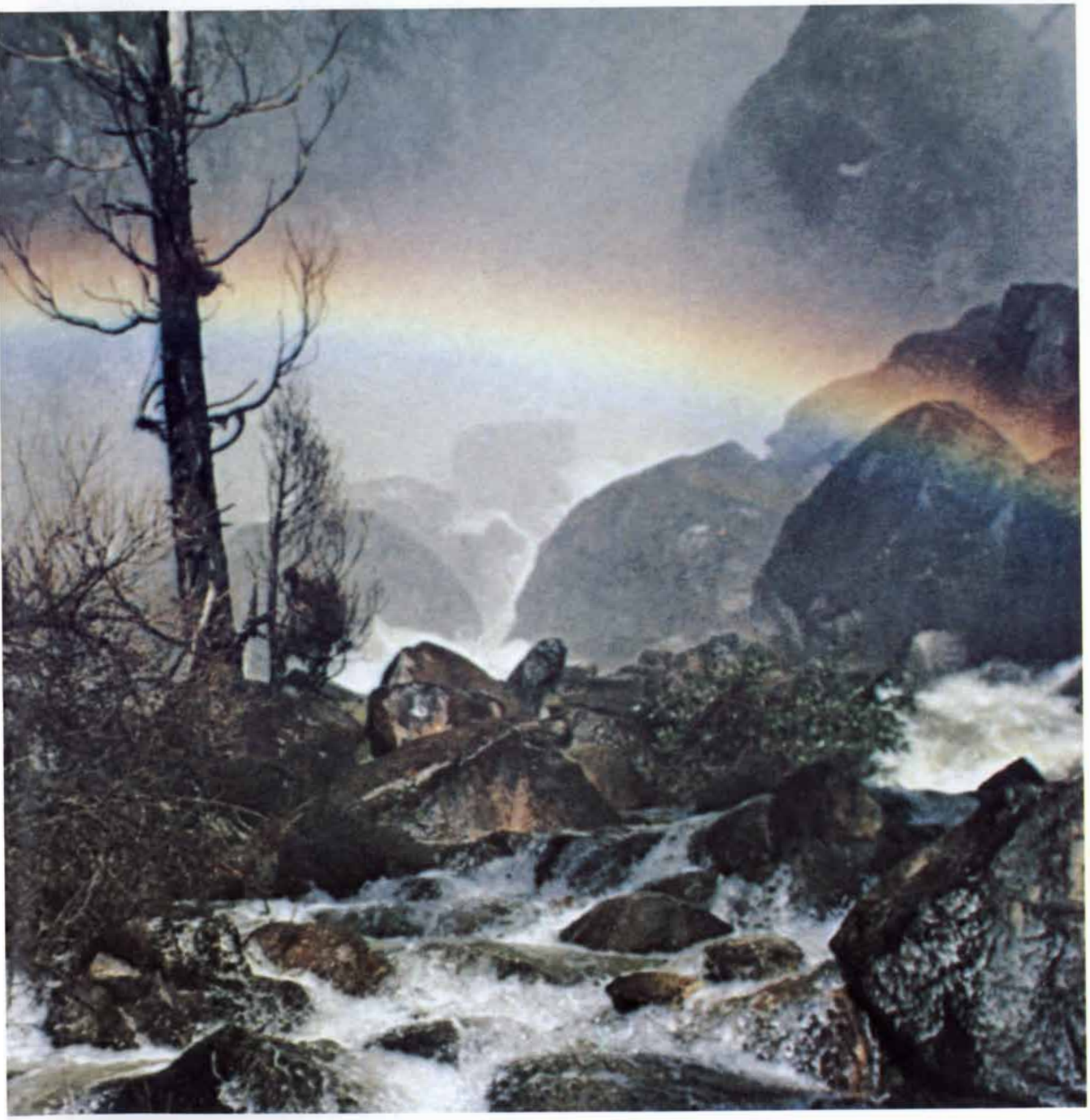
# SCIENTIFIC AMERICAN



THEORY OF THE RAINBOW

$1.50

*April 1977*

# The April Honda: lowest priced car in America.†

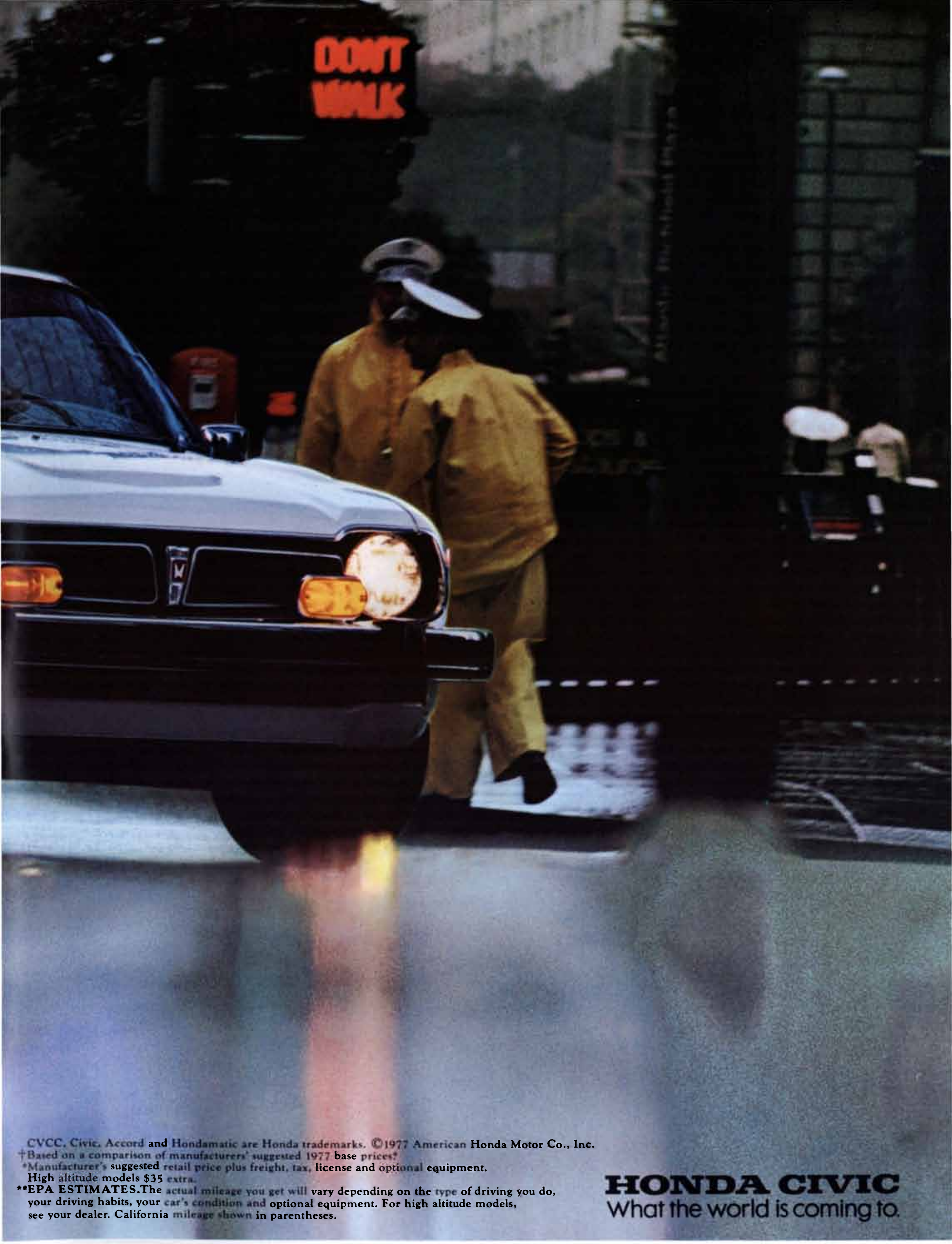
$2,779.* You won't find a car in the country priced lower than the Honda Civic Sedan. But it isn't the price itself that's so amazing. It's that the Honda Civic Sedan gives you so much for so little.

For example, the Civic gives you great handling, even in wet weather. The chassis layout, with front-wheel drive and transverse-mounted engine up front, provides excellent traction. And the Civic has power-assisted front disc brakes and rack and pinion steering as standard equipment.

Now that you've priced the Honda Civic, we hope you'll find out more about it from your Honda dealer. The Honda Civic isn't just a great buy. It's a great car.

| Civic 1237cc (not available in Calif. and high altitude counties) | | Price* | EPA Mileage Estimates** | |
|---|---|---|---|---|
| | | | Highway | City |
| Sedan | 4-Speed | $2779 | 43 | 28 |
| Hatchback | 4-Speed | $3049 | 43 | 28 |
| | Hondamatic | $3199 | 29 | 23 |
| **Civic CVCC 1488cc** | | | | |
| Sedan | 4-Speed | $2999 | 50 (46) | 39 (35) |
| Hatchback | 4-Speed | $3299 | 50 (46) | 39 (35) |
| | Hondamatic | $3449 | 37 (34) | 32 (28) |
| | 5-Speed | $3599 | 54 (51) | 41 (34) |
| Wagon | 4-Speed | $3549 | 41 (37) | 30 (28) |
| | Hondamatic | $3699 | 32 (32) | 27 (25) |
| **Accord CVCC 1600cc** | | | | |
| Hatchback | 5-Speed | $4145 | 48 (47) | 38 (33) |
| | Hondamatic | $4295 | 31 (32) | 26 (25) |

| APRIL | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FRI | SAT | SUN | MON | TUES | WED | THURS | FRI | SAT | SUN | MON | TUES | WED | THURS | FRI |




| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAT | SUN | MON | TUES | WED | THURS | FRI | SAT | SUN | MON | TUES | WED | THURS | FRI | SAT |

Established 1845 SCIENTIFIC AMERICAN April 1977 Volume 236 Number 4

ARTICLES

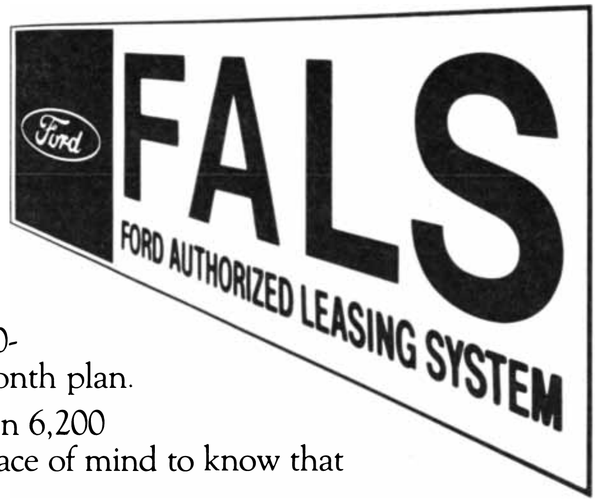


DEPARTMENTS

# Why more people are turning to Maintenance Leasing

## at Ford Authorized Leasing System dealers.

**It is economical** because **F**ord **A**uthorized **L**easing **S**ystem dealers are specializing in a new maintenance lease program that can possibly offer you savings.




**Volume counts.** As a group, **F**ord **A**uthorized **L**easing **S**ystem dealers do more leasing business than any other organization, therefore, they are well qualified to advise you on individual or fleet leasing.

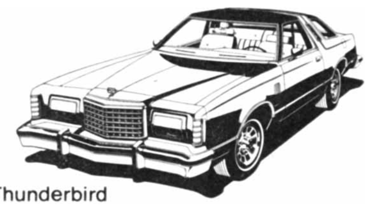
**It's flexible.** You can choose from two different maintenance leasing plans. Pick the 24,000-mile/24-month plan or the 40,000-mile/36-month plan.

**Service is available nationwide** at more than 6,200 Ford Motor Company Dealers. It gives you peace of mind to know that service is always close by.

**You can use coupons, not cash, for service.** You can make a one-time payment for the convenient coupon book or have the cost prorated in your monthly rate. Insurance can be included, too.

**Above all, FALS Maintenance Leasing is convenient.** Your car is cared for by Ford-trained professionals using the latest diagnostic equipment and tools. And they have genuine Ford parts in stock for fast service. Remember, too, your **F**ord **A**uthorized **L**easing **S**ystem dealer offers the kind of plan you need—open-end, closed-end, cars or trucks, individual or fleet. And if low cost is important to you, he's the man you want to see.

Ford LTD



Thunderbird

Granada

FORD

FORD DIVISION

***Consult your Yellow Pages for your nearest Ford Authorized Leasing System dealer.***

## THE COVER

The photograph on the cover shows a rainbow formed in spray over Bridal Veil Creek in Yosemite National Park in California. The rainbow is unusual for the brilliance and purity of its colors. It is also distinguished by its apparent scale: the bow seems to be viewed at close range. Actually a rainbow cannot have a well-defined location. It is merely a cone of light rays with the observer at the apex (see "The Theory of the Rainbow," by H. Moysés Nussenzveig, page 116).

## THE ILLUSTRATIONS

Cover photograph by Richard Rowan, Photo Researchers, Inc.

| Page | Source |
|---|---|
| 23 | General Dynamics Corporation |
| 24–29 | Allen Beechel |
| 31 | EROS Data Center |
| 32–33 | Andrew Tomko |
| 34–35 | Alan D. Iselin |
| 36–37 | Andrew Tomko |
| 38 | Alan D. Iselin |
| 39 | EROS Data Center |
| 40 | V. P. Solonenko |
| 41 | EROS Data Center |
| 43 | Ben Rose |
| 44–50 | Jerome Kuhl |
| 57 | Jet Propulsion Laboratory, California Institute of Technology |
| 64–80 | Lorelle M. Raboni |
| 82 | Shinya Inoué, University of Pennsylvania |
| 83 | Douglas Faulkner |
| 84 | Patricia J. Wynne |
| 86 | Shinya Inoué, University of Pennsylvania |

| Page | Source |
|---|---|
| 89 | Patricia J. Wynne |
| 90–92 | Kayo Okazaki and Kent McDonald |
| 97 | Kitt Peak National Observatory |
| 98–99 | George V. Kelvin |
| 100–101 | George V. Kelvin (*drawings*); Helmut A. Apt, Kitt Peak National Observatory (*spectra*) |
| 102 | George V. Kelvin |
| 107–113 | Tom Prentiss |
| 114 | D. W. Phillipson, British Institute in Eastern Africa |
| 117 | Sharon Beals, Black Star |
| 118–126 | Gabor Kiss |
| 129–134 | Ilil Arbel |
| 140B | *Man Discovers the Galaxies.* © 1976, Science History Publications (*top*); Kitt Peak National Observatory (*bottom*) |
| 142 | *The Works of Isambard Kingdom Brunel: An Engineering Appreciation.* © 1976, Institution of Civil Engineers |
| 144 | Richard Lee/Anthro-Photo |

# LETTERS

Sirs:

The following verses are written in response to Martin Gardner's invitation ["Mathematical Games," February] to write poems based on transposing etaoin shrdlu, with or without an added letter:

SHARON DILUTE

Old hunter, as I
dash in to rule,
insult a horde
and so lie hurt,

no adult heirs,
riled, shout an
oath: "Sin lured,
had soul inert."

I share untold
tales. Our hind
had true loins,
a tin shoulder.

(Lutherans do. I
should retain
a ruined sloth,
his aunt—older.)

Saturn held Io,
hurled oats in
hot laundries;
lost hair, nude.

(Do the urinals
in Utah solder
the dual irons
thin? Also rude?)

Sail round the
lush, rationed
Hanoi strudel:
Oh, true island!

Thus, dear lion,
run! Hie, lads, to
unlit Hades, or
South Ireland!

WALTER G. LEIGHT

Bethesda, Md.

Sirs:

In "Mathematical Games" Martin Gardner states that no one has made a dictionary word out of etaoin shrdlu, although "outlanderish" comes close.

The word certainly does come close. "Outlandisher" appears in *New Standard Unabridged Dictionary*. It follows "outlander," meaning foreigner. "Outlandisher" is not further defined but is apparently a variant of "outlander."

WESTON HARE

Oxford, Mass.

Sirs:

Regarding the scrambling of the words in the poem by Oscar Wilde, with anagrammatic authorship ["Mathematical Games," February], how about the following?

Vilest poison in prison air
Withers man, and well.
Is it only there that the good
Wastes like weeds?
What bloom is in deeds?
—OSWALD C. IRE

What is prison?
It is that man wastes like weeds.
Only in poison air well bloom
vilest deeds,
And there in withers the good.
—ROD I. CLAWSE

In vilest prison is man well?
Only there bloom poison weeds,
And the good wastes in that air.
What is it like?—withers deeds!
—ERIC O. L. DAWS

What good is in prison?
The vilest deeds
Bloom well in there,
And poison the air.


**Scientific American,** April, 1977; Vol. 236, No. 4. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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

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

**Offprint correspondence** and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 35 cents.

**Subscription correspondence** should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name

New Address

Old Address





*This is Delta's Wide-Ride℠ L-1011 TriStar, a $21 million superjet. Cabins are almost 19 feet wide. All seats are two-on-the aisle.*

Man it is that wastes,
Withers like weeds.
—DORA S. WILEC

Man withers in prison; vilest deeds
Bloom well, and only there
It is that good wastes like weeds.
What is in the poison air?
—ROSA W. CLIDE

This is the first time that the English major of the family has had more fun with the "Mathematical Games" section than the more mathematically oriented member of the family. I am sure you will agree that the anagrammatic authors deserve the inattention fame has paid them, but I did have fun making the verses up.

CATHARINE M. HOFF

Skokie, Ill.

Sirs:

In the interesting article on the antibody combining site by J. Donald Capra and Allen B. Edmundson [SCIENTIFIC AMERICAN, January] it was stated that Bence-Jones urinary proteins were first observed by Henry Bence-Jones, a London chemical pathologist. Although it is true that Bence-Jones's papers of 1847 and 1848 were concerned in some detail with the properties and analysis of the protein that has come to bear his name, he was not the first to describe its characteristics. That honor belongs to William MacIntyre, who participated as a consulting physician in the first recorded case of the condition now commonly known as multiple myeloma. In particular MacIntyre was the first to note the key and rather unusual property the protein had of first precipitating out of the urine at about 50 degrees Celsius and then redissolving as the temperature increased. This behavior was so characteristic that it was used for more than a century as a test, albeit one prone to ambiguities and misinterpretations, for the disease.

Although Bence-Jones in his papers gives credit to MacIntyre for the discovery of the unusual properties of the urine, it is Bence-Jones who is associated with the protein. This is probably because MacIntyre delayed publication of his observation in a paper under his own name until 1850, a fact that perhaps should be noted by modern researchers.

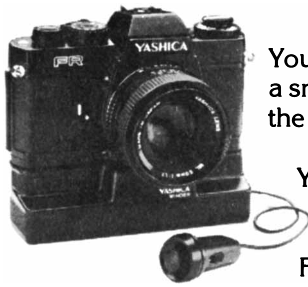
More detailed information on this controversy can be found in an informative article by John R. Clamp that appeared in the December 23, 1967, issue of *The Lancet*.

S. D. KRAMER

Oak Ridge National Laboratory
Oak Ridge, Tenn.

# Any ordinary camera could have taken this photograph.

# The Yashica FR is no ordinary camera.




Your chances of getting more than a snapshot jump sky high with the extraordinary Yashica FR system.

For a lot of reasons.

You see, an ordinary camera is like a local train. It has to stop at every shot. Hook up the Yashica FR to its compact, lightweight winder and you've got the express that's powered with 6 batteries, not 4, for constant high speed shooting. Just focus, set exposure and you're off and "clicking" photo sequences at speeds up to 2.5 frames per second. And that's 20% faster than our competitors' fastest winders.

Only the Yashica FR has an advanced electromagnetic release system for instant response and a unique focal plane shutter that assures shutter-speed accuracy by avoiding the "inertia lag" found in conventional designs. You get sharper prints and slides shot after shot after shot.

Even the cable release is out-of-the-ordinary. Electromagnetics let you fire instantly from up to 30 feet away. The line is so flexible you can wrap it around your fist.

And then, if you can't be tied to the camera at all, just pick up the infrared remote control, point at the camera and shoot. Anything but ordinary. Likewise, the lenses. The Yashica bayonet mount takes on both Yashica ML and Carl Zeiss T-Star (T*) multi-coated lenses.

The extraordinary Yashica FR: the fastest winder around, the accuracy of the electromagnetic release system and rugged compact design. No other camera can keep up with it. Visit your dealer. Or write to us: Yashica, Inc., 411 Sette Drive, Paramus, New Jersey. 07652. In Canada: Yashica Canada Ltd., Ontario L4T IL2.

YASHICA FR

# 50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

APRIL, 1927: "'Hello, London. New York is calling.' Three months ago these words wafted across the sea over an invisible 'talk bridge' and officially opened a telephonic service between the Old World and the New. When the international circuit was closed after its first day of operation, scores had chatted back and forth across the ocean at a cost of $75 for three minutes and $25 for each additional minute. It was a big day for radio."

"The twin laws of the conservation of matter and the conservation of energy are as useful as ever, for they still serve to clarify our conceptions and to guide our experimentation. Yet neither law is now regarded as absolute in itself, and it seems that we shall have to substitute a general law that will include the two and allow for the transformation of matter into energy and vice versa. Einstein has worked out the formula for the equivalence of matter and energy, so that we can now calculate how much heat will be produced if a certain mass of matter is annihilated. This idea has been welcomed by the astronomers, who have long been hard put to it to devise means of keeping up the fires of the sun."

"Atmospheric nitrogen is quite useless to either the explosive maker or the farmer until it can be forced to combine with hydrogen or oxygen to form ammonia or nitrates. When it is thus combined, it becomes quite invaluable to both, for either ammonia or nitric acid will serve the farmer, and ammonia can be converted into the nitric acid required by the explosive maker. The synthesis of ammonia by various processes is rapidly gaining ground in the United States, and at the present time we have a potential capacity for ammonia synthesis equivalent to 30,330 tons of nitrogen per year. The by-product coke ovens of the world are also producing immense quantities of ammonia for agricultural use. In the United States in 1925 there was produced ammonia from coke ovens equivalent to 123,600 tons of nitrogen. Over the past 25 years the world has changed from practically 100 per cent dependency on the natural niter beds of Chile for its inorganic nitrogen supply to 30 per cent dependency."

"With a car driven by a 500-horsepower motor running at 2,200 revolutions per minute Captain Malcolm

# "The most complete and most scholarly dictionary of the English language"

*—The Christian Science Monitor*

As an introduction to membership in the

## BOOK-OF-THE-MONTH CLUB®

## yours for only

## $17.50 [Publisher's list price $95]

THE SUGGESTED TRIAL: You simply agree to buy four Club choices within a year at substantial savings on most books you choose.

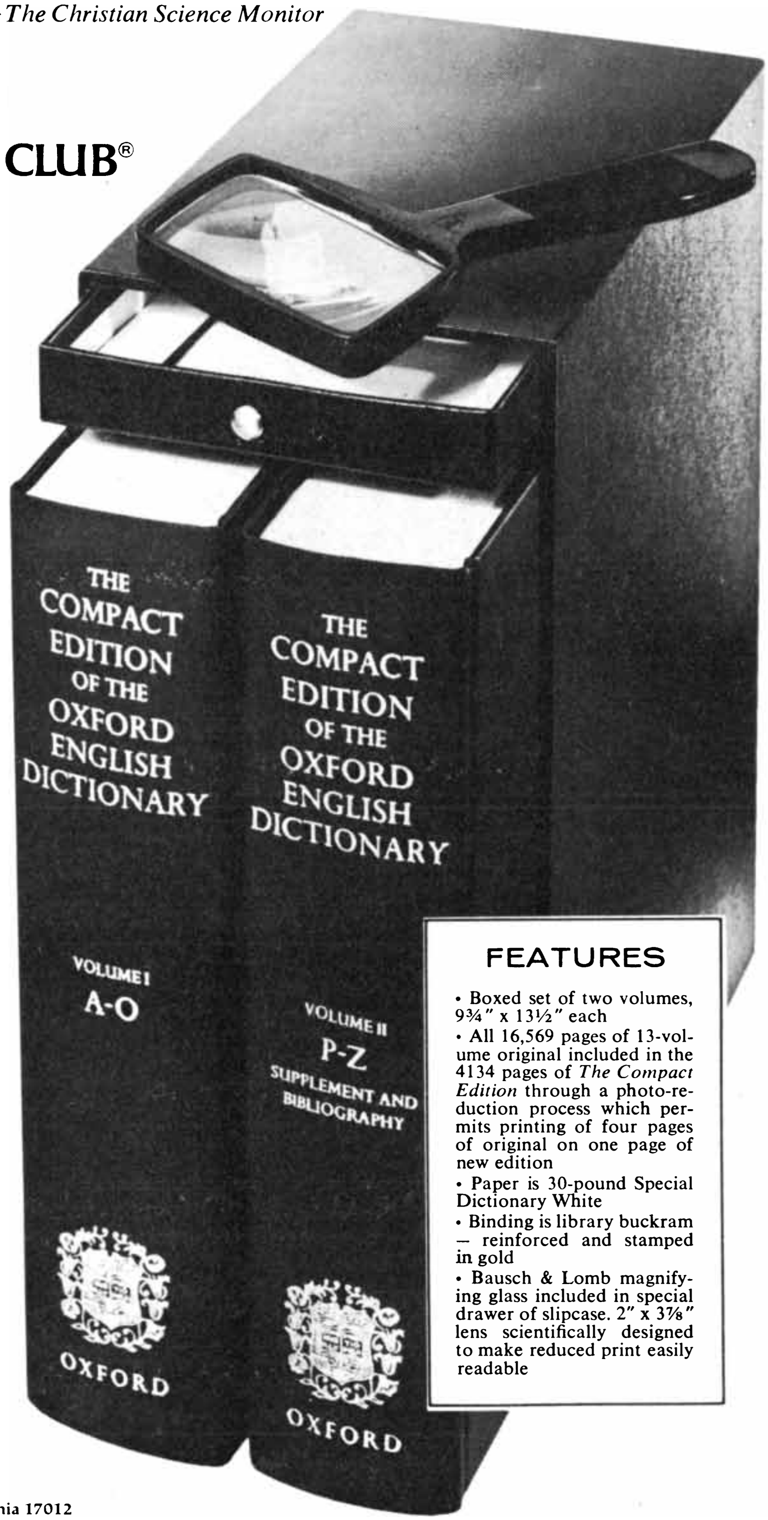
The Oxford English Dictionary is generally regarded as the final arbiter of the English language. Until recently, it has been available only as a thirteen-volume set, currently priced at $395. Now, through the combination of an ingenious method of micrographic reproduction and a fine Bausch & Lomb optical lens, every single one of its 16,569 pages, 50 million words and close to 2 million illustrative quotations appears, in easily readable form, in *The Compact Edition.*

> Book critic Christopher Lehmann-Haupt of *The New York Times* has said of this edition: "It is something of a miracle.... *The Compact Edition* is easier to work with than the original with its thirteen separate volumes."

Even more extraordinary, as a trial member of the Club you may obtain the two-volume set for only $17.50. And as long as you remain a member, you will receive the *Book-of-the-Month Club News,*® a literary magazine announcing the coming Selection and describing other important books, known as Alternates, most of which are available at substantial discounts—up to **40%** on more expensive volumes.

If you continue after your trial membership, you will earn at least one Book-Dividend® Credit for every Selection or Alternate you buy. These Credits enable you to acquire for your library a wide variety of books, called Book-Dividends, at astonishing savings—*at least 70% of publishers' list prices.*

**All Book-of-the-Month Club books, including Book-Dividends, are identical to the publishers' editions in content, format, size and quality.**




### FEATURES

- Boxed set of two volumes, 9¾" x 13½" each
- All 16,569 pages of 13-volume original included in the 4134 pages of *The Compact Edition* through a photo-reduction process which permits printing of four pages of original on one page of new edition
- Paper is 30-pound Special Dictionary White
- Binding is library buckram — reinforced and stamped in gold
- Bausch & Lomb magnifying glass included in special drawer of slipcase. 2" x 3⅞" lens scientifically designed to make reduced print easily readable

BOOK-OF-THE-MONTH CLUB, INC., Camp Hill, Pennsylvania 17012

# IT TAKES A VERY SPECIAL CASSETTE DECK TO GET SO MUCH BEAUTIFUL MUSIC OUT OF SOMETHING THIS LITTLE.


The recording tape in a cassette is only an eighth of an inch wide.

Crammed into that eighth of an inch may be as many as 64 tracks mixed down to two. A hundred musicians. Countless overdubbings. Not to mention the entire audible frequency range.

Any cassette deck can reproduce part of what's been put down on that eighth of an inch.

But the Pioneer 9191 was designed to reproduce all of it. Superlatively. Without dropouts, unacceptable tape hiss, or noticeable wow and flutter.

Take our tape transport system, for example.

Since the tape in a cassette moves at only 1-7/8 inches per second, even the most minuscule variation in tape speed will make a major variation in sound.

To guard against this, where most cassette decks give you one motor, the 9191 comes with two. The first is used only for fast forward and rewind, so the second can be designed exclusively to maintain a constant speed for play and record.

All of our tape drive components—the capstan, belt, and flywheel—are finished to incredible tolerances. Which give the 9191 the kind of wow and flutter figures that no deck in our price range can match.

Of course, having a great tape transport system means nothing if you don't have great electronics to back it up.

We do.

The 9191 comes with an advanced three stage direct coupled amplifier that extends high frequency response and minimizes distortion. The built-in Dolby system can reduce tape hiss by as much as 10 decibels in high frequencies.

Our multiplex filter lets you record

Campbell recently broke the world's record for a mile from a flying start: 174.883 miles per hour. The same engine could drive a bombing plane at a speed of 232 miles per hour. The car cost $40,000 and required two years to make."

"We have long been wont to think of our United States as having a more nearly universal use of modern conveniences than any nation on earth. So we fondly dreamed. But now come figures issued by our Department of Commerce showing that in the proportion of homes equipped for electric service we are not first or even second but sixth. The list runs: Switzerland, 96.5 per cent; Japan, 73.4 per cent; Denmark, 72.0 per cent; Canada, 62.3 per cent; New Zealand, 59.0 per cent; United States, 56.0 per cent. To explain that the countries of the Old World are more densely settled than ours and that Canada and New Zealand have most of their populations in relatively small areas is poor consolation."

SCIENTIFIC AMERICAN

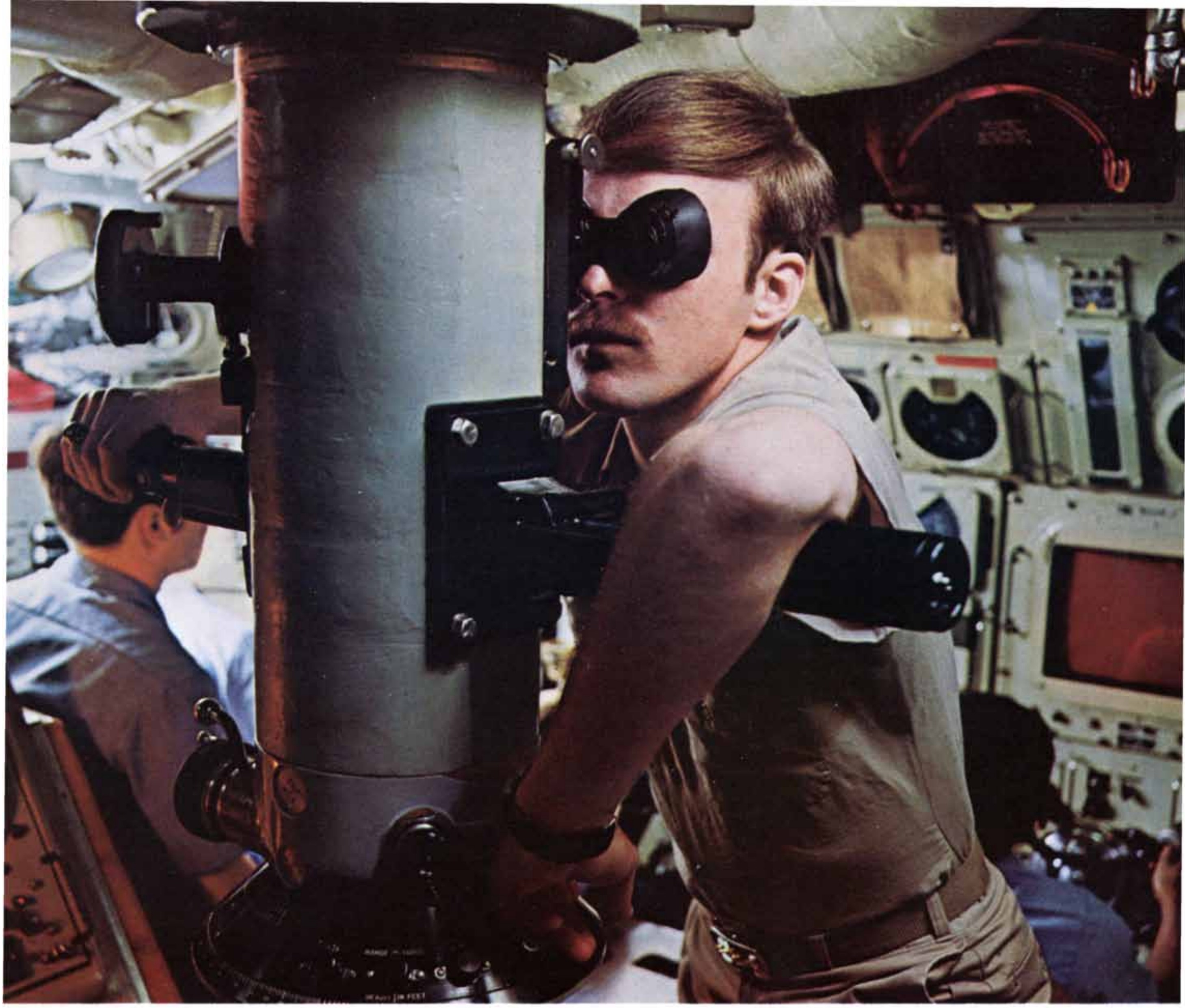
APRIL, 1877: "The coal fields of Colfax County, to the north of Las Vegas, are the most valuable and extensive deposits within our territorial limits. After the purchase of the so-called Maxwell Land Grant by English capitalists a few years ago the new owners engaged the services of able engineers, who after surveying the grant reported they had traced seams of coal averaging more than six feet in thickness over a tract of more than 100 square miles. More than 60,000 acres of coal! Without taking into account the thousands of hearths that could be supplied with fuel, what a vast trade this would give to a railroad extended south from Colorado! What an immense amount of steam power this would create for the manufacturer and other enterprises!"

