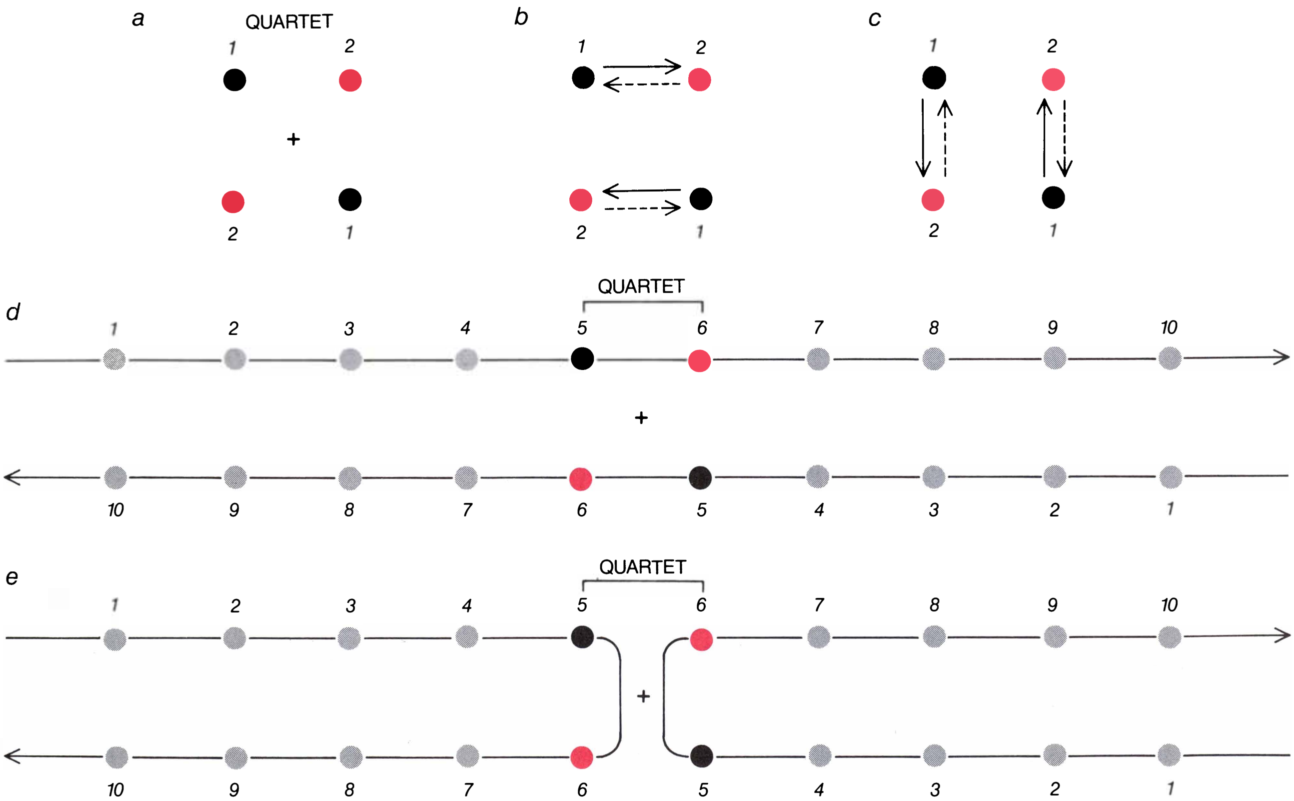
NECKER CUBE, named for the Swiss naturalist Louis A. Necker, can be seen to oscillate between two alternative perspectives.

all points on a moving object are assumed to move in synchrony. Imagine a leopard leaping from a branch of one tree to a branch of another. According to the rule of rigidity, the viewer who picks out any salient feature of the leopard, such as its basic shape (or even the splash of light shading, or low spatial frequency, of its coat), and finds the same feature in a second frame does not need to also compare every black spot on the animal. Without actually perceiving each leopard spot, the person assumes that all spots—indeed, all parts of the leopard—move in synchrony with the salient feature; correspondences suggesting that the leopard's spots can fly off in all directions are not even considered.

An experiment that demonstrated the rule of rigidity involved two uncorrelated random-dot patterns we alternated in a continuous cycle, exposing each picture for half a second [*see top illustration on opposite page*]. Viewers saw random incoherent motion, much like "snow" on an untuned television set. Now we added a narrow strip of dots to the left of and abutting the edge of each image. The "grain" of the dots in the strips was the same as the grain in the images to which they were added, but the strip added to the second image was wider than the strip added to the first one, so that the left margins of the new images did not align. When the images were again alternated, the left margin appeared to shift from side to side. Strikingly, the entire display suddenly seemed to move in synchrony with the margin as a single solid sheet. We call this effect motion capture. Apparently unambiguous motion, such as that seen at the left edge of the images, "captures" ambiguously moving fragments because the visual system tends to presume that all moving parts are fragments of a single object whose surface features move in synchrony.

A further experiment also demonstrated the phenomenon of captured motion, and particularly the ability of low spatial frequencies to effect such capture. We superposed blurred vertical bars of low-contrast lightness and darkness, called sine-wave gratings, on a pair of alternating and uncorrelated random-dot patterns [*see bottom illustration on opposite page*], so that ripple-like shadows seemed to move smoothly across the pattern. The moving shadows caused all the dots in the display to appear to move as a uniform sheet in step with the shadows.

The phenomenon of captured motion now enables us to explain how it is that an oscillating square can emerge



BISTABLE QUARTET, a square matrix of four dots (*a*), is a key component of an experiment demonstrating that the visual system tends to see moving objects follow a straight path. Numbers indicate the order of presentation of dots on a screen; subjects are told to fix their gaze on the central cross. When dots at opposite corners of the quartet (*black*) are flashed and then replaced by identical dots (*color*), viewers are as likely to see the first dots move horizontally (*b*) as they are to see them move vertically (*c*). If two parallel rows of dots (*d, e*) are flashed in sequence (with two of the dots visible at a time), viewers can in theory see one of two trajectories (*arrows*) when the dots in the central, bistable quartet are flashed: horizontal "streaming" (*d*) or vertical "bouncing" along a U-shaped path (*e*). In practice, when the distances between rows and columns of dots are equal, viewers invariably perceive the dots as streaming.

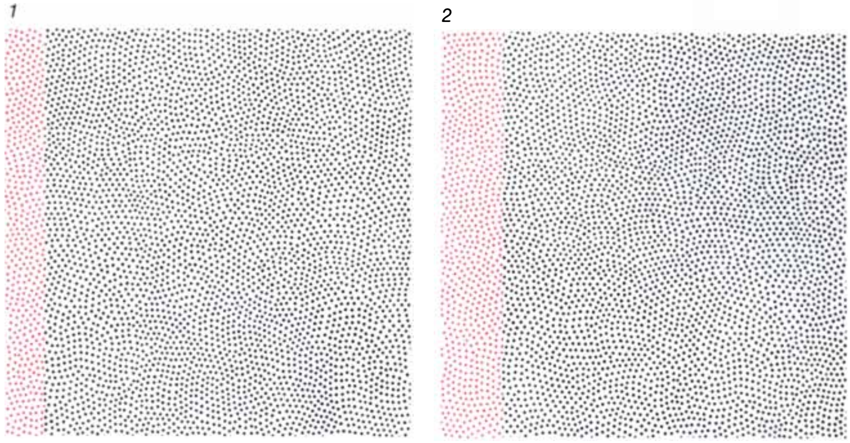
when two images that individually do not appear to include a discrete central square are alternated [*see bottom illustration on page 104*]. Proponents of the computer analogy would contend that the illusion is a product of point-to-point matchings. It seems more plausible to suppose a viewer's visual system extracts a salient cluster of dots from the first display, finds it again in the second display and then assumes that all other "jumping" dots move in synchrony with the salient cluster. Such a short cut would result in faster detection of correspondence than would comparing each point with every other point in successive images. A strategy of this kind would be particularly helpful in the real world, where additional salient features are usually found.

A third rule applied by the visual system, and something of a corollary to the other two, is that a moving object will progressively cover and uncover portions of a background. In other words, when matter, which is normally opaque, temporarily occludes a background, the background still exists; it does not disappear.

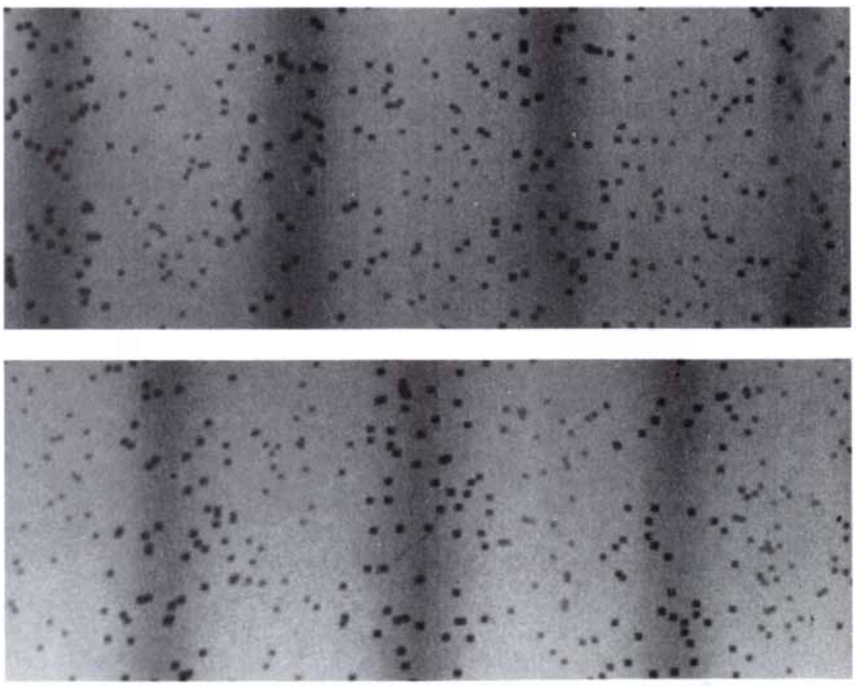
Consider a display in which a triangle and a square below it are presented and then are replaced by another square adjacent to the triangle and directly to its right [*see top illustration on next page*]. One might expect to see the triangle and first square move toward the second square and fuse with it, or to see the first square alone move obliquely toward the second square while the triangle just blinks on and off. In practice one sees something quite different: the triangle appears to move horizontally and to hide behind the obliquely moving square, which now appears to occlude a triangle that is not in fact being displayed. Clearly the brain turns to the real-world property of occlusion to explain the otherwise mysterious disappearance of the triangle. The continued existence of objects is accepted as a given by the visual system, even if the brain sometimes has to invent evidence to fulfill this expectation!

In a related experiment two dots of light in one frame were replaced in the second frame by a single dot, shifted to the right and parallel to the top dot. The images in the first frame seemed to converge at the image in the second frame. On the other hand, when a patch of tape or cardboard was added below the dot in the second frame, a new illusion was produced. Now observers saw the two dots move in parallel, with the bottom one hiding behind the patch, which was perceived to be an occluder. Once again the visual system tended to perceive the motion it was likely to find in the real world.

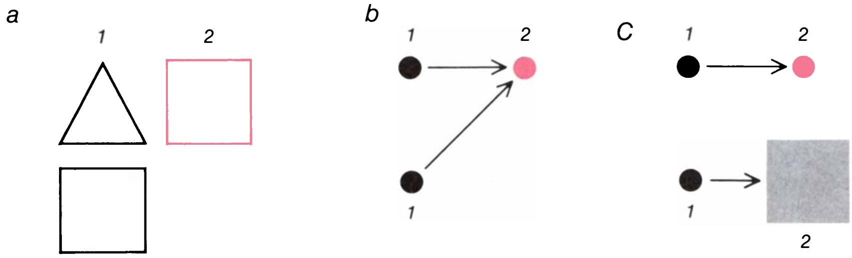
Yet another experiment demonstrated the power of the expectation that one object can occlude another. One of us (Ramachandran) showed viewers an image containing two clusters of four disks each [*see middle illustration on next page*]. In one cluster a pie-shaped wedge was removed from each disk and in the other the disks were complete. We alternated this image with one in which the clusters were transposed. Subjects could in theory see four robotlike shapes, something like those in the game Pac-Man, fac-



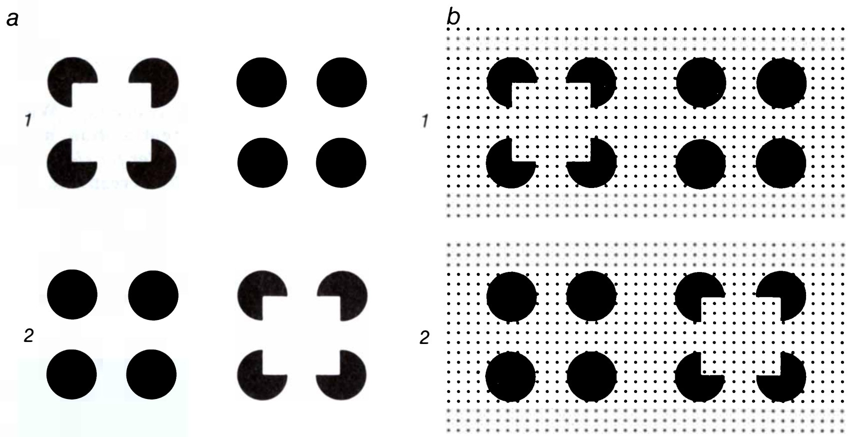
UNCORRELATED RANDOM-DOT PATTERNS form the basis of these displays. When the patterns shown in black are alternated rapidly, viewers see incoherent motion, much like "snow" on a television set. The addition of a strip of dots (*shown in color for clarity*) to the left edge of the images, resulting in these displays, totally changes the perception. The strip in the image at the right (*2*) is wider than the strip in the image at the left (*1*). When the displays are alternated, viewers see the left margin oscillate horizontally and also see the entire display move in synchrony with the margin, a phenomenon known as motion capture. The finding suggests that the visual system tends to see uniform motion and to assume that all parts of an object move in synchrony with any salient part of the object.