"Dr. Kühne, Professor of Physiology in Heidelberg, announces his having been able to obtain actual images on the retina that correspond to objects that had been looked at during life. Kühne took a rabbit and fixed its head and one of its eyeballs at a distance of a meter and a half from an opening 30 centimeters square in a window shutter. The head was covered for five minutes by a black cloth and then exposed for three minutes to a midday sky. The animal was then instantly decapitated, the eyeball that had been exposed was rapidly extirpated in yellow light, then opened and instantly plunged in a 5 per cent solution of alum. On the following morning the milk-white and now toughened retina of the eye was carefully isolated, separated from the optic nerve and tanned. It then exhibited on a beautiful rose-red ground a nearly square image with sharply defined edges. Professor Bunsen was among the witnesses to this beautiful experiment."

"At the Physical Society, London, Prof. Foster showed experimentally the polarization of heat rays, employing Nicol's prisms of $2^5/_8$ in. aperture and a thermopile surrounded by a double jacket and connected with a Thomson galvanometer. When the prisms were placed between the thermopile and a heat source and the prisms were at 90° to each other, only a slight movement of the galvanometer was observed. As the angle was diminished the amount of deflection increased steadily up to about 60 divisions on the galvanometer's scale."

"The Central Pacific Railroad Company has lately arranged to have 40,000 trees of the species *Eucalyptus globulus* set out along the 500 miles of the right-of-way of the company. This is only the first installment, as it will require about 800,000 of the trees for the 500 miles of valley where they are to be cultivated. The immediate object of the plan is to increase the humidity of the region and lessen the liability to drought. It is an established fact that the destruction of our forest trees over large tracts of the country is having a direct effect on the climate, and we are glad to know that this company is replacing, at least in part, the forests that have been destroyed."

"A criminal lately gave a reporter of the New York *Herald* the following mode of introducing powder into a safe for the purpose of blowing open the doors. 'What tools did you use in drilling the holes?' asked the reporter. 'Good cracksmen don't use tools,' answered the burglar. 'I'll show you how to blow open any safe in New York without any tools. Just take me to a safe.' There happened to be a safe in Judge Kilbreth's private room, and the writer acquainted the magistrate with the prisoner's proposal. 'By all means,' said Judge Kilbreth, 'let us learn,' and in a moment the room was filled with spectators. The prisoner knelt beside the safe. Said he: 'Look at this door. It fits so tight that no instrument can be introduced in the cracks and powder cannot be inserted. The burglar simply sticks putty along all the cracks except in two places, one at the top of the door and the other at the bottom, where he leaves a space about an inch long uncovered by the putty. At the lower place he puts a quantity of powder, and he sucks out the air from the upper place, either with a suction pump, which is the better way, or with his mouth. The vacuum created in the safe draws in the powder through the small crack below. The entire job does not take more than five minutes.'"



# Volkswagen celebrates the millionth Rabbit with cars that look like a million.

Imagine! One million Rabbits hopping around all over the world!

No one ever introduced a car so successfully before.

So to mark the occasion, we're offering our stunning "Champagne Edition" Volkswagens —all decked out and dolled up like you've never seen VWs before.


The limited *Champagne*


Bally
FIREBALL!
PLAYERS
1 2 3 4
SCORE
BALL IN PLAY
100
100
Bally

# THE AUTHORS

ELISABETH DRAKE and ROBERT C. REID ("The Importation of Liquefied Natural Gas") are respectively manager of the Hazard Assessment and Control Unit at Arthur D. Little, Inc., and codirector of the Liquid Natural Gas Research Center at the Massachusetts Institute of Technology. Drake received her S.B. and Sc.D. degrees in chemical engineering from M.I.T. and has been with Arthur D. Little since 1958. Her background in cryogenic technology, safety research with hazardous materials and methodology for assessing risks led her into the analysis of the safety and reliability of facilities where hazardous substances, including liquefied natural gas, are handled. Reid received his doctorate in chemical engineering from M.I.T., later becoming a member of the faculty. His primary interests are in thermodynamics and the physical-chemistry aspects of engineering. He has worked on numerous research projects concerned with the safety of handling industrial chemicals.

PETER MOLNAR and PAUL TAPPONNIER ("The Collision between India and Eurasia") are respectively assistant professor in the department of earth and planetary sciences at the Massachusetts Institute of Technology and attaché de recherche of the French National Scientific Research Council (C.N.R.S.) at the University of Montpellier. Molnar was graduated from Oberlin College in 1965 with a bachelor's degree in physics. An avid interest in the outdoors led him to do his graduate work in geophysics at the Lamont-Doherty Geological Observatory of Columbia University. After receiving his Ph.D. from Columbia in 1970, he spent another year there and two years at the Scripps Institution of Oceanography. He has been at M.I.T. since 1974. Tapponnier, born and raised in the town of Annecy in the French Alps, majored in physics and mathematics at college in Lyon. In 1967 he entered the École Nationale Supérieure des Mines in Paris, where he studied geology and geophysics. In 1971 he began work at Montpellier on tectonics. From September, 1973, through February, 1975, he was at M.I.T. studying rock formation under William F. Brace. It was then that he and Molnar met and started their work on the geology of Asia. Tapponnier reports: "I am now trying to use some of the physical insights we got into continental deformation in Asia to understand more about the Alpine chains of the Mediterranean and Middle East."

DEREK C. BURKE ("The Status of Interferon") is professor of biological sciences at the University of Warwick. After receiving his B.Sc. and Ph.D. in chemistry from the University of Birmingham, in 1953 he went to Yale University, where he spent two years studying the nucleosides of Caribbean sponges. In 1955 he returned to England to work at the National Institute for Medical Research with the late Alick Isaacs on the nucleic acid of influenza viruses. Burke began research on interferon within a few months of its discovery by Isaacs and Jean Lindenmann in 1957. He has continued his work on it since then, sometimes as his sole research interest and sometimes along with investigations on the biochemistry of viruses. In 1960 he joined the faculty of the University of Aberdeen, where he remained for nine years. He has been at Warwick since 1969.

DONALD E. KNUTH ("Algorithms") is professor of computer science at Stanford University. He received his bachelor's and master's degrees in mathematics at the Case Institute of Technology and his Ph.D. in mathematics from the California Institute of Technology in 1963. He remained at Cal Tech for another five years, eventually becoming full professor of mathematics. In 1968 he joined the computer science department at Stanford. Knuth has received numerous awards for his research, including the A. M. Turing Award of the Association for Computing Machinery in 1974. The Turing Award cited the seven-volume series of books he is writing titled *The Art of Computer Programming*. He plans to have the first five volumes completed by 1980. When the entire series is finished, Knuth wants to turn to writing music. His home is built around a two-story pipe organ that he designed. He also designed a baroque-style pipe organ that was built for the Bethany Lutheran Church in Menlo Park, of which he is an associate organist. Knuth has written a novel about mathematics titled *Surreal Numbers,* which was published by Addison-Wesley in 1974.

SHINYA INOUÉ and KAYO OKAZAKI ("Biocrystals") are respectively professor of biology and director of the program in biophysical cytology at the University of Pennsylvania and associate professor of biology at the Tokyo Metropolitan University. Both Inoué and Okazaki received their early research training at the Misaki Marine Biological Station in Japan. Inoué went to Princeton in 1948 for graduate study and has subsequently become a resident of the U.S. Before he joined the faculty at Pennsylvania he taught at the University of Washington, the Tokyo Metropolitan University and the University of Rochester and was chairman of the department of cytology and anatomy at the Dartmouth Medical School. Currently he is a trustee of the Marine Biological Laboratory at Woods Hole, Mass. Okazaki is one of the few women to hold the rank of associate professor in Japan. Her main research has been pioneering work on the embryology of sea urchins.

HELMUT A. ABT ("The Companions of Sunlike Stars") is an astronomer at the Kitt Peak National Observatory in Arizona. After completing his undergraduate work at Northwestern University, in 1952 he obtained his Ph.D. from the California Institute of Technology. For a year he did research at the Lick Observatory of the University of California, and from 1953 through 1963 he was at the Yerkes Observatory of the University of Chicago. In 1955 and 1956 he did much of the fieldwork in searching for a site for what became the Kitt Peak Observatory, and he has been at Kitt Peak since 1963. His main interests are problems concerned with stellar motions—the speed of a star through space, its rotational velocity and its orbital motion—and why some of those motions affect the star's composition and evolutionary history. Since 1971 Abt has been managing editor of *The Astrophysical Journal.*

D. W. PHILLIPSON ("The Spread of the Bantu Language") is assistant director of the British Institute in Eastern Africa, based in Nairobi. As an undergraduate at Gonville and Caius College at the University of Cambridge he studied archaeology and anthropology, being graduated in 1964. From that year until 1973 he was the director of the National Monuments Commission of the Zambia government. In addition to his continuing research on the iron age of Africa he is currently investigating the evidence for farming techniques in the northern parts of East Africa before the iron age.

H. MOYSÉS NUSSENZVEIG ("The Theory of the Rainbow") is professor of physics at the University of São Paulo. Born in São Paulo in 1933, he received his B.Sc. in physics at the university there in 1954 and his Ph.D. in physics from the same institution in 1957. Thereafter he moved to the Brazilian Center for Physics Research in Rio de Janeiro, where he was appointed full professor in 1962. The following year he came to the U.S. After a year as a visiting member of the Courant Institute of Mathematical Sciences at New York University and another year at the Institute for Advanced Study in Princeton he joined the faculty of the University of Rochester. There he remained until 1975, when he returned to Brazil and his native São Paulo.

Established 1845 SCIENTIFIC AMERICAN April 1977 Volume 236 Number 4

# The Importation of Liquefied Natural Gas

*The shortages of the past winter have dramatized the increasing uncertainty of U.S. gas supplies. The importation of liquefied natural gas by ship may ameliorate the situation, but is it safe?*

by Elisabeth Drake and Robert C. Reid

The consumption of energy in the U.S., which reached a peak of 75 quads (quadrillion British thermal units) in 1973 and then declined slightly, is now on the rise again and is expected by the Federal Energy Administration to be at a level of 85 quads in 1980 and 95 quads in 1985. About 30 percent of the energy comes from gas, and the incentives for maintaining that proportion as the demand for energy rises are strong. Yet domestic production of natural gas has fallen off sharply since it reached a peak in 1973 and appears likely to decline even more rapidly in the future unless new sources are found and developed. One way to supplement the dwindling supplies of natural gas is to import liquefied natural gas on a large scale from areas overseas that have an abundance of natural gas and little or no market for it. It is relevant to ask how promising this option is, that is, whether the arguments for the importation of liquefied natural gas outweigh those against it.

The obvious problem in transporting large amounts of natural gas by any means other than a pipeline is that at ordinary temperatures a gas occupies a large volume. If natural gas is cooled to a temperature of −162 degrees Celsius (−259 degrees Fahrenheit), however, it becomes a liquid with a six-hundredth the volume of the gas. The problems of transporting the fuel to places that cannot be reached by pipeline and of storing it in large quantities thereby become manageable. With liquefied natural gas the main problems are to find materials that will hold such a cold substance reliably and to determine what safety measures must be adopted against the possibility that the liquid might escape.

The arguments in favor of the large-scale importation of liquefied natural gas can be summarized quickly, since they are not complex. The U.S. needs additional sources of energy. Gas is particularly desirable as a fuel because it burns cleanly and is easy to distribute to consumers. Gas heating is now installed in many homes, factories and public buildings. Natural gas, much of which is produced as a by-product of the petroleum with which it is associated, is still wasted in many areas by being vented into the air or burned, since there is no local market for it. The technology for liquefying natural gas, for transporting it on the high seas and for storing it on land is in hand. Moreover, that technology is less complex and less expensive than the technology for making synthetic gas out of oil or coal. A considerable international trade in liquefied natural gas already exists: Japan imports about 80 percent of its gas in liquid form, Western Europe 5 percent and the U.S. somewhat less than .1 percent. Indeed, the U.S. is at present a net exporter of liquefied natural gas: exports from Alaska to Japan are about three times greater than imports from Algeria to a terminal in Massachusetts. (In last winter's gas-supply emergency the shipment of liquefied natural gas from Alaska to Massachusetts by way of the Panama Canal was authorized for the first time.) In 1976 the figures were exports of 32 trillion B.t.u.'s to Japan and imports of 10.8 trillion B.t.u.'s to Massachusetts. Put in more familiar terms, the annual imports were 450,000 cubic meters (three million barrels or 10 billion standard cubic feet); in terms of the size of the ships currently fitted out to transport liquefied natural gas the imports amounted to about 12 shiploads.

The arguments against the large-scale importation of liquefied natural gas focus on the question of the safety of the undertaking. Liquefied natural gas, like gasoline and natural gas, can be dangerous if it is handled carelessly or if large amounts are released in an accident against which insufficient safeguards have been provided. Allowance must also be made for the fact that the liquefied gas is at a very low temperature.

A facility for liquefying natural gas and storing the liquid was built in Cleveland in 1941. In 1944 one of the storage tanks failed and a disastrous accident resulted. No dike had been built around the tank to contain all the liquid in the event of a leak, so that the liquefied gas flowed unimpeded into the surrounding area. The cold liquid boiled rapidly, and the vapors soon reached a source of ignition that touched off a huge fire in which 128 people died.

This accident set the fledgling liquefied natural gas industry back many years. Nearly two decades passed before the advantages of this form of the fuel were explored again. The possibility of doing so arose in part from work in the National Aeronautics and Space Administration and the Department of Defense on rocket fuels; that work yielded improved materials for storing cryogenic liquids and new safety precautions for handling them.

In the 1960's several municipal gas


**TANKER FOR LIQUEFIED NATURAL GAS** is under construction at the Quincy Shipbuilding Division of the General Dynamics Corporation. Here a crane lowers one of the five metal spheres that will hold the liquefied natural gas into a space in the hull that has been prepared for it. The sphere has a capacity of 25,000 cubic meters of liquefied natural gas, and the ship will carry five such spheres. The spheres are protected by the white hemispherical structures. The spheres are built separately and shipped to the tanker by barge.

companies designed and built facilities that employed liquefied natural gas for "peak shaving," that is, for supplying fuel at times of unusually high demand. A company would buy, liquefy and store pipeline gas during periods when the demand was low and the price was below normal. Later, during cold spells when the capacity of the pipeline was strained and the price of gas was higher, the liquefied natural gas could be revaporized and fed into the mains serving the company's customers. More than 60 such plants are now operating in the U.S. and Canada.

The peak-shaving plants have proved to be economically successful and have had an excellent safety record. They played an important role in maintaining gas service to residential customers during the unusually cold periods last winter. On this scale, then, the use of liquefied natural gas as a fuel appears to be a sound concept. With the depletion of gas reserves in the region of the 48 contiguous states of the U.S., however, the incentive to import liquefied natural gas for base-load operations is increasing. A base-load terminal would receive liquefied natural gas on a regular basis from a fleet of tankers, store it in large tanks and revaporize it continuously to send it to the mains as fuel.

The tankers projected for operations on this scale would each have a capacity of up to 165,000 cubic meters, which is enough liquefied natural gas to cover a regulation American football field to a depth of about 40 meters (125 feet). The storage tanks at the receiving terminal would be from two to four times the size of the 50,000-cubic-meter structure typical of present peak-shaving facilities. Contemplating these prospects, many people have voiced concern about what would happen if a tanker carrying liquefied natural gas were involved in a collision or an airplane hit a storage tank or an earthquake occurred in an area occupied by a tank or a facility were sabotaged. The question to be faced is whether the safety provisions that can be devised would be sufficient to allow the large-scale importation of liquefied natural gas to proceed with acceptable risks to the public.

One cause for the concern about liquefied natural gas is an accident that took place in 1973 at a storage facility on Staten Island, which is part of New York City. The accident involved a tank that had been used to store liquefied natural gas. From outside to inside the tank consisted of a concrete wall, an insulating layer of polyurethane foam and a container for the liquid. The container was a membrane made of aluminized Mylar plastic film.

After the tank had been in operation for about two years a leak developed in the Mylar film. The tank was shut down and the liquefied natural gas in it was pumped out and vaporized for local consumption. Nitrogen gas was then blown into the tank to warm it and to eliminate any residue of flammable vapor. Eventually air was admitted to the tank and workmen entered to begin repairs.

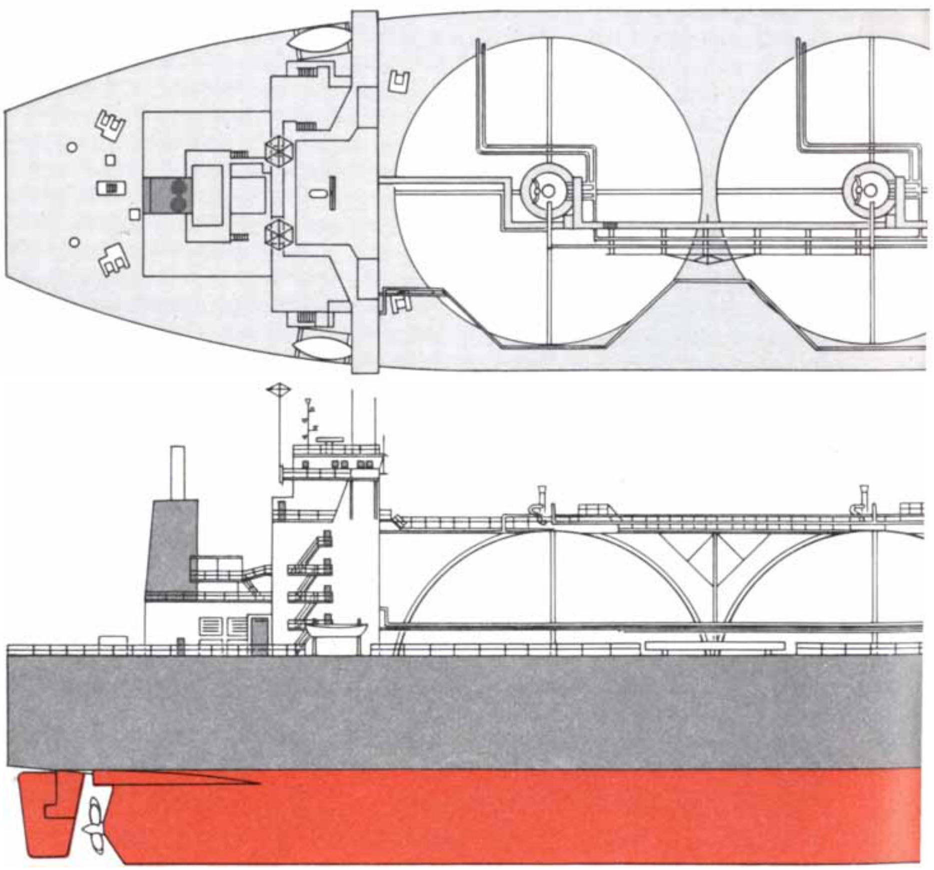
While the workmen were in the tank the insulation somehow caught fire. Hot gases from the combustion raised the internal pressure. The roof rose, as it had been designed to do if excessive pressure developed in the tank; in settling it collapsed into the tank, killing 40 workers. Even though the accident did not involve any liquefied natural gas or endanger the public safety, it aroused local opposition to a terminal that was under construction nearby for the importation of liquefied natural gas. Another source of concern was the fact that the terminal would bring ships carrying liquefied natural gas into waters where marine accidents had occurred. The issue of whether or not the terminal is to be operated remains unresolved.

Most of the questions concerning the safety of liquefied natural gas relate to tankers and storage tanks. We shall discuss some typical designs before returning to the safety question.

The 35 liquefied natural gas tankers now in operation have an average capacity of 46,000 cubic meters of liquid; 41 other tankers that are under construction or in the design stage have a mean capacity of 124,000 cubic meters. As we have mentioned, plans are being laid for vessels that will carry 165,000 cubic meters.

The tanks on a ship carrying liquefied natural gas must have double walls and insulation more than a meter thick to reduce the vaporization rate. Accordingly a collision that might lead to a loss of cargo would have to be considerably severer than it would if the ship were built like a petroleum tanker, where the tank is the hull of the ship itself. Accidents such as ramming and grounding are highly unlikely to cause the release of cargo from a ship carrying liquefied natural gas.

The tanks on a given liquefied natural gas ship have one of three basic designs. The first is freestanding tanks, which are separate from the hull except for support members. The outside of each tank is insulated, and a clearance may be left between the insulation and the hull. The second design is membrane tanks, which receive support from the



**SPECIALIZED SHIP of the type that is under construction by the General Dynamics Corporation is designed to carry 125,000 cubic meters of liquefied natural gas in five spherical tanks.**

hull through load-bearing insulation. The third design is spherical tanks that are built outside the ship and then lowered into prepared spaces during the final stages of its construction. With all three types of tank the primary container for the liquid is made of welded metal. In addition the freestanding tanks and the membrane tanks are required by the U.S. Coast Guard to have a secondary cryogenic barrier. (The requirement does not apply to spherical tanks because they are built under conditions where the welds can be pretested for integrity.)

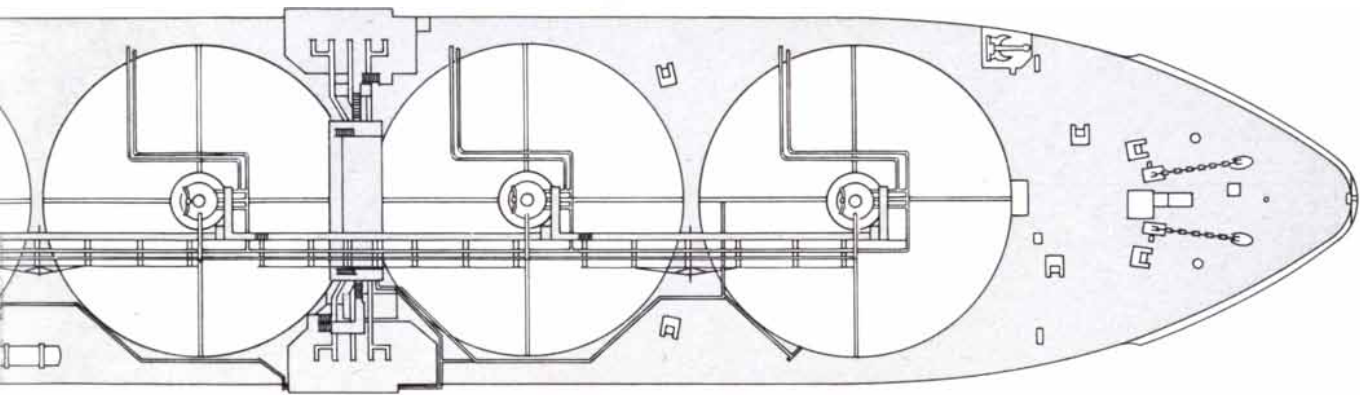
At sea a tanker carrying liquefied natural gas presents an unusual profile. The density of the cargo is less than half the density of seawater, so that the ship rides high in the water. A typical ship has a draft of only 10 meters and a main deck that is from 20 to 25 meters above the waterline. Ballast tanks, which are filled with water, are required to provide stability on return voyages. At such times the tanker nonetheless carries a residual "heel" of liquefied natural gas; part of it is allowed to vaporize slowly to provide fuel for the ship, and the remainder serves to keep the temperature in the tanks low so that they do not have to be cooled before being loaded again. Since it is expensive to build and operate such a large and highly specialized vessel, its operators seek to minimize the amount of time the ship is idle. This approach has led to the development of elaborate and efficient schemes for loading and unloading. Pipes carry liquefied natural gas between a ship and a tank on shore at a rate as high as some four cubic meters per second (50,000 gallons per minute).

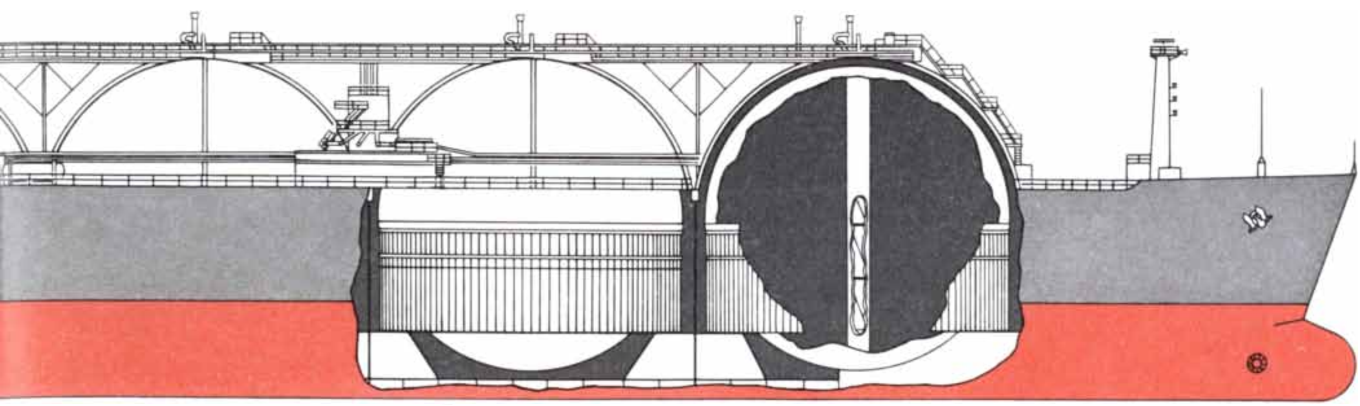
The storage tanks must be built of materials that retain their strength at $-162$ degrees C. Carbon steel, which is the basic material for most steel construction, is not satisfactory because it gets brittle at low temperatures. Certain materials, however, have been extensively tested with liquefied natural gas and found to be suitable for the construction of the inner tank, the one that actually holds the cold liquid. The materials include high-nickel steels, some aluminum alloys and prestressed concrete. The insulation that is a requisite for every tank holding liquefied natural gas is put outside the structure that contains the liquid and is held in place by an outer tank. The outer tank is what an observer sees. The insulation is usually the noncombustible material known as perlite.

Storage tanks for liquefied natural gas at an import terminal are never small. A typical tank has an inner diameter of 55 meters, an outer diameter of 58 meters and a height of 55 meters. It holds some 90,000 cubic meters (550,000 barrels) of liquefied natural gas.

Each tank is surrounded by an impoundment dike that must be capable of containing at least the entire contents of a full tank. In many cases the dike is built at some distance from the tank and is therefore fairly low (from two to 10 meters high). Some recent designs have put the dike close to the tank, partly because in the event of a fire the flames would be confined to a smaller area. Such a dike is likely to be almost as high as the tank itself. In addition to the diking requirement, Federal regulations stipulate that a buffer zone wide enough to prevent another disaster of the Cleveland type must be provided between the dike and the boundary of the facility.

In the early stages of the development of facilities for liquefied natural gas there were plans to store the material in roofed holes in the ground. The liquefied natural gas was expected to create a sealed cavity by freezing the adjacent soil. Experience with such tanks in the U.S. showed, however, that stresses





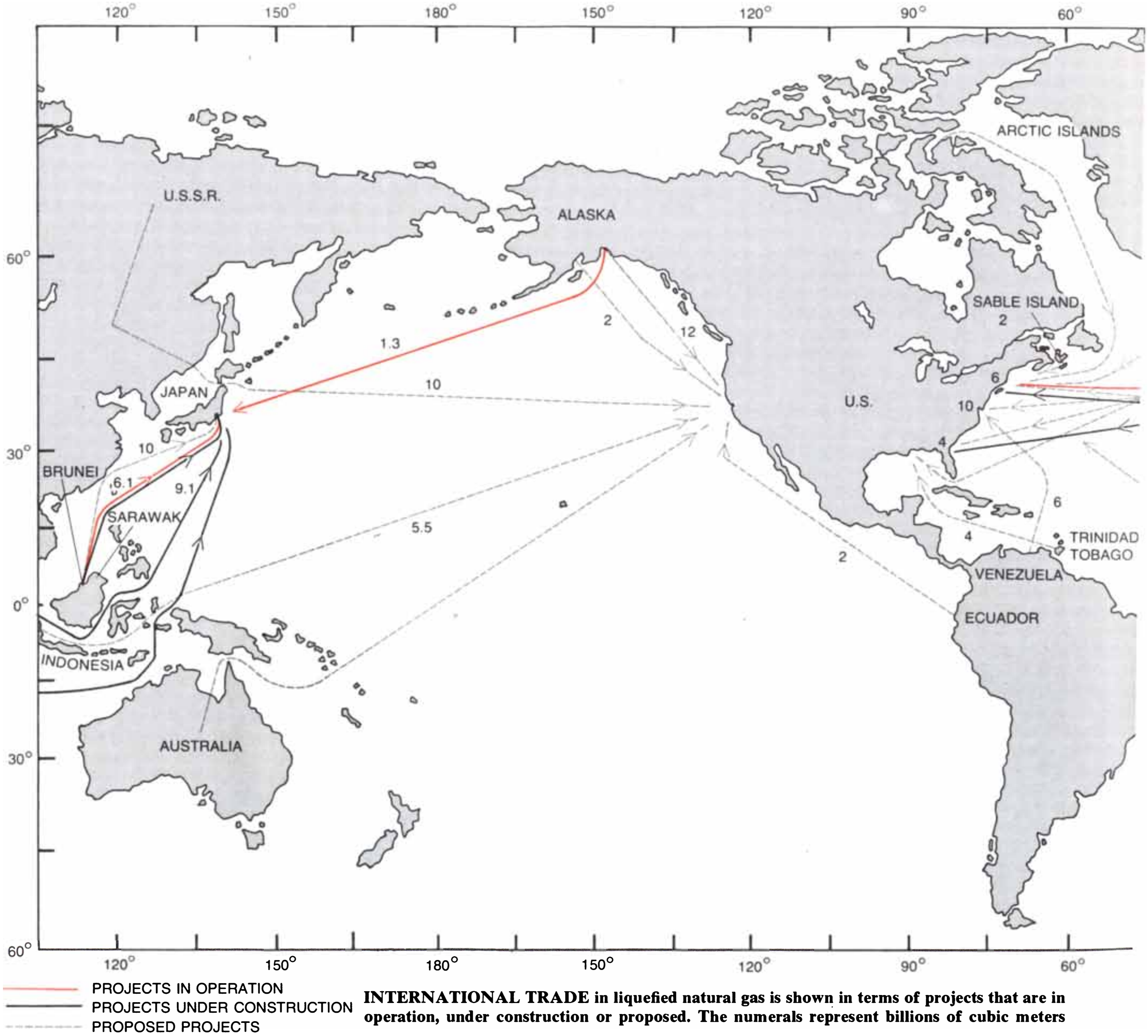
**Here the vessel is depicted in plan and elevation views. Each tank has an inside diameter of 120 feet and is heavily insulated to minimize vaporization of the cold liquid. Vapor that does form is put to use as supplemental fuel for the ship's engines, which usually burn oil.**

build up in the frozen soil and cracks develop. The increase in the amount of contact between the liquefied natural gas and the soil leads to an increased rate of heat transfer into the cavity, and the economics of the facility become marginal. Four such tanks in the U.S. have been abandoned; a few remain in operation in other countries, but they too are being phased out of service.