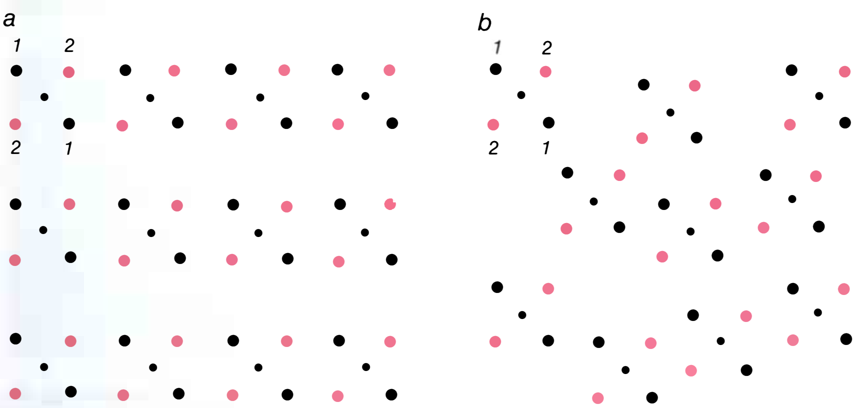
BLURRED LOW-CONTRAST BARS, components of a so-called sine-wave grating, are shown superposed over random-dot displays that produce incoherent motion when they are themselves superposed and alternated in the absence of a grating. The addition of a grating that moves across the screen without ambiguity captures the motion of all the dots and causes them to move with the grating as a single sheet. This effect was studied by one of the authors (Ramachandran) together with Patrick Cavanagh of the University of Montreal.



ILLUSIONS OF OCCLUSION are produced by these images. To explain the mysterious disappearance of an object, the visual system will often assume that the object has been occluded, or hidden, by a larger one. In one experiment (*a*) a triangle and a square are presented simultaneously in one frame (*black*) and are then replaced by a single square displaced to the right (*color*). Numbers indicate the order of presentation. Subjects usually perceive the triangle as "hiding" behind a square that has moved to occlude it. In another experiment (*b*) two spots presented in the first frame (*black*) usually appear to move and fuse with the single spot displaced to the right in the second frame (*color*). If an opaque strip of paper is then pasted on the screen below the second dot, as is shown in image *c*, a new illusion of occlusion results: the lower spot appears to move horizontally and to hide behind the paper occluder. The tendency of viewers to apply the rule of occlusion in resolving perceptual ambiguities has also been emphasized by Irvin Rock of Rutgers University.



DISK-SHAPED IMAGES are elements in computer displays that produce further illusions of occlusion and motion capture. In the images at the left (*a*) pie-shaped wedges are missing from four of eight black disks, first from the cluster of disks at the left (*1*) and then from the cluster at the right (*2*). When the two images are superposed and then alternated, viewers see a white square moving right and left, occluding and uncovering disks in the background, rather than robots opening and closing their mouth. The images at the right (*b*) are identical with the first set but are presented against a background of stationary dots. When these images are alternated, viewers see the dots jump along with the oscillating square.



CLUSTERED BISTABLE QUARTETS are shown. The central dots are fixation points, which are static and continuously visible. When quartets are displayed simultaneously, each quartet is seen to have the same axis of motion (horizontal or vertical) as every other one, regardless of whether the quartets are arranged in regular rows (*a*) or are scattered randomly (*b*). This finding suggests that, in the absence of unambiguous cues to the contrary, the visual system tends to perceive all objects in a given field as moving in the same way.

ing into the center with their "mouths" opening and closing; or viewers could imagine that the white space between the wedges formed a single oscillating square that first partially occluded and then uncovered four disks. It turns out that the visual system interprets the images as an oscillating square, probably because in the three-dimensional world one is more likely to see a square shape occluding a background than to see four identical robots opening and closing their mouths. The property of occlusion overrides any tendency to see movement between the closest similar objects.

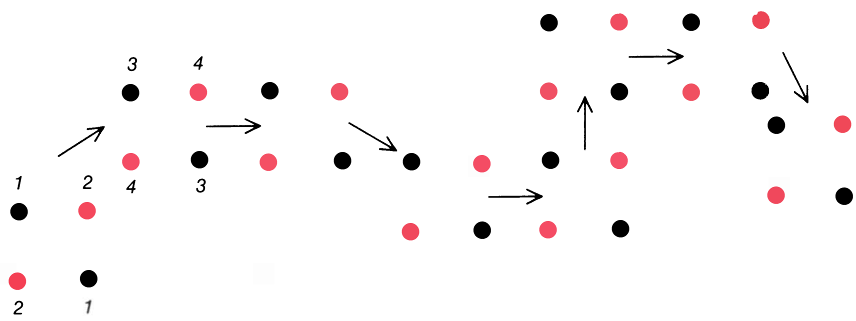
A slightly modified version of this stimulus illustrates the visual system's ability to combine strategies, in this case a predisposition to see both occlusion and rigidity in moving objects. When we superposed the alternating disk images on a background of stationary dots, viewers saw the illusory square oscillate as before, but now they also perceived a sheet of dots oscillating along with the square. The stationary dots were perceived to be a part of the square and therefore were "captured" by its apparent movement. Amazingly, the visual system sees all of this solely as the result of a change in just four tiny pie-shaped wedges.

Having found that the visual system does indeed take short cuts to detect correspondence between images of a single object, we wondered what strategy the system would adopt when faced with many objects in apparent motion. Would it analyze each object independently or would it again take short cuts? Our studies suggest that the visual system tends economically to perceive all objects in a field as moving in the same way unless there are unambiguous cues to the contrary. Gestalt psychologists would call this a tendency to see "global field effects."

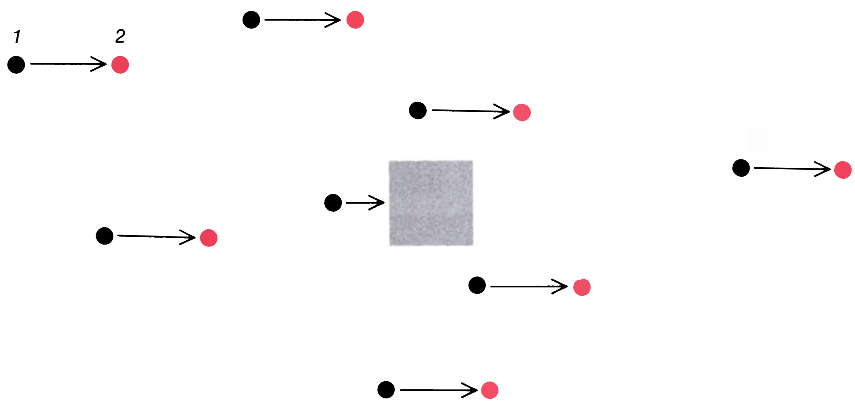
In two related experiments we rapidly and simultaneously displayed many bistable quartets, each of which could be perceived to be in vertical or horizontal oscillation. One experiment had the quartets in three neat rows, whereas the second experiment presented the quartets more randomly. We found that observers perceived all the quartets in each experiment as locking together so that they all had the same axis of motion [*see bottom illustration at left*]. If the visual system did not prefer to see an entire field behave uniformly, and if it processed each quartet independently, our viewers would have seen a mixture of horizontally and vertically oscillating dots.

The unified perception of the clustered quartets suggests that field effects may often be the result of generalizing from a particular instance. That is, the motion seen in one region of a visual field may be significantly influenced by such contextual cues as motion perceived in another part of the field. One way to test this is to cause a bistable quartet to take a "random walk" across the screen [*see upper illustration at right*]. After showing three or four cycles of alternating dots in one bistable quartet, we switched off the display for about half a second before making it reappear elsewhere on the screen. Each of six individuals who viewed the display reported that the motion axis always remained the same even when the square moved to a new location. Once any particular motion axis was seen, the perception apparently acted as a template that created an enduring tendency to perceive similar motion in all other regions.

We recognized that subjects may have interpreted the four-dot display as a single object moving through space. To simplify the test of field effects further, one of us (Ramachandran) alternated images of eight randomly positioned dots with a set of identical dots shifted to the right [*see lower illustration at right*]. Next we masked one of the dots in the second image. Normally when viewers are shown a single dot that is flashed on and off next to an apparent occluder, they see no oscillation. In the context of an array of oscillating dots, however, the perception changed: viewers saw the unpaired dot as oscillating horizontally behind the occluder. They saw what we call entrained motion; that is, motion in one part of the field caused the viewer to see the identical axis of motion in all other parts of the visual field. (The presence of the occluder strengthened the illusion, but the solitary dot also oscillated weakly when no occluder was shown.)



SEVEN QUARTETS that are displayed to subjects sequentially are presented here simultaneously. Arrows indicate the direction of movement from one bistable quartet to the next and numbers indicate the order of presentation of the dots. The dots in color are the ones flashed second in each quartet. Once viewers see the first quartet as having a vertical or a horizontal axis of motion, they almost always see the same axis in quartets presented later and perceive the quartets to be just a single quartet "walking" across the display screen.



RANDOMLY PLACED DOTS are the basis of an illusion studied by one of the authors (Ramachandran) and his student Victor Inada. The display results in a phenomenon known as entrained motion, in which motion seen in part of a field controls the motion seen elsewhere. In a continuous cycle, eight scattered dots (*black*) are flashed on the screen and are then replaced by eight identical dots (*color*) shifted to the right. Viewers see the dots move horizontally (*arrow*). When one dot in the second image is eliminated and replaced with a patch on the display screen (*square*), as is shown here, the partner of the eliminated dot appears to move behind the patch as though it were entrained by the motion in the field.

Our evidence indicates that in perceiving motion a viewer's visual system rapidly extracts salient features and applies built-in laws of motion when processing the features. It also responds to contextual clues in the rest of the field. Of course, even if one believes in the existence of such mechanisms and rejects the concept of laborious point-to-point matchings, an obvious—and much debated—question remains: How does the visual system apply all these strategies? Does it have neurons that are "hard-wired" with the strategies from birth? Or does the perception of motion require some higher level of cognition?

As we mentioned above, the experiments described in this article were designed to eliminate the effects of high-level cognition; specifically, we flashed images at speeds too rapid to allow the brain to make thoughtful decisions about what it was seeing. Our results therefore suggest that low-level processes can, on their own, control the perception of apparent motion during the early stages of visual processing.

Some other evidence also favors this notion over theories requiring the participation of intellect in early, as well as late, stages of motion perception. For instance, an illusion can be seen even when an individual knows an image is an illusion. Neurobiological evidence has been adduced in the past decade by David H. Hubel and Margaret S. Livingstone of the Harvard Medical School, by David C. Van Essen and John M. Allman of the California Institute of Technology and by Semir Zeki of University College London. They have found in monkeys that nerve cells sensitive to the motion of images with low spatial frequencies are distinct from the cells that are sensitive to color, line terminations, angles and other sharp features. This is consistent with our finding that the brain's motion-detecting system pairs off objects sharing low spatial frequencies faster than it pairs off objects sharing sharp features, and it suggests that neuronal activity may be sufficient to account for the initial detection of correspondence by the viewer.

The cellular events that mediate early visual processing in human beings are still very much a mystery, but in time the neurobiological approach should combine with the psychological to elucidate the processes by which the visual system detects correspondence. Our findings suggest, meanwhile, that new advances in the construction of motion-detecting vision machines might be made if investigators who design those machines would attempt to substitute the tricks we have described here for the point-to-point schemes that are currently in vogue.

Modern Windmills

The principles are ancient but the materials and the technology have changed. The chief function of modern mills is to generate electricity, thereby conserving fuel for conventional power plants

by Peter M. Moretti and Louis V. Divone

The windmill, which is one of the oldest means of producing mechanical power, might by now have become a curiosity but for the energy crisis of the 1970's. Because of the interest then in alternatives to petroleum as a source of energy, the windmill came in for a good deal of technological improvement. As a result a number of individual windmills and windmill farms, or arrays of mills, are contributing to the local electric-power system in several countries. The contribution is small now, but one can predict that it will become significant with the inevitable eventual tightening of the petroleum supply and the improving technology of windmills.

The obvious attraction is that wind is free. The equipment is not free, however, and the technology is complex. As any sailor knows, harnessing the wind is not an easy feat. The wind is probably unique among power sources in its random variability. Even though various averaging procedures show daily and seasonal patterns in wind strength, the person who would harness the wind must deal with gusts, lulls and shifts in direction that come unpredictably.

Unpredictability from moment to moment and place to place is not the only problem. The wind's variability also covers a wide range of velocities. Moreover, the effect of velocity is enlarged by the fact that wind force (the amount of push) varies with the square of velocity, whereas the power (the rate at which the wind can be made to do work) varies with the cube of velocity. Hence when a gust reaches eight times the average wind speed, as sometimes happens in a storm, the force of the wind is 64 times the normal value. When the wind speed changes by a factor of two, as it often does, the power varies by a factor of eight.

Wind variability confronts the operator with the technical problems of matching the load to the available power and protecting his mill from gales. The original solution to the first problem, centuries ago, was to make the load adapt to the power source. Threshing or milling took place when wind was available and was suspended when the wind died down. Today's mills are combined with other energy producers, such as an electric utility's generators, to overcome this intermittent operation. The original solution to the problem of gales was to provide for dismantling some of the mill's aerodynamic surfaces, as one would furl sails, in the face of an approaching storm. Modern mills incorporate a variety of mechanisms for letting storm winds blow past without harm.

Wind has a second major characteristic in addition to variability: its diffuseness. It is not a concentrated source of energy. Its drag force on a square meter of surface is quite small at ordinary wind velocities (although it can be very large during storms), and the power of the wind passing through a square meter of area is modest—about on a level with the energy flow of the sun's radiation. When the wind is blowing at a brisk rate of six meters (20 feet) per second, or about 13.4 miles per hour, the energy flux is only about 130 watts per square meter (.016 horsepower per square foot), and not all of that can be extracted by a windmill. When the objective is to generate electricity, these are negligible numbers compared with the levels of power represented by a flowing brook or a domestic fire, not to speak of a fossil-fuel power plant. In order to generate a significant amount of power a windmill must therefore harvest a large cross-sectional area of wind.

The earliest windmills known were in western Asia. They looked like large paddle wheels partially exposed to and partially shielded from the wind. One ancient mill found on the Persian plateau was built around a vertical axis, suggesting that it developed from a mill operated by draft animals pulling the end of a beam along a circular path. The need for large surfaces to harvest the low specific forces of the wind was met by building up windmill sails out of bundled or woven fibrous reeds.