In Japan, where land is scarce and earthquakes are a severe hazard, a number of buried concrete tanks for liquefied natural gas have recently been built. A tank of this kind typically has a foamed polymer inside the concrete as insulation. The insulation supports a membrane that functions as the seal for the liquefied natural gas.

Liquefied natural gas has a number of interesting characteristics arising from the fact that methane, which is its main constituent, is mixed with small amounts of other compounds such as ethane, propane and nitrogen. One characteristic is the possibility of a "flameless vapor explosion" when liquefied natural gas comes in contact with water. The phenomenon was first reported in 1970 by David S. Burgess and his colleagues at the U.S. Bureau of Mines, who were measuring the boiling rate of liquefied natural gas spilled on water. In the 56th test of an otherwise uneventful series a sharp explosion destroyed the experimental tank containing the water. Later a similar but larger explosion took place after about a quarter of a cubic meter of liquefied natural gas had been spilled on an open pond. In neither case did the material catch fire.

To determine the cause of these explosions Torr Enger and David E. Hartman of the Shell Pipe Line Corporation carried out an extensive program of testing. They concluded that the explosions resulted from the fact that a thin layer of the liquefied natural gas on the water had become so superheated that homogeneous nucleation of vapor occurred very rapidly. A flameless vapor explosion can happen only when the liquefied natural gas incorporates significant fractions of ethane and propane. The range of composition is quite different from that of typical liquefied nat-




**INTERNATIONAL TRADE** in liquefied natural gas is shown in terms of projects that are in operation, under construction or proposed. The numerals represent billions of cubic meters ural gas, which incorporates only small percentages of ethane and propane.

Even when a flameless vapor explosion takes place, the damage is slight. At the Liquid Natural Gas Research Center of the Massachusetts Institute of Technology we measured pressures in the liquid close to the surface where the explosion occurs and found they were well below 690 kilonewtons per square meter (100 pounds per square inch). Since the pressure wave would attenuate rapidly with distance, if such an explosion took place on water adjacent to a tanker, it would have little effect on the hull of the ship. Pressures in the air above the explosion are of course even lower.

A second interesting effect is that mixtures of liquefied natural gas with differing composition and density can stratify. This phenomenon was first recognized in 1971 after an incident at an import terminal in Italy. About 18 hours after a cargo of liquefied natural gas had been put into a storage tank the pressure in the tank began to rise. Soon the safety valves were actuated, and in the course of an hour some 150,000 kilograms of natural gas were vented. Fortunately the vented gas did not ignite and the pressure in the tank did not go high enough to cause any serious damage.

An analysis of the incident revealed that the new load of liquefied natural gas had a higher density than the material that was already in the tank. It was also warmer. Because of the difference in density the new material, which had been loaded into the bottom of the tank, remained there, creating a two-layer system. Over the following 18 hours heat flowed into the tank through the bottom and the side walls, since even though the liquefied natural gas in the bottom layer was warmer than that in the top layer, it was still much colder than the surroundings. Such a flow of heat is normal, and in an unstratified storage tank it drives a convective circulation of liquefied natural gas that maintains the entire contents of the tank well mixed and at essentially the same temperature.

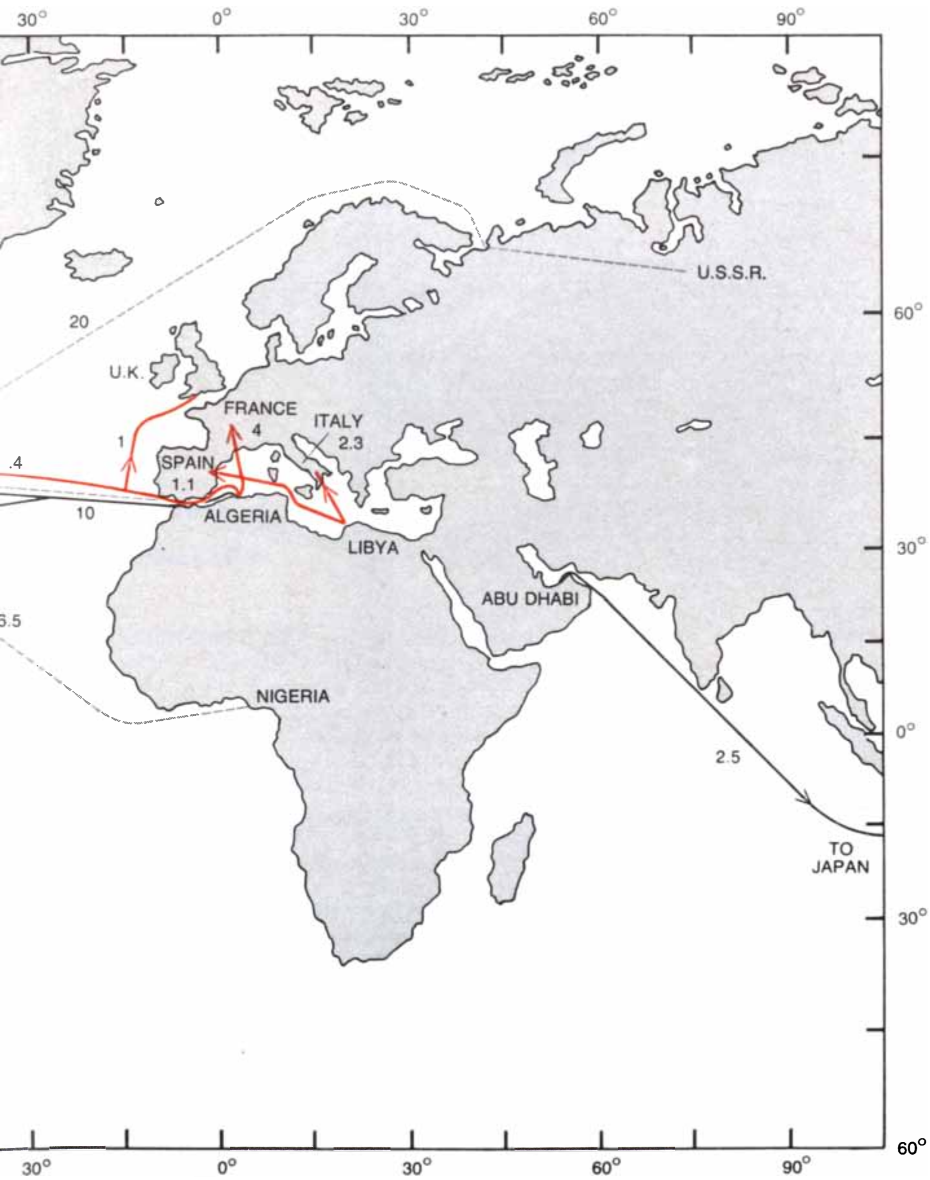
In a two-layer system the liquid in each layer still flows convectively, but the denser fluid is not sufficiently buoyant to enable it to penetrate the less dense upper layer. Thus energy can be stored in the denser layer. In time the transfer of heat and mass between the layers tends to equalize the difference in density. Then the layers may mix rapidly. As the material attempts to achieve thermal equilibrium, the intrusion of the warm bottom layer leads to a rapid generation of vapor to release the excess energy that was accumulated while there were two layers. The contents of the tank have suddenly mixed. The phenomenon is commonly called rollover.

Modern facilities have standard procedures for preventing rollover. The less dense liquid may be loaded into the bottom of the tank and the denser into the top. This procedure tends to prevent stratification. Another option is to always load the tank from the top; then, even if stratification occurs, the time required for rollover is much longer than the residence time of liquefied natural gas in a storage tank at a typical import terminal.

Problems other than flameless vapor explosions and rollover have received considerable attention. If liquefied natural gas is spilled, it boils rapidly. Although the vapor is not toxic, it may in high concentration cause asphyxiation by excluding oxygen. Moreover, the low temperature of the material may result in frostbite for anyone in the immediate vicinity of a spill. Both of these potential hazards are too localized to be of concern outside the boundaries of the facility.

The major hazard of liquefied natural gas is fire. If a spill is ignited, it burns much like a pool of gasoline. If it is not ignited quickly, the flammable vapor may be carried by the wind until a source of ignition is encountered. Experiments in the U.S., France and Japan have shown that once the vapor has been ignited a flame front burns back through the vapor toward the source from which the vapor came.

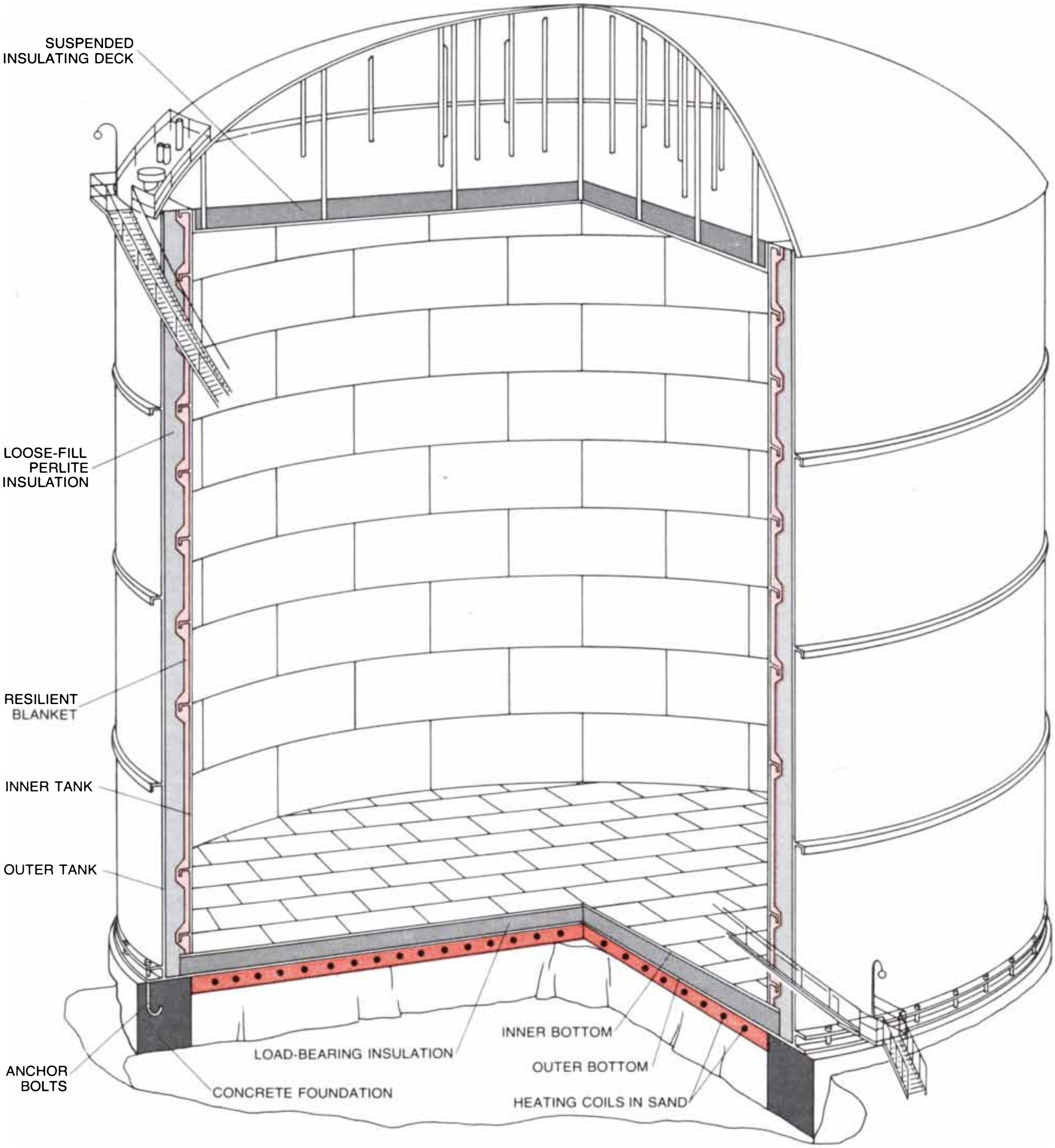
Another serious question is whether a




**of natural gas before liquefaction; the amount of liquefied natural gas moving in international trade or planned to be moved is about a six-hundredth the volume of the gaseous material.**

fire in a cloud of vaporized liquefied natural gas might develop into a detonation capable of producing a damaging blast wave. The consensus among workers who have studied the problem is that the probability of such an explosion in an unconfined space is small. The detonation of a vapor in open air has not been confirmed even when a high-explosive charge served as the initiator. Charles D. Lind of the Naval Weapons Center at China Lake, Calif., has been unable to obtain detonations even with charges of up to two kilograms.

Lind carried out his tests with mixtures of methane and air enclosed in thin polyethylene hemispheres up to 20 meters in diameter. Ignition sources or explosive charges were placed in the center at ground level, and the results were followed with high-speed photography. The polyethylene enclosure was destroyed by either the detonation of the high explosive or the resulting blast wave, but in no case did the combustion wave of the methane accelerate to become a detonation. Explosions of methane-air mixtures have been demonstrated in rigid enclosures, but normally the combustion wave must travel 30 meters




**STORAGE TANK designed to hold liquefied natural gas after it has been delivered by tanker to an import terminal has a capacity of 90,000 cubic meters. The tank is well insulated so that the liquefied natural gas, which is stored at slightly more than atmospheric pressure, does not vaporize too rapidly. Heating coils under tank serve to prevent frost heaves caused by temperature variations in the ground.**

or more before it can become a self-supporting detonation wave.

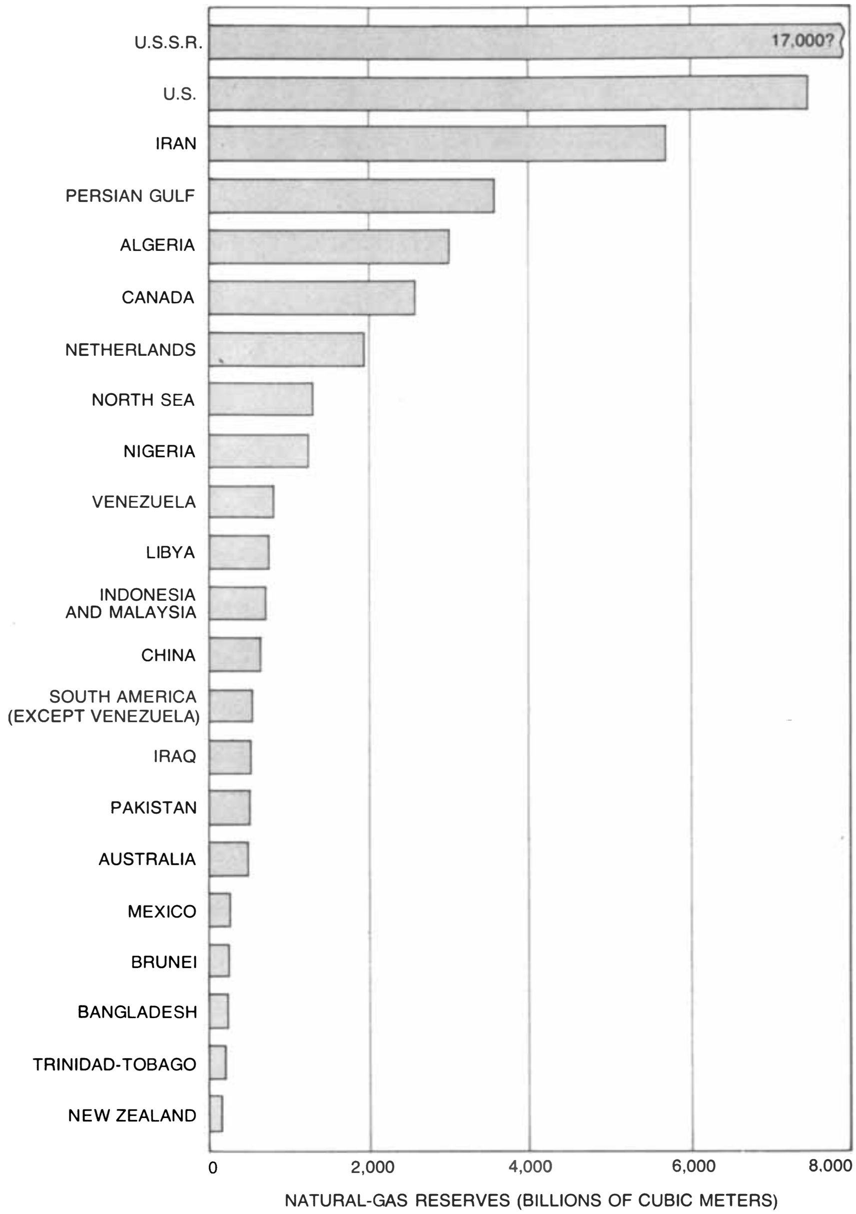
Liquefied natural gas inside a storage tank is innocuous; it cannot burn unless it is vaporized and mixed with air. Only if it is released can there be a hazard. Accordingly tankers and storage facilities for liquefied natural gas are designed and operated to prevent accidental release.

Typically tanks for liquefied natural gas are designed to meet standards that specify the severity of wind, earthquakes and so on that the tank must withstand. The standard usually is the severest condition experienced within the past 50 or 100 years. Safety factors, which are also laid down in building codes, ensure that a tank will be able to withstand an even severer event. Moreover, in a storage facility it must be demonstrated that the public would not be endangered if the largest pipe connection to the tank should fail at the maximum rate of flow of liquefied natural gas.

Similar analyses of hazard are made for a liquefied natural gas tanker entering a port. Data are collected on the traffic density and the type, speed and length of other vessels in nearby waters. Since the probability of collision should approach zero if the rules of the road are followed, analysts frequently make the conservative assumption that all ships move in a random manner. In any event, not all collisions would be severe enough to penetrate the double hull of a ship carrying liquefied natural gas deep enough to release any of the cargo. Therefore the analysis is carried further to estimate the minimum momentum and angle of impact for a striking ship to cause a major release of liquefied natural gas.

Only after careful analyses can one estimate the potential hazards to the public of an operation involving liquefied natural gas. As an example, the risk associated with the proposed importation of liquefied natural gas to Staten Island is estimated to be about one fatality every 10 million years for people living or working along the approach route to the harbor. That level of risk is about 10 times less than the risk of dying from a fire at home and about the same as the risk of being struck by lightning.

Nevertheless, even though a facility can be designed to provide a high degree of safety in the event of the kind of accident envisioned in the design standards, one may still postulate severer accidents resulting, say, from a massive earthquake or from a sophisticated act of sabotage. The sudden destruction of a tank containing liquefied natural gas would be expected to lead to a major fire with potentially disastrous consequences. Can society tolerate such a risk?




**GAS RESOURCES OF THE WORLD are listed. The figures are for gaseous natural gas. Much of it is unlikely to be liquefied and exported, but many of the countries have little or no domestic market for natural gas and are therefore likely to consider exports of liquefied gas.**

The general problem is not unlike the problems faced daily by industrial engineers and governmental regulatory agencies. For example, should a big jet airplane be allowed to fly over a crowded public facility such as an athletic stadium? There is a low but always finite possibility of a crash with catastrophic results.

Similar examples reveal that society tolerates certain highly improbable risks—risks that cannot feasibly be eliminated by design—in order to reap the benefits of energy, consumer products, transportation, central water-supply systems and so on. On the basis of the studies that have been made of the hazards of liquefied natural gas and the experience that has been gained in several countries in handling the fuel on a large scale, it seems reasonable to say that the hazards are similar to those involved in handling other fuels such as gasoline, propane (liquefied petroleum gas) and gaseous natural gas. Liquefied natural gas therefore appears to be a promising alternative source of energy. Moreover, it can be expected that further experience in handling liquefied natural gas and further research by the industry and by regulatory agencies will raise the level of safety in its use even higher.

# The Collision between India and Eurasia

*For the past 40 million years the Indian subcontinent has been pushing northward against the Eurasian land mass, giving rise to the severest earthquakes and the most diverse land forms known*

by Peter Molnar and Paul Tapponnier

The region of the earth that exhibits the greatest diversity of geology, topography and climate, together with a strong susceptibility to major earthquakes, is the part of Eurasia that lies east of the Ural Mountains and north of the Ganges, embracing northern India, Pakistan, Afghanistan, the Tibetan plateau, Mongolia, most of China and a large part of the eastern U.S.S.R. This profoundly disarranged region is roughly equal in area to all of North America from the Rio Grande to 60 degrees north latitude, the parallel that crosses the northern tip of Labrador on the east and Cook Inlet in Alaska on the west. The world's highest mountains, the Himalayas, rise abruptly from the flat, densely populated Ganges Plain in northern India and shield the Tibetan plateau from the seasonally shifting monsoon winds of southern Asia. Tibet supports a population of less than 1.5 million in an area somewhat larger than Texas, Oklahoma and New Mexico combined. Tibet's average elevation is 5,000 meters, higher than any point in the 48 contiguous states of the U.S. In contrast, eastern China has abundant rainfall and supports a population of close to a billion. (As long ago as 1556 a single earthquake near Sian, then the capital of China, is known to have killed 830,000 people.) The broad, high mountainous zone that includes the Himalayas and the Tibetan plateau presents such a barrier to travel that throughout history the populations of India and China have had remarkably little contact. North of Tibet the Gobi Desert presents another formidable barrier to migration and communication. The Tarim Basin, part of the Gobi Desert, is one of the driest and most inhospitable regions on the earth. Mountains as high as 5,000 meters surround it on three sides. Nearly constant winds blow across the basin with such force that they pile up sand dunes as much as 150 kilometers long with wavelengths of three to five kilometers, clearly visible in satellite pictures. North of the Gobi Desert, Lake Baikal, at 1,800 meters the deepest lake in the world, fills the Baikal rift zone, a huge crack in the earth's crust similar to the one in East Africa.

Although the geology of Asia seems to present a chaotic jumble of land forms, much of the deformation of the surface, when it is viewed as a whole with the help of satellite photographs, seems to fall into a simple, coherent pattern attributable to a single cause: a geological collision between the Indian subcontinent and the rest of Eurasia. This collision is still in progress. As an example of its current effects, we believe the great earthquake that devastated the industrial city of Tang Shan last summer was the result of forces originating in the collision area 2,500 kilometers to the southwest.

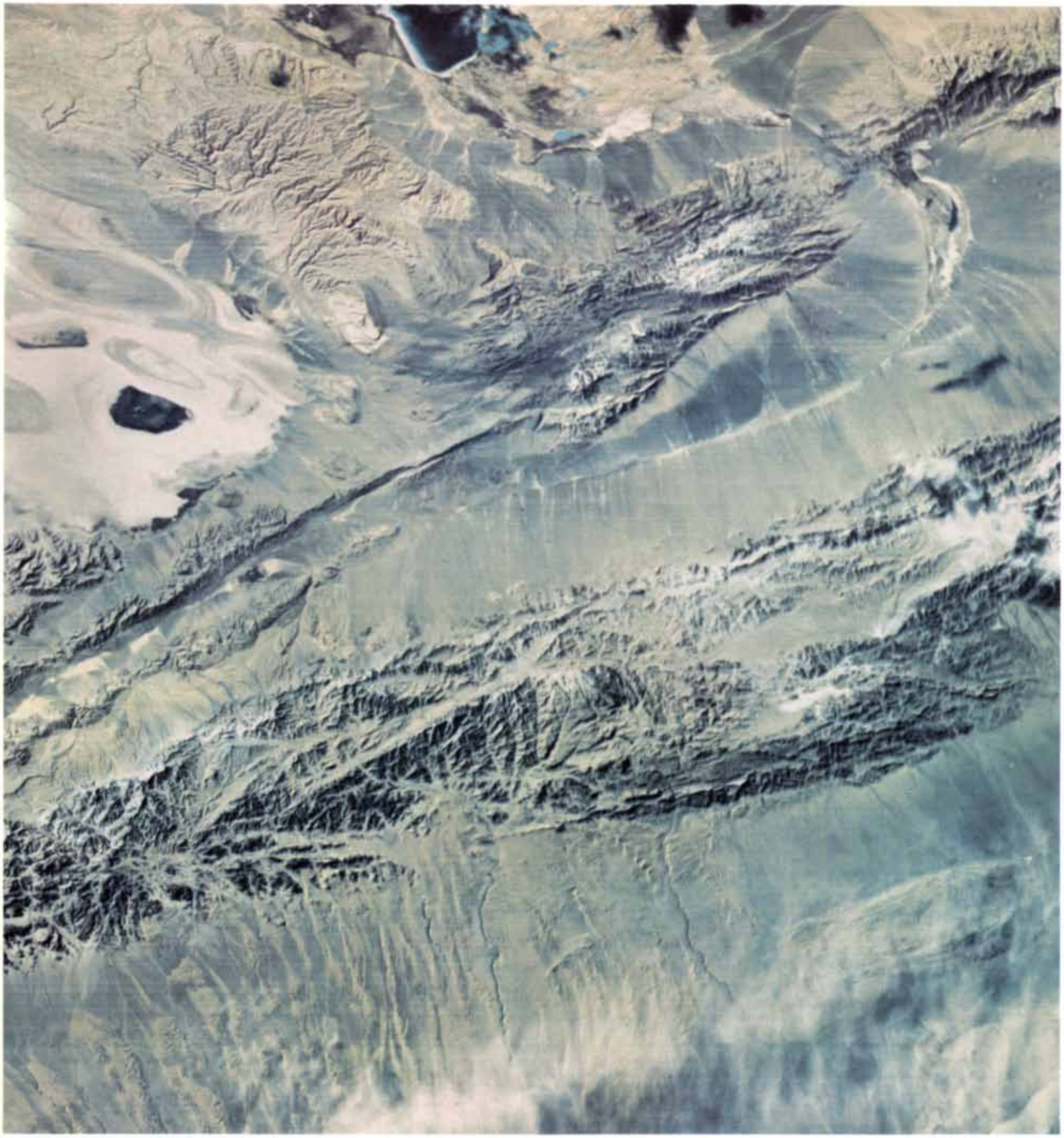
The possibility of quantitatively analyzing the collision between India and Eurasia has developed only in the past 10 years. The hypothesis of sea-floor spreading was first advanced in the mid-1960's to explain the mid-ocean ridge, the great mountain range on the ocean floor that runs for 40,000 kilometers across the world's ocean basins. At a rift in the crest of the ridge molten rock wells up from the mantle below and fills the gap that is left as the ocean floor on each side of the rift moves outward. It was also recognized that the ocean floor is magnetized in stripes of opposite polarity, depending on the magnetic polarity of the earth itself at the time the molten rock crystallized. (For reasons still not known the earth's magnetic polarity reverses at intervals of several hundred thousand years.) The study of magnetic reversals in the floors of the Indian and Atlantic oceans showed that India had traveled 5,000 kilometers with respect to Eurasia before the two collided. Since the beginning of the collision India has continued to move northward another 2,000 kilometers with respect to Eurasia. How can one account for the vast area of land that was displaced by the collision? We shall describe our hypothesis to account for the displacement and present the evidence supporting it.

The recognition of sea-floor spreading, which confirmed earlier conjectures of continental drift, quickly led to the much broader concept of plate tectonics, a concept that has inspired the renaissance in the earth sciences. Plate tectonics provides a physically simple mechanism for large-scale horizontal motions of separate portions of the earth's crust and makes it possible to be quantitatively precise in describing the kinematics of continental drift. One of the central concepts of plate tectonics is that a small number of large plates of the lithosphere, the high-strength outer shell of the earth, move rigidly with respect to one another at rates of one centimeter to 20 centimeters per year over the hotter, low-strength asthenosphere under them. Some 100 kilometers thick, this outer shell consists of the earth's crust and the upper part of the mantle. Thus in plate tectonics the earth's crust, which is both lighter than and chemically different from the mantle under it, is visualized as being carried passively as part of a lithospheric plate.

It is ironic that Alfred Wegener, who was probably the most important early proponent of continental drift, based much of his argument on the difference between the continental crust and the oceanic crust. The continental crust, which stands high above the oceanic crust, has deep roots going down about 35 kilometers, whereas under the oceans the crust is only about six kilometers thick. Wegener viewed the continents as sturdy ships sailing majestically through the much weaker crust and mantle under the oceans. In actual fact, as is recognized in plate tectonics, the oceanic crust and upper mantle—the lithosphere—are extremely strong. They seem to be deformed only at the boundaries of plates, so that the relative motion of the plates can be described as the motion of rigid bodies.

The continents usually act as rigid structures when they lie entirely within one plate. When a boundary between two plates passes through a continent, however, it is generally more diffuse than the boundary between two oceanic plates or the boundary between an oceanic plate and a continental one, and it is accordingly much more difficult to define. Such diffuseness is apparent throughout the Mediterranean area, and it is particularly evident in Asia, where the India and Eurasia plates are converging.

Before proceeding further we should like to emphasize the important role played in plate tectonics by the difference between continents and oceans. Oceanic crust is formed as a part of the lithosphere at spreading ocean ridges. The oceanic lithosphere cools, shrinks and gets denser as it moves away from



**PORTION OF ALTYN TAGH FAULT, photographed from an altitude of 950 kilometers by the Earth Resources Technology Satellite (ERTS) in February, 1973, is some 1,200 kilometers north of Mount Everest in the Sinkiang region of China. The fault, which slants to the upper right corner from the lower left, may be the greatest active continental strike-slip fault in the world. The term strike-slip means that the two sides of the fault are sliding past each other in opposite directions. The southern block of the Altyn Tagh fault is evidently being driven to the right, or eastward, with respect to the northern block as a consequence of the collision of India with the southern margin of Eurasia. The collision began some 40 million years ago and is continuing. About 180 kilometers of fault appears here. Entire fault can be traced for more than 2,500 kilometers if one includes Kansu fault, with which it merges at its eastern end (*see map on pages 36 and 37*). This ERTS photograph and those on pages 39 and 41 are reproduced with south at the top, so that topographical features are illuminated from the top. Otherwise shadows would point upward and the relief would tend to reverse so that the valleys would appear to be ridges.**

the spreading center. An amount of crust equivalent to the amount added at a spreading ridge must return continuously to the mantle. This happens at subduction zones, where the oceanic lithosphere plunges into the asthenosphere, pulled down because the descending lithosphere is cooler and therefore slightly denser than the asthenosphere surrounding it. Although the oceanic crust is about 15 percent less dense than the mantle, it is sufficiently thin to be carried down as part of the lithosphere. This appears to happen in many places, typically forming deep trenches adjacent to island arcs such as those rimming the western Pacific.

Consider, however, what happens when a continent riding on the downgoing slab of lithosphere tries to plunge into the asthenosphere, traveling behind the consumed oceanic lithosphere. The continental crust, some 35 kilometers thick and almost 20 percent less dense than the mantle, is both too thick and too light to be carried down into the asthenosphere. Thus the buoyancy of the continental crust keeps it from being subducted and probably explains the fact that whereas the oceanic crust recycles itself about every 200 million years, large areas of the continents have existed for billions of years.