The antiquity of these windmills is uncertain. One early reference, which is highly ambiguous, suggests a date of about 1700 B.C., more than 1,000 years after the first use of sails on ships. Even if that estimate is correct, windmills must have been scarce during the next 1,000 years. By A.D. 1100, however, crusaders and travelers found windmill technology well established in the Middle East. It is clear that windmills were not in wide use even at that time; the problems presented by the variability and the diffuseness of the wind had not yet been adequately solved.

The variability of the wind could be accepted in applications that allowed for storage of the raw materials and the products, for example in grinding grain and sawing logs. Pumping water for irrigation or swamp drainage also offered some flexibility. Nevertheless, water-driven mills offered more reli-

MODERN WINDMILL based on an unconventional concept is the Darrieus turbine, named for the French engineer J. G. S. Darrieus. He invented the concept in the 1920's, but it was forgotten; workers at the National Research Council in Canada reinvented it in the 1970's. The Darrieus rotor turns on a vertical axis, whereas conventional windmills have a propellerlike, horizontal-axis rotor. This array of Darrieus turbine mills is in California.


FLOWIND

able power where they were feasible, and so windmills were likely to be found primarily where water power was not available.

The diffuseness of wind power and the resulting low output of windmills presented a problem that affected all applications and called for new technological solutions. Those solutions had already been developed on sailing ships. Sailors had long since discovered that a vessel will proceed faster when the wind comes from the side than when it comes from behind. The reason is the difference between lift and drag. A considerable aerodynamic lift develops along a sail when wind from the side sweeps across it with a large relative velocity. The push of a tail wind, on the other hand, is reduced to a small relative velocity by the motion of the ship.

Similarly, a windmill based on lift forces can produce several times the output of paddle-wheel blades that are simply dragged along by the same cross-sectional area of wind. Moreover, a mill taking advantage of lift needs less blade material than a mill based on drag. The reason is that lifting blades do not have to cover as much area with their surface; they merely sweep across the wind repeatedly. Finally, the blades rotate faster; this is an advantage in a mill that has to gear up the shaft speed to enable the load (usually an electric generator in such cases) to run efficiently.

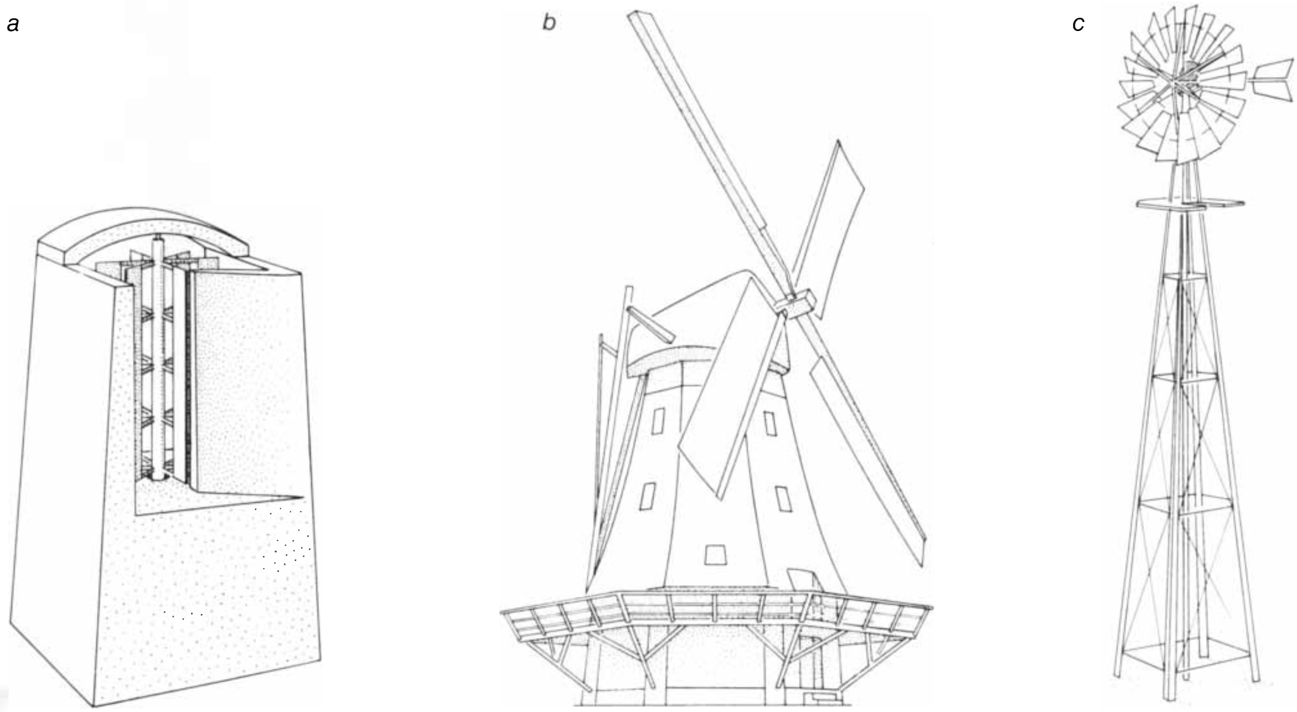
The era of propellerlike windmills utilizing lift forces probably began as mill technology moved westward and met the technology of the sailing ship in the Persian Gulf and the Mediterranean Sea. The loose-sailed windmill rotors still seen on Mediterranean islands bear a clear resemblance to the sail of a small boat.

As windmill technology moved west and north from the Mediterranean, regions of higher winds were encountered. That brought the development of sturdier blades made of wood latticework to support the sailcloth. The evolution of these mills is evident in medieval English drawings and the paintings of the Dutch masters.

The basic form of the rotor remained the same; the gradual refinement was evident in the changing proportions. The supporting spar, which was initially at the center of the lattice, was moved toward the leading edge and finally ended up at about the quarter-chord point. (The chord is a line joining the leading edge and the trailing edge; the quarter-chord point is where the center of pressure of an airfoil is usually found.) Through experience and observation the builders had learned that this location of the supporting spar reduced the twisting distortion of the blades when they are under load.

A major structural change during the same medieval period was in the design of the tower. On post mills, the earliest of the northern windmills, the entire tower was turned by hand whenever the wind shifted. As a result the base of the tower could not be anchored to a solid foundation, and so the tower was rather weakly supported. In later designs the tower was stationary except for a top section that could be turned to face the wind. This development allowed the size of the windmill to increase considerably over that of the earlier machines.

Another gradual improvement, beginning about 400 years ago, was the recognition of the fact that a blade works better if it is shaped with a twist. The builders realized, apparently by intuition alone, that a twist would compensate for the fact that the blade moves with differing speeds at different points along its radius, slowest near the hub and fastest at the tip. (The technical term is that the blade has differing tangential velocities at different radiuses.) Without twist the angle of



EVOLUTION OF WINDMILLS is illustrated by three old machines spanning several centuries. The ancient paddle-wheel mill (*a*) operated on a vertical axis but did not take advantage of the lift that wind can provide. It relied on drag forces, like a sailboat with the wind directly astern. Mills of the propeller type (*b*) came into fashion when it was discovered that lift forces from 10 to 50 times greater than drag forces could be achieved with wing-shaped surfaces. The automatic, self-furling, multiblade water pumper for farms (*c*) was developed in the U.S. in the middle of the 19th century and served widely. The features of the mill included a pivoting, spring-loaded tail that normally kept the rotor pointed into the wind but automatically swung it away from the wind in a storm.

the blade with respect to the wind is far from ideal and power is lost.

The most sophisticated improvements were effected in the cross-sectional shape of the blades. The addition of a canted leading-edge board in the windmills of Holland in about 1600 provided camber: a curve along the chord. Camber gives the blade more lift and less drag. By the end of the 19th century the leading edge of English windmill blades had developed not only this form of curvature but also an aerodynamically contoured shape with a thick leading edge. The blade had become a true airfoil.

This slow technological evolution took place largely in the absence of any theory or clear understanding of aerodynamics. Trial and error and the accumulation of generations of experience brought about better and more effective configurations. At the same time economic forces led to the development of larger windmills.

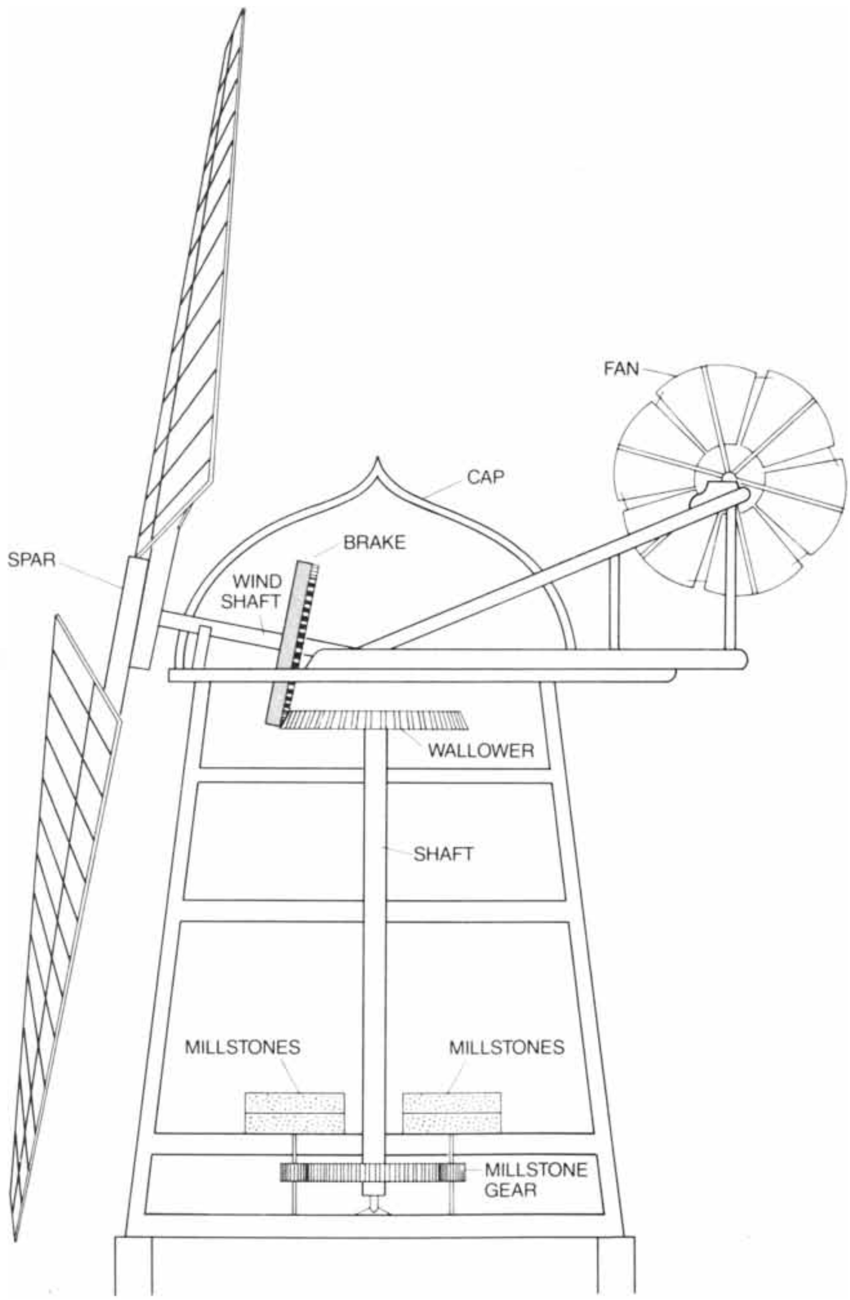
A large windmill can be more cost-effective than a small one because, among other things, it can do more work with the same operating staff. On the other hand, there are practical limits that constrain the maximum size. For the wood-and-canvas mills of the 18th century the upper limit worked out empirically to a rotor diameter of about 30 meters and a mechanical power output of about 40 kilowatts in a very brisk wind.

Other economic factors, in particular the recent emphasis on generating electricity with windmills (more often called wind-turbine generators in this application) rather than grinding grain or pumping water, have continued to push the size of windmills upward. Today's experimental wind turbines have reached about 100 meters in diameter. Even larger mills would be desirable for generating electricity in significant amounts, but increasing size brings structural costs that eventually outpace the economic gains. One limitation at present is that metal-fatigue stresses at the root of the blade, caused by the pull of gravity on the long structure, inhibit further gains in size. Another problem is that as the size increases, the area of wind that is harvested goes up by only the square of the increase, whereas the mass of material needed goes up by the cube. Hence there is clearly a point of diminishing returns and therefore some optimal size, both of which depend on the application intended for the mill.

In conception the windmill has changed little over the centuries. A rotor of the propeller type is supported on a tower; a main shaft is geared up or down to drive some load, and mechanisms are provided to control the number of revolutions per minute, preventing excessive speed in high winds, and to keep the rotor facing the wind.

Today's windmill designers, under pressure to construct mills that generate electricity efficiently, have had to deal with the same problems that confronted early builders and have moved in the same directions. The most obvious results are the development of improved airfoil profiles and the changes in materials made possible by metallurgy and polymer science.

A more subtle development has been the shift from torque-maximizing designs to energy-maximizing designs. The old mills, pumping water and grinding grain, had to start under load, and to do so they required a strong



WINDMILL COMPONENTS have remained virtually unchanged for 500 years. A tower supports a rotor of the propeller type that generates aerodynamic forces, similar to those on the wing of an airplane, in order to turn a main shaft. Gears in the power train drive millstones, a pump or (more commonly in modern mills) a generator at an appropriate speed. Controls such as the fan turn the rotor as the wind shifts, adjust power and prevent excessive running speed. Only the technology for carrying out these functions has changed.

torque, or twisting force, from the blades in order to turn the grindstones or lift the water. Modern transmission technology in general and electric generation in particular have made it possible to start a wind turbine under minimal load and then to operate it at some optimal speed. Hence the blades no longer must be large enough to enable the mill to start under a heavy load at low wind velocities; instead they can be designed for good full-speed operation.

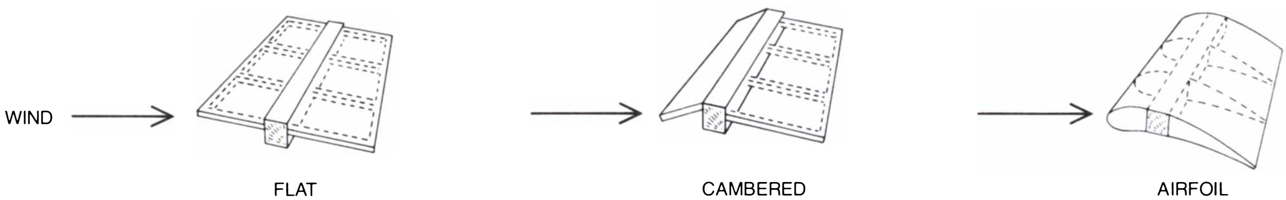
A windmill that is optimized for capturing energy will usually have slender, fast-moving blades; the airfoil will exhibit a high lift-to-drag ratio. The theoretically ideal windmill would have a large number of narrow blades. Requirements of structure, on the other hand, call for only a few wide and thicker blades—as many as three or as few as one. (A single blade has to be counterbalanced in some way.)

The high speed of the blades is necessary in order to minimize torque, which sets up a swirling of air behind the blades that carries energy away instead of converting it to generate power. High speed, however, has its limits: drag ultimately catches up with a fast-moving airfoil, and the tendency of the blade to compress the air through which it moves causes loss of efficiency and excessive noise when the speed of the tip reaches from a third to half the speed of sound.

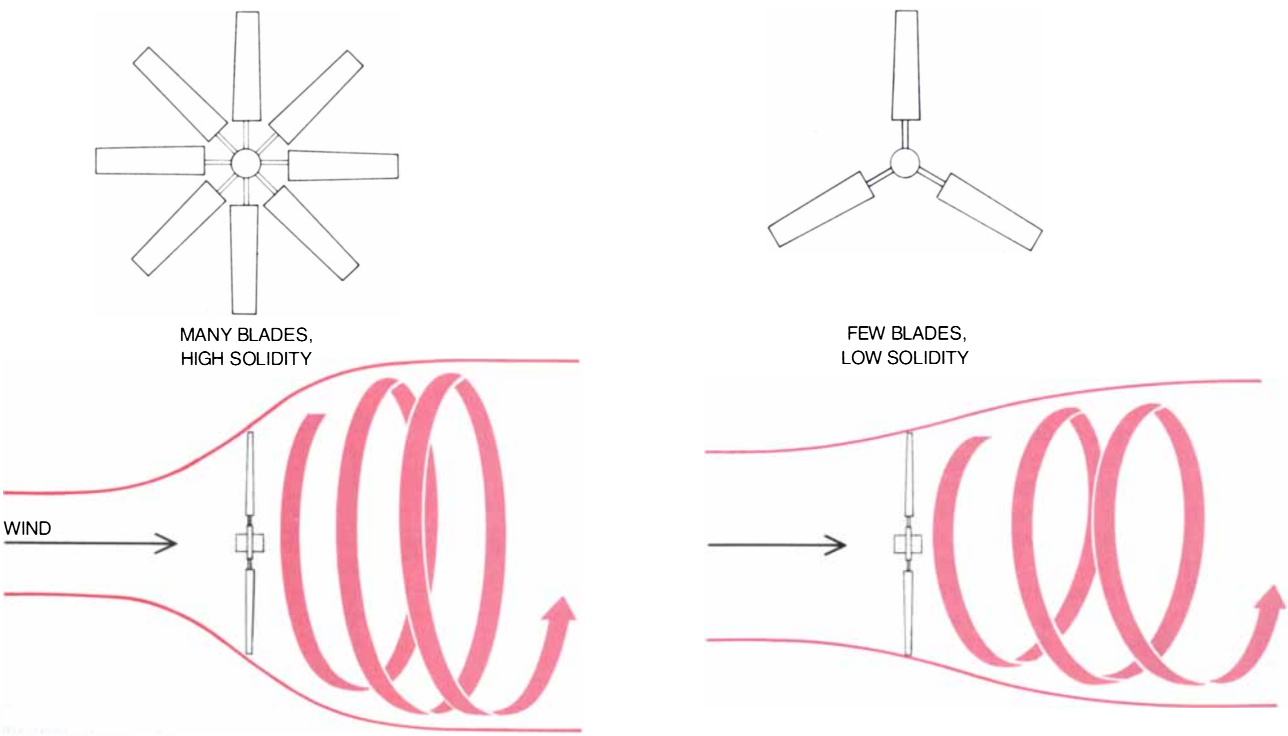
Economic factors also tend to raise operating speeds. The cost of machinery is closely tied to its weight. High torque necessitates heavier equipment. Power is the product of torque times revolutions per minute. If one can get more speed from a prime mover, one can get more output from machinery of a given size. As a result modern generators, including those driven by windmills, all run much faster than those of a few decades ago.

The wind turbine is one of the few prime movers that intrinsically turn slower than the design speed (either 1,200 or 1,800 r.p.m.) of a typical generator. Therefore it requires gearing up, rather than the usual gearing down, between the prime mover and the load. This puts added pressure on the designer to increase the operating speed and reduce the torque. The tip of a modern blade may move at up to 300 miles per hour in a wind of 20 m.p.h. In other words, the ratio of the tip speed to the wind speed would be 15. (The relation is usually called the tip-speed ratio.)

Now, the optimal tip-speed ratio is



CONFIGURATION OF BLADE has changed gradually as designers took advantage of aerodynamic forces to extract more power from the wind. The original blades were flat. Cambering, which increases the lift force, was introduced in the 17th century. A modern windmill blade has an airfoil configuration that develops even better lift forces with less drag. Each blade is shown with part of its supporting spar. Early windmill blades were covered with sailcloth; fiberglass, metal or wood veneer serves now for covering blades.



SOLIDITY OF ROTOR, meaning the ratio of the space taken up by the blades to the area swept by the rotor, has tended to decline in modern windmills. A rotor with broad blades or a large number of blades tends to deflect too much of the wind around rather than through the rotor disk. This high-solidity configuration also operates with high torque, or twisting force, and low speed of rotation, so that a significant part of the wind's energy is dissipated in the form of swirling (*color*) in the rotor's wake. A low-solidity rotor, which has few blades, is more efficient. The diameter of the stream tube, or affected airstream, is indicated by the top and bottom lines.

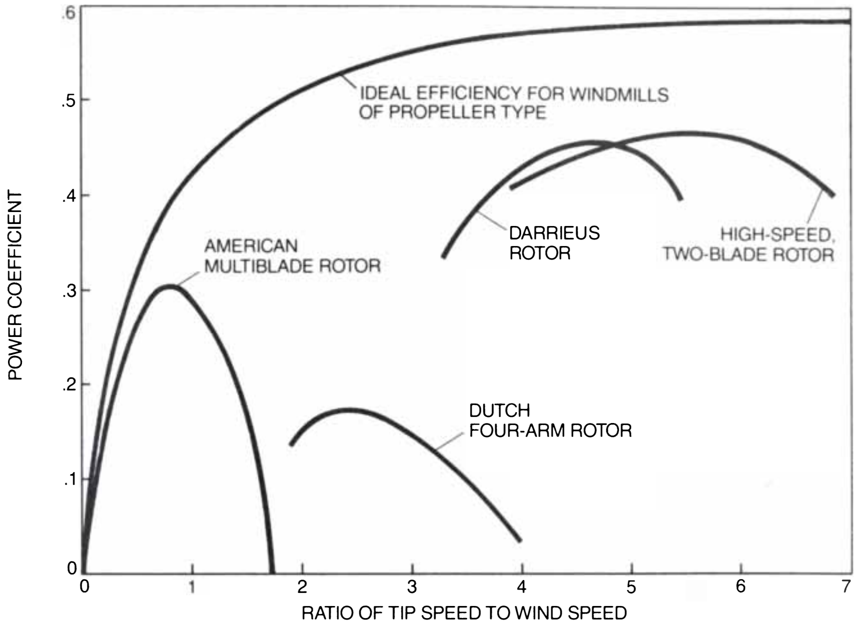
the same for large and small rotors alike, whereas the tip speed is the product of the radius times the r.p.m. For this reason a small wind turbine appears to spin quite fast and a large one to turn at a somewhat leisurely rate, yet they may have the same tangential speed at the tip and the same tip-speed ratio. Small or large, however, a contemporary windmill looks rather spidery and insubstantial compared with an old one.

Research is continuing on the shape of the airfoil and the aerodynamics of the blade. The aim is to improve the power coefficient of the rotor: the ratio of power extracted to the latent power in the wind over an area equal to that of the rotor. Even greater changes are taking place in the control of wind turbines. There are four key problems. First, the machine should shut down if the wind is so weak that spinning the equipment will absorb more power in friction losses than the wind can provide. Second, the turbine should extract the maximum amount of power in medium-to-brisk winds. If the rotor slows the wind excessively, too much of the wind will flow around the rotor; if the rotor does not slow the wind enough, too much energy will pass through unused. (An optimum is achieved when the wind is slowed by two-thirds, resulting in a theoretical power coefficient of .593. A typical modern wind turbine achieves a power coefficient of about .4.) Third, in strong winds the rotor should spill or waste excess power the generator cannot handle. Fourth, the system must protect itself from severe storms, usually by shutting down.

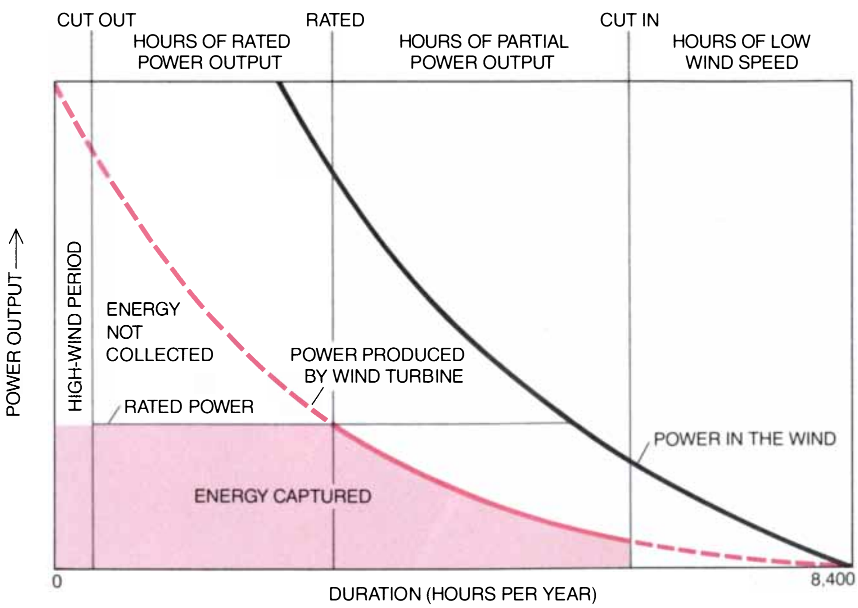
The control system must carry out all those functions by means of devices such as blades of variable pitch, spoilers, stalling airfoils, yaw control and brakes. The system must be "intelligent" enough to handle the wind's constant variability without overreacting to a momentary condition. This objective is reached with the use of inexpensive microprocessors for control. They are a far cry from the flyball governors, the mechanical springs and the skill of the miller that constituted the control system in mills of the past.

A windmill of the propeller type is not the only possible configuration. Many alternatives exist. During the oil crisis of a decade ago proposals for unconventional types of windmill came in to the U.S. Department of Energy at a rate of several per week. A few of them show a potential for major advances. Notable are clever proposals for furling the apparatus during storms, elaborate ones for matching the load on the rotor to the wind and proposals for various types of vertical-axis machines.

Among the last the most successful design is based on the rediscovery of a concept put forward by the French engineer J. G. S. Darrieus in the 1920's. The machine looks like an eggbeater standing on end. Because of its mechanical simplicity and ease of maintenance (all the mechanical and control equipment is at ground level), it is proving to be competitive with propeller-driven mills. Two companies are now producing Darrieus systems for



EFFICIENCY OF ROTORS with various types of blade is compared in terms of the power coefficient v. the tip-speed ratio. The power coefficient is the ratio of power extracted from the wind to the power available in it over the area covered by the diameter of the rotor. The tip-speed ratio compares the speed at the tip of the blade with the speed of the wind. The trend in the design of wind turbines that generate electricity is toward fewer and narrower blades, so that the turning speed of rotors has increased. High-speed, low-solidity rotors have reduced the loss of power due to swirling in the wake of the rotor and have also reduced the weight of structural and mechanical components and the cost of mill operation.



ENERGY CAPTURED over the course of a year by a typical modern wind turbine is indicated by the colored area of this graph. The black curve represents the amount of power in the wind. The turbine designer seeks to maximize not only the amount of energy captured by the mill but also the amount of time during which the machine is generating power.

windmill farms in California. Several other vertical-axis mills are now under development in Canada and the U.K.

Since most people in developed countries get electricity from fossil-fuel, nuclear or hydroelectric plants, one wonders how much of a role windmill power can have. At present it serves mainly to supplement conventional systems, but it may come to play a more important role, at least in developing countries.

What is taking place now is a steady growth in the number of wind turbines connected to established power grids. The wind turbines or arrays of turbines are ordinarily distributed widely enough to average out some of the short-term variations in the output of individual machines; the conventional power plant of the grid responds to longer variations. The wind turbines act as fuel savers for the power company, reducing the load on the fossil plant when the wind blows.

For the longer future one can predict several trends in wind-turbine development. Foremost among them are gradual decreases in cost and increases in mechanical reliability as machines of all sizes move through generations of design development and are produced in greater numbers. Machines in all categories are moving upward in capacity. Residential turbines, typically rated at one kilowatt only a few years ago, are moving toward 10 kilowatts. Machines in windmill farms are advancing from 25 kilowatts toward 200 or more. Research and development machines have grown from a few hundred kilowatts to more than three megawatts.