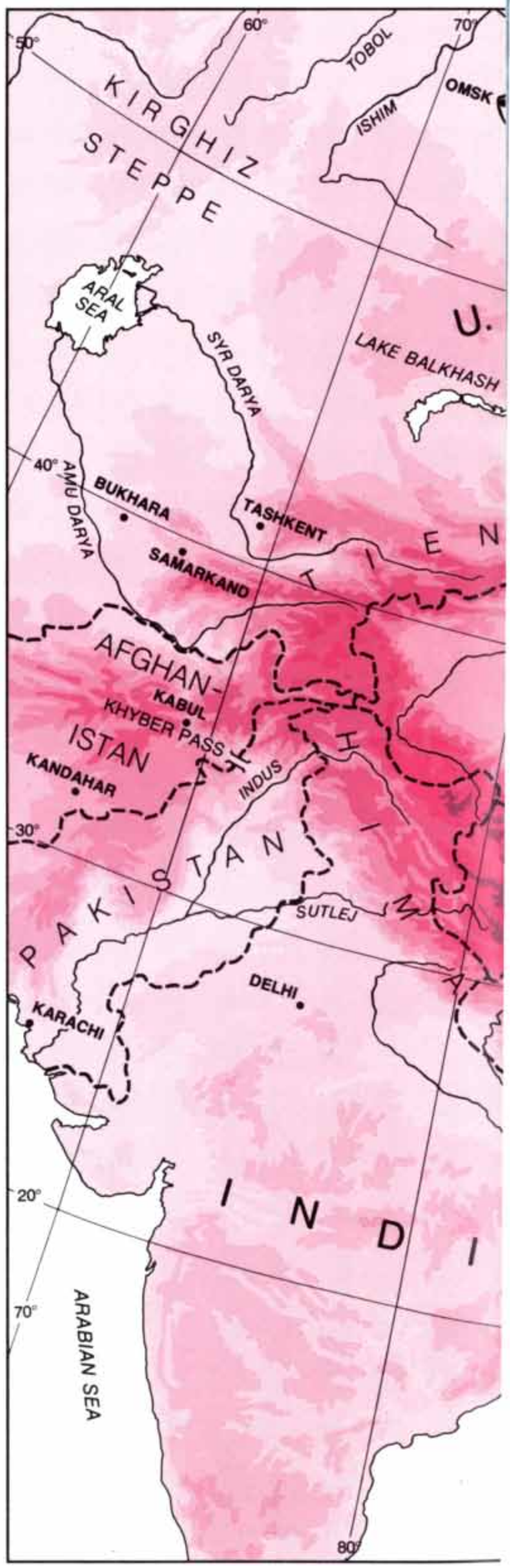
When two continents collide, they suture themselves together to form a larger continent. Either the relative motion of the two plates on which the continents ride will change or a new boundary will form between the plates at a different place. Geologic history provides several examples. Some 250 million years ago Europe and Siberia were sutured at the Ural Mountains, and at about the same time North America and Africa were sutured at the southern Appalachians, before the present-day Atlantic Ocean formed. When one looks at a geological map of Asia, several ancient sutures are apparent. A few years ago Peter N. Kropotkin, the first well-known Russian geologist to accept continental drift (and the great-nephew of the famous anarchist of the same name), described Eurasia as a "composite continent" to which fragments have been successively sutured by collision over the past 800 million years. Most of the sutures in Eurasia appear to be older than 200 million years; only the suture between the Indian subcontinent and the rest of Eurasia appears to be younger. Here we shall consider only this most recent collision and its consequences.

Although we argue that typical plate tectonics does not seem to apply to the portion of Asia that lies between India and Siberia and is tectonically active today, plate tectonics can be used to place constraints on the history of the relative motion of the India and Eurasia plates. We can exploit the fact that if we know the history of the relative motion of plate *a* and any two other plates, *b* and *c*, we can calculate the history of the relative motion of *b* and *c*. From geological studies in the North Atlantic by Jean Francheteau of the Centre Océanologique de Bretagne in France and by Walter C. Pitman III and Manik Talwani of the Lamont-Doherty Geological Observatory of Columbia University we know how both Eurasia and Africa moved with respect to North America, and thus we can calculate how they moved with respect to each other. Similarly, from the work of Robert L. Fisher of the Scripps Institution of Oceanography, D. P. McKenzie of the University of Cambridge and John G. Sclater of the Massachusetts Institute of Technology we know how India and Africa moved with respect to each other and therefore we can calculate where India was with respect to Eurasia at different times in the past.

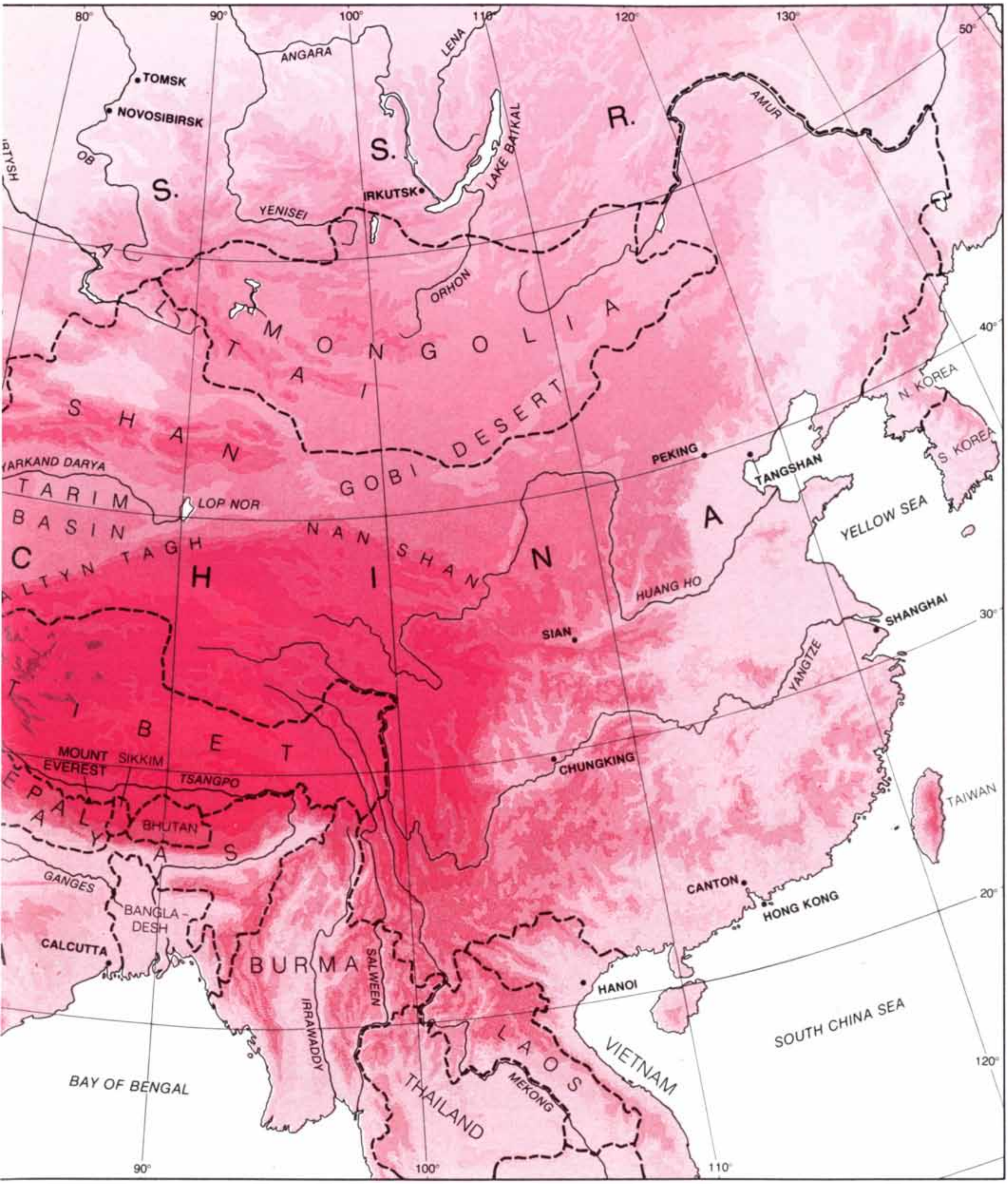
We do not know where the northern margin of the original Indian continent lies with respect to India today, because that margin has been much deformed in the creation of the Himalayas. For the purposes of calculating how much the India and Eurasia plates moved with respect to each other in different intervals, however, our ignorance on this point does not matter. Moreover, although we do not know how far north of India the old margin lies, we can determine from the geology of Asia the position of the boundary between rocks of the present Indian subcontinent and those that were part of Eurasia long before the collision.

The primary evidence used to delineate a suture between two continents is the existence of the sequence of rocks known as an ophiolite suite. Ophiolites have three distinguishing characteristics. They show a particular sedimentary sequence incorporating bedded cherts, which are characteristic of deep-ocean sedimentary deposition. They contain remnants of pillow basalts: igneous rocks of lumpy form that are typical of basaltic lava extruded under water, as at spreading ocean ridges. And they incorporate dense, dark rocks low in silicon dioxide known as ultramafics, which are thought to be typical of the mantle. Ophiolites are interpreted as being a slice of the oceanic crust and upper mantle; accordingly their presence implies the former existence of an ocean basin.

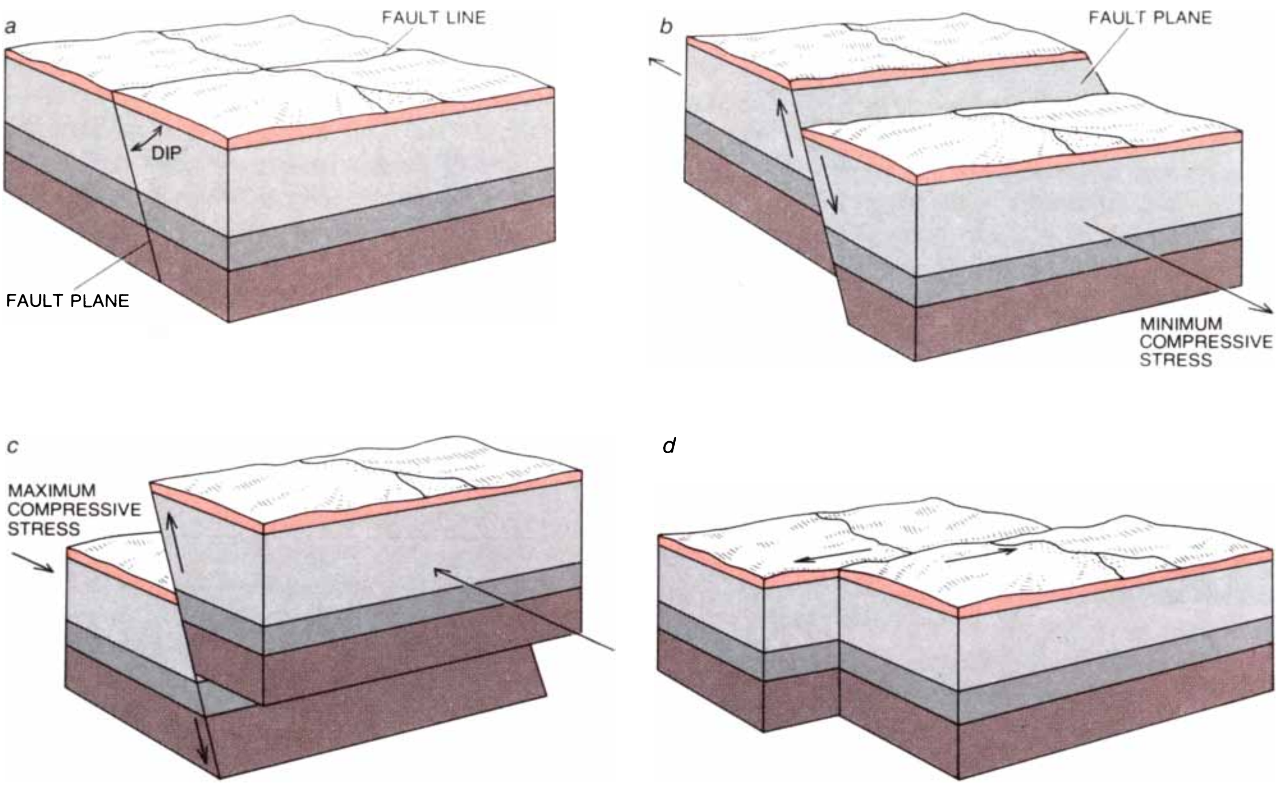
A belt of ophiolites follows the Indus and Tsangpo valleys in southern Tibet north of the Himalayas; it appears to mark the boundary between the sutured continents. Somewhat farther north there are volcanic rocks typical of those found at subduction zones, as in the Andes of South America. Just as oceanic




**COLLISION between India and Eurasia appears to have compressed and distorted the earth's crust from the Himalayas to Siberia and from Afghanistan to the coast of China. Immediately north of the Himalayas, Tibet has been raised to an average elevation of 5,000 meters. The Tibetan plateau is drained**




**to the east by a series of great rivers that arise in nearly parallel channels but eventually fan out into rich deltas along an arc extending from the Bay of Bengal to the Yellow Sea. To make the geography and distances more familiar one can imagine a map of North America traced on top of this one so that the Rio Grande is roughly aligned with the Ganges, which would place Brownsville, Tex., close to Calcutta. (Actually Calcutta is about four degrees farther south than Brownsville.) Los Angeles would then coincide approximately with Kandahar in Afghanistan, Miami would be near Hanoi, and Portland, Me., would be close to Peking. The Himalayas would sweep in a great arc from Nevada-Utah border to Mississippi, crossing Arizona, New Mexico, Texas and Louisiana. Tibetan plateau would include Colorado, Nebraska, Kansas, Oklahoma, Iowa and Missouri. Far to the north Lake Baikal would cut across northernmost tip of Quebec.**




**DURING AN EARTHQUAKE one block of the earth's crust slips with respect to an adjacent block along a fault plane. In a normal fault (*b*) the blocks act as if they were being pulled apart. The overlying block slides down the dip of the fault plane. In a thrust fault (*c*) the overlying block is forced up the dip of the fault plane because the maximum compressive stress is horizontal and perpendicular. In a strike-slip fault (*d*) the two blocks slide past each other. Lateral slippage can be combined with normal or thrust faulting. From an analysis of the seismic waves generated by an earthquake, called a fault-plane solution, one can tell what kind of fault motion has occurred.**

crust now plunges to the east under the Andes, presumably the old ocean floor between India and Eurasia plunged to the north under Tibet. In contrast the Himalayas, south of the suture, consist of slices of the old northern portion of India that have been stacked one on top of another to form the mountains. Patrick LeFort of the Centre de Recherches Pétrographiques et Géochimiques in France and Maurice Mattauer of the University of Montpellier have shown that there is a progression from north to south in the piling up of such slices, so that the oldest thrust is to the north. The next step will probably be for a new fault to form farther south on the Ganges Plain and for material to the north of it to slide up and over the plain to the south.

Although the geology of the Himalayas and Tibet is not well enough known to enable us to reconstruct the position of the northern margin of the original Indian continent, it does place important constraints on when India and Eurasia could have collided. Four observations imply a date of between 40 and 60 million years ago. First, from geological investigation of the ophiolites in the area, Augusto Gansser of the Swiss Federal Institute of Technology found "exotic blocks" of the late Cretaceous (about 70 million years ago) within the suite. Thus the ophiolites must have been part of an ocean floor until after that time.

Second, Gansser describes a sequence of sedimentary rocks on the northern edge of the Himalayas that is typical of sequences found on continental shelves and slopes and that begins in the Cambrian (about 500 million years ago) and continues until the early Eocene (about 55 million years ago). Hence there appears to have been an ocean between the converging continents until approximately that time.

Third, there is no known fossil record of mammals in India before about 50 million years ago. Ashok Sahni and Vimal Kumar of Lucknow University in India report that the oldest mammals in India date from the middle Eocene (about 45 million years ago). The first Indian mammals are similar to those found in Mongolia. Thus although mammals had evolved on other continents, it appears that they did not evolve independently in India, indicating that the continent remained isolated until about 45 million years ago. The collision enabled a horde of Mongolian mammals to sweep into India. Fourth, Gansser infers that major mountain building in the Himalayas began in the Oligocene (about 35 million years ago).

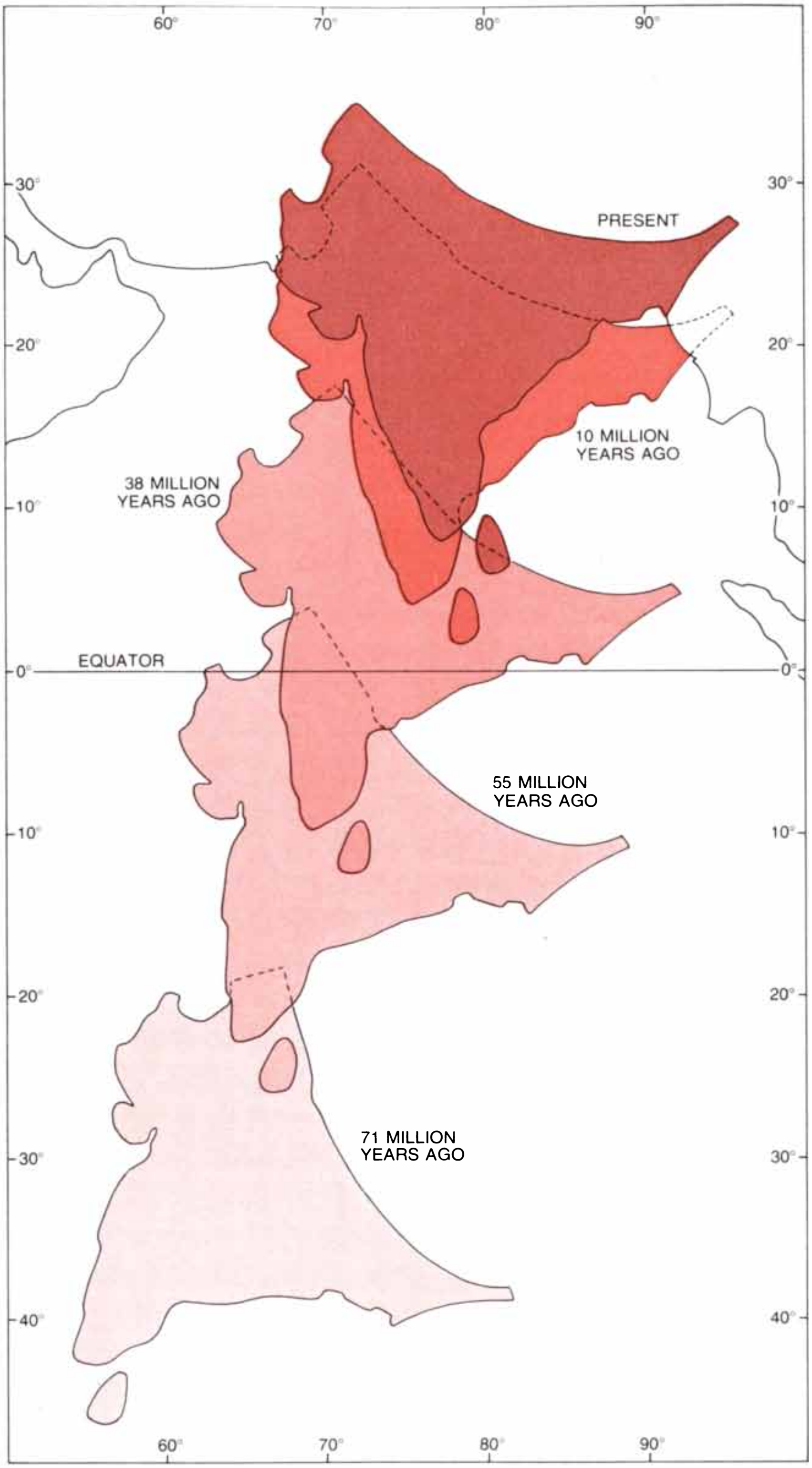
These observations do not enable us to determine precisely when India and Eurasia collided, but we think the event probably occurred 45 million years ago, give or take a few million years. It is highly improbable, however, that the old margins of India and Eurasia met flush along their full length. It is more likely that they first made contact when peninsulas met and that with the passage of time the zone of contact grew until the ocean basin between the continents was swallowed up. Thus we consider it likely that the initial contact may have been a few million years earlier than intimate contact.

From the history of sea-floor spreading in the Indian Ocean one can calculate the relative positions of India and Eurasia over some tens of millions of years. One can see immediately that between 70 and 40 million years ago India moved about twice as far with respect to Eurasia as it has since then. A plot of the distance of the northeast and northwest corners of India at different times in the past from their present positions shows that the rate of convergence between them changed by a factor of two about 40 million years ago. Given the uncertainties in the data the change in rate could have come a few million years later or as much as 10 million years earlier. In any event we interpret the change in the rate at which India approached its present position as an indication that the first stages of collision came at about the same time as the change in rate, and that the buoyant continental crust of India, instead of being subducted, put a brake on the northward motion of the India plate.

Although the reconstructions provide support for the view that continental crust cannot be subducted, they leave us with what may be a more difficult problem. If we conclude that India and Eurasia collided 40 million years ago, we must also conclude that since then India has traveled northward about 2,000 kilometers with respect to Eurasia. If the continents collided earlier, the distance covered is even greater. Bearing in mind that continental crust cannot be subducted, we are faced with the problem of accounting for the displacement of a piece of crust that has an area the width of India and is 2,000 kilometers long.

The continuing northward motion of India at a rate of about five centimeters per year is probably responsible for the widespread tectonic activity in Asia. For example, seismic activity is detected over an area extending some 3,000 kilometers north and east of the Himalayas. Among the 22 greatest earthquakes listed by Beno Gutenberg and Charles F. Richter of the California Institute of Technology for the period 1897–1955, seven occurred in central and eastern Asia, four of them north of the Himalayas. The deformation of the surface of the earth that accompanied some of these earthquakes was huge. The 1957 Gobi-Altai earthquake in Mongolia, which came after the Gutenberg-Richter study and would probably have been too small to have been included, caused displacements of as much as 10 meters along the main fault associated with the earthquake. Such large displacements along faults seem to have been characteristic of several of the great earthquakes in Asia. In any case it is clear that the region as a whole does not act as a rigid plate.

Asia is known for high mountains not only in the Himalayas and Tibet but also farther north and east, where in the Tien Shan and Nan Shan ranges there are peaks of up to 6,000 meters. Ordinarily high mountains are rapidly worn down by erosion, so that their existence implies large crustal movements in recent geologic times. Studies conducted by Russian geologists (V. N. Krestnikov, A. V. Goryachev, S. A. Zakharov and others) show that the area of the Tien Shan range was nearly flat from 200 million years ago to 30 or 40 million years ago, and that it has been elevated since then. Although other reports we have seen are less definitive, they do suggest that the relief in Mongolia and China




**INDIA'S NORTHWARD DRIFT has been reconstructed from magnetic reversals in the floors of the Indian and Atlantic oceans. As molten rock welled up into the rift in the ocean floor and hardened it became magnetized according to the prevailing polarity of the earth's magnetic field. At infrequent and irregular intervals the earth's polarity changes, leaving a record that can be dated. This "time lapse" reconstruction shows that India traveled some 5,000 kilometers northward with respect to Eurasia in the 20 to 30 million years before its collision with Eurasia. Over the past 70 million years the northeastern tip has actually traveled some 7,000 kilometers. Velocity for continent as a whole was about 10 centimeters per year for the first 30 million years and about five centimeters per year for the next 40 million. In this reconstruction it is arbitrarily assumed that the boundary of Eurasia is fixed in its present location.**

today is comparably new. Hence the deformation of a large part of the crust of Asia appears to have begun after the collision with India. We infer that the penetration of India into Eurasia caused the deformation.

In trying to explain how India could have traveled 2,000 kilometers after it began to collide with Eurasia we obviously have greater freedom in accounting for the material displaced if we imagine its being absorbed over several million square kilometers rather than in a narrow zone near the suture. The formation of the Himalayas can account for only a fraction of the material displaced, so that it remains necessary to explain where the rest of it has gone. One method that has proved to be particularly useful for deciding how material is displaced in oceanic regions at present is the determination of "fault-plane solutions" of earthquakes. Such solutions reveal the direction of relative movement along the fault. From the study of the waves radiated by an earthquake one can determine both the type of faulting that took place during the earthquake and the direction of motion of one side of the fault with respect to the other.

Geologists classify faults in three main categories. A normal fault results when stresses directed horizontally cause one side of the fault to sink with respect to the other. A thrust fault is produced when compressive stresses drive one side of a fault over the other side. In a strike-slip fault the two sides of a fault slide horizontally past each other. If the opposite side of the fault moves to the left, as viewed from either side, the movement is termed left-lateral displacement. Movement in the opposite sense is right-lateral displacement. Working with Thomas J. Fitch, who is now at the Lincoln Laboratory of M.I.T., and Francis T. Wu of the State University of New York at Binghamton, we have compiled fault-plane solutions for some 75 earthquakes in Asia.

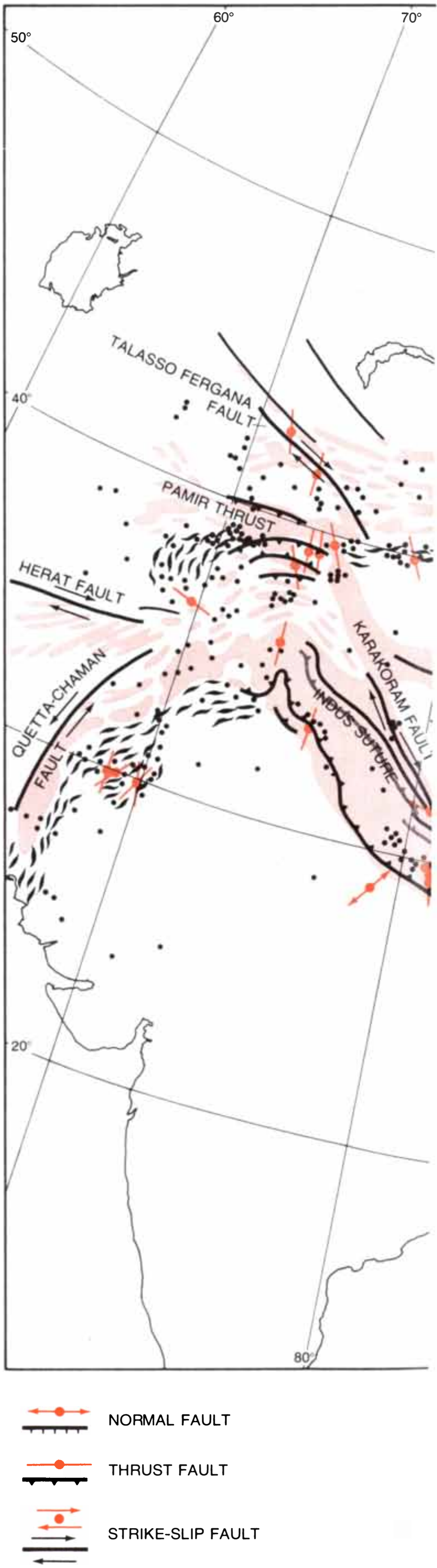
Studies of earthquakes in the Himalayas corroborate geological observations and show that India is continuing to thrust under the Himalayas in a northerly to northeasterly direction. Over the past 50 years several investigators have suggested that the northern margin of India was also thrust under the Tibetan plateau and that Tibet is at a higher altitude because of it. We and many other geologists now find this very unlikely. We see no evidence of such underthrusting going on now, and mechanically such a phenomenon seems highly contrived.

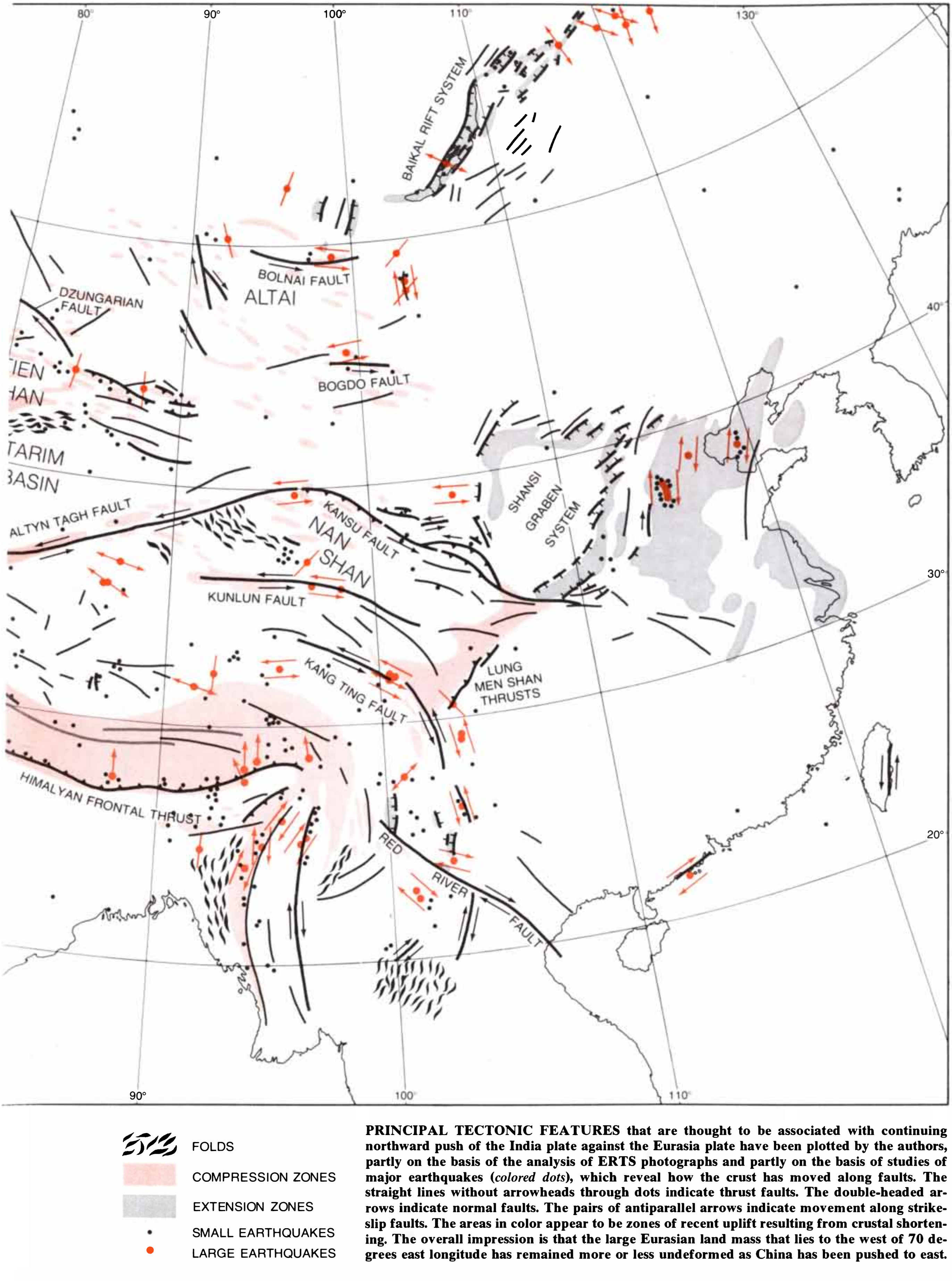
More recently John F. Dewey and Kevin Burke of the State University of New York at Albany have suggested that the Tibetan plateau is a result of a shortening of the crust in the entire Tibetan-Himalayan region, with the Tibetan crust behaving like an accordion, contracting horizontally and expanding vertically. Although we are not convinced that this concept is inapplicable to Tibet, we see no persuasive evidence for accepting it. For one thing, studies of earthquakes show that Tibet is not shortening but is stretching in an east-west direction. Moreover, from a comparison of satellite photographs of Tibet with those of other mountainous areas in Asia where shortening is currently taking place, the surface deformation of the Tibetan plateau appears to be much less, and if folding of the crust has occurred, it is less recent. Nevertheless, even if the thickness of the crust has doubled in Tibet as a result of pressure from the south, only 600 or 700 of the 2,000 kilometers can be explained in this way.

In the Tien Shan area the pattern of deformation is again dominated by thrust faulting, with shortening in a northerly direction. The Russian seismologist V. I. Ulomov estimates that the amount of shortening in the western part of the area is about 300 kilometers, a figure arrived at by imagining what would happen if the thickened crust there was flattened out (in his words "with a rolling pin") to normal thickness. For most of the rest of Asia fault-plane solutions of earthquakes indicate the overall predominance of strike-slip faulting. Thrust faulting is characteristic in only limited areas; normal faulting is fairly uncommon. For example, in Mongolia most earthquakes are associated with strike-slip faulting. In any case earthquakes in Mongolia exhibit a fairly consistent northeast-southwest orientation of the maximum compressive stress. Although the pattern east of Tibet is more complex, it is similar to the pattern in Mongolia.