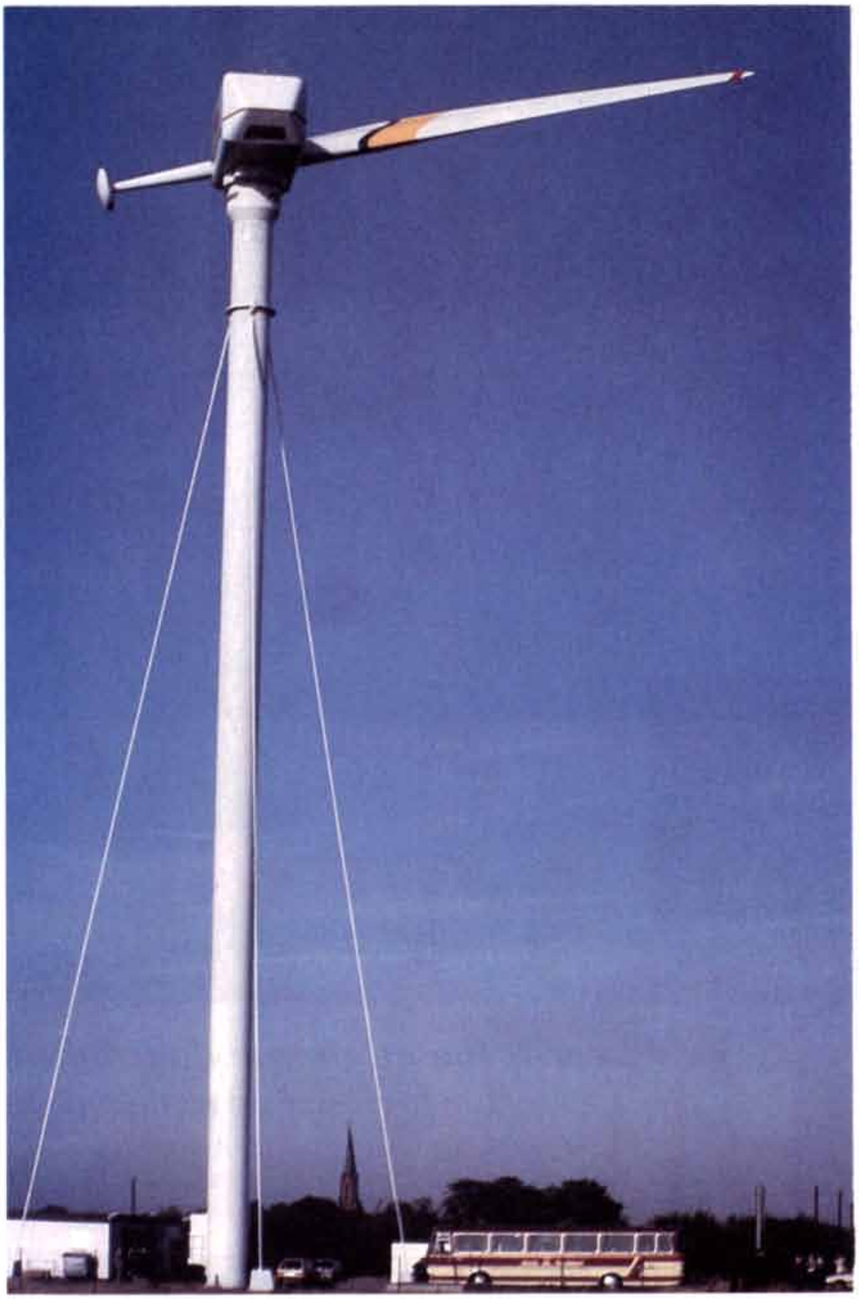
The most promising developments, however, entail neither greater size nor new inventions but reduced costs for materials and manufacturing. As wind turbines become larger, a greater percentage of the construction cost is absorbed by the rotor. The development of an effective and inexpensive construction process therefore becomes crucial to the financial success of the machine. Materials being considered, chiefly for blades, include stressed-skin metal, fiber-reinforced plastic and weather-resistant wood impregnated with epoxy. In addition, alternative rotor designs of greater structural simplicity are being explored. Among them is a rotor airfoil in which only a few panels are variable and can be adjusted to meet wind conditions and to control the mill; the entire airfoil does not have to be installed on a variable-pitch hub.

Technically sophisticated systems for control are increasingly replacing

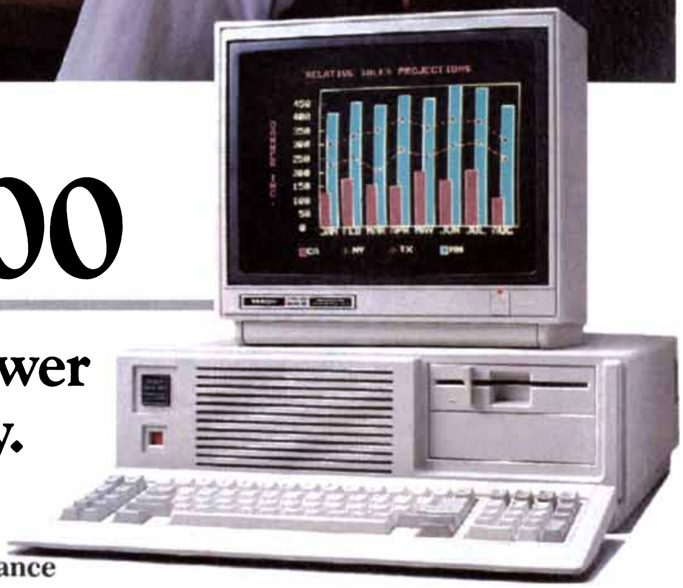
WINDMILL FARM in California consists of an array of wind turbines contributing to the local power grid. More than 13,000 turbines have been installed in California since 1982; by last summer they had produced the energy equivalent of a million barrels of oil.

the "simple and cheap" systems that had been the convention. The cost of the control system is becoming small in relation to the gains in performance achieved when the mill operates more efficiently.

The energy needs of the developing countries, now heavily oriented toward oil, will constitute a fertile area for the development of wind power, particularly since all but the larger mills are fairly easy to assemble. Even in the U.S. the current glut of oil cannot be expected to continue indefinitely, and the generation of power by wind is likely to gain gradually on conventional methods. Within just the past three years enough wind turbines have been installed in California to equal 2 percent of the state's installed capacity; they produce 1 percent of its energy. Although the percentages sound small, they represent a power output of more than 600 million kilowatt-hours. Perhaps in a generation or so windmills will again represent a significant source of power, as they did centuries ago.



SINGLE-BLADE WIND TURBINE called Monopterous is an experimental structure near Bremerhaven in West Germany. It represents the kind of novel configuration that turbine designers have tested in the quest for higher speed, reduced blade area and lower cost.



THE AMATEUR SCIENTIST

Mirrors make a maze so bewildering that the explorer must rely on a map

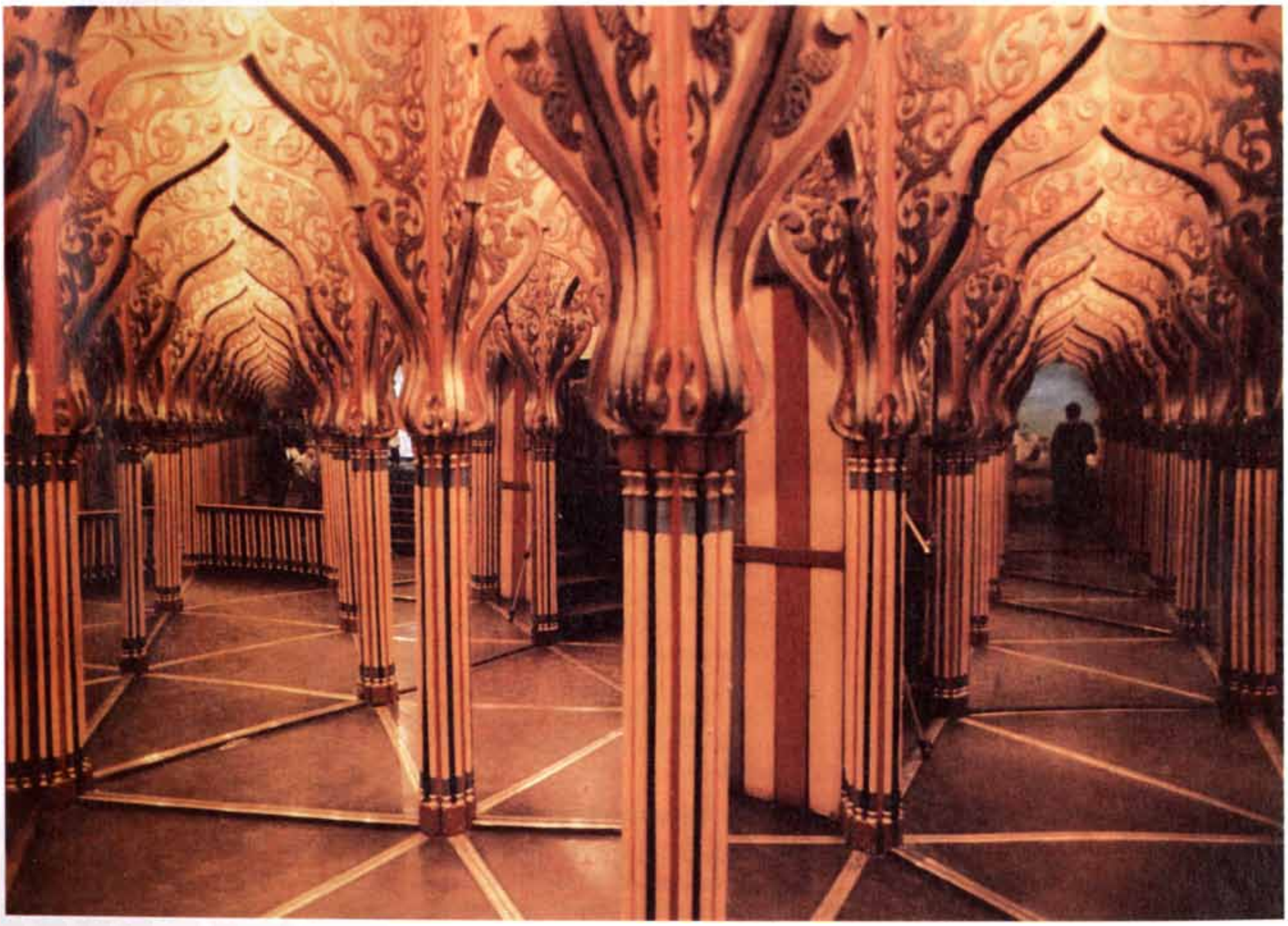
by Jearl Walker

In the Glacier Garden in Lucerne there is a mirror labyrinth built originally for the Swiss National Exhibition of 1896 in Geneva. I entered this Hall of Mirrors expecting it to be a short maze of the kind I solved easily at state fairs as a child. Almost immediately, however, I became so disoriented that I had to begin feeling my way by holding a brochure in front of me to ascertain when I was getting near a mirror.

The maze was laid out on an array of equilateral triangles marked on the floor. Where there were mirrors they were positioned upright on the borders of the triangles (one or two per triangle) and were outlined with elaborate furnishings. In most directions I saw a jumble of images: sections of the furnishings or scenes intended to imitate the Moorish style of the Alhambra in Granada. In certain directions I saw apparent hallways, each seeming to consist of a series of doorways. The door frames were actually the furnishings around mirrors elsewhere in the maze. At the end of each hallway there was one of the Moorish scenes. I set out along one of the hallways but soon found my way blocked by a mirror. Turning, I followed another apparent hallway and again found my progress blocked by a mirror.

Eventually I began to come on the scenes. One, behind a transparent glass wall at a dead end in the maze, was a flower arrangement between two mirrors forming a 60-degree corner. The mirrors made images of the flowers fill my field of view, as if I were looking into a giant kaleidoscope. Along a different route I passed a sultan staring stony-faced at my wanderings. In due course I came to a door that opened into a small room containing furniture. When I faced back into the maze, I saw the sultan at the end of one hallway and a blank wall at the end of another hallway, even though I was then far from the sultan and even farther from the blank wall. The trip back to the entrance was just as confusing as the trip to the room.

Determined to understand the maze,



The Hall of Mirrors in Lucerne

I explored it again, this time making a map of the mirrors. Mary Golrick, my wife, ventured several triangles ahead of me so that I could determine which hallways led into the maze. Once we had reached the small room we retraced our steps by following the map. We no longer had to feel for the mirrors with the brochure.

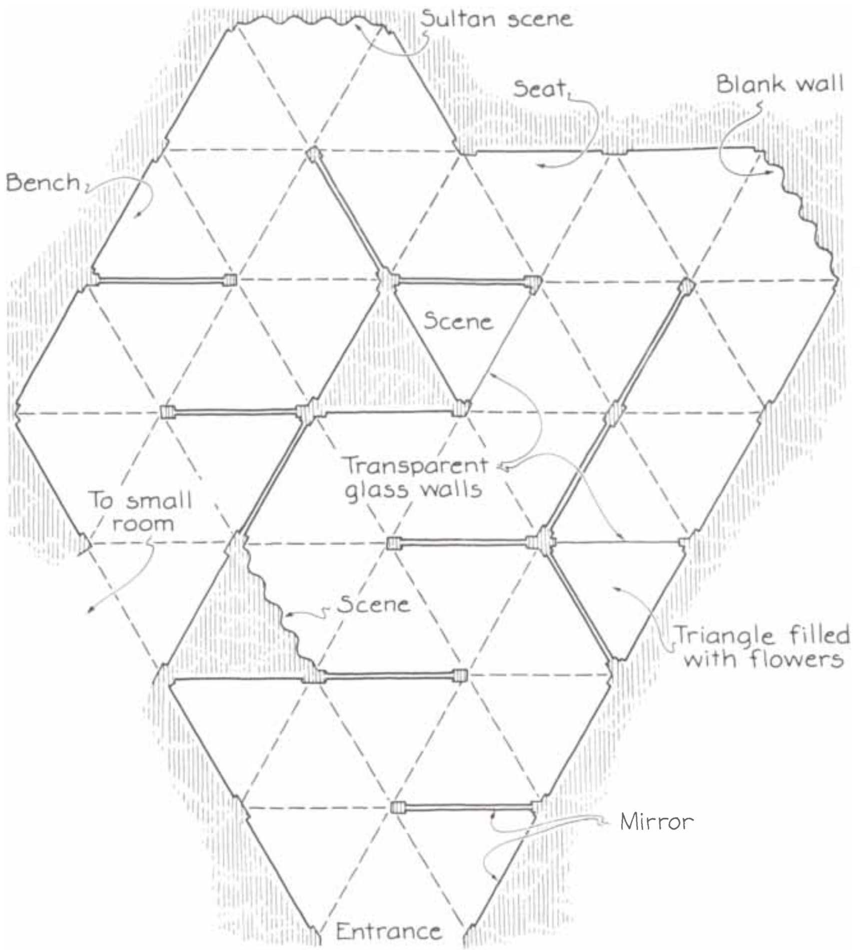
I left Glacier Garden with many questions. Does one build an optical maze such as the Hall of Mirrors by planning it on paper? Can the images one sees inside a maze be predicted from a floor plan of the setup? Do different mazes have any common properties? What produces the hallway illusion? What determines which scene lies at the end of a hallway?

I began to study the properties of a typical optical maze by drawing features of it on triangular-coordinate paper made by the Keuffel & Esser Co. The graph paper is ruled with equilateral triangles. I drew mirrors in one color and the paths of light reflecting from them in another color.

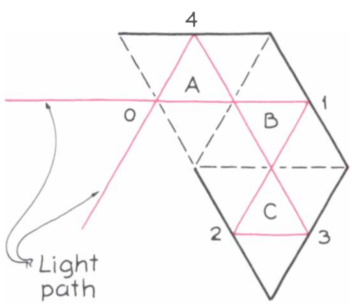
I also studied small mazes by building models of them with inexpensive mirrors from Jerryco, Inc. (601 Linden Place, Evanston, Ill. 60202). The mirrors are pushed into a bed of sand for stability and oriented with the aid of a protractor. I adjust the alignment by looking at the reflections. The mirrors, which are eight centimeters along a side, are large enough so that I can peer into the entrance of a maze. In order to examine the view as if I were inside the maze, I hold a small mirror at an angle of about 45 degrees to the horizontal and point it toward whatever I want to look at. The mirror reflects that scene back to me. The mirrors do not give clear images, particularly after many reflections of the light, because the weak reflection from the front of the glass muddles the stronger reflection from the rear surface. Still, the mirrors serve well enough for small mazes.

I first studied a maze with a single opening. I soon found I could predict many of the maze's features by drawing the floor plan. In one four-mirror maze, for example, the mirrors are on the borders of equilateral triangles [*see illustration at bottom left on this page*]. From my studies I realized the light rays creating a hallway image all are reflected from the mirrors at 60 degrees. Hence they always travel parallel to one side of a triangle. To avoid cluttering the floor plan I drew one line, which I call the light path, to represent the rays creating a single hallway image. I numbered the reflections of the light in sequence. The light enters the maze at the point labeled 0.

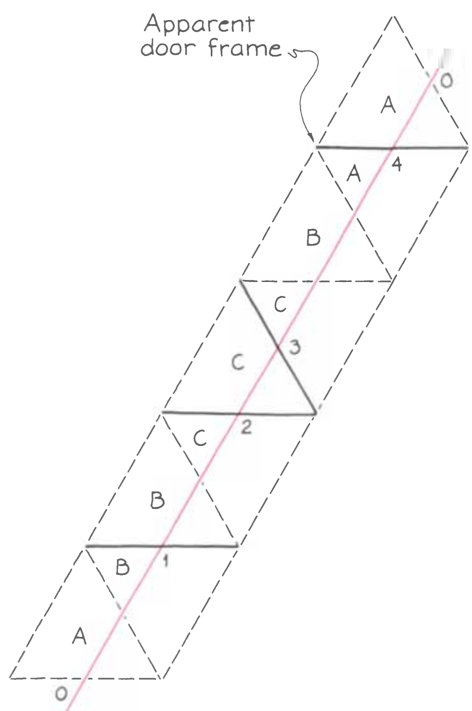
Looking into this maze along either



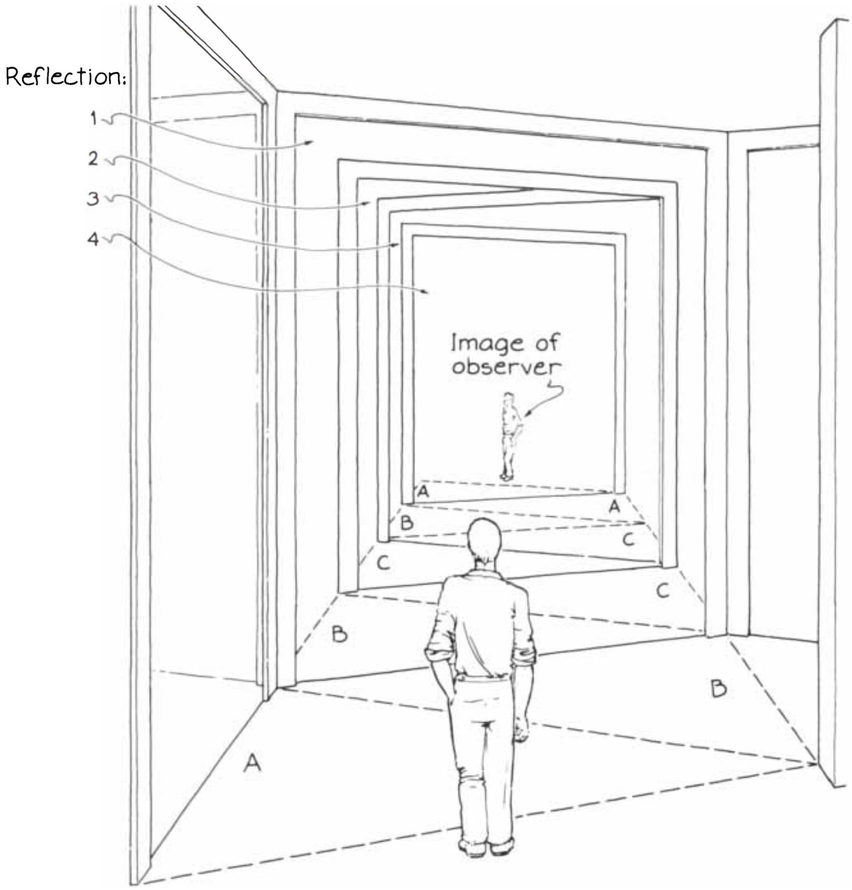
How the Hall of Mirrors is laid out



Light in the maze



An "apparent hallway"



An apparent hallway with four door frames

of the pathway lines extending from it, you perceive a hallway stretching away from you [*see illustration above*]. Suppose the mirrors are framed as the mirrors at Lucerne are. The hallway appears to have four door frames. The floor of the hallway consists of the triangles crossed by the light path. At the end of the hallway is an image of whatever lies at the other end of the light path leaving the maze. If you are standing at point 0, the image at the end of the hallway is your own, angled 60 degrees from the front. The light that forms the image travels from you at point 0, through the reflection sequence 4 through 1 and in the end back to 0, at which point it is intercepted by your eyes.

Another way to depict the maze is to draw a floor plan of the hallway you perceive as you look into the maze [*see illustration at bottom right on preceding page*]. Point 0 lies at both ends of the hallway. The triangles and the door frames passed by the light path lie between the two ends.

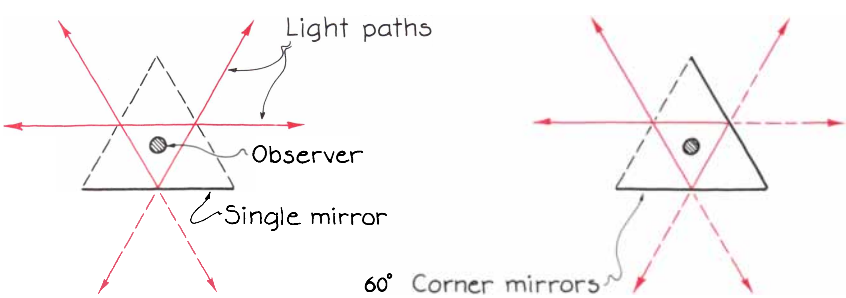
If you stand at point 0 and look into the maze, you see two hallways: one of them extends from reflection 1 and the other from reflection 4. To see either of them you must face in the proper direction; all other directions yield a jumble of images. If you stand within the entrance triangle, you will see four hallways by turning around. Two of them consist of a single door frame only. If you are in a triangle farther in the maze and out of direct view of the entrance, six hallways extend from you regardless of whether the triangle is bordered by one mirror or two mirrors. Each hallway is parallel to a side of the triangle. If you look in any other direction, you again see a confusing jumble of images.

I soon realized I would never discover the common properties of mazes by drawing them without a plan. I needed to construct them systematically from some basic starting design and under specified rules. The starting design had to be the simplest of all the mazes that produce a hallway image: a single triangle bordered by two mirrors. The light path for the hallway images in this maze involves one reflection from each mirror. If you stand at the entrance of the maze, you see two hallways, each with two door frames.

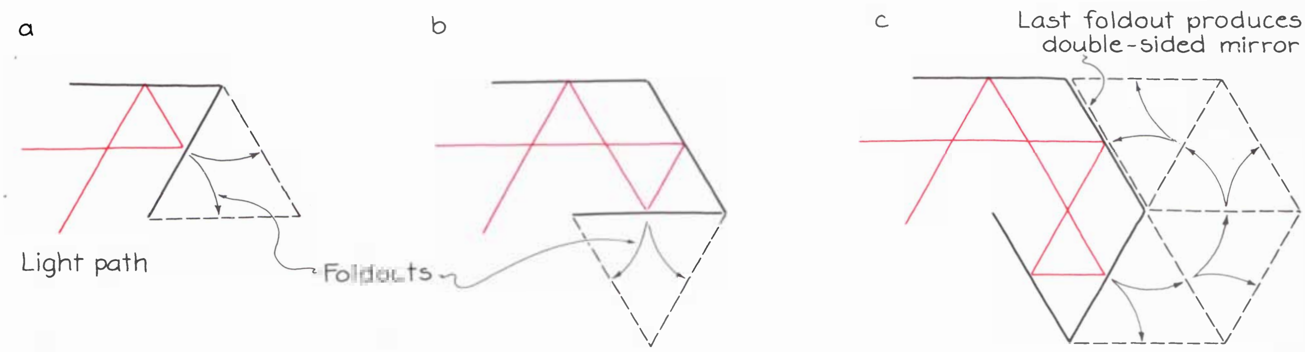
From the simplest maze I built more complex mazes by an operation I call a foldout. It entails replacing a mirror on the outer boundary of a triangle in the maze with two new mirrors on an adjacent triangle outside the maze. The maze grows larger because the foldout adds a triangle to it. Sometimes a foldout produces a double-sided mirror that projects into the maze from the perimeter.

The advantage of building a maze by repeated foldouts is that each foldout leaves unchanged the light path in the rest of the maze. After you add a foldout the light path passes through the new triangle with one reflection from each new mirror and then continues through the rest of the maze as before. This fact allows for some generalizations about a maze built from the simplest maze by repeated foldouts. No matter how large or complex the maze becomes, the light path enters the maze, is reflected once from each mirror and then leaves. Hence the hallway you see when you look into the maze at point 0 must have a door frame for each mirror and an image of you at the other end. Within the hallway you see the entire interior of the maze. When you are inside the maze, each of the six hallways that extend from you has its entrance at its far end.

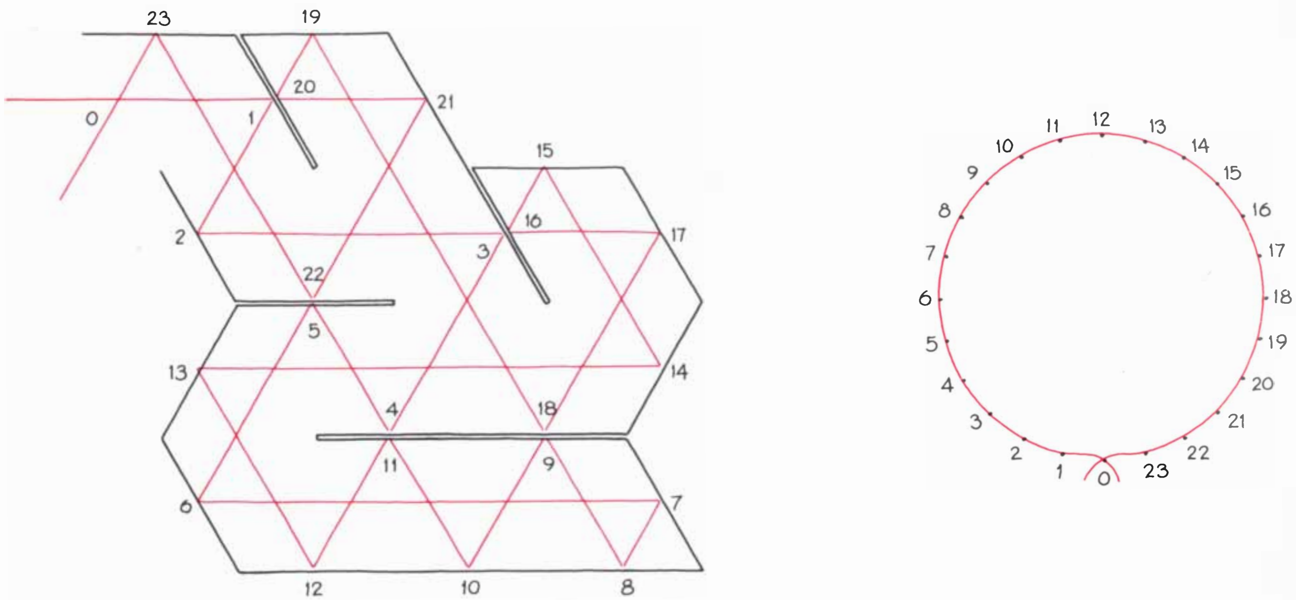
I constructed a fairly complex maze from the simplest one by repeated foldouts [*see middle illustration on opposite page*]. Note that the procedure left several double-sided mirrors in the maze. The reflections of the light path are numbered in sequence on the basis of an arbitrary choice of direction for the path. (The sequence could be reversed by choosing the opposite direc-