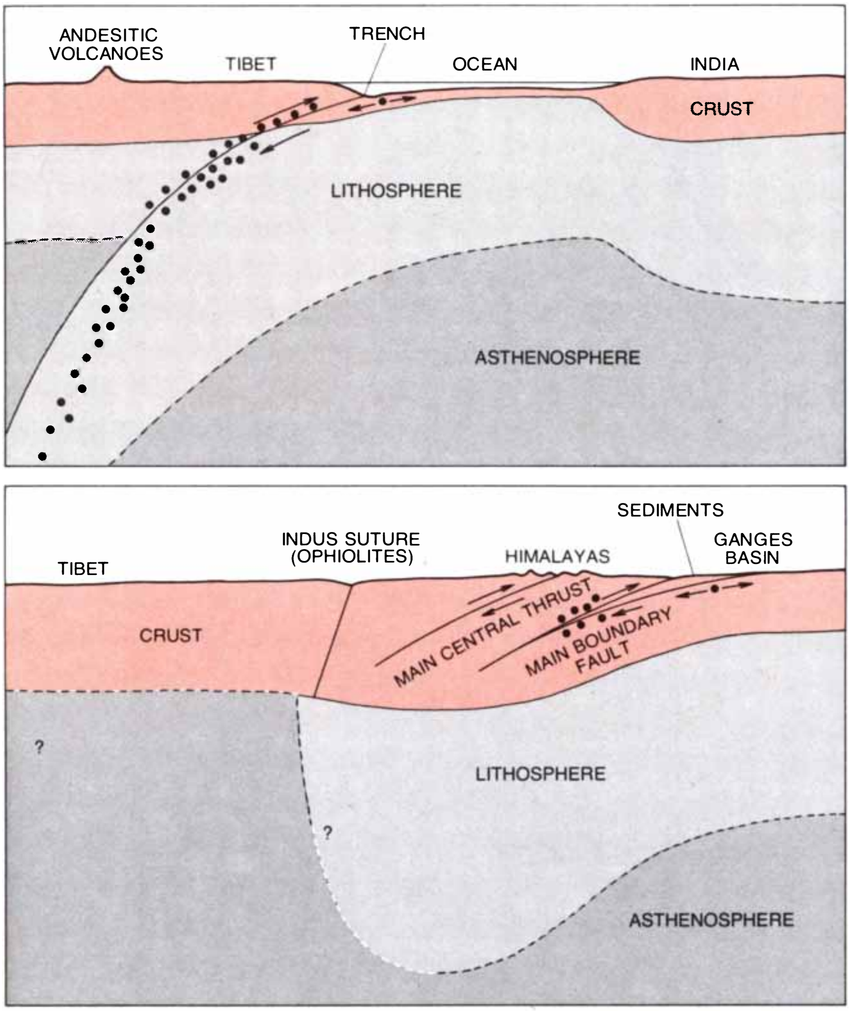
Thus the faulting associated with earthquakes indicates that much of Asia is being squeezed in a direction lying between north-south and northeast-southwest, a pattern that is compatible with India's northward motion. The squeezing causes the crust in parts of Asia to shorten and thicken. Here again, however, even if one adds a possible 700 kilometers of crustal shortening in Tibet to 300 kilometers in the Tien Shan area, one can still account for only 1,000 kilometers of shortening, or half of India's total travel toward Eurasia since the beginning of the collision. Another explanation must be sought for the remaining 1,000 kilometers of displacement even if we are wrong in doubting the "accordion" shortening in Tibet.

The satellite pictures of Asia play a central role in leading us to an alternative explanation for the 2,000 kilometers of convergence between India and Eurasia. Perhaps the most striking fea-

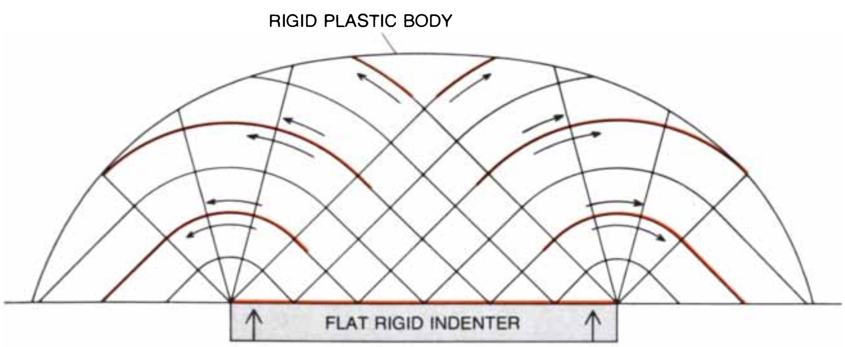





**PRINCIPAL TECTONIC FEATURES** that are thought to be associated with continuing northward push of the India plate against the Eurasia plate have been plotted by the authors, partly on the basis of the analysis of ERTS photographs and partly on the basis of studies of major earthquakes (*colored dots*), which reveal how the crust has moved along faults. The straight lines without arrowheads through dots indicate thrust faults. The double-headed arrows indicate normal faults. The pairs of antiparallel arrows indicate movement along strike-slip faults. The areas in color appear to be zones of recent uplift resulting from crustal shortening. The overall impression is that the large Eurasian land mass that lies to the west of 70 degrees east longitude has remained more or less undeformed as China has been pushed to east.




**CROSS SECTION OF THE COLLISION between India and Eurasia plates is shown schematically. The upper diagram shows a cross section through the lithosphere and the asthenosphere about two million years before actual contact, when the land masses were still separated by about 200 kilometers of ocean. At that time the lithospheric plate carrying India was plunging under the Eurasia plate as today the Pacific plate is plunging under the South America plate, creating the Andes along the west coast of South America. The black dots show how earthquakes tend to cluster along the boundary between plates and within the descending plate. The lower diagram depicts the situation today. The suture line, the Indus suture, is marked by the presence of ophiolites: sequences of rocks containing ocean sediments and showing other characteristics of having been formed in a suboceanic environment. Earthquakes are more diffusely distributed and shallower than they were before the collision. The Himalayas are slices of old Indian crust that have overthrust rest of India to the south, creating new faults that migrate southward. Active faulting seems to occur on main boundary fault. Crust under Tibet appears to be unusually hot; lithosphere there may be so thin that bottom may lie within crust.**




**GEOMETRY OF SLIP LINES observed when an indenting tool made of a hard material such as steel is pressed into a softer material such as brass bears a striking resemblance to the distribution and directional sense of strike-slip faults in Asia. The slip lines are probably less symmetrical in Asia than they are in idealized case because of asymmetry of boundary conditions.**

tures seen in aerial photographs of any kind are strike-slip faults. They usually appear as sharply defined linear features in the topography. In Asia the strike-slip faults are among the clearest in the world. Although some of these faults were known from displacements during earthquakes or from geological studies, it appears that until recently some were not known, at least to most geologists.

The Talasso Fergana fault in the Tien Shan area is well known to Russian geologists and also stands out clearly in satellite pictures. The sharp bends in the ridges near the fault trace distinctly show that the sense of motion for the fault is right lateral. The Bolnai, or Khangai, fault in northern Mongolia ruptured in 1905 during one of the greatest earthquakes of the century. Although the Bolnai fault is little known outside Mongolia and the U.S.S.R., it too shows up clearly in satellite photographs. Detailed fieldwork by the Russian geologists S. D. Khilko, N. A. Florensov and their colleagues shows that left-lateral displacement of a few meters extended for a distance of 370 kilometers during the 1905 earthquake. Perhaps the longest and greatest strike-slip fault in the world is the Altyn Tagh fault just north of Tibet. As far as we can determine it was unknown as an active fault, at least outside China, until the satellite pictures became available.

For us the recognition of major strike-slip faults in Asia was an exciting moment because it was immediately clear that there is a fairly simple pattern to them. By combining fault-plane solutions of earthquakes in Asia with the observed displacements on the earth's surface following some of the largest earthquakes and with the tectonic features linked to the faults that are visible in the satellite pictures we have been able to determine the sense of motion associated with the principal faults. In China and Mongolia the strike-slip faults generally run approximately east-west and consistently with left-lateral motion. Faults in the Tien Shan area, and some in or near Mongolia, run north-south or northwest-southeast and consistently with right-lateral motion. This is the pattern one would expect from a squeezing of the region in a direction lying between north-south and northeast-southwest, a squeezing that the collision between India and Eurasia would probably cause. Moreover, the predominance of strike-slip motion suggests that the northward motion of India is accommodated essentially by material moving out of the way of the impinging continents.

In general the clearest, and therefore presumably the most important, faults are left-lateral faults that run east-west. Thus most of the lateral displacement of China out of the way of India would be to the east. This is understandable; material displaced to the west would have to push against thousands of kilometers of the Eurasian land mass. The motion of China to the east is easily accommodated by its thrusting over the oceanic plates along the margins of the Pacific. Anyone who squeezes a tube of toothpaste experiences a homely analogy to this type of displacement. The thumb and fingers correspond to India and the rest of Eurasia. The closed end of the tube is analogous to the vast continental region of Eurasia that essentially blocks large-scale movement toward the west. The open end of the tube is analogous to the subduction zones of the western Pacific, with China and Mongolia acting as the toothpaste.

The existence of major strike-slip faults also helps to account for some other apparent peculiarities in the geology of Asia. Although most of Asia appears to be experiencing horizontal compression in a direction between north and northeast, there are two notable exceptions. Lake Baikal occupies a part of the Baikal rift system that is an expression of a northwest-southeast extension. Similarly, Deng Qidong and his colleagues at the Geological Institute in Peking describe the Shansi graben in eastern China as a rift system created by another extension with the same orientation. We interpret both of these systems as being partly the result of their proximity to strike-slip faults and as being comparable to the tension cracks that develop at the ends of shear cracks.

On this interpretation, after India collided with Eurasia the relative velocity of motion decreased, but India continued to drive into Eurasia at a rate of five centimeters per year. In so doing it

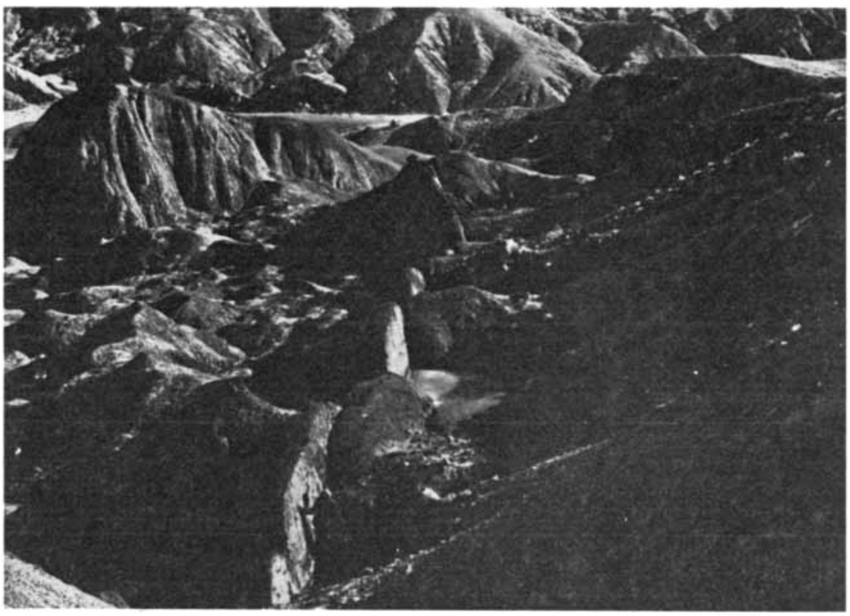
**CRUSTAL THICKENING AND SHORTENING appear to have taken place in the region of the Tien Shan range southeast of Lake Balkhash, as is indicated in this ERTS photograph centered on 40.5 degrees north latitude and 78.5 degrees east longitude. The picture shows folded sedimentary formations on the south side of the Tien Shan resulting from thrust faults in the crust under the sediments and dipping to the north under the Tien Shan. Seismic studies show that the crust is 20 to 30 kilometers thicker to the north of the Tien Shan than it is in stabler adjacent areas. This fact suggests that the earth's crust in the eastern portion of the Tien Shan has been shortened, or compressed, by as much as 300 kilometers, presumably because of the northward thrust of India, more than 1,000 kilometers away.**

squeezed up mountains in front of it, but more important it pushed material out of its way. The flow of that material may have opened up rift valleys in a direction perpendicular to the direction of shortening, so that in a sense India has wedged Eurasia apart. The overall pattern of deformation is physically similar to that caused by the indentation of a plastic medium by a blunt tool. We think the solutions developed by mechanical engineers for indentation problems may enable us to treat the state of stress in Asia quantitatively and may lead to a better understanding of the large-scale deformation of continents.

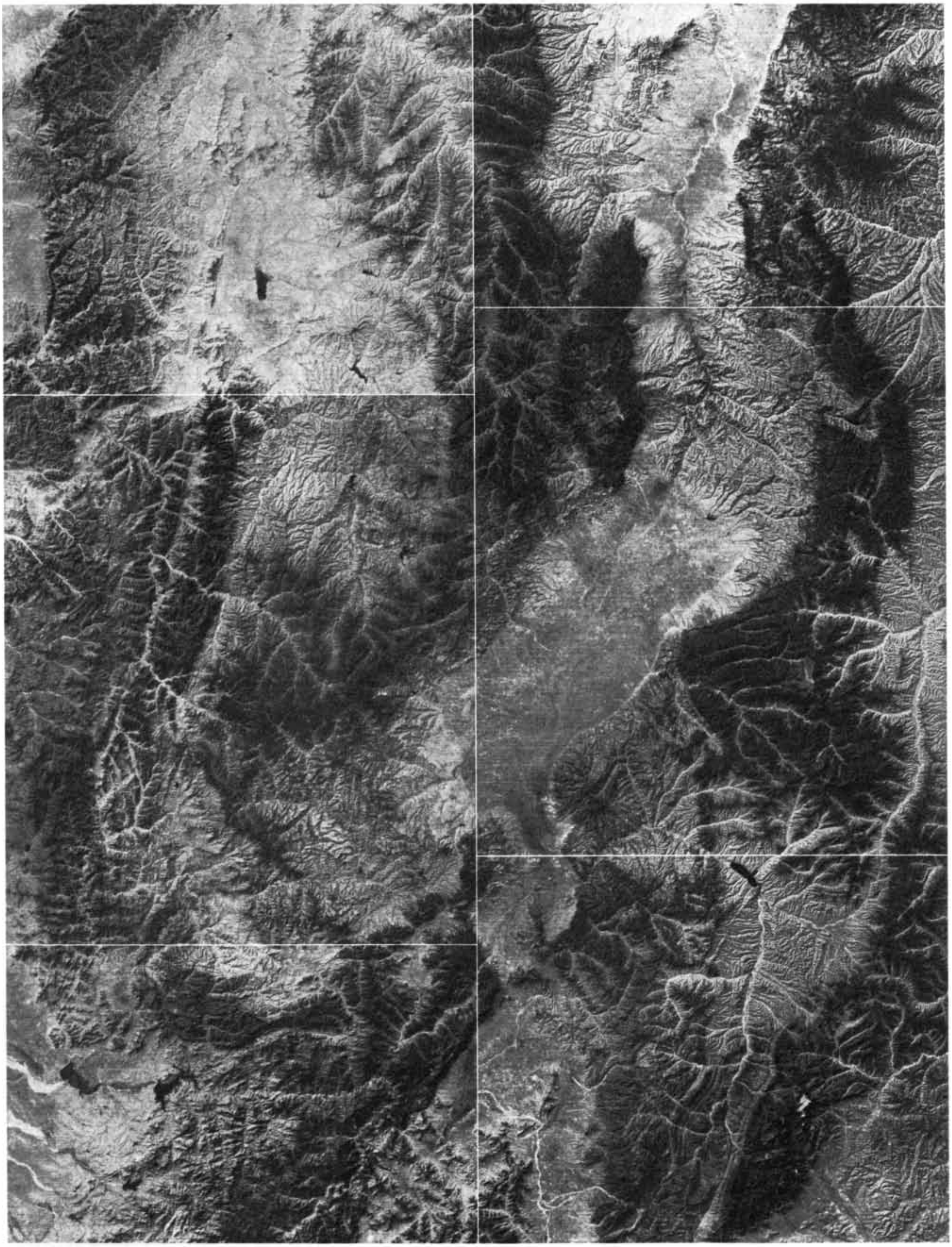
The basic unproved assumption in the scheme we have been describing is that large horizontal displacements have actually occurred on the strike-slip faults of Asia. To the best of our knowledge the most thoroughly studied of these faults is the Talasso Fergana fault in the U.S.S.R. There is controversy among Russian geologists over how much displacement has taken place on the fault, but V. S. Burtman suggests that it may amount to 200 to 250 kilometers of right-lateral motion. He suggests, however, that much of the displacement occurred before India collided with Eurasia. He is unable to estimate how much has taken place since then. As for the other faults, we can only say that most of them are as prominent on the satellite photographs as the San Andreas fault in California is, which has undergone 300 kilometers of right-lateral displacement in the past 25 million years. If the displacements along the highly visible Asian faults are comparable to the displacement along the San Andreas fault, as we suspect, one can conclude that most of the 2,000 kilometers of convergence between India and Eurasia can be attributed to the lateral displacement of China. On the other hand, our hypothesis would be fatally wounded if it could be shown that the displacement along the major faults has amounted to only a few kilometers or at most a few tens of kilometers over the past 40 million years.

In any case it is virtually certain that strike-slip faulting plays a key role in the process of suturing continents together. McKenzie noted some years ago that in the Middle East, the other important region where continents are actively colliding, the motion of the Arabian subcontinent toward Eurasia is forcing part of Turkey to move to the west in a direction perpendicular to that of the converging continental blocks. Moreover, it is clear from detailed studies by Mattauer and his colleagues of portions of some old mountain belts such as those in France and Spain and in Morocco that strike-slip motion was important long after the continents had collided. It seems quite likely that as other ancient mountain belts are studied, evidence for large-scale horizontal movement along strike-slip faults will be found. Such analyses of continental tectonics are clearly not a direct application of plate tectonics to continents. To apply plate tectonics to the deformation of Asia would call for so many plates that the concept's utility would be lost. We suspect that the same will be found to be true for older continental collisions.

At the same time we view the tectonics of Asia as being a direct consequence of plate motions. The earthquakes and great faults of India, China, Mongolia and the U.S.S.R. may be attributed to a simple phenomenon: the northward motion of the Indian subcontinent riding on the India plate toward the Eurasia plate. What is perhaps most interesting about this interpretation is that it indicates that the movement of India caused the deformation of a region more than 3,000 kilometers away. Since the mountains were created by movement along the faults, and since the climate of the region is in turn profoundly influenced by the topography, environmental conditions throughout much of Asia, including the harsh climate of the Himalayas, the Tibetan plateau, the Gobi Desert, Mongolia and parts of China, can also be attributed to a collision that has been in progress for 40 million years.



**GOBI-ALTAI EARTHQUAKE OF 1957, which occurred along the Bogdo fault in the Gobi Desert, gave rise to the crustal displacements shown in these photographs made by the geologist V. P. Solonenko. The earthquake, whose magnitude was 7.9 on Richter scale, produced a left-lateral strike-slip motion, which means that far block, as viewed from either side, moved to left. Motion reached 10 meters, among largest strike-slip displacements of recent times.**



**PORTION OF SHANSI GRABEN SYSTEM** separates the eastern end of the forbidding Ordos plateau, part of the Gobi Desert, from the fertile, heavily populated valley of the Huang Ho. A graben is a sunken region where crustal blocks are being pulled apart. This mosaic of six ERTS photographs shows the surface features in an area 260 kilometers by 345 kilometers near Taiyuan, about 500 kilometers southwest of Peking. The pictures in the mosaic were made at different times of the year, so that in some of them snow and ice appear on ground and rivers. Shansi graben and the Baikal rift system resemble tension cracks that appear at oblique angles near strike-slip faults.

# The Status of Interferon

*The protein that defends cells from viruses was discovered 20 years ago. It has been difficult to purify it and learn how it acts, but it still promises to be valuable against serious viral diseases*

by Derek C. Burke

It is now 20 years since Alick Isaacs and Jean Lindenmann discovered a substance that is released by cells exposed to a virus and that subsequently protects other cells against viral infections. They named the substance interferon, because it seemed clearly to be the agent of viral interference, the phenomenon they were investigating at the National Institute for Medical Research in London. It had been recognized since the 1930's that the infection of animal cells in culture, or of an experimental animal itself, with one virus would for a time "interfere" with infection by other viruses. What Isaacs and Lindenmann found was that the first virus did not simply block other viruses from access to the same cells; it induced the infected cells to release something into their medium such that when the medium, free of any cells, was added to another cell culture, it made the new cells resistant to viruses.

Isaacs and several of us then working in his laboratory quickly established some of the properties of interferon. It appeared to be a protein, and one with a fairly low molecular weight. Its action was not virus-specific. That is, it did not react selectively with a particular virus as an antibody reacts with a particular antigen, and it did not react with a free virus particle as an antibody reacts with an antigenic substance. It reacted instead with cells, and the affected cells then became resistant to a large number of different viruses.

On the other hand, interferon was ordinarily species-specific: interferon made in chick cells was active in other chick cells but not in duck cells, and so on. In other words, interferon was a cellular agent, apparently part of the animal cell's natural first line of defense against viruses. Because it was liberated from cells and could be harvested, because it was effective against many viruses and because it was a natural cell product, unlikely to harm cells or to provoke resistance in viruses exposed to it, interferon gave promise of having great clinical significance. If it could be isolated, it could presumably be administered to prevent or combat viral infection in much the same way that antibiotics are deployed against infection by bacteria.

After two decades that promise has still not been fulfilled. It has taken a long time to find out just what interferon is, how it is made and how it functions in a cell. For those reasons, and because it is difficult to prepare large quantities of interferon, it is hard to establish just how effective the substance will be in clinical medicine. Within the past few years, however, small amounts of what seems to be pure interferon have been prepared, so that the task of characterizing and analyzing it can proceed. We now have some indication of how it is induced and how it interferes with viral infection. And meanwhile a number of clinical trials have yielded evidence that interferon may be valuable in combating certain serious viral diseases.

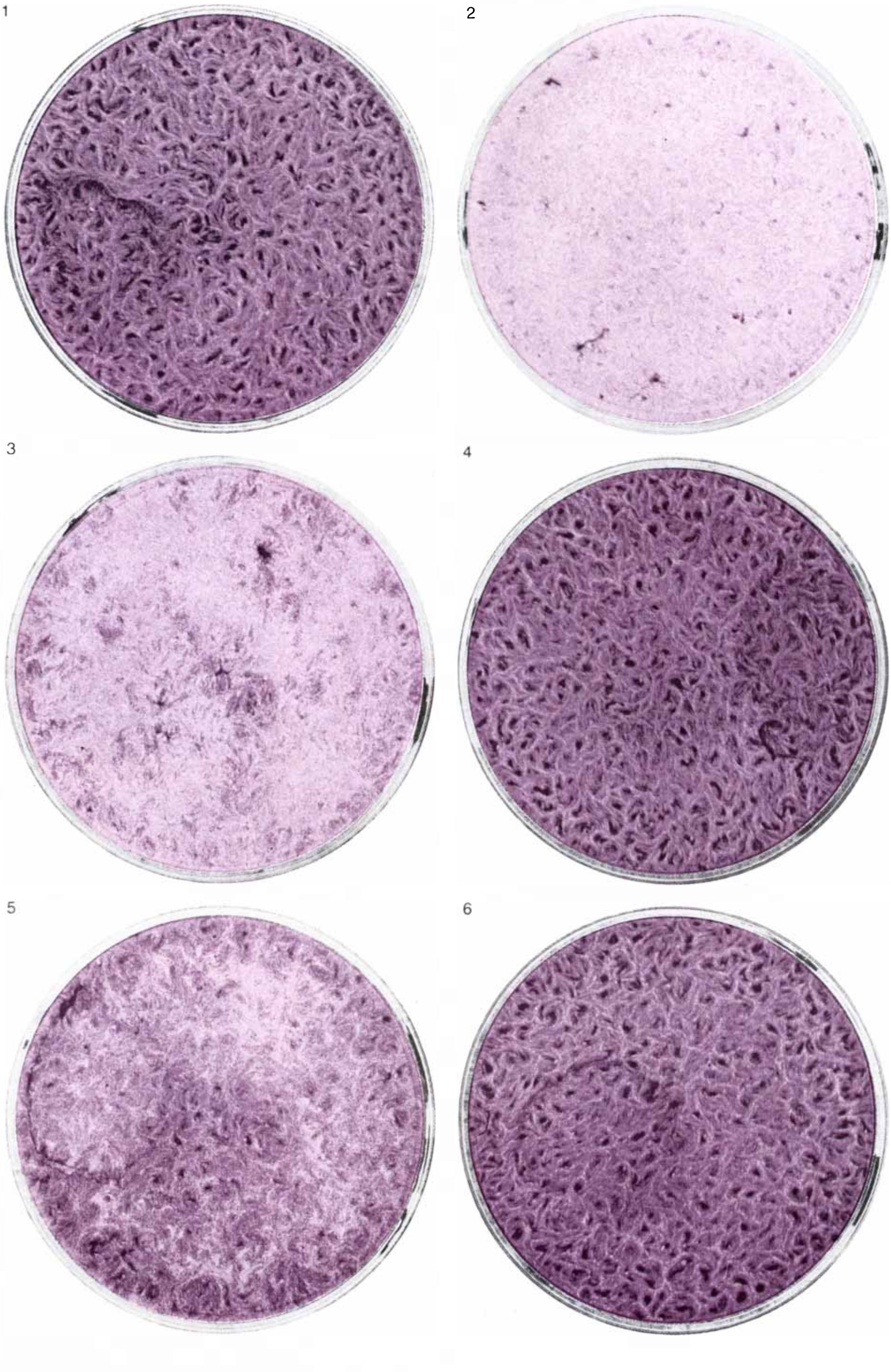
Since interferon is a protein, it should have been only a matter of time before it was purified by the standard multistage procedures for purifying any protein: "salting out," ion-exchange chromatography, gel electrophoresis and so on. A number of workers tried those approaches. Disappointingly, although the standard methods did purify the substance to a considerable degree, they did not yield a homogeneous product that could be described as a single protein. The difficulty of isolating the pure protein has been a major problem.

One reason for the difficulty is the very high specific activity of interferon. Activity is measured in arbitrary units, one unit being the amount of interferon that is required to reduce the yield of virus by 50 percent. The specific activity of pure interferon is apparently about a billion ($10^9$) units per milligram of the protein. The starting material for purification, the crude interferon derived from a virus-treated cell culture, typically has an activity of only 1,000 or 10,000 units per milligram of protein, so that a very large volume of the crude material is required to obtain a very small amount of purified interferon. It is expensive to prepare such large volumes, and each step of the purification requires painstaking biological-assay procedures.

There is a more fundamental difficulty: interferon molecules tend to stick to other molecules—in particular to other protein molecules—present in the crude solution. As a result many investigators, myself included, have sometimes purified what turned out to be an unwanted protein that was merely contaminated by a small amount of interferon.

Recently these difficulties have been largely overcome by two new developments. The first is the application of affinity chromatography [*see top illustration on page 45*]. A substance for which interferon has a strong affinity is bound to a solid matrix, which is packed into a

**EFFECT OF INTERFERON is measured by assaying its ability to protect cells in a laboratory culture. These photographs indicate the degree of purification of human fibroblast interferon attained by Ernest Knight, Jr., of E. I. du Pont de Nemours & Company. A monolayer of human fibroblast cells is grown in a dish (*1*). Vesicular stomatitis virus is added to the medium; after 24 hours the cells are dead (*2*). If unpurified interferon is applied to the culture for 24 hours before virus challenge at a concentration of about .02 microgram per milliliter, about half of the cells are protected (*3*); in other words, that concentration of the crude interferon provides one biological unit of activity. Ten biological units of the same material protect all the cells (*4*). When the purified interferon is applied, much smaller quantities are required to give the same degree of protection: only .005 nanogram per milliliter to provide a unit of activity, or 50 percent protection (*5*), and .05 nanogram to protect all the cells (*6*). Purification has been 4,000-fold; the specific activity of the purified protein is about 200 million units per milligram.**


1
2
3
4
5
6

chromatography column. When the crude interferon solution is poured through the column, the interferon is retained on the matrix particles, whereas many of the impurities pass through. The interferon is thereupon eluted, or washed out, with a suitable solvent. Kurt Paucker and his colleagues at the Medical College of Pennsylvania have achieved a large degree of purification by using antibody to interferon as the high-affinity substance. William A. Carter and his colleagues at the Roswell Park Memorial Institute in Buffalo capitalized on one of the ordinarily troublesome interactions with a protein (albumin) to obtain a 2,000-fold purification in a single step. Another type of affinity chromatography exploits the fact that interferon is a glycoprotein: its molecule has one carbohydrate side chain or more ending in a sialic acid group. Such a group binds very specifically to the plant substances known as lectins. That interaction has been exploited, first by Rudolf Weil and his colleagues at the Sandoz Laboratories in Vienna and later by other workers.

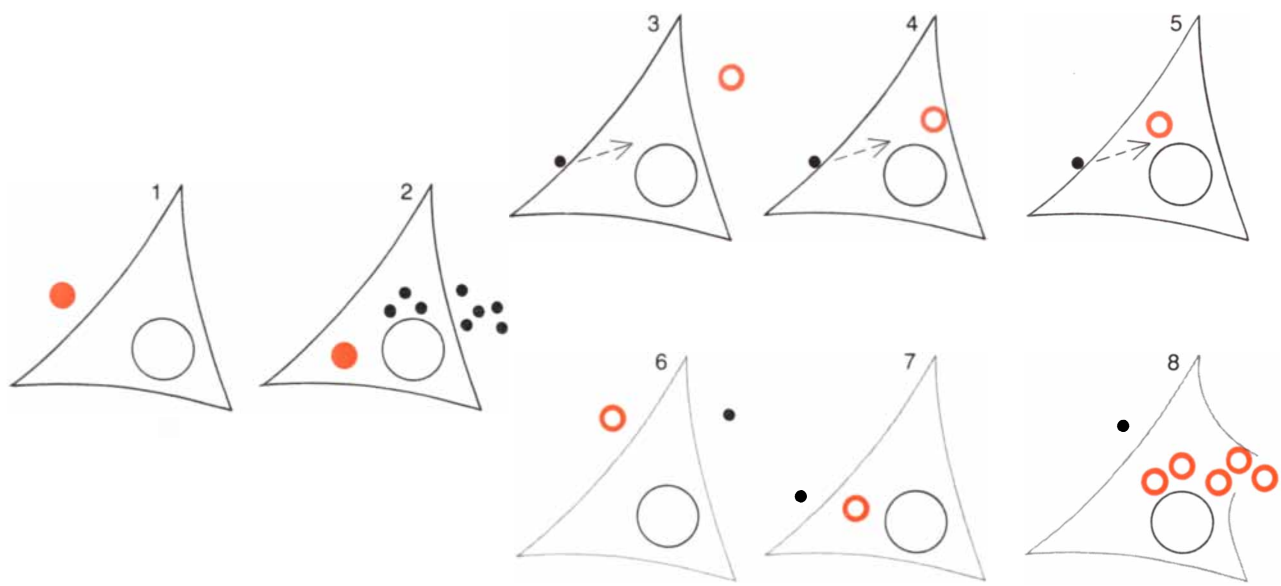
The second development that has advanced the task of purification is the discovery of a way to block the interaction of interferon with other proteins without destroying interferon's biological activity. William E. Stewart, working at the Catholic University of Louvain in Belgium, boiled interferon prepared from human leukocytes (white blood cells) in a strong detergent, sodium dodecyl sulfate. Such treatment denatures a protein, that is, it unfolds the protein molecule's long chain of amino acids and makes the molecule inactive. Stewart found that after cooling and the removal of the detergent the interferon recovered its original biological activity: it was renatured. This stability of interferon, under severe denaturing conditions that destroyed the usual protein-protein interactions, made it possible to separate interferon molecules from interferon that was adsorbed to some other molecular species.

Moreover, Stewart found that if the interferon preparation was boiled in the detergent in the presence of a reducing agent, it could be only partially renatured [*see bottom illustration on opposite page*]. This, he suggested, was because the crude preparation contained two types of interferon, one of which was not stable under reducing conditions. It thus became possible to separate two types of human-leukocyte interferon. The major component, designated *Le*, has a molecular weight of 15,000 and is not renatured after detergent treatment under reducing conditions. The other component, which accounts for less of the total interferon, is designated *F;* it has a molecular weight of about 21,000 and is active whether or not a reducing agent is present during the detergent treatment. The two interferons differ in their degree of antiviral activity in cells from other species, the *Le* type having only about 5 percent of the protective effect on rabbit cells that it has on human cells, whereas the *F* type is equally effective with rabbit and human cells. The *F* type is a glycoprotein; the *Le* type is probably not. The *F* type has also been shown to be very similar to the interferon produced by human fibroblasts, the precursors of connective-tissue cells.

The *Le* type is itself separable into two proteins with molecules of rather similar size. We therefore have a remarkable situation: human fibroblasts manufacture one type of interferon, whereas human leukocytes manufacture two types and a trace of a third. No wonder purification has been so difficult! We do not know how these types of interferon differ chemically, but their molecules probably have different amino acid chains, suggesting either that they are coded for by different genes or that a single gene product is processed in different ways in the two kinds of cell.