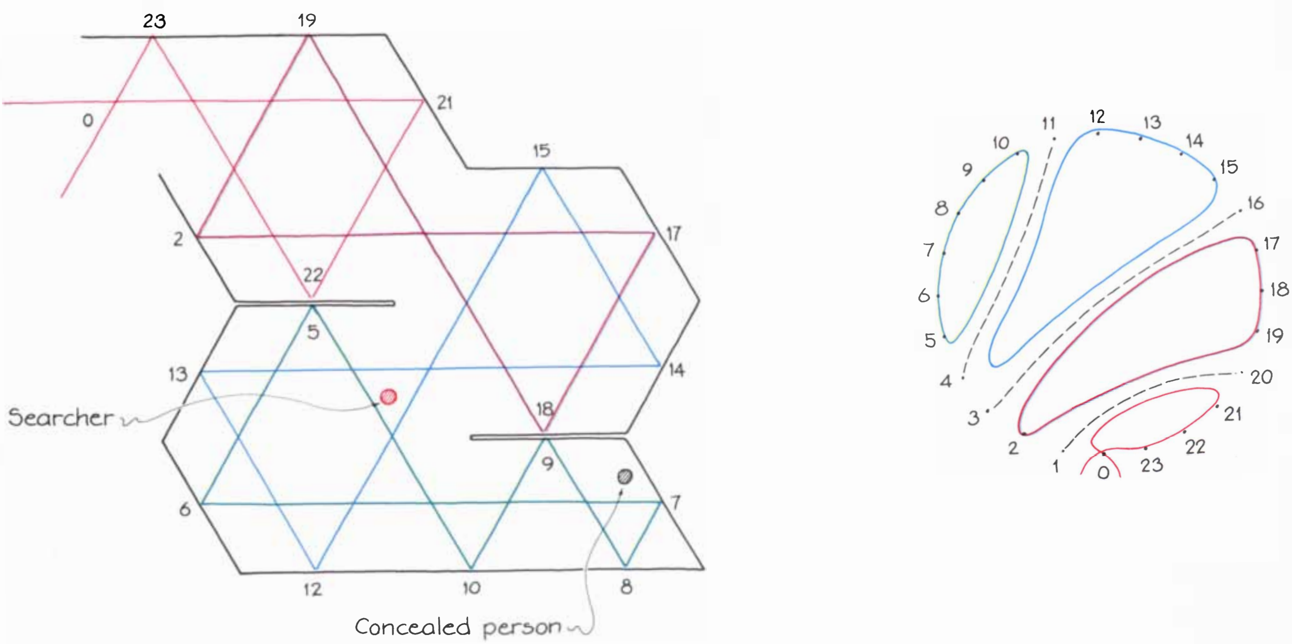
Arrangements of an apparent hallway



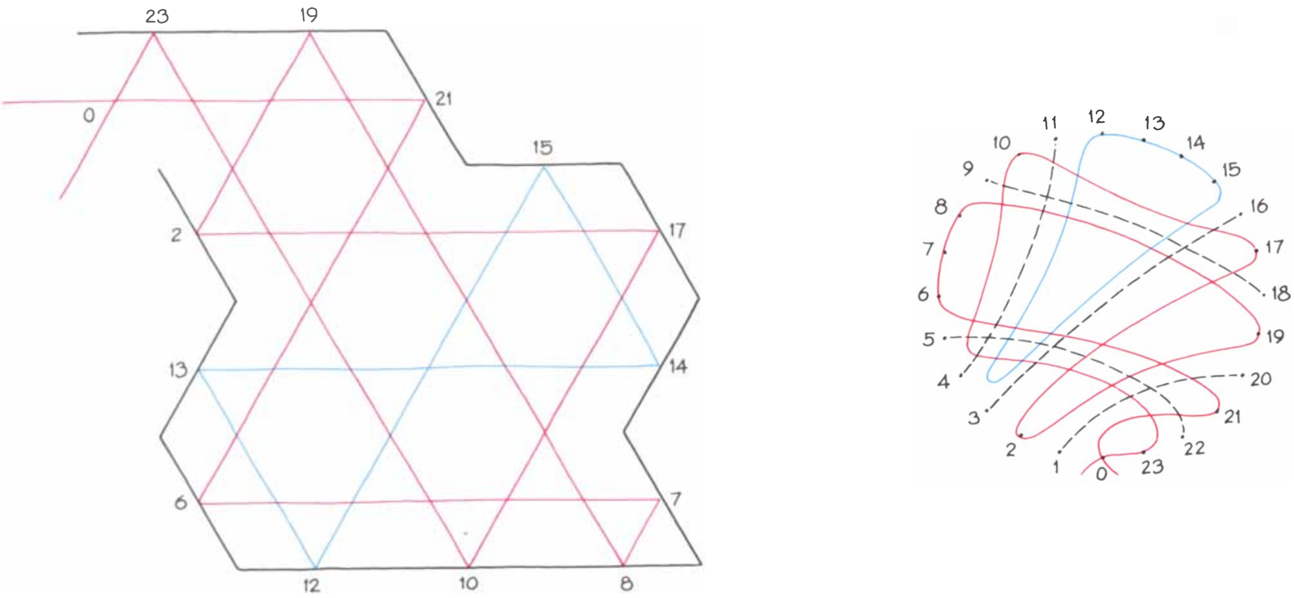
Building a maze by means of foldouts



A maze and its loop representation



A maze with four hallway systems



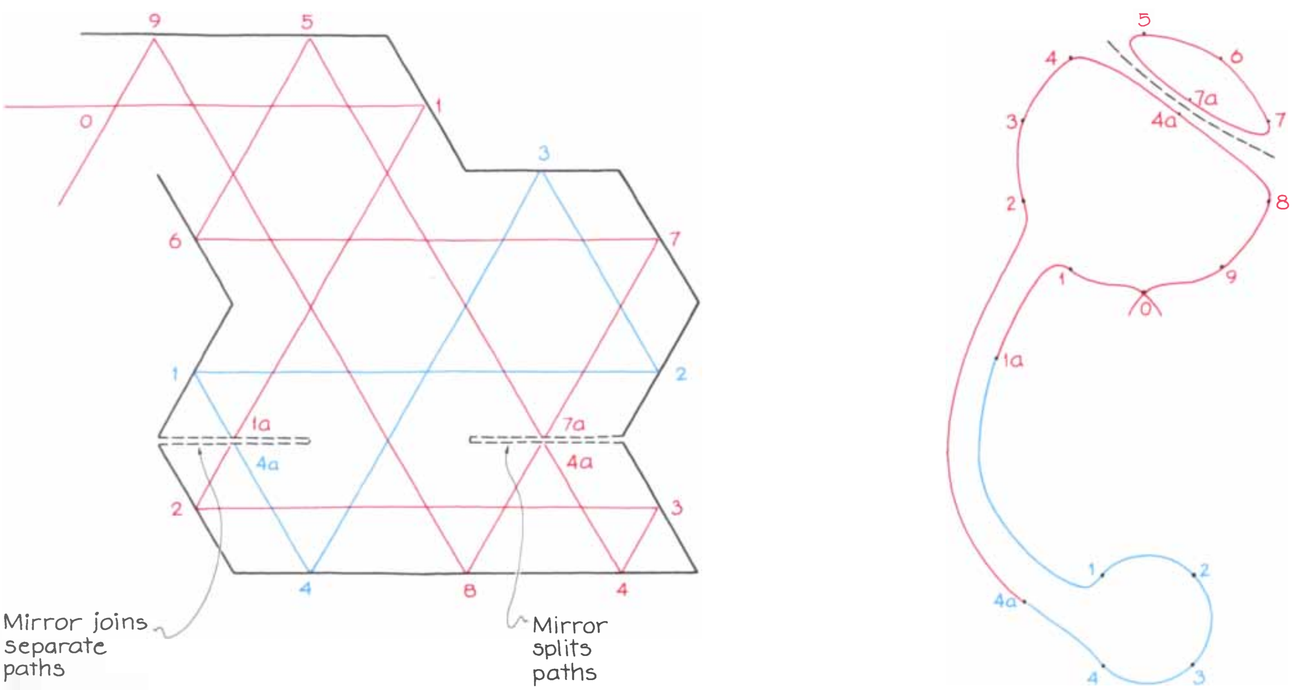
A complex-loop representation of a maze

tion.) Even without tracing the light path through the maze, I can guarantee that the path is reflected once from each mirror because the maze was built by foldouts.

The illustration also includes a way of representing the sequence of reflections as a loop. The line for the light path enters the loop at point 0 and runs past all the numbers representing reflections until it leaves the loop at 0. All mazes built by foldouts have similar loops, each loop with a single line running past all the numbers.

Your position within a maze can be represented by a point on the loop. The sections of the loop between your point and the entrance point 0 represent two of the hallways you can see. Since from any triangle deep within a maze you can see three pairs of hallways, your position can be represented by three points on the loop. If the triangle has no mirrors, the points are well separated on the loop. If the triangle has a single mirror on its perimeter, two of the points are adjacent. If the triangle consists of a 60-degree corner formed by two mirrors, all three points must be adjacent.

Suppose one of the mirrors on the perimeter of the maze is removed to provide a second opening to the maze. For example, if you mentally remove the mirror associated with reflection 17, the loop is altered. Under this arrangement a line enters at 0, passes 1 through 16 and leaves the loop at 17. Another line enters the loop at 17, passes 18 through 23 and leaves at 0. In brief, the removal of a perimeter mirror splits the loop and produces



The effect of inserting mirrors

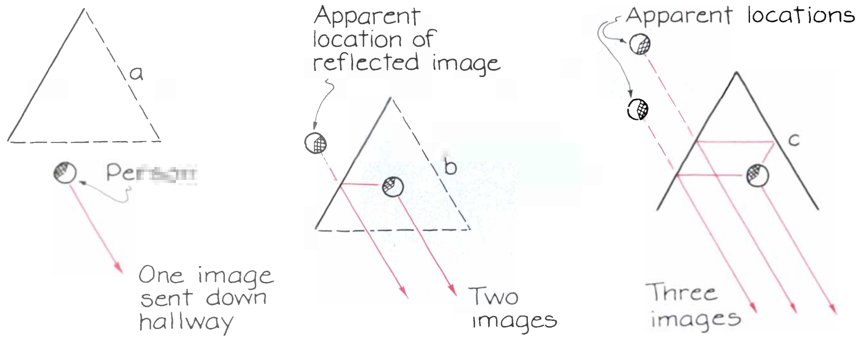
within the maze two light paths that are independent.

At entrance point 0 you will still see two hallways, but they are now unrelated. In one direction you see a hallway with door frames corresponding to reflections 1 through 16. In another direction the hallway has door frames corresponding to the reflections from 23 through 18. At the end of each hallway you see whatever lies outside the maze at the new opening. When you are within the maze, some of the hallways extending from you have the original opening at their ends, whereas other hallways have the new opening at their ends. (You can alter the maze in a similar way if instead of making a new opening you insert a nonreflecting scene at position 17.)

What happens to the light path if, the mirror having been returned to position 17, a double-sided mirror is removed from the maze? I removed the mirror for reflections 1 and 20, the one for 3 and 16 and the one for 4 and 11. These moves break the light path into four parts [*see bottom illustration on page 123*]. If you stand at the entrance of the maze, you see a hallway consisting of only three door frames (reflections 21, 22 and 23). At the end of it there is an image of yourself.

You can see deeper into the maze only by walking into it as far as the second triangle. There you intercept another light path consisting of reflections 2, 17, 18 and 19. Since this light path is closed (unbroken by an entrance or a nonreflecting scene), the hallways it generates are in principle infinite in extent. For example, if you face toward reflection 2, you see a hallway that stretches away from you with an infinite number of door frames. Your own image appears periodically along it. (With real mirrors the distant door frames and images of you become too muddled to resolve.)

Much of the maze remains hidden even when you move several more triangles deeper into it. Imagine someone hiding in the maze at the position of the black dot just in front of the mirror that makes reflection 7. When will you first see him in a hallway? He will appear only when you enter the closed light path in which he is standing. You first reach that light path at the colored dot after rounding the corner of the mirror that makes reflections 5 and 22. Before then you might catch tantalizing glimpses of him in the jumble of hallway images, but they give no clue to his location. Even at the colored dot you can be misled because the hallways he is in extend toward reflections 5 and 10. If you follow the hallway toward reflection 10, rounding the corner of the mirror that makes reflec-



Multiple images of a person in a hallway

tions 9 and 18, you will finally see the person directly.

The removal of the double-sided mirrors can be studied with a loop representation of the maze. The loss of a double-sided mirror creates a "bridge" in the loop. For example, the loss of the mirror that makes reflections 1 and 20 creates a bridge between those positions in the loop. Similarly, the loss of the 3–16 and 4–11 mirrors creates two more bridges. The line in the loop that enters at 0 and heads toward reflection 1 is immediately redirected by a bridge to reflection 21. Then it passes 22 and 23 and leaves through 0. Another line passes from reflection 2 through reflections 17, 18 and 19, being trapped between two bridges. The third line, which is also trapped between a pair of bridges, consists of reflections 12 through 15. The light path crosses a bridge from 15 toward reflections 3 and 4 but is turned back by another bridge. The final line, which is bounded by the bridge between 4 and 11, consists of reflections 5 through 10.

In this way you can study how the loss of a double-sided mirror alters the light paths in a maze without building or even redrawing the maze. Begin with the loop of the maze, add a bridge between the numbers corresponding to the reflections the mirror had and then draw lines within the loop. The first line begins at 0. When you come to a bridge, take it across the loop. Sometimes you may find another bridge immediately on the other side of the loop.

In one experiment I removed all the double-sided mirrors from the maze [*see top illustration on page 124*]. The loop representation is complex but consists of only two lines, indicating that the maze has only two light paths. One line enters at 0, follows a bridge to 21, takes another bridge to 6, 7 and 8 and then passes to 19 by another bridge. From there it takes bridges to 2, 17 and 10. Next it follows a bridge toward 4 but is immediately turned back onto another bridge to 23. Finally it passes from 23 through 0 and out of the maze.

The other line is simpler; it corresponds to a closed light path. It consists of reflections 12 through 15 and is constrained by two bridges. This maze would not be fun for a game of hide-and-seek even though the second light path is far from the entrance to the maze. When you enter the maze, you can see most of the interior. The layout needs no hallways.

The action of removing a double-sided mirror can be generalized with the aid of a loop representation of the maze. Imagine a loop already sectioned by bridges from previous losses of mirrors. If the new removal adds a bridge within a section, that section is split into two sections. The removal therefore splits a light path within the maze into two light paths. If instead the removal adds a bridge between two sections of the loop, the sections are joined. This removal joins two light paths within the maze.

Another way to modify a maze is to add double-sided mirrors. I studied how this step would modify the empty maze I have described [*see bottom illustration on page 124*]. For convenience I renumbered the reflections and redrew the loops as two separated sections. The larger section consists of point 0 and reflections 1 through 9. The smaller section is closed and consists of reflections 1 through 4.

In the example there are two types of insert. On the left side of the maze a mirror is inserted at a point where the two independent light paths cross. The new mirror connects them and hence connects their loop sections. I designated the reflection on one side of the new mirror 1*a* (because it is intermediate between 1 and 2 in the initially longer light path) and the reflection on the other side 4*a* (because it is intermediate between reflections 4 and 1 in the initially shorter light path). These reflections lie on bridges joining the two loop sections.

On the right side of the maze the inserted mirror is at a place where one of the light paths was crossing itself. I labeled the new reflections 4*a* and 7*a*. In the loop section the 4*a* reflection is on a bridge between 4 and 8 and 7*a* is on a bridge between 5 and 7. This type of insert splits a light path into two paths. As with the removal of a double-sided mirror, the addition of such a mirror can either split a loop section and a light path or it can connect two sections and two light paths.

In principle you can construct any maze by beginning with the simple two-mirror version. Expand it with foldouts. So far there is one light path, and you can see the entire interior by looking along a hallway at the entrance. The hallway exhibits one door frame for each mirror; your image appears at the end of the hallway. When you are in the maze, all six hallways extending from you show an image of the entrance at their end.

If you change the number of double-sided mirrors, make a new opening or replace any mirror with a nonreflecting scene, you split the light path. Your view along a hallway from the entrance or anywhere within the maze is then limited. As you walk through the maze you pass into different light paths and find new hallways with different scenes at their end or ones that extend indefinitely. You might enjoy tracing the light paths and making the loops for the Hall of Mirrors. Imagine walking into that maze. Where will you first see each scene? Where can a person hide from you?

Suppose you see in an unclosed hallway an image of someone who is not in your direct view. The number of images you see depends on the arrangement of mirrors near him and his position in the hallway. Suppose the person stands near one end of a mirror, outside the mirror's triangle [*see illustration on this page*]. The light path along which you view him extends toward the lower right. Suppose the person faces in that direction. Then you see a single image of his front that is due to light rays moving directly along the light path.

If he moves into the mirror's triangle, you see an additional image of him that is due to light rays starting from his right rear, reflected by the mirror and traveling down the hallway to you. If he is in a 60-degree corner, you see a third image of him that is due to rays starting from his left rear, reflected once by each mirror and traveling down the hallway.

Can you see multiple images of the person with other arrangements of the mirrors? What is the maximum number of images you can see of that person in a single hallway that is not closed? What is the maximum number of images of yourself that you can see in such a hallway?

DODGE ARIES K
5/50 PROTECTION, STANDARD.

It's true. In fact, for three years running Aries 2-door, 4-door and wagon have been the best six-passenger cars you can buy. That's because Aries offers the highest standard EPA mileage rating, Aries is still the best six-passenger car you can buy. One reason for this continued success is that we keep improving and refining Aries. And that includes options—like this year's new electron- And over 30 other features. In short, the kind of hard content that has traditionally made Aries such a great value.

So sure, when we say Aries is the best six-passenger car you can buy, it's a pretty bold claim. But it's also true,

THE BEST SIX.

BIBLIOGRAPHY

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

MATHEMATICAL GAMES

STEINER MINIMAL TREES. E. N. Gilbert and H. O. Pollak in *SIAM Journal of Applied Mathematics,* Vol. 16, No. 1, pages 1–29; January, 1968.

MINIMAL SURFACES REDISCOVERED. Rita M. Ehrmann in *The Mathematics Teacher,* Vol. 69, No. 2, pages 146–152; February, 1976.

STEINER TREES FOR LADDERS. Fan Rong K. Chung and R. L. Graham in *Annals of Discrete Mathematics,* Vol. 2, pages 173–200; 1978.

SMART SOAP BUBBLES CAN DO CALCULUS. Dale T. Hoffman in *The Mathematics Teacher,* Vol. 72, No. 5, pages 377–385, 389; May, 1979.

BEYOND THE ERA OF MATERIALS

THE LIMITS TO GROWTH. Donella H. Meadows, Dennis L. Meadows, Jørgen Randers and William W. Behrens III. New American Library, 1972.

WORLD DEMAND FOR RAW MATERIALS IN 1985 AND 2000. Wilfred Malenbaum. E/MJ Mining Information Services, McGraw-Hill, Inc., 1978.

THE WORLD ECONOMY: HISTORY AND PROSPECT. W. W. Rostow. University of Texas Press, 1978.

MATERIAL CONSUMPTION PATTERNS AND INDUSTRIAL ENERGY DEMAND IN INDUSTRIALIZED COUNTRIES. E. D. Larson, R. H. Williams and A. Bienkowski. PU/CEES 174, Center for Energy and Environmental Studies, Princeton University, 1984.

ENERGY-INTENSIVE MATERIALS AND THE DEVELOPING COUNTRIES. Alan M. Strout in *Materials and Society,* Vol. 9, No. 3, pages 281–330; 1985.

FIBRONECTINS

FIBRONECTINS: MULTIFUNCTIONAL MODULAR GLYCOPROTEINS. Richard O. Hynes and Kenneth M. Yamada in *The Journal of Cell Biology,* Vol. 95, No. 2, Part 1, pages 369–377; November, 1982.

CELL ATTACHMENT ACTIVITY OF FIBRONECTIN CAN BE DUPLICATED BY SMALL SYNTHETIC FRAGMENTS OF THE MOLECULE. Michael D. Pierschbacher and Erkki Ruoslahti in *Nature,* Vol. 309, No. 5963, pages 30–33; May 3, 1984.

BIOLOGICALLY ACTIVE SYNTHETIC PEPTIDES AS PROBES OF EMBRYONIC DEVELOPMENT. Jean-Claude Boucaut, Thierry Darribère, Thomas J. Poole, Hirohiko Aoyama, Kenneth M. Yamada and Jean Paul Thiery in *The Journal of Cell Biology,* Vol. 99, No. 5, pages 1822–1830; November, 1984.

MOLECULAR BIOLOGY OF FIBRONECTIN. R. O. Hynes in *Annual Review of Cell Biology,* Vol. 1, pages 67–90; 1985.

IS NATURE SUPERSYMMETRIC?

THE PARTICLE CONNECTION. Christine Sutton. Simon & Schuster, 1984.

THE SEARCH FOR SUPERSYMMETRY: PROBING PHYSICS BEYOND THE STANDARD MODEL. Howard E. Haber and G. L. Kane in *Physics Reports,* Vol. 117, Nos. 2–4, pages 75–263; January, 1985.

HOPE GROWS FOR SUPERSYMMETRY. John Ellis in *Nature,* Vol. 313, No. 6004, pages 626–627; February 21, 1985.

SUPERSYMMETRY—LOST OR FOUND? R. Michael Barnett, Howard E. Haber and Gordon L. Kane in *Nuclear Physics B,* Vol. 267, Nos. 3–4, pages 625–678; April 21, 1986.

EL NIÑO

RESOURCE MANAGEMENT AND ENVIRONMENTAL UNCERTAINTY: LESSONS FROM COASTAL UPWELLING FISHERIES. Edited by Michael H. Glantz and J. Dana Thompson in *Advances in Environmental Science and Technology,* Vol. 11; 1981.

EQUATORIAL PACIFIC OCEAN CLIMATE STUDIES: PROGRESS AND PLANS, OCTOBER, 1985. Stanley P. Hayes. NOAA Environmental Research Laboratories Technical Report, 1986.

THE HEART AS A SUCTION PUMP

DYNAMIC DETERMINANTS OF LEFT VENTRICULAR FILLING: AN OVERVIEW. Edward L. Yellin, Edmund H. Sonnenblick and Robert W. M. Frater in *Cardiac Dynamics,* edited by Jan Baan, Alexander C. Arntzenius and Edward L. Yellin. Martinus Nijhoff Publishers, 1980.

THE STRUCTURAL BASIS AND IMPORTANCE OF RESTORING FORCES AND ELASTIC RECOIL FOR THE FILLING OF THE HEART. E. H. Sonnenblick in *European Heart Journal,* Vol. 1, Supplement A, pages 107–110; 1980.

SKELETAL FRAMEWORK OF MAMMALIAN HEART MUSCLE: ARRANGEMENT OF INTER- AND PERICELLULAR CONNECTIVE TISSUE STRUCTURES. Thomas F. Robinson, Leona Cohen-Gould and Stephen M. Factor in *Laboratory Investigation,* Vol. 49, No. 4, pages 482–498; October, 1983.

SEXUAL SELECTION IN BOWERBIRDS

BOWER-BIRDS: THEIR DISPLAYS AND BREEDING CYCLES. Alan J. Marshall. Oxford University Press, 1954.

BIRDS OF PARADISE AND BOWER BIRDS. E. Thomas Gilliard. Published for the American Museum of Natural History by Natural History Press, 1969.

SEXUAL SELECTION: THE INTERFACE OF THEORY AND EMPIRICISM. S. Arnold in *Mate Choice,* edited by P. Bateson. Cambridge University Press, 1983.

THE EVOLUTION OF BOWER-BUILDING AND THE ASSESSMENT OF MALE QUALITY. Gerald Borgia, Stephen G. Pruett-Jones and Melinda A. Pruett-Jones in *Zeitschrift für Tierpsychologie/Journal of Comparative Ethology,* Vol. 67, Nos. 1–4, pages 225–236; January, 1985.

THE PERCEPTION OF APPARENT MOTION

APPARENT MOVEMENT WITH "SUBJECTIVE" CONTOURS. V. S. Ramachandran, V. M. Rao and T. R. Vidyasagar in *Vision Research,* Vol. 13, pages 1399–1410; July, 1973.

APPARENT MOTION. S. M. Anstis in *Handbook of Sensory Physiology: Vol. 8,* edited by R. Held, H. Leibowitz and H. L. Teuber. Springer-Verlag, 1978.

PERCEPTUAL ORGANIZATION IN MOVING PATTERNS. V. S. Ramachandran and S. M. Anstis in *Nature,* Vol. 304, No. 5926, pages 529–531; August 11, 1983.

MODERN WINDMILLS

WINDMILLS & WATERMILLS. John Reynolds. Praeger Publishers, 1970.

WIND-CATCHERS: AMERICAN WINDMILLS OF YESTERDAY AND TOMORROW. Volta Torrey. The Stephen Greene Press, 1976.

WINDMILLS. Suzanne Beedell. David & Charles Limited, 1979.

PUTNAM'S POWER FROM THE WIND. Revised by Gerald W. Koeppl. Van Nostrand Reinhold Company, 1982.

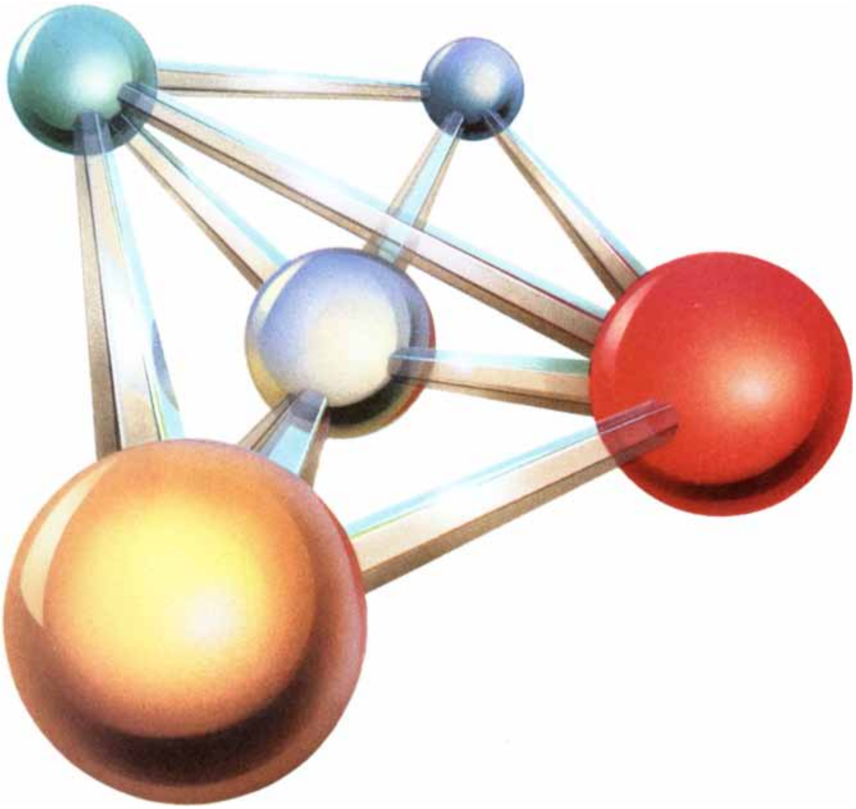
THE AMATEUR SCIENTIST

MULTIPLE IMAGES IN PLANE MIRRORS. Thomas B. Greenslade, Jr., in *The Physics Teacher,* Vol. 20, No. 1, pages 29–33; January, 1982.

'strat·ə·jē

Strategy

- *A structured means of achieving a goal, such as providing long-term growth in shareholder value.*
- *A carefully devised plan; e.g., balanced diversification between government and commercial businesses.*
- *Skillful management to attain an end; e.g., leveraging advanced technology to sustain a competitive advantage.*



A key Rockwell International strategy is maintaining leadership positions we have achieved in all of the principal markets we serve. By offering our customers greater value and quality in the products we sell. By building our strong global position to effectively service world markets. And by committing more than $2.5 billion to capital expenditures over the past five years to improve facilities, reduce costs and enhance our competitive position.

Above all, Rockwell is an $11.3 billion company committed to excellence and performance. This commitment is fulfilled by the dedicated work of 123,000 employees worldwide, 25,000 of them scientists, engineers and supporting technical personnel. To learn more about us, write Rockwell International, Department 815S-238, 600 Grant Street, Pittsburgh, PA 15219.

Rockwell International

...where science gets down to business

Aerospace / Electronics / Automotive
General Industries / A-B Industrial Automation


ABSOLUT
Country of Sweden
VODKA®
This superb vodka
distilled from grain grown
fields of southern Sweden
It has been produced at the famous
old distilleries near Åhus
in accordance with more than
400 years of Swedish tradition
Vodka has been sold under the name
Absolut Since 1879.
80 PROOF
PRODUCED AND BOTTLED IN SWEDEN 1,0 LITER (33,8 FL.OZ.
IMPORTED
IMPORTER AND SOLE DISTRIBUTOR FOR THE U.S.
CARILLON IMPORTERS LTD., TEANECK, N.J.
ABSOLUT TREASURE.
FOR GIFT DELIVERY ANYWHERE CALL 1-800-CHEER-UP (EXCEPT WHERE PROHIBITED BY LAW)
80 AND 100 PROOF/100% GRAIN NEUTRAL SPIRITS (ABSOLUT COUNTRY OF SWEDEN™) © 1986 CARILLON IMPORTERS LTD., TEANECK, NJ.