By capitalizing on affinity chromatography and on the detergent treatment several investigators have made considerable progress in purification. For example, Ernest Knight, Jr., of the central research and development department of E. I. du Pont de Nemours & Company has obtained human fibroblast interferon in what he thinks is probably a homogeneous form. It is a rather small protein, with a molecular weight of 20,000. If enough of it can be accumulated, the determination of its amino acid sequence could proceed, and then artificial synthesis would be a real possibility.

Until such synthesis is achieved we must depend on cells to make interferon. There are two possible ways to arrange that. One can induce cells in tissue culture to produce interferon and



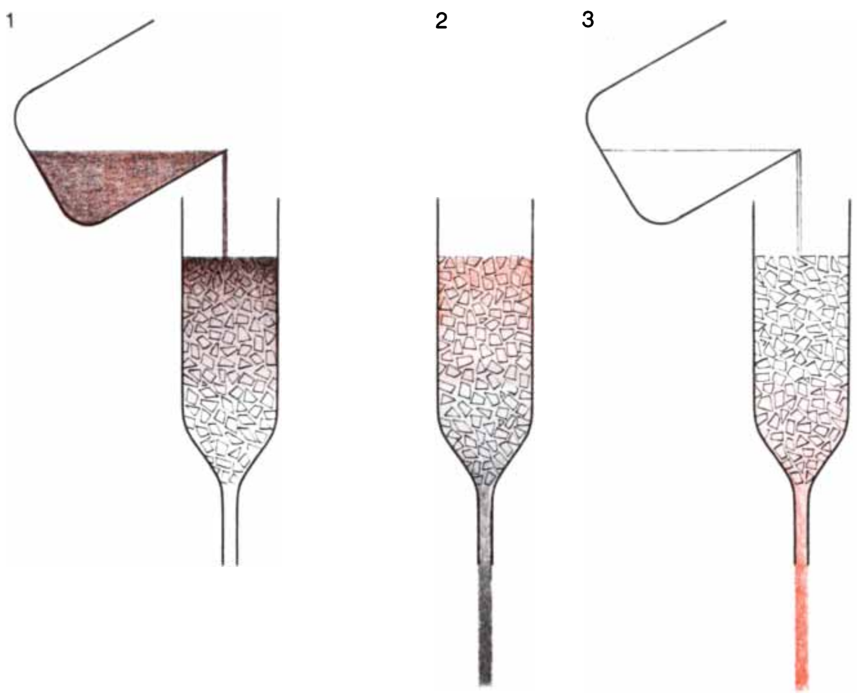

**INTERFERON (*black dots*) produced by a cell infected by a virus (*colored disk*) is liberated into the medium (*1, 2*). The interferon somehow protects another cell of the same species (*3*) from a different virus (*open colored circle*), which is prevented from proliferating (*4, 5*). The interferon does not protect a cell of a different species (*6*), in which the virus is able to multiply, eventually killing the cell (*7, 8*). This representation of the observed effects of interferon reflects the state of knowledge in about 1960, before mode of action was known.**

then harvest the protein, purify it and administer it to patients. Or one can administer an inducer directly to the patient, stimulating the patient's cells to manufacture endogenous interferon. In either case we need to know how cells can best be induced to make interferon.

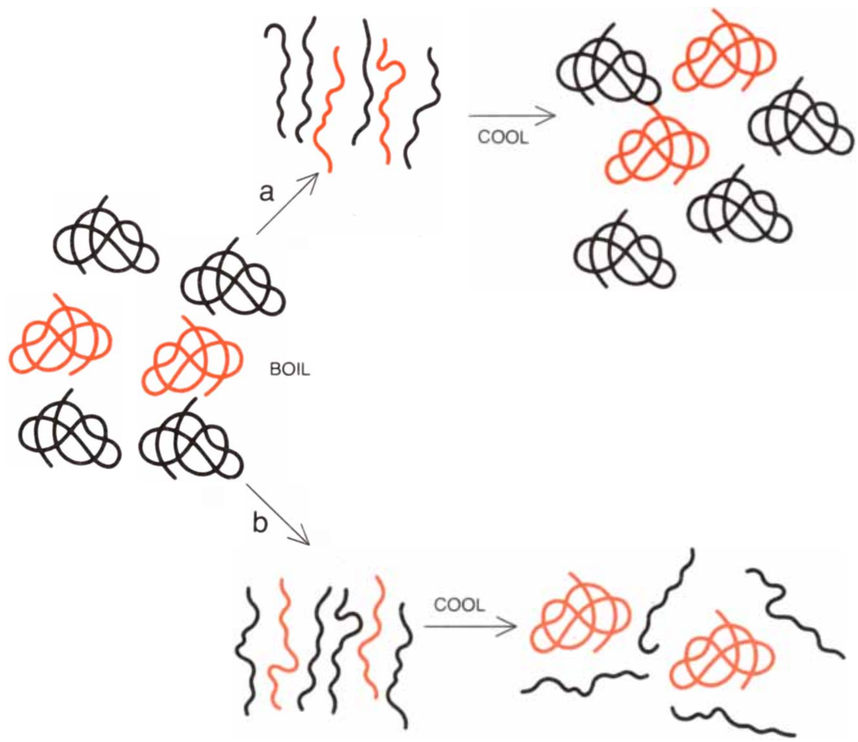
Interferon was originally discovered by treating chick cells with influenza virus, but it was soon apparent that many other viruses would induce the production of the substance. The virus did not even have to be infectious; attenuated viruses would do. Then it was found that agents other than viruses would induce interferon. The first such agents to be reported were two antibiotics, helenine and statolon, that had been extracted from molds. Investigators at the Merck Sharp & Dohme Research Laboratories undertook to find the active principle in helenine. First, however, they noted a report to the effect that antibody responses in animals were enhanced by synthetic polynucleotides: analogues of DNA or RNA that are made in the laboratory by linking the nucleotide subunits of nucleic acids in arbitrary ways. They found that poly *rI:rC*, a synthetic RNA consisting of paired strands of riboinosinic acid and ribocytidylic acid subunits, was an extremely active inducer of interferon. Then they found that the active principle in helenine was also a double-strand RNA and showed that it came from a virus; the same thing turned out to be true for statolon.

The Merck Sharp & Dohme workers went on to screen a variety of synthetic and natural nucleic acids by injecting them in rabbits and testing the rabbits' blood serum for the ability to protect a cell culture against viral infection. A number of double-strand RNA's were found to be effective inducers, with poly *rI:rC* particularly active. DNA was not effective, nor was single-strand RNA; the investigators suggested that the single strand of RNA that is the genetic material of most animal viruses must replicate within the cell to produce a double-strand RNA before it can trigger the production of interferon.

Further research showed that viruses and double-strand RNA's are not the only substances that induce interferon (although they are the only ones that are active with typical tissue-culture cells). In particular, lymphoid cells, which make antibody and are involved in cellular immunity, were found to manufacture interferon when they were treated with the appropriate antigen, a mechanism that may have a role in the immune system. The antigen-induced interferon is readily distinguishable, both immunologically and by its physical properties, from the interferon produced by human cells in response to viruses or RNA's. This confirms the impression that there are a number of interferons, which have




**AFFINITY CHROMATOGRAPHY is one method of purifying interferon. The crude interferon is poured into a column containing a solid matrix (a coarse powder) to which is attached a molecule that interacts strongly with interferon (*1*). The interferon binds to the molecule and is retained in the column, whereas the impurities in the crude solution tend to pass through (*2*). Then the interferon (*color*) is washed out by passing a suitable solvent through the column (*3*).**



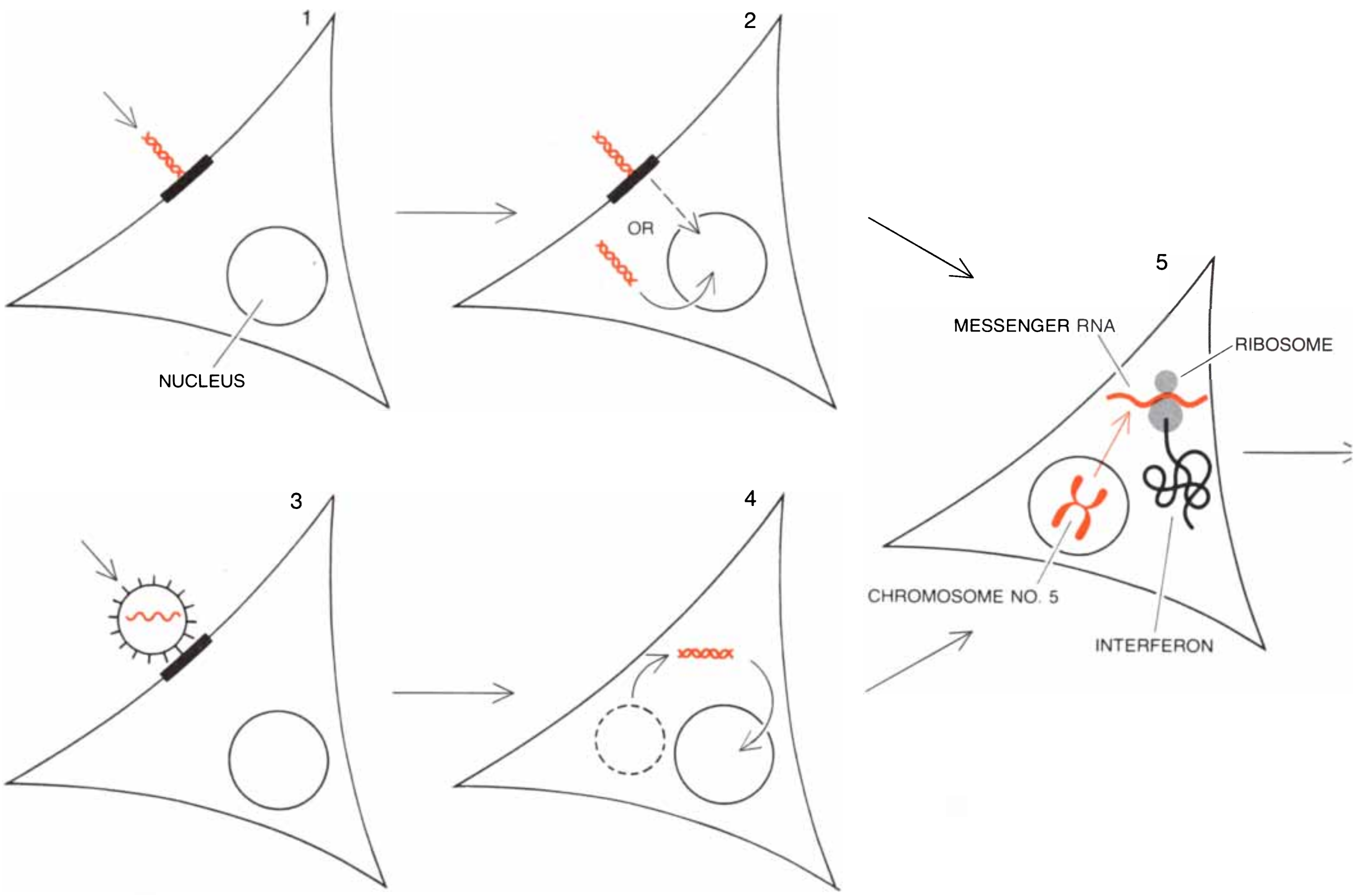

**BOILING IN DETERGENT sodium dodecyl sulfate (*a*) denatures human leukocyte (white blood cell) interferon, that is, it unfolds the protein chain and inactivates the molecule; on cooling, the protein is renatured and regains its activity. On cooling after being boiled with the detergent and a reducing agent, 2-mercaptoethanol (*b*), however, the interferon is only partially renatured. Apparently leukocyte interferon has two components, one of which (*black*) is not renatured after being boiled under reducing conditions. The boiling treatment destroys most protein-protein interactions, so that it serves to separate interferon from some contaminants.**

antiviral activity in common but differ in other ways. In the whole animal interferons have been induced by a variety of substances. Genetic analysis of interferon production in inbred mice shows that the interferons manufactured in response to different inducers are controlled by different host genes—another indication of the complexity of the interferon system.

The question of just how induction is accomplished is better investigated with cells in tissue culture than in the whole animal (since one can study a single population of cells) and with double-strand RNA's as the inducer. It seems clear that these RNA's interact with a receptor on the cell surface. That would explain why very particular structural requirements must be met by any nucleotide inducer: it must be double-stranded; both strands must have RNA's typical ribose ring, with a hydroxyl (OH) group at the position of carbon atom No. 2; it must have a melting temperature of more than 60 degrees Celsius, which is to say that it must be in the helical form—not untwisted—at the temperature of the induction experiment, and it must have a certain minimum molecular weight.

When the experiment is conducted at four degrees C., the RNA interacts with the receptor but is apparently only superficially bound, since treatment with the enzyme ribonuclease, which digests RNA, prevents the formation of interferon. If the preparation is warmed to 37 degrees (body temperature), ribonuclease no longer has any effect on interferon formation. There are two possible reasons: either the RNA has entered the cell at that temperature and is therefore no longer subject to attack by the enzyme or the RNA is able at the higher temperature to send into the cell, without entering it, a signal that triggers the induction process. We cannot yet distinguish between these two possibilities; it is known that the double-strand RNA's can enter cells, but no experiment has yet shown that they have to do so. On the other hand, it is clear that when a virus induces interferon, it does enter the cell and make new viral RNA inside the cell; even the inactivated viruses that induce interferon do so. That may be, as I mentioned above, because single-strand viral RNA must replicate and form double-strand RNA, which then serves as an inducer, but this has not been proved.

What happens next, after the inducer has either entered the cell or sent a message into it? The species specificity of interferon indicated that its formation is controlled by the host cell. This was confirmed by the fact that interferon formation is inhibited by actinomycin *D* (which means that it requires RNA synthesis directed by DNA, presumably that of the host cell) and also by cycloheximide (which means that it requires protein synthesis). It would appear, then, that the inducer somehow triggers the "derepression" of a gene in the host cell, which is thereupon transcribed into messenger RNA, which in turn is translated into a protein. Such a scheme has been confirmed in several ways. Y. H. Tan and Frank H. Ruddle of Yale University showed, by means of experiments with human-mouse hybrid cells,



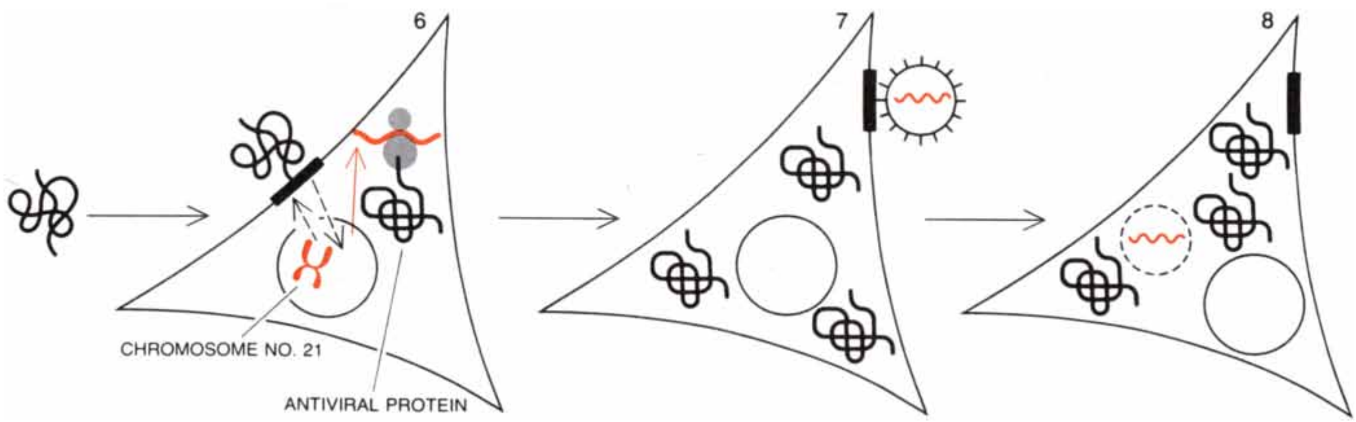

**INDUCTION AND MODE OF ACTION of interferon are now understood in some detail. A cell can be induced to make interferon by a double-strand RNA that binds to a cell-surface receptor (*1*) and then either enters the cell or sends a message to the cell nucleus from the surface (*2*). Or interferon can be induced by a virus that binds to a receptor (*3*) and then enters the cell, where it makes new RNA (which may need to be double-stranded) that serves as inducer (*4*). In any case the inducer triggers human chromosome No. 5 to produce a mes-**

that human chromosome No. 5 is essential for interferon formation. The control of interferon synthesis by genes in the nucleus was supported by my own finding that the removal of the host cell's nucleus prevents interferon formation.

The RNA-to-protein pathway has been demonstrated also by the extraction of messenger RNA from interferon-producing cells and its subsequent translation to form interferon. Translation was first accomplished in cells of a second species by Jacqueline De Maeyer-Guignard, Edward De Maeyer and Luc Montagnier of the Fondation Curie–Institut du Radium at Orsay in France. They added RNA from interferon-producing mouse cells to chick cells, which were thereby persuaded to make mouse interferon. Later Paula M. Pitha and her collaborators at the Johns Hopkins School of Medicine found that when mouse-interferon messenger was injected into oöcytes (precursor egg cells) of the frog *Xenopus*, it too was translated. Finally the interferon messenger was translated in a cell-free protein-synthesizing system (a mixture of cell extracts in the test tube); the interferon was identified not only by its biological activity but also by precipitation with interferon-specific antibody to yield a polypeptide with the expected molecular weight. These last results imply, incidentally, that interferon does not depend on its carbohydrate side chains for its biological activity, because a cell-free system cannot attach the chains. All these experiments leave unanswered, however, the central problem of how the inducer triggers the host-cell gene to manufacture interferon messenger RNA.

Once interferon is synthesized, how does it act in other cells against viruses? First it must interact with the other cells. The interaction appears to involve binding to a cell-surface receptor, and again that binding must initially be superficial: interferon binds at four degrees, but at that low temperature its antiviral effect can be destroyed by treating the cells with enzymes that digest protein. The antiviral effect is achieved by warming the system to 37 degrees, when it is no longer susceptible to the protein-digesting enzymes. What happens at 37 degrees is not clear, although recent investigations indicate that the interferon need not enter the cell: it is active even when it is bound to a solid support, which must keep it outside. Ruddle and others have suggested that synthesis of the receptor is controlled in human cells by chromosome No. 21, because only those human-mouse hybrid cells that retain human chromosome No. 21 respond to human interferon, and the response is stronger in cells that have an extra copy of the chromosome.

Interferon is not itself the antiviral agent. Its action can be inhibited by treating cells with actinomycin *D* or cycloheximide, suggesting that it acts by triggering a process involving both DNA-directed RNA synthesis and protein synthesis. Moreover, the cell nucleus must be present for interferon to act against viruses. The most obvious interpretation is that interferon, acting at the cell membrane, induces the synthesis of a new cellular messenger RNA that is in




**senger RNA that is translated into the protein interferon (*5*). The interferon leaves the cell and binds to another human cell (*6*) at a receptor coded for by chromosome No. 21. That chromosome or another one is thereby caused to make a messenger RNA that is translated into an antiviral protein, which protects the cell: when a new virus appears and binds to the cell (*7*), it is able to enter the cell (*8*) but is not able (because the antiviral protein interferes with RNA synthesis or protein synthesis or both) to multiply and thus to kill the cell.**

turn translated to yield another protein that actually has the antiviral effect. The translation of such a messenger can yield a large number of antiviral-protein molecules, so that such a scheme would provide an amplification mechanism: a single molecule of interferon can trigger the synthesis of a lot of antiviral protein. That would explain the very high biological activity of interferon, only a few molecules of which seem to be required to protect a cell.

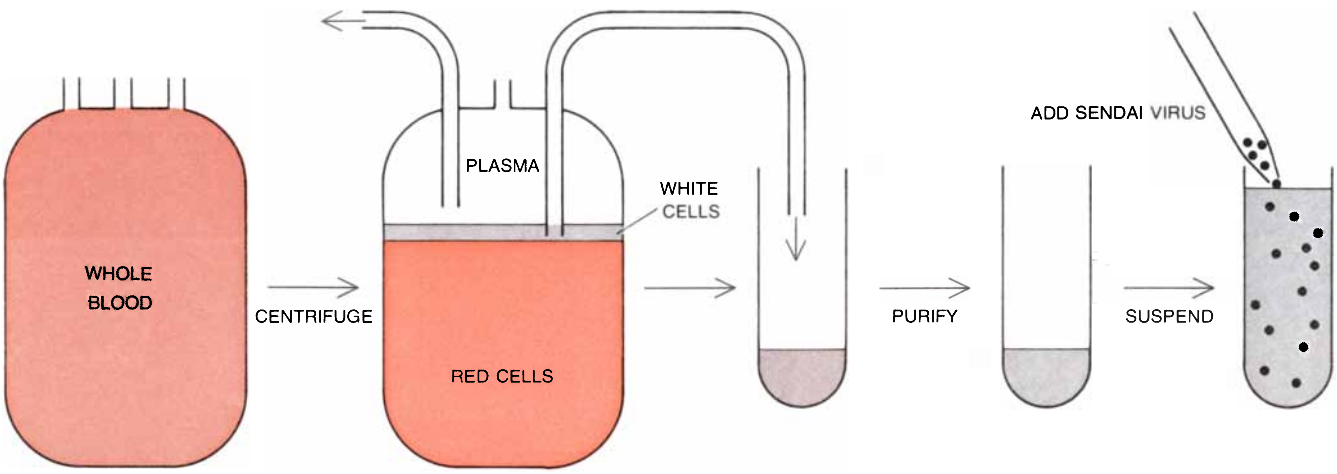
The question of how the antiviral protein actually inhibits the multiplication of viruses has been investigated in infected cells treated with interferon and in cell-free systems incorporating components from treated cells. Experiments with cells have established that interferon treatment does not keep the virus particles from entering the cell. There is good evidence that what it does is inhibit the synthesis of viral nucleic acid or viral proteins or both; different combinations of viruses and cells give different answers as to which step is inhibited. The step seems to be viral protein synthesis in cells infected with vaccinia (cowpox) virus or with one of the small RNA viruses, but there is equally good evidence for an effect on viral RNA synthesis in cells infected with the small DNA-containing tumor virus SV40.

As for cell-free systems, those prepared from interferon-treated cells seem to be affected in several ways, all of which have to do with the formation of new proteins. The initiation of protein synthesis is slowed down, perhaps partly because interferon treatment inhibits the necessary modification, called capping, of the end of the messenger-RNA molecule where translation begins. In addition extracts from interferon-treated cells are more sensitive to the effects of double-strand RNA, a known inhibitor of the initiation of protein synthesis. The rate of growth of a new protein molecule is also depressed by treatment with interferon. That may be because the treatment depletes the supply of a particular transfer RNA required for protein synthesis or because it induces the formation of ribonucleases that degrade the new messenger RNA's. Obviously the impact of interferon treatment on protein synthesis is complex, suggesting that interferon acts more like a hormone, which may have many effects, than like a metabolic inhibitor with a single effect at a precisely defined stage of a particular synthetic pathway.

If interferon can affect both nucleic acid and protein synthesis, why does it not kill cells as well as preventing virus multiplication? There are several possibilities. First, virus multiplication is an exponential process, so that a small effect on an early event can have a magnified effect on subsequent stages. Second, it may be that interferon does inhibit cellular processes as well as viral ones but that it is active only in cells that are infected by viruses; the infected cells would die but the vast majority of cells would not be affected. Third, there are indications that viral messenger RNA's have a higher avidity for components of the cell's protein-synthesis machinery than cellular messenger RNA's and that they may therefore be more strongly affected by interferon treatment than cells are. In any case interferon's discrimination between viruses and cells is not absolute: interferon does slow the growth of cells, although at much higher concentrations than are required to block the growth of viruses. A variety of other effects have been reported for interferon. For example, it enhances several different cellular defense mechanisms, mediated by different kinds of cells, and it changes the way certain cell-surface antigens are expressed. It is increasingly seen as being a versatile agent with effects on both viruses and host cells, but with a much greater effect on viruses.

As early as 1960 a collaborative research project was organized in Britain by the Medical Research Council and three pharmaceutical companies (Imperial Chemical Industries, Glaxo Laboratories and the Wellcome Foundation) aimed at determining interferon's value as an antiviral agent in human beings. The first problem, and a continuing one, was to prepare enough interferon that was both active in human cells and safe to administer to human subjects. (The first trial was with monkey interferon, but that proved to be extremely expensive, and subsequent trials were with human interferon.) Another problem was to devise test situations in which an adequate amount of interferon could be administered efficiently and a controlled and measurable response to viral infection could be observed. Norman B. Finter's group at Imperial Chemical Industries did experiments with mice that provided the basic information for the design of clinical trials.

In the first trial, in 1961, monkey interferon was injected into one site on the upper arm of volunteers who had never been vaccinated against smallpox and a placebo was injected into another site. The next day both sites were challenged with vaccinia virus. Vaccinial lesions developed at 37 of 38 control sites and at only 14 of the 38 sites pretreated with interferon, a highly significant difference. Such a trial was deliberately set up to favor a positive result: the interferon was injected locally into a limited number of cells and the same cells were challenged with virus. Successful results were achieved in a similar trial of the effects of interferon on vaccinia virus in the eye of rabbits.




**LARGE-SCALE PRODUCTION of human leukocyte interferon is achieved by a method developed by Kari Cantell in Helsinki. Whole blood from blood banks is centrifuged and the thin layer of leukocytes (white cells) is drawn off; the leukocytes are "contaminated" with some red cells, which must be eliminated. The purified white cells are suspended in a culture medium that includes some human**

It was much more difficult to demonstrate protection against less localized infections. Two trials in which interferon was administered to subjects with viral respiratory infections, in 1965 and 1970, were negative. In retrospect it is clear that not enough interferon was given to protect the cells of the upper respiratory tract, which are very efficient at removing foreign material—including a potentially protective agent—through the wavelike motion of their cilia. The discouraging result served, however, to discredit the exogenous-interferon approach. We turned for a time to the other possible approach: administering poly *rI:rC* and other nucleotides as antiviral drugs with the hope of inducing the volunteer's own cells to produce endogenous interferon. The same approach was taken in a number of other laboratories in the late 1960's [see "The Induction of Interferon," by Maurice R. Hilleman and Alfred A. Tytell; SCIENTIFIC AMERICAN, July, 1971]. Unfortunately, although the double-strand RNA's stimulated the production of large amounts of circulating interferon in experimental mice and other rodents, they gave rise to much less of the protein in monkeys and in human beings, probably because they were destroyed by an enzyme in the serum. They also tended to cause high fever and had other toxic effects in human volunteers. It was concluded that at least in their unmodified form the double-strand nucleotides would not be effective and safe clinical agents in man. Recently, however, Hilton B. Levy and his colleagues at the National Institutes of Health and Carter's group at Roswell Park have been able to increase the activity-to-toxicity ratio of some double-strand RNA's by modifying them chemically, and clinical trials of the modified agents are planned.

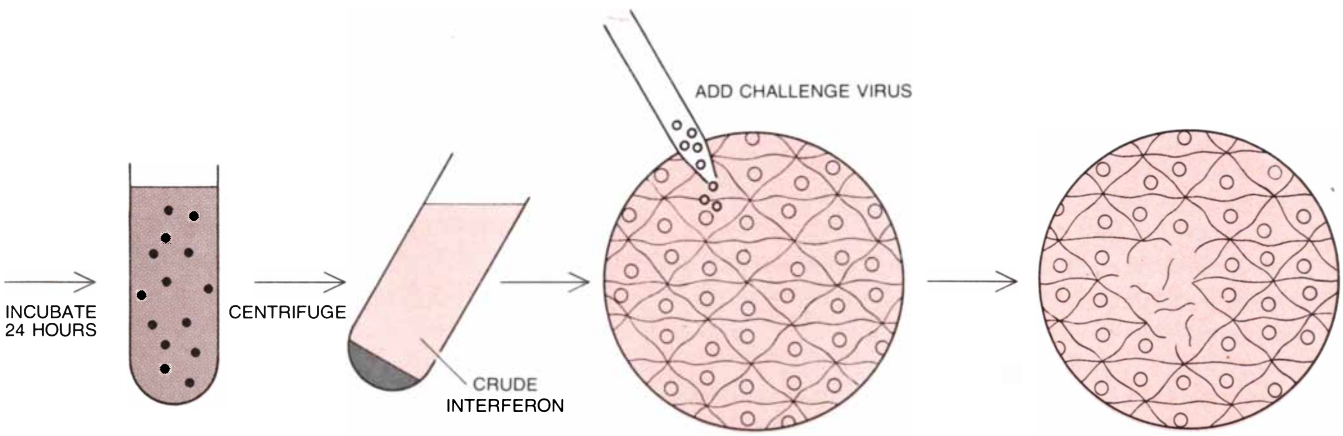
When it seemed that the double-strand nucleotides would not serve as inducers of endogenous interferon, the attention of clinical investigators turned back to the exogenous protein. Kari Cantell of the State Serum Institute in Helsinki had developed a method for the large-scale production and purification of human interferon from leukocytes obtained from blood banks. Interferon produced by Cantell's group has made possible a number of important clinical trials.

In 1972 a group at the Medical Research Council's Common Cold Unit in Salisbury tested interferon against influenza and common-cold infections. A total dose of 800,000 units of human interferon per volunteer only slightly delayed the onset of infection by the influenza virus. Much larger amounts (14 million units per patient given in 39 doses over a four-day period) resulted in a statistically significant decrease in the symptoms of common-cold infection and in the presence of virus in material washed out of the nose. This was the first successful trial of interferon in common-cold infection, and it was significant that no toxic effects were noted. It is not practical, however, to give interferon to prevent colds, if only because the cost of a million units of interferon is now about $50.

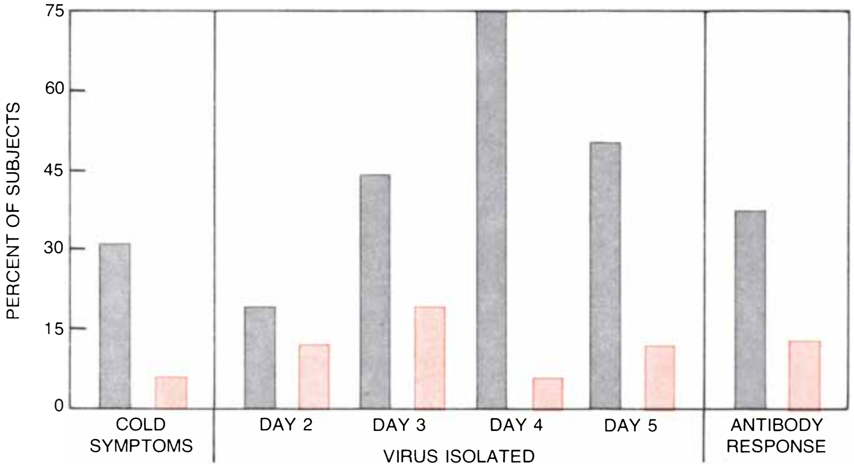
There are people, however, for whom a viral infection can be life-threatening. Patients who have a chronic immunological disorder or whose immune system has been depressed by drugs given to prevent the rejection of an organ transplant or to treat cancer are unusually susceptible to a wide range of viral infections. Thomas C. Merigan of the Stanford University School of Medicine has found that interferon doses of between 2.5 and 10 million units per day will diminish both the spread and the pain of varicella-zoster infection in patients with malignancies; another trial is now under way with 20 million units a day. Merigan has reported that the same high dosage appears tbe required to reduce the excretion of virus from newborn infants (who have an immature immune system) chronically infected with cytomegalovirus.

In what may be the most significant clinical result so far a group headed by Merigan and William S. Robinson at Stanford has reported a striking effect of interferon on a chronic infection associated with serum hepatitis, or hepatitis *B*. In the U.S. about 10 percent of the patients hospitalized with hepatitis ***B*** become chronically infected with the virus. It has been estimated that worldwide some 100 million people suffer from the chronic condition; they are subject to repeated active hepatitis attacks, with associated liver damage, and they can also transmit the disease to other people. Treatment with interferon doses of between 500,000 and 10 million units per day caused a rapid decrease in the level of several hepatitis indicators in the blood: two antigens and an enzyme associated with the particles that are believed to be the virus responsible for the disease. The effect was transient when the interferon was given for 10 days or less, but it persisted as long as small doses were continued, and even for a time thereafter. In other words, the interferon appears to be interfering with the multiplication of the putative virus particles. The trial is being extended to study the effect of interferon on active hepatitis episodes.

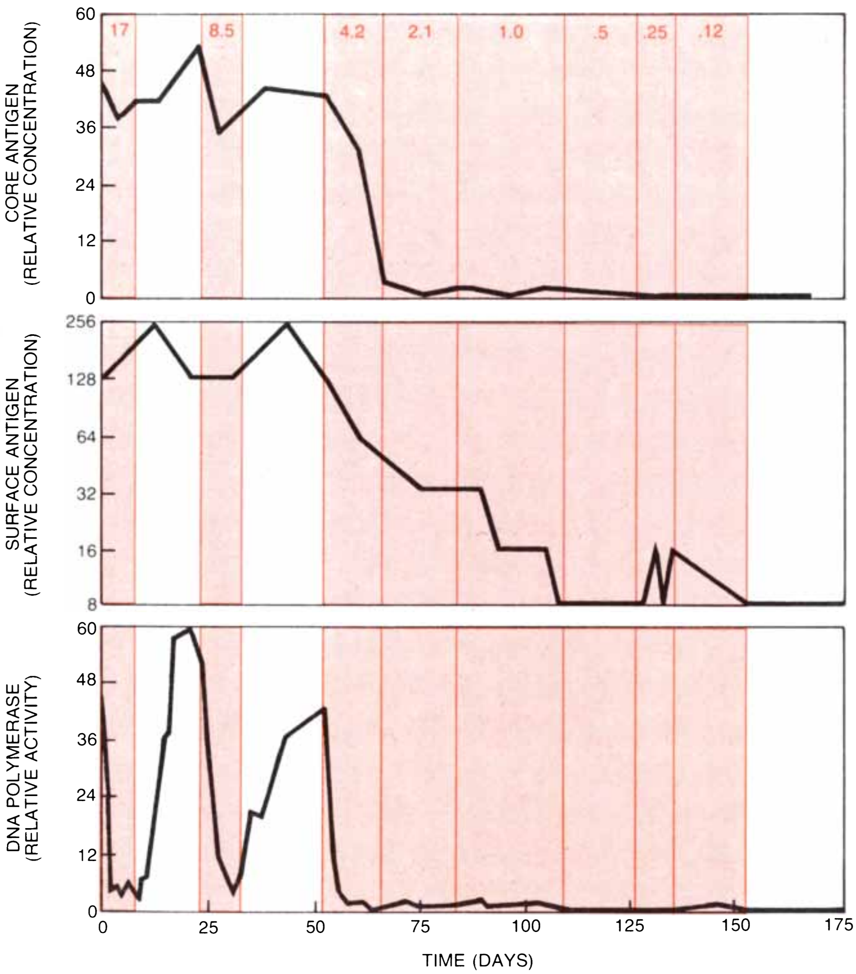
A most interesting trial is being conducted by Hans Strander and his colleagues at the Karolinska Institute in Stockholm on the effects of interferon on osteogenic sarcoma, a rare bone can-




**or animal blood serum. Then the inducer, Sendai virus, is added; the cell suspension is incubated with the virus for 24 hours, in the course of which the cells produce interferon (*color*). The suspension is centrifuged to remove the cells and viruses; the supernatant is a crude interferon preparation. It is assayed by testing its ability to reduce the areas of killed cells in a culture exposed to a challenge virus.**




**COMMON-COLD VIRUS (rhinovirus) was sprayed into the nose of volunteers, half of whom were treated (for one day before virus challenge and three days thereafter) with interferon, also by nasal spray. The treated subjects developed fewer common-cold symptoms (*color bars*) than the untreated control subjects did (*gray bars*). Less virus could be isolated from the nose of the treated subjects on the days indicated, and they also developed less antibody to the virus.**




**CHRONIC INFECTION with hepatitis *B* was treated with interferon in a trial reported by Harry B. Greenberg, then at the Stanford University School of Medicine, and his colleagues. In the case illustrated here three separate courses of interferon were given by injection; the numbers at the top give the dosage in multiples of 10,000 units per kilogram of body weight per day. The curves trace the effect of the interferon on the level in the patient's blood of two antigens and an enzyme, DNA polymerase, that are associated with "Dane particles," which are present in the blood of hepatitis *B* patients and are thought to be the hepatitis *B* virus. Short courses of large interferon doses had a clear effect, but a transient one; a sustained course of small doses appeared to result in a prolonged suppression of the putative virus's activity.**

cer that may be initiated by a virus. It is a serious disease, with a two-year survival rate of only 20 percent. In the current trial the patients were first treated surgically and in some cases with X-ray therapy to reduce the tumor load. Then between two and three million units of interferon were injected three times a week for 18 months. Twenty-one patients have now been treated without any side effects of the interferon being observed. In comparison with similar cases recorded earlier there appears to be a reduced incidence of metastasis (spread of the cancer to other tissues), which is the usual cause of death. The results are encouraging and the trial is continuing. If interferon is indeed influencing the course of this disease, it is hard to say how. It could be acting against some unidentified virus involved in the spread of the cancer. Or it could be functioning more generally as a cellular control agent retarding the growth of the cancer cells; the loss of the control of growth and proliferation is one characteristic of cancer cells. Now Martin S. Hirsch and his colleagues at the Massachusetts General Hospital have begun testing the effect of interferon on various cancers of the lymphatic tissues in patients who have received kidney transplants.

The results of the trials I have cited clearly call for continued clinical investigation of interferon's effect on malignant diseases as well as on viral infections. Such trials require large amounts of interferon, and so it is good news that the National Institutes of Health has approved the expenditure of $1 million for the manufacture of human and mouse interferons. For the long term, however, what is needed is much cheaper interferon, and several ways of achieving that goal are being investigated. One possibility is to grow very large numbers of human lymphoblasts, the precursors of immune cells, and then to induce them to manufacture interferon; another is to alter some line of human cells by genetic manipulation so that they yield more interferon. Someday it may be possible through genetic manipulation to introduce the genes for interferon production into bacteria, vast quantities of which could then be grown to serve as interferon factories.

Interferon has by now been firmly established as a specific and highly potent inhibitor of virus multiplication. There are indications that it is actually a regulator of various cell functions and that antiviral activity is just one of its properties. We have increasingly firm grounds for believing it can be a valuable clinical agent—not for protecting the population as a whole from colds and influenza but for protecting the substantial number of patients for whom viral infections are potentially lethal.

# Crinoids are beautiful even in black-and-white.

## Actually, you wouldn't be able to see here the difference between Kodachrome film and the new Kodak Ektachrome films.

D. B. Macurda, Jr. of the University of Michigan Museum of Paleontology shot more than 350 36-exposure rolls of Kodachrome 25 film underwater in the Pacific and Indian Oceans in a program to learn more from the fossil record left by the ancestors of these beautiful echinoderms and to convey the story to specialists, students, and interested public.

For most purposes, the slightly finer grain in Kodachrome 25 film—imperceptible except to a user like Professor Macurda—is outweighed by the speed advantages of Kodachrome 64 film. Graininess is best judged not from the fine detail but from the large uniform areas in a picture. Stay with Kodachrome film for finest grain. Both Kodachrome films can resolve 100 line pairs per mm for a test object of 1000:1 contrast, about half that for a more realistic 1.6:1 test object.

If you are one to go by numbers, we give you 125 line pairs per mm resolving power (for a 1000:1 test object) in the new Kodak Ektachrome professional films for process E-6. Process E-6 is swift and simple at the expense of more complexity we build into the film compared with Kodachrome film, for which the processing is not at all simple. With Ektachrome film you can be using the slides in a talk within the same hour they were shot, and if there is anybody in the audience who can see a difference in image quality from your other slides on Kodachrome film, you should suspect that person of being a Kodak employee. The directions that come with the Kodak Ektachrome film processing kit, process E-6, even tell how to "push-process" for enhanced speed and contrast. If you prefer to send your processing out, you can order "push-processing" if you need it.

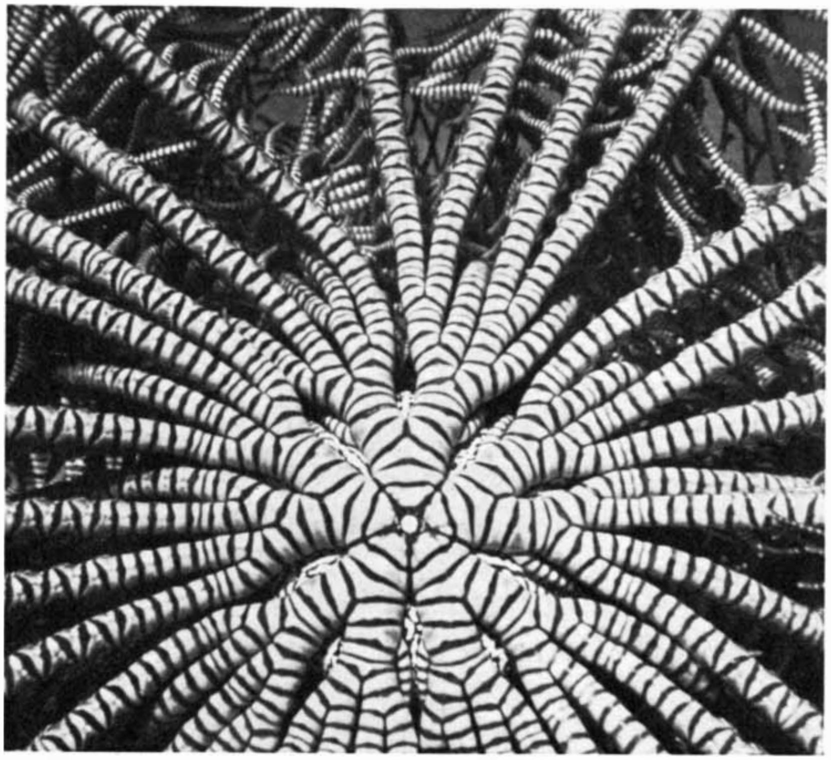
Kodak Ektachrome professional film for 35 mm slides comes in four versions: nominal speed 64 and 200 for daylight, 50 and 160 for tungsten. "Professional" signifies that a statement of actual measured speed is packed with the film and that refrigeration until use and processing promptly thereafter are assumed.*

If you wish to exaggerate color contrast more than can be done by "push-processing" (and if film speed 16 is fast enough), order Kodak photomicrography color film 2483 from your supplier and give it process E-4.†

*Details in Kodak Pamphlet E-37, "KODAK EKTACHROME Professional Films (Process E-6)."

†Details in Kodak Publication P-302, "KODAK Photomicrography Color Film 2483."

Either available from Dept. 55W, Kodak, Rochester, N.Y. 14650.



Kodak

### Showing and telling *without* slides

At the Second International Mycological Congress (Tampa, Fla., 8/27-9/3), all contributed "papers" are to be given in poster sessions, where from a visual display one decides whether to stop and exchange thoughts person to person. We are honored that the congress circular refers those unfamiliar with the format to an advertisement of ours. No need to dig it out. Just scribble "POSTER SESSIONS" on a postcard giving your name and address to Dept. 55W, Kodak, Rochester, N.Y. 14650.

# SCIENCE AND THE CITIZEN

## *Banning the Test Bomb*

A few days after taking office President Carter said he was "in favor of eliminating the testing of all nuclear devices instantly and completely" and would "proceed quickly and aggressively" to negotiate a comprehensive test-ban treaty with the U.S.S.R. The present Threshold Test Ban Treaty places a limit of 150 kilotons on underground nuclear explosions on the assumption that it is possible to determine whether any seismic event larger than that is natural or man-made. That treaty, which went into effect just a year ago (March 31, 1976), and a companion one placing the same limit on the yield of nuclear explosives for peaceful purposes (for example canal excavation) are still to be ratified by the U.S. Senate. Although U.S. experts have not been able to find a justifiable use for peaceful nuclear explosives, the Russians profess a continuing interest in their potential. The peaceful-use treaty allows multiple explosions of devices smaller than 150 kilotons, with three provisions: they must be separately distinguishable, their total yield must not exceed 1,500 kilotons and the country conducting the explosions must allow observers from the other country to assure themselves that the event has no military significance. In the 21 months between the signing of the Threshold Test Ban Treaty in July, 1974, and the date a year ago when the treaty took effect both countries accelerated testing of weapons with yields above the 150-kiloton threshold. The U.S. conducted 25 underground tests of larger yield; the U.S.S.R. conducted an undetermined number, at least 10 of which exceeded 500 kilotons.

The same groups in and out of the Senate that voiced opposition to the nomination of Paul C. Warnke as chief negotiator for the strategic-arms-limitation talks may oppose ratification of the two threshold test-ban treaties and would almost certainly oppose banning all tests. An example of the arguments that can be marshaled for and against a comprehensive test ban appears in the current issue of *International Security.* The article warning of the dangers of a total test ban is written by Robert W. Helm and Donald R. Westervelt of the Los Alamos Scientific Laboratory, where nuclear weapons are designed. A critique of their position is written by George Rathjens and Jack Ruina, arms-control specialists at the Massachusetts Institute of Technology.

Although Helm and Westervelt are not opposed to ratification of the threshold test-ban treaties as such, they argue that the agreements probably make no contribution to the goal of limiting the spread of nuclear weapons. The treaties, they think, could exacerbate dissatisfactions already expressed by certain non-nuclear states, in that the treaties not only allow the two major nuclear powers to continue testing weapons of formidable power but also encourage a distinction between the military and peaceful uses of nuclear explosives. Thus the treaties will allow the U.S. and the U.S.S.R. "to maintain a basic competence in nuclear-explosive design technology ... denied [to] the non-nuclear-weapon states [that] are parties to the [Nonproliferation Treaty]." This asymmetry is already being cited by some countries, such as India, as a reason for their not signing the Nonproliferation Treaty.

The deeper concern of Helm and Westervelt, however, is that a comprehensive test ban, by forbidding tests so small as to be indistinguishable from natural seismic events, would ultimately rest on good faith, even though it might contain provisions for the parties to challenge whether or not a particular seismic event was natural or nuclear. "This arrangement," they say, "would be unfavorable to the United States, where a watchful Congress, news media and public would interpret a [comprehensive test ban] in the most restrictive way possible and make sure the United States conducted no tests whatever. The Soviet Union, however, with fewer domestic constraints, could conduct undetectable low-yield tests of substantial military significance." Helm and Westervelt also believe the pending treaties, by legitimating the distinction between military and peaceful explosives, make it impossible either to outlaw or to supervise peaceful explosions in any complete test-ban agreement. They contend that the on-site inspection feature in the present peaceful-use treaty is workable only if it is coupled to a threshold limit high enough to be verifiable. They do not believe this threshold can ever go below "some tens of kilotons."

In their commentary on the Helm and Westervelt position Rathjens and Ruina concede that they "know of United States/Soviet nuclear confrontation scenarios where the outcome depends strongly on details of weapons technology. [But they] lack military and political credibility, resting as they do on highly questionable assumptions.... No fine tuning of military technology [can alter] the basic truth: neither nation can defend itself against the other's destructive intent and no military technology now in the cards will change this because of both the size and the destructive capability of American and Soviet nuclear forces.... [The] technical characteristics of United States and Soviet nuclear warheads per se are secondary to such technical aspects as those associated with the vulnerability, accuracy and retargeting capability of our nuclear forces and, most important, with communications, command and control that determine the flexibility possible in the use of these forces."

Rathjens and Ruina conclude: "An agreement at long last by the Soviet Union and the United States to terminate all nuclear tests would increase both internal and external pressure on other nations to follow suit and could discourage other countries [that] have not yet initiated nuclear tests, particularly those [that] are most concerned with enhancing their world image by going nuclear."

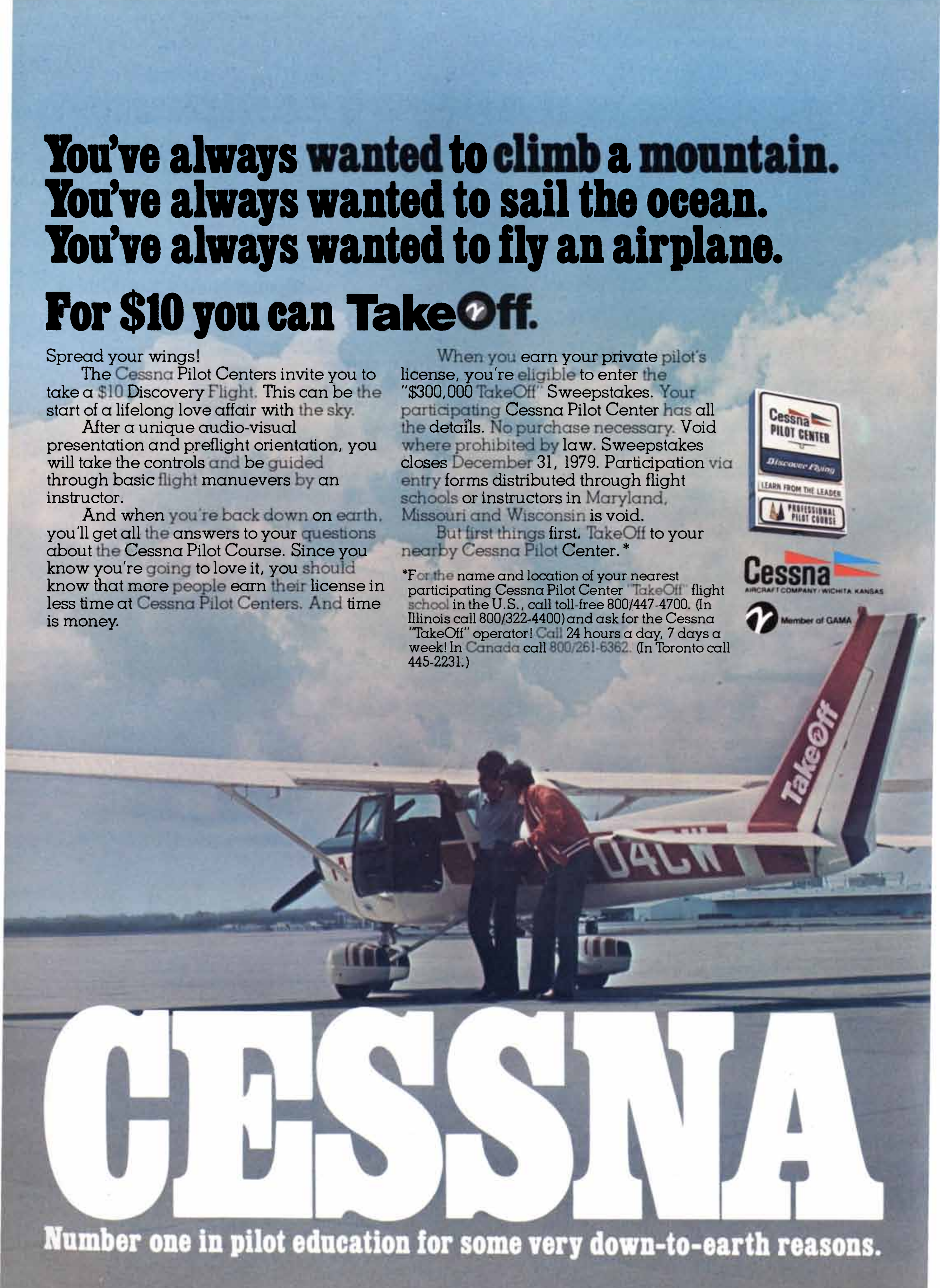
Although Rathjens and Ruina do not discuss peaceful nuclear explosives specifically in their commentary, they have stated on other occasions that an objective examination of the evidence should persuade other countries of the correctness of the U.S. opinion that their utility is negligible. In any case the U.S. and the U.S.S.R. could agree to supply non-nuclear countries (or even each other) with special nuclear explosives that could be used for engineering purposes under joint U.S.-U.S.S.R. technical supervision.

## *City Medicares*

A decade of Medicare and Medicaid has moved the U.S. toward an increasingly expensive medical-care system that is "privately owned and publicly financed" without adequate control in the public interest, according to a report on health-care spending in New York City by Columbia University's Center for Community Health Systems. The study documents escalating expenditures and a massive shift from private to public support between 1966 and 1975. The city now pays proportionately less of the medical bill than in 1966 and the Federal Government pays more, but the city nonetheless spent four times as much in 1975 as it did a decade earlier and has far less control over the cost and the quality of the services for which it must pay.

The authors of the report, Nora Piore, Purlaine Lieberman and James Linnane, estimate that $6.7 billion was spent for hospital, physician, nursing-home and other health services in New York City in the fiscal year 1975, more than two and a half times as much as in 1966, just before Medicare and Medicaid went into effect. Public support (spending by the city, state and Federal governments) accounted for 60 percent


KA 599 FP


You've always wanted to climb a mountain.
You've always wanted to sail the ocean.
You've always wanted to fly an airplane.
For $10 you can TakeOff.
Spread your wings!
The Cessna Pilot Centers invite you to take a $10 Discovery Flight. This can be the start of a lifelong love affair with the sky.
After a unique audio-visual presentation and preflight orientation, you will take the controls and be guided through basic flight manuevers by an instructor.
And when you're back down on earth, you'll get all the answers to your questions about the Cessna Pilot Course. Since you know you're going to love it, you should know that more people earn their license in less time at Cessna Pilot Centers. And time is money.
When you earn your private pilot's license, you're eligible to enter the "$300,000 TakeOff" Sweepstakes. Your participating Cessna Pilot Center has all the details. No purchase necessary. Void where prohibited by law. Sweepstakes closes December 31, 1979. Participation via entry forms distributed through flight schools or instructors in Maryland, Missouri and Wisconsin is void.
But first things first. TakeOff to your nearby Cessna Pilot Center.*
*For the name and location of your nearest participating Cessna Pilot Center "TakeOff" flight school in the U.S., call toll-free 800/447-4700. (In Illinois call 800/322-4400) and ask for the Cessna "TakeOff" operator! Call 24 hours a day, 7 days a week! In Canada call 800/261-6362. (In Toronto call 445-2231.)
Cessna PILOT CENTER
Discover Flying
LEARN FROM THE LEADER
PROFESSIONAL PILOT COURSE
Cessna
AIRCRAFT COMPANY / WICHITA KANSAS
Member of GAMA
TakeOff
04CW
CESSNA
Number one in pilot education for some very down-to-earth reasons.

of the 1975 total compared with only 29 percent in 1966. The shift to public funding was noted in all components of health care. Public funds paid for 74 percent of all hospital and nursing-home care in 1975 compared with 51 percent in 1966, and for 68 percent of the care in all the city's hospitals as opposed to 42 percent in 1966. In the nominally private sector government money subsidized 55 percent of the care in voluntary and proprietary (profit-making) hospitals compared with only 15 percent earlier; public funds accounted for 40 percent of the income of physicians in private practice in 1975, up from only .7 percent in 1966.

The city has been hit hard even though its share of public funding dropped from 45 to 30 percent while the Federal Government's share increased from 17 to 44 percent. (Federal money accounts for 70 percent of the public health-care spending in the U.S. as a whole.) The city's bill rose from $350 million in 1966 to $1.3 billion in 1975, from 13 percent of its total revenue to 17 percent. And whereas 78 percent of the public support for hospitals had gone to the municipal institutions in 1966, 53 percent of it went to voluntary and proprietary hospitals in 1975. As the city was forced increasingly to subsidize private hospitals and practitioners instead of financing its own municipal institutions, it increasingly lost control of its expenditures.

Medicare and Medicaid have undoubtedly provided better and more secure health care for old people and poor people, according to the report. The sharp income standards established for Medicaid have discriminated against the working poor, however, and have encouraged fraud on the part of patients; poor control over public payments has led to slipshod care and fraud by practitioners. The authors note, incidentally, that Medicaid, which is assumed to be a program for the very poor, has tended increasingly to fill the gap left by lack of adequate insurance for unusual medical expenses: 41 percent of the New York City Medicaid dollar was spent in 1973 for the "medically needy": those faced with paying for long-term care or catastrophic illness out of proportion to their ability to pay, who constituted only 9 percent of all Medicaid beneficiaries.

The authors of the report draw several conclusions. One is that major cities such as New York are finding it increasingly difficult to pay for health care for their populations out of shrinking revenues. Another is that Medicare and Medicaid are pumping public money into what remains a haphazard private health-care system. The report calls for "the design of instrumentalities capable of assuring equitable access to medical services [and] more prudent use of health-care resources [allocated] in accordance with public priorities."

## *Zooming in on Phobos*

In February this magazine published an article by Joseph Veverka of Cornell University on the two moons of Mars, featuring photographs of them made by *Mariner 9* and the *Viking 2* orbiter through September, 1976. The latest photographs of the inner Martian satellite Phobos, obtained during the latter half of February by the *Viking 1* orbiter, have now revealed the nature of that satellite in unprecedented detail.

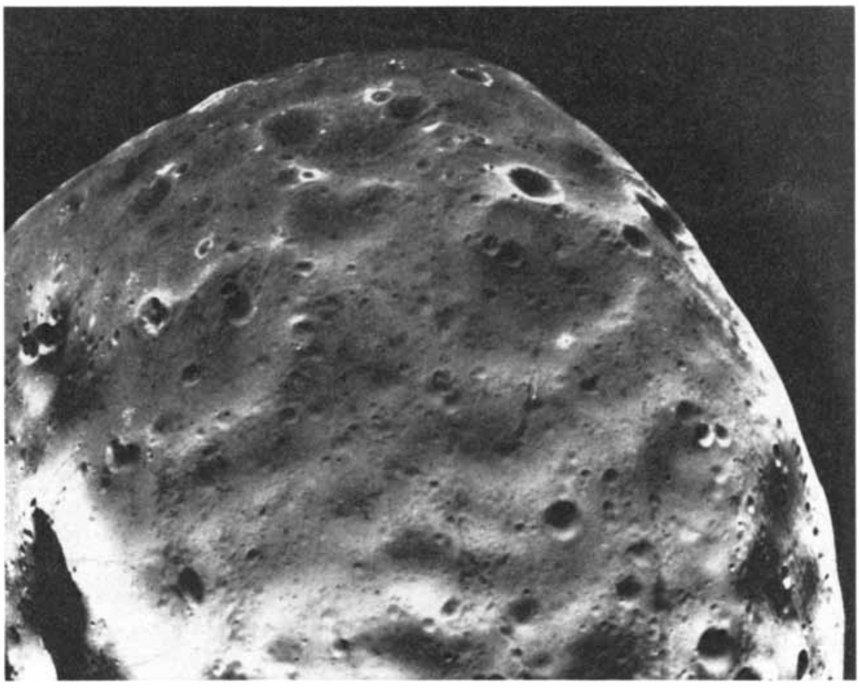
Whereas the best of the earlier photographs of the Martian moons were made from a distance of 880 kilometers, the new pictures were obtained at distances ranging from 660 kilometers to as little as 100 kilometers. The effective resolution of the best images is sufficient to distinguish objects smaller than 10 meters across. Moreover, the *Viking 1* orbiter passed so close to Phobos that the spacecraft's path was measurably perturbed by the satellite's feeble gravitational field. The measurements of the perturbations are being analyzed for information about the object's mass and mean density, which will bear on its composition.

Although the earliest *Mariner 9* photographs of Phobos showed that its surface is saturated by primary craters excavated by the impact of large meteorites, the new high-resolution images have further revealed that the surface is also covered with numerous chains of irregularly shaped craters similar to chains of secondary-impact craters on the moon. Surprisingly, the chains of craters on Phobos are preferentially oriented in one direction: they lie parallel to the plane of the satellite's orbit around Mars.

The high-resolution images have also provided more precise information on the nature of the enigmatic grooves discovered on the surface of Phobos by the *Viking 2* orbiter. These earlier images showed that the grooves seemed to be concentrated near the satellite's north pole. The latest pictures reveal that the grooves are actually chains of craters, oriented parallel to the direction in which Phobos travels in its orbit. The grooves could have been created if Phobos had overtaken a cloud of debris composed of objects as large as hundreds of meters across. If the objects had been somewhat north of the center of Phobos' direction of travel, they would have struck Phobos obliquely near the north pole, skidding and perhaps even rolling across the satellite's surface. Over the next few months analyses of the 100 or so new pictures sent back by the *Viking 1* orbiter should further disclose the nature of Phobos.

## *Plugging the Leaks*

About 33 percent of the energy produced in the U.S. is consumed in buildings: heating them, cooling them and running the appliances in them. The Energy Research and Development Administration (ERDA) sees a "great potential for saving energy and money" by improving the insulation of buildings and increasing the efficiency of appliances. To that end the agency has set in



*Phobos is seen from the* **Viking 1** *orbiter at a distance of 480 kilometers* motion an extensive program that will not only seek ways of conserving energy but also try to see that information about the new schemes reaches the consumers of energy.

A major target is windows. According to ERDA, "about one-fourth of all energy used for heating and cooling... buildings is lost through windows." The loss is equivalent to 1.7 million barrels of oil per day, averaged over a year. ERDA is studying a plan that would combine an optical shutter and a heat mirror to make windows perform better. The shutter would be a gel-like substance attached to a second pane that would be added to an existing window. In summer when the rays of the sun reached certain levels of light and heat, the shutter would change the pane automatically from a transparent state to a cloudy, white state in which the window would reflect heat and light. The pane would also incorporate the heat mirror, consisting of a film capable of reflecting back into a room the long-wave infrared radiation supplied by the heating system of the building. In winter the sun's rays would pass through the window to help heat the room but the outward movement of the long-wave radiation would be inhibited by the heat mirror.

ERDA has several ideas about appliances. One is the development of a long-lasting fluorescent light bulb that would screw into a socket like an incandescent bulb. The fluorescent bulb would supply the same amount of illumination but would consume 70 percent less energy. By 1985, according to the agency, the saving could amount to the equivalent of 165,000 barrels of oil per day.

ERDA is also interested in developing tandem arrangements for appliances. The idea is that waste heat from one appliance would be used to operate the other. For example, exhaust heat from a refrigerator or an air conditioner could be put to work in a water heater. At the same time ERDA has found that adding insulation to a water heater is a good way to conserve energy. A kit for this purpose is already on the market. ERDA estimates the saving at 30,000 barrels of oil per day by 1985 if all or most of the water heaters in the country were reinsulated.

## *Particle Refrigerator*

Most of what is known of the structure of matter has been learned by observing violent collisions between the constituent particles of atoms. The most violent collisions—and hence perhaps the most revealing ones—are those that take place when particles of equal energy meet head on. In electron-positron storage rings, for example, electrons and their antiparticles circulate in opposite directions inside a single toroidal chamber. When an electron and a positron collide, they annihilate each other and all their energy is converted into other forms of matter. The machine is particularly simple because the same impulses can be employed to accelerate and steer both the matter and the antimatter beams.

An obvious extension of this idea would be a similar machine in which protons and their antiparticles, the antiprotons, orbit in opposite directions. No device of this kind has been built, primarily because it has seemed too difficult to generate a sufficiently intense beam of antiprotons that could be confined long enough to yield a worthwhile rate of collisions. Now there are two proposed ways in which such a beam might be created and stabilized, and plans have been discussed for converting large proton accelerators into proton-antiproton storage rings. If the plans are adopted, the converted accelerators might provide collision energies 40 times greater than those available today.

The difficulty of stabilizing a particle beam can best be understood by considering the particles in their own frame of reference, as if one were traveling with the beam at just under the speed of light. It is then apparent that the particles do not move in unison: some go a bit faster or slower than the rest, and all of them have complicated motions transverse to the beam direction. The entire collection of particles can be regarded as a gas, and the random motions are indicative of the gas temperature. If that temperature is too high, particles will strike the walls of the accelerator and the beam will be dissipated.

Electrons (and positrons) have a system of natural cooling: each time an electron is accelerated it emits the radiant energy known as synchrotron radiation. In this way the disruptive thermal energy of the beam is radiated away. Protons and antiprotons, however, emit much weaker synchrotron radiation, and on their own they cannot rapidly damp their thermal motions. It has thus been proposed to refrigerate them.

One method of reducing random fluctuations was suggested by Simon Van der Meer of the European Organization for Nuclear Research (CERN). The cross section of the beam, like the spread of a shotgun pattern, is measured at one section of the ring, and the "center of mass" of the pattern is determined. Corrections to the trajectory are then quickly calculated and applied by electric fields farther around the ring. The correction signals are able to reach the far side of the ring in time only because they are sent across a chord of the ring while the particles follow the circumference. The technique has been given the name stochastic cooling, since the circulating beams are pushed by a statistical process toward the target orbit. It was tested at CERN in 1975.

The second technique is called electron cooling, and it was first suggested a decade ago by Gersh Budker of the Institute of Nuclear Physics at Novosibirsk in the U.S.S.R. The idea was first tested (at Novosibirsk) only last year. It requires a straight section in a particle-storage ring carrying protons or antiprotons. Electrons are passed through the straight section in the same direction and at the same speed as the heavier particles circulating in the ring. Collisions between the two kinds of particles are at thermal velocities, and in those collisions most of the excess momentum is transferred to the electrons. Since the electrons are almost 2,000 times lighter than protons or antiprotons, they can effectively cool by convection. After many passes through the "refrigerated" section of the ring, the beam has been confined to a narrower spread of energies and a narrower cross-sectional area.

The most important application of beam cooling would be the creation of a proton-antiproton storage ring from a large proton accelerator at CERN or at the Fermi National Accelerator Laboratory in Illinois. The proposal for Fermilab has been described by David B. Cline of the University of Wisconsin and Carlo Rubbia and Peter M. McIntyre of Harvard University. Antiprotons would be created in collisions powered by the main proton synchrotron at Fermilab, which now operates at an energy of from 400 to 500 GeV (billion electron volts). Antiprotons with a comparatively wide range of energies would be collected in a small storage ring, where they would be formed into a stabilized beam, probably by electron cooling. Accumulating a sufficiently intense beam might take several hours. Finally the antiprotons would be dumped back into the main ring, where particles and antiparticles would be accelerated simultaneously. Collisions might liberate an energy of 800 GeV. To achieve the same effective energy with protons striking a fixed target would require a beam energy of 320,000 GeV.

What do physicists expect to see at 800 GeV? That energy corresponds to the temperature of matter prevailing in the first few moments of the universe. It is far enough removed from energy regimes that are now accessible that no prediction can be reliable. Moreover, since matter and antimatter annihilate each other in the collision, there are few a priori restrictions on what might be created. One possibility is that the collision might knock loose a quark, one of the supposed constituents of protons and related particles. A more likely observation, however, is the intermediate vector boson, or *W* particle, which is believed to transmit the "weak" nuclear force. The *W* particle has been ardently sought for 10 years as the most important confirming evidence for an ambi-


SONY
4 SPEAKER

tious theory linking the weak force and electromagnetism. It is now generally conceded that the mass of the *W* is beyond the range of existing accelerators. It should be well within the range of a proton-antiproton storage ring.

## *Cloudburst*

On Saturday, July 31, 1976, a flash flood swept down the Big Thompson River in northern Colorado, killing 139 people. An analysis of the flood and of the violent thunderstorm that triggered it has now been published by two meteorologists, John F. Henz and Vincent R. Scheetz, and a geologist, Donald O. Doehring of Colorado State University. Writing in the journal *Weatherwise,* they point out that the river, which rises in Rocky Mountain National Park, is confined to a narrow canyon for a distance of 20 miles between Estes Park and Loveland. This circumstance brought the Big Thompson above its banks in less than an hour after the rains began; a 20-foot crest moved downstream at a speed of 25 feet per second and literally scoured out the canyon.

The stage was set for the disaster the day before, when a large mass of cool Canadian air began to spill southwestward from Ontario and Manitoba. The Canadian air brought relief from 100-degree heat in Nebraska and Wyoming before coming in contact with masses of moist tropical air moving up from the south on Saturday.

One of the tropical air masses, from the Gulf of California, had by then crossed Arizona on a northeasterly track; the other, traveling northwestward from the Gulf of Mexico, had passed over Texas and the Oklahoma panhandle. The warm and the cool air masses came together over Colorado; by Saturday afternoon their interaction covered the eastern face of the Rockies with clouds at low, intermediate and high altitudes.

When storm clouds build up over the Rockies, the prevailing westerly winds usually carry them away from the heights and out across the plains to the east. On this Saturday in July, however, the strong westward movement of the Canadian air mass held the growing thunderheads pinned against the eastern slopes. Between 6:00 and 6:30 P.M. several large thunderstorms developed over Boulder and Larimer counties, northwest of Denver. The moisture-laden air from the south began to mushroom upward until, within an hour, the tops of the thunderheads had reached an altitude of 60,000 feet in the evening sky, darkening the land before sunset.

One thunderstorm in particular remained motionless from 6:30 to 10:30 P.M. over a 250-square-mile area including the eastern border of Rocky Mountain National Park, the adjacent resort town of Estes Park and the Big Thompson Canyon communities of Glen Comfort and Drake, between Estes Park and Loveland to the east.

When the rainfall began in the Big Thompson area, between 6:30 and 7:00, it was heavy. Moreover, because of the position of the storm more than half of the rain fell within a 150-square-mile portion of the watershed of the Big Thompson and its main tributary, the North Fork. The heaviest fall was near a North Fork community, Glen Haven: some 14 inches in three and a half hours. At the height of the storm the rate of precipitation held steady at five inches per hour for 30 to 45 minutes. With so much of the rainfall confined to a single watershed, the flash flood developed almost instantaneously.

Floodwaters first began to rise a few miles upstream from Drake at about 7:00 P.M. By 7:30 rocks and debris carried off the slopes above U.S. 34 had blocked the highway. The river was soon 20 feet above its normal level in some places, and its rate of discharge had risen to more than 31,000 cubic feet per second. The velocity of the water in some feeder streams exceeded 30 miles per hour.

By 9:30 P.M. the storm had weakened, and by 10:00 it began to move off to the northeast. Even before that the floodwaters had begun to subside. The Big Thompson peaked at Glen Comfort at 8:00 P.M., at Drake, some 10 miles downstream, at 9:00 and at the mouth of the canyon at 11:00. A late burst of rainfall upstream caused one tributary, Dry Gulch, to peak at 10:30. Fortunately Dry Gulch has a small watershed; its flow eroded only some 6,000 cubic yards of fill from the base of the earth dam that holds Lake Estes before returning to a normal level. If the dam had failed, the catastrophe would have been virtually a cataclysm.

## *The Physics of Swinging*

Everyone knows how to set a playground swing in motion by "pumping" it, but few are familiar with the physical principles involved. In *The American Journal of Physics* Stephen M. Curry of the University of Texas at Dallas analyzes the mechanics of swinging and compares it to the functioning of a parametric amplifier, an oscillating system in which gain is produced by varying a single parameter, such as the spring constant of a spring or the capacitance of an electrical circuit.

Curry begins his analysis with a child pumping a swing by the method of alternately standing and squatting. This up-and-down motion periodically changes the length of the swing, or the distance from the pivot point to the center of mass. Because even at rest the swing is not perfectly vertical but has some slight initial energy, the child's pumping increases its amplitude by some small increment at each cycle of its motion. As the amplitude of the swing grows, the centrifugal force increases and the child has to apply a larger amount of power at every oscillation. The sum of the kinetic and the potential energy generated by his mechanical work is therefore periodically compounded like interest in a bank and increases exponentially with time.

Contrary to intuition, weight is not an advantage in swinging; although a light child does less work than a heavy one against the gravitational and centrifugal forces, proportionately less energy is required to get him swinging. Height, on the other hand, is a distinct advantage: the greater the distance is over which the child is able to vertically shift his center of mass, the larger will be the change in amplitude at each oscillation and the greater will be the rate of energy growth.

According to theory, a swing cannot be made to oscillate from rest if its initial energy is exactly zero. This apparent paradox can be understood intuitively on grounds of symmetry: if the swing is fixed in the vertical position, there is nothing that indicates any specific direction of lateral motion, and the up-and-down components of the child's pumping cancel out. Such a condition of zero initial energy cannot, however, be realized in practice. Even if one reduces the temperature to absolute zero, the swing will still have a very small energy owing to the quantum-mechanical impossibility of knowing the position and momentum of the center of mass exactly. Curry shows that in this worst possible case the minute initial energy of the swing can be amplified to a macroscopic level in a surprisingly short time. For a typical three-meter swing with a periodic length change of 40 centimeters the stored energy will reach a few hundred joules (a factor of increase of $10^{36}$) in about six minutes.

The mechanics of swinging while sitting is more complex and requires a different kind of analysis. In order to pump, the child leans back suddenly so that his body acquires a net angular momentum around his center of mass, displacing it from the initial position. This periodically applied torque rapidly starts the swinging motion. Unlike the pumping in the standing position, where the input power increases with amplitude, the torque applied in the sitting position remains approximately constant, and the total energy of the system increases linearly with time rather than exponentially. The best strategy for pumping a swing from rest, Curry concludes, is to pump in the sitting position until a certain optimum energy is reached and then resort to the standing position to increase further the amplitude of motion.


IMPORTED BY BROWNE VINTNERS, NEW YORK AND SAN FRANCISCO. AD PREPARED BY TTSP INC.
B&G
Never buy a Beaujolais by the bottle.
Buy a soft, fruity Beaujolais by the label.
B&G
1976
SAINT-LOUIS
BEAUJOLAIS
Appellation Beaujolais contrôlée
PRODUCE OF FRANCE
BOTTLED BY
BARTON & GUESTIER
BARTON & GUESTIER — FRANCE
NÉGOCIANTS A BEAUNE (COTE-D'OR)
The B&G label.
B&G. 250 years of winemaking history from all the great wine regions of France: Beaujolais, Bordeaux, the Loire Valley, Côtes du Rhône and Burgundy.
Now B&G brings you this exceptional Beaujolais.
Soft, fruity Beaujolais Saint-Louis.
Buy it by the label. B&G.
Letters of recommendation from France

# Algorithms

*An algorithm is a set of rules for getting a specific output from a specific input. Each step must be so precisely defined it can be translated into computer language and executed by machine*

by Donald E. Knuth

Ten years ago the word "algorithm" was unknown to most educated people; indeed, it was scarcely necessary. The rapid rise of computer science, which has the study of algorithms as its focal point, has changed all that; the word is now essential. There are several other words that almost, but not quite, capture the concept that is needed: procedure, recipe, process, routine, method, rigmarole. Like these things an algorithm is a set of rules or directions for getting a specific output from a specific input. The distinguishing feature of an algorithm is that all vagueness must be eliminated; the rules must describe operations that are so simple and well defined they can be executed by a machine. Furthermore, an algorithm must always terminate after a finite number of steps.

A program is the statement of an algorithm in some well-defined language. Thus a computer program represents an algorithm, although the algorithm itself is a mental concept that exists independently of any representation. In a similar way the concept of the number 2 exists in our minds without being written down. Anyone who has prepared a computer program will appreciate the fact that an algorithm must be very precisely defined, with an attention to detail that is unusual in comparison with the other things people do.

Programs for numerical problems were written as early as 1800 B.C., when Babylonian mathematicians at the time of Hammurabi gave rules for solving many types of equations. The rules were stated as step-by-step procedures applied systematically to particular numerical examples. The word algorithm itself originated in the Middle East, although at a much later time. It comes from the last name of the Persian scholar Abu Ja'far Mohammed ibn Mûsâ al-Khowârizmî, whose textbook on arithmetic (about A.D. 825) had a significant influence for many centuries.

Traditionally algorithms were concerned solely with numerical calculation. Experience with computers has shown, however, that the data manipulated by programs can represent virtually anything. Accordingly the emphasis in computer science has now shifted to the study of various structures by which information can be represented, and to the branching, or decision-making, aspects of algorithms, which allow them to follow one or another sequence of operations depending on the state of affairs at the time. It is precisely these features of algorithms that sometimes make algorithmic models more suitable than traditional mathematical models for the representation and organization of knowledge. Although numerical algorithms certainly have many interesting features, I shall confine the following discussion to non-numerical ones in order to emphasize the fact that algorithms deal primarily with the manipulation of symbols that need not represent numbers.

## Searching a Computer's Memory

In order to illustrate how algorithms can fruitfully be studied, I shall consider in some depth a simple problem of retrieving information. The problem is to discover whether or not a certain word, $x$, appears in a table of words stored in a computer's memory. The word $x$ might be the name of a person, the number of a mechanical part, a word in some foreign language, a chemical compound, a credit-card number or almost anything. The problem is interesting only when the set of all possible $x$'s is too large for the computer to handle all at once; otherwise one could simply set aside one location in the memory for each word.

Suppose $n$ different words have been stored in the computer's memory. The problem is to design an algorithm that will accept as its input the word $x$ and will yield as its output the location $j$ where $x$ appears. Thus the output will be a number between 1 and $n$, if $x$ is present; on the other hand, if $x$ is not in the memory, the output should be 0, indicating that the search was unsuccessful.

It is, of course, easy to solve this problem. The simplest algorithm is to store the words in locations 1 through $n$ and to look at each word in turn. If $x$ is found in location $j$, the computer should output $j$ and stop, but if the computer exhausts all $n$ possibilities with no success, it should output 0 and stop. Such a description of the search strategy is probably not precise enough for a computer, however, and so the procedure should be stated more carefully. It might be written as a sequence of steps in the following way:

Algorithm A; sequential search.

A1. [Initialize.] Set $j \leftarrow n$. (The arrow here means that the value of variable $j$ is set equal to $n$, the number of words in the table to be searched. This is the initial value of $j$. Subsequent steps of the algorithm will cause $j$ to run through the sequences of values $n$, $(n-1)$, $(n-2)$ and so on until it reaches either 0 or a location containing the input word $x$.)

A2. [Unsuccessful?] If $j = 0$, output $j$ and terminate the algorithm. (Otherwise go on to step A3.)

A3. [Successful?] If $x = \mathrm{KEY}[j]$, output $j$ and terminate the algorithm. (The term KEY[$j$] refers to the word stored at location $j$.)

A4. [Repeat.] Set $j \leftarrow j - 1$ (decrease the value of $j$ by 1) and go back to step A2.

This algorithm can be depicted by a flow chart that may help a person to visualize the steps [*see illustration on page 65*]. One reason it is important to specify the steps carefully is that the algorithm must work in every case. For example, the informal description given first might have suggested an erroneous algorithm that would go directly from step A1 to step A3; such an algorithm would have failed when $n = 0$ (that is, when no words at all were present), since step A1 would set $j$ to 0 and step A3 would refer to the nonexistent KEY[0].

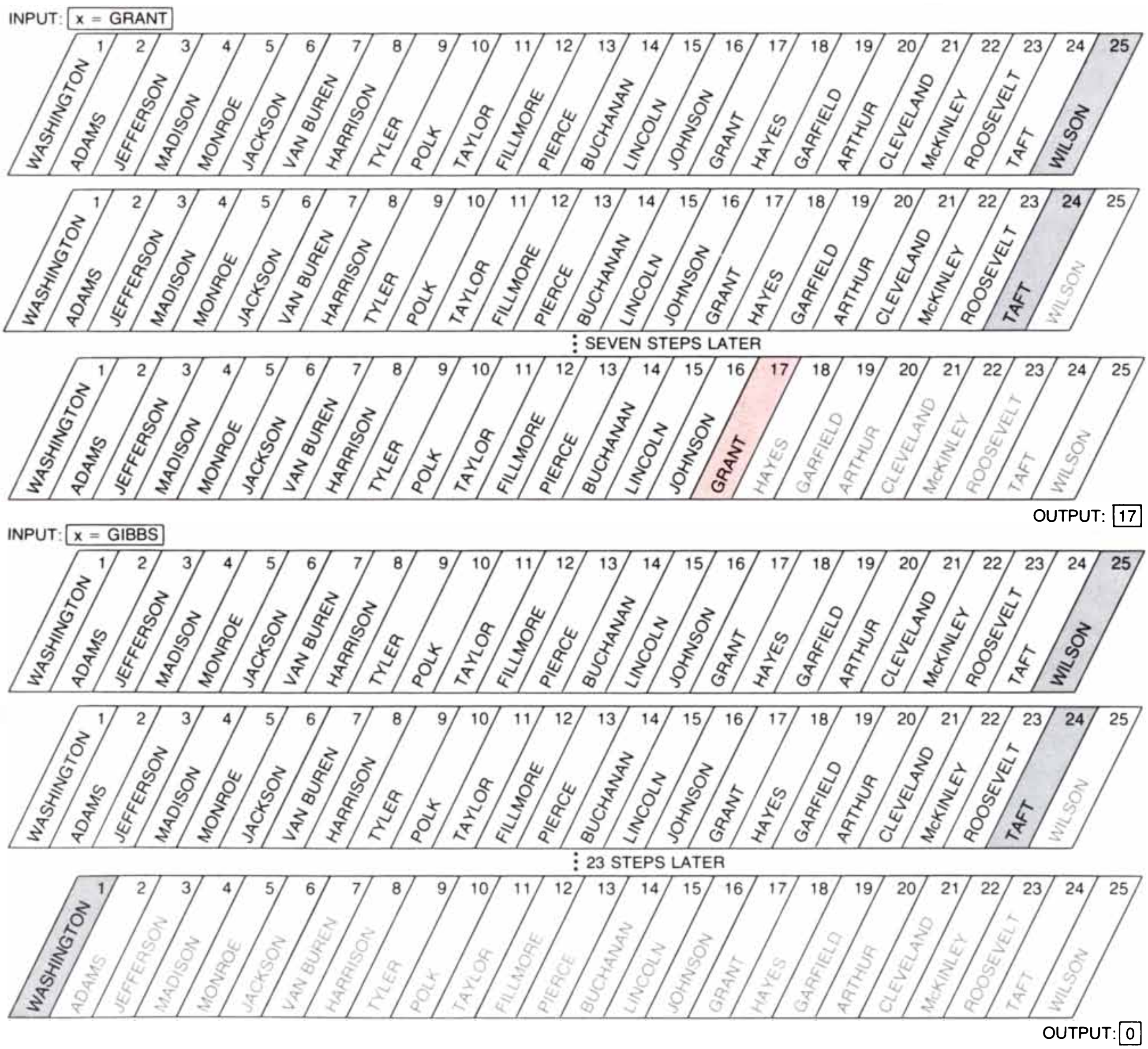
It is interesting to note that Algorithm A can be improved by giving meaning to the notation KEY[0], allowing a word to be stored in "location 0" as well as in locations 1 through $n$. Then if step A1 sets $\mathrm{KEY}[0] \leftarrow x$ as well as $j \leftarrow n$, step A2 can be eliminated and the search will go about 20 percent faster on many machines. Unfortunately for programmers, the most commonly used computer languages (standard FORTRAN and COBOL) do not allow 0 to be employed as an index for a memory location; thus Algorithm A cannot be so easily improved when it is expressed as a program in those languages.

Algorithm A certainly solves the problem of searching through a table of words, but the solution is not very good unless the number of words to be searched is quite small, say 25 or fewer. If *n* were as large as a million, a simple sequential search would usually be an unbearably slow way to look through the table. We would hardly go to the expense of building such a large table unless we expected to search it frequently, and we would not want to waste any time during the search. Algorithm A is the equivalent of looking for someone's telephone number by going through a telephone directory page by page, column by column, one line at a time. We can do better than that.

## The Advantage of Order

It is, in fact, instructive to consider a telephone directory as an example of such a large table of information. If one were asked to find the telephone number of someone who lives at 1642 East 56th Street, there would really be no better way than to do a sequential search equivalent to Algorithm A, since a standard telephone directory is not organized for searches according to address. On the other hand, when one looks up someone's name, it is possible to take advantage of alphabetical order. Alphabetical order is a substantial advantage indeed, since a single glance at almost any point in the directory suffices to eliminate many names from further consideration.

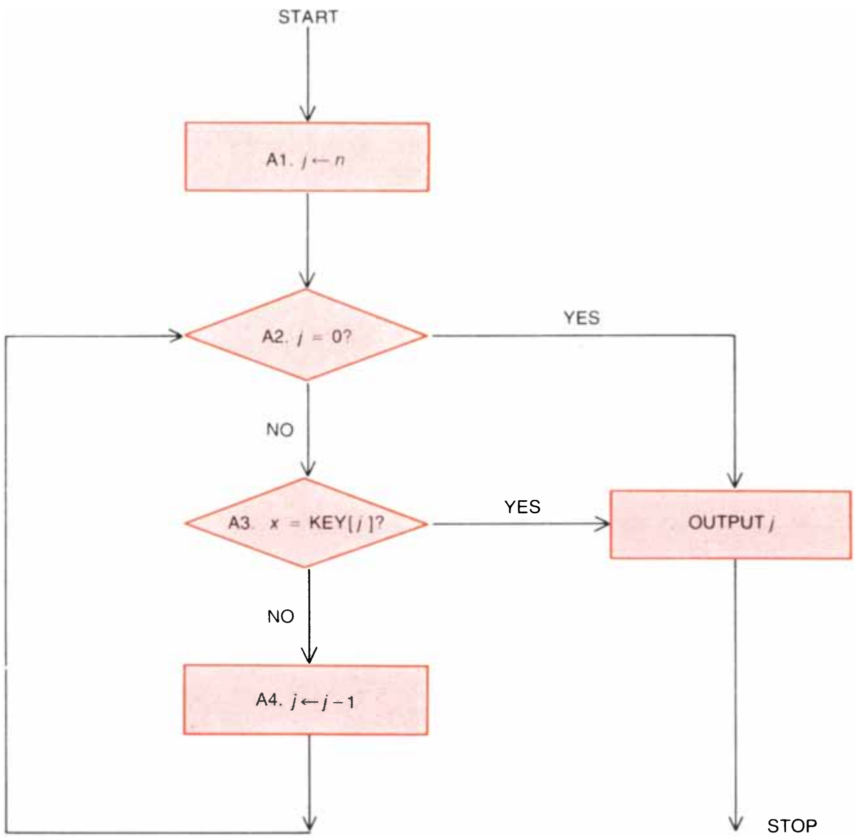
If the words of a table appear consistently in some order, there are several ways to design an efficient search procedure. The simplest procedure starts by looking first at the entry in the middle of




**SEQUENTIAL-SEARCH ALGORITHM (Algorithm A in the text of this article) looks for an input word in a table where the entries have not been arranged in any particular order. This table has 25 entries, or keys: KEY[1], KEY[2] and so on up to KEY[25]. Each key is a person's name. Suppose the input word is the name "Grant". Algorithm A searches for "Grant" by comparing it first with KEY[25], which is "Wilson", then with KEY[24], which is "Taft", and so on. Here "Grant" is found to be KEY[17], so that the algorithm outputs "17" (*top*). If input had been "Gibbs", Algorithm A would have compared "Gibbs" with all keys and output would have been 0 (*bottom*).**

the table. If the desired word $x$ is numerically or alphabetically less than this middle entry, the entire second half of the table can be eliminated; similarly, if $x$ is greater than the middle entry, one can eliminate the entire first half. Thus a single comparison yields a search problem that is only half as large as the original one. The same technique can now be applied to the remaining half of the table, and so on until the desired word $x$ is either located or proved to be absent. This procedure is commonly known as a binary search.

Although the ideas underlying binary search are simple, some care is necessary in writing the algorithm. First, in a table that has an even number of elements there is no unique "middle" entry. Second, it is not immediately clear when to stop in the case of an unsuccessful search. Teachers of computer science have noticed, in fact, that when students are asked to write a binary-search procedure for the first time, about 80 percent of them get the program wrong, even when they have had more than a year of programming experience! The reader who feels that he understands algorithms fairly well but has never before written a binary-search algorithm might enjoy trying to construct one before reading the following solution.




**FLOW CHART FOR ALGORITHM A illustrates the logical path by which the brute-force sequential search looks for an input word $x$ in a table of $n$ keys. The algorithm searches for $x$ by comparing it first with KEY[$n$], then with KEY[$n-1$], then with KEY[$n-2$] and so on. If $x$ matches some KEY[$j$], the algorithm outputs $j$, the location at which $x$ was found. If $x$ is not in the table, the output of the algorithm is 0. Arrow in step A1 ($j \leftarrow n$) means "Set $j$ equal to $n$" in that step. Step in each box is explained in detail in the fuller form of Algorithm A in third column of text on page 63. On the average Algorithm A must search half of the table to find $x$. In the worst case, if $x$ is at KEY[1] or if $x$ is not present, Algorithm A must search the entire table.**

Algorithm B; binary search. This algorithm employs the same notation as Algorithm A. Moreover, it is assumed that the first word, KEY[1], is less than the second word, KEY[2], which is less than the third word, KEY[3], and so on all the way up to the last word, KEY[$n$]. This condition can be written KEY[1] < KEY[2] < KEY[3] < ··· < KEY[$n$].

B1. [Initialize.] Set $l \leftarrow 0$, $r \leftarrow n + 1$. (The letter $l$ stands for the left boundary of the search and $r$ stands for the right boundary. More precisely, KEY[$j$] cannot match the given word $x$ unless the location $j$ is both greater than $l$ and less than $r$.)

B2. [Find midpoint.] Set $j \leftarrow \lfloor (l + r)/2 \rfloor$. (The brackets $\lfloor\ \rfloor$ mean "Round down to the nearest integer." Thus if $(l + r)$ is even, $j$ is set to $(l + r)/2$; if $(l + r)$ is odd, $j$ is set to $(l + r - 1)/2$.)

B3. [Unsuccessful?] If $j = l$, output 0 and terminate the algorithm. (If $j$ equals $l$, then $r$ must be equal to $l + 1$, since $r$ is always greater than $l$; therefore $x$ cannot match any key in the table.)

B4. [Compare.] (At this point $j > l$ and $j < r$.) If $x = \text{KEY}[j]$, output $j$ and terminate the algorithm. If $x < \text{KEY}[j]$, set $r \leftarrow j$ and return to step B2. If $x > \text{KEY}[j]$, set $l \leftarrow j$ and return to step B2.

A play-by-play account of Algorithm B as it searches through a table of 25 names is shown in the illustration on the next page.

It seems clear that binary search (Algorithm B) is much better than sequential search (Algorithm A), but how much better is it? And when is it better? A quantitative analysis will answer these questions.

## Quantitative Analysis

First let us analyze the worst cases of algorithms A and B. How long can it possibly take each algorithm to find word $x$ in a table of size $n$? The answer is easy for Algorithm A. If $x$ equals KEY[1], or if $x$ is not in the table at all, it will take $n$ executions of step A3; that is, the desired word $x$ must be compared with all $n$ entries in the table before the search stops. Furthermore, the algorithm will never execute step A3 more than $n$ times. When sequential search is applied to a table with a million entries, a million comparisons will be made in the worst case.

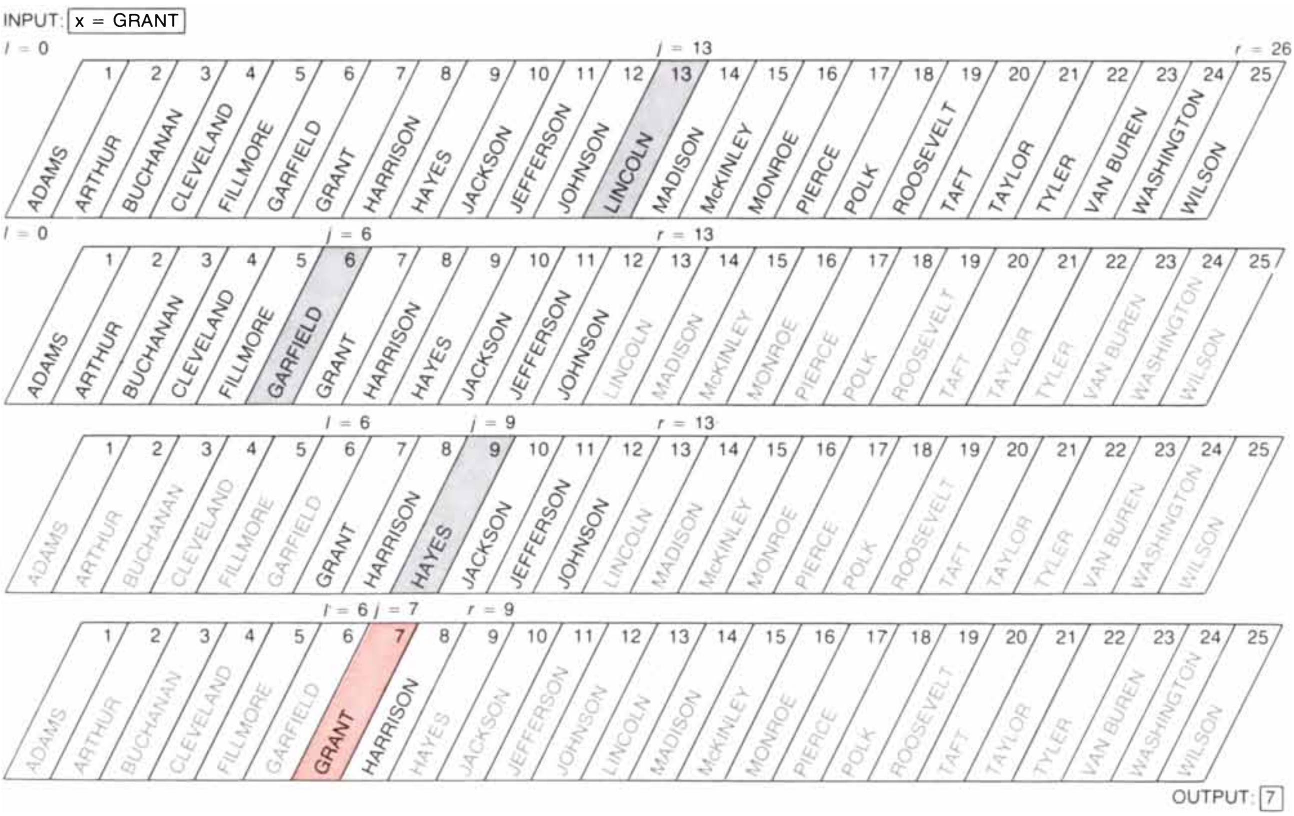
The answer is only slightly more difficult for binary search. Since Algorithm B discards half of the table remaining after each execution of step B4, it first deals with the entire table, then half of the table, then a quarter of the table, then an eighth of the table and so on. The maximum number of executions of step B4 will be $k$, where $k$ is the smallest integer such that $2^k$ is greater than $n$. For example, when binary search is applied to a table with a million ($10^6$) entries, $k$ will be equal to 20, since $2^{20}$ is greater than $10^6$ but $10^6$ is greater than $2^{19}$. Thus if a table with $10^6$ entries is searched using Algorithm B, at most only 20 of those entries will ever be examined in any particular search.

From the standpoint of worst-case behavior, one can go further and say that Algorithm B is not only a good way to search; it is actually the best possible search algorithm that proceeds solely by comparing $x$ to keys in the table. The reason is that a comparison-based algorithm cannot possibly examine more than $2^k - 1$ different keys during its first $k$ comparisons. No matter what strategy is adopted, the first comparison always selects a particular key of the table and the second comparison will be with at most two other keys (depending on whether $x$ was less than or greater than the first key); the third comparison will be with at most four other keys; the fourth comparison will be with at most eight other keys, and so on. Therefore if a comparison-based search algorithm makes no more than $k$ comparisons, the table can contain no more than $1+2+4+8+\cdots+2^{k-1}$ distinct keys, and this sum equals $2^k-1$.

The familiar game of Twenty Questions can be analyzed by reasoning in a similar way. In this game one player thinks of a secret object, the name of which he conceals on a folded piece of paper. The other players try to guess what the object is by asking as many as 20 questions that must be answered only by "Yes" or "No." The other players are also initially told whether the secret object is animal, vegetable or mineral, or a combination of those supposedly well-defined attributes. By arguing as I have in the preceding paragraph, one can prove that the other players cannot possibly identify more than $2^{23}$ different objects correctly, no matter how clever their questions are. There are only $2^3$ (or eight) possible subsets of the set of attributes animal, vegetable and mineral, and there are only $2^{20}$ possible outcomes of the 20 yes-no questions. Thus the total number of objects one can possibly identify is $2^{23}$. The argument holds even when each question asked depends on the answers to the preceding questions.

Stating this conclusion another way, if more than $2^{23}$ different objects must be identified, 20 questions will not always be enough. The search problem is similar but not quite the same, since an algorithm for searching does not simply ask yes-no questions. The questions asked by algorithms of the type we are considering have three possible outcomes, namely $x <$ KEY[$j$] or $x =$ KEY[$j$] or $x >$ KEY[$j$]. When a table contains $2^k$ or more entries, the above reasoning proves that $k$ comparisons of $x$ with keys in the table will not always be enough. Therefore every algorithm that searches a table of a million words by making comparisons must in some instances examine 20 or more of those words. In short, binary search has the best possible worst case.

The worst-case behavior of an algorithm is not the whole story, since it is overly pessimistic to base decisions entirely on one's knowledge of the worst that can happen. A more meaningful understanding of the relative merits of algorithms A and B can be gained by analyzing their average-case behavior. If each of the $n$ keys in a table is equally likely to be looked up, what is the average number of comparisons that will be needed? For sequential search (Algorithm A) the answer is the simple average $(1+2+3+\cdots+n)/n$, which is equal to $(n+1)/2$. In other words, to find $x$ with Algorithm A one will on the average have to search through about half of the table. To determine the average number of comparisons needed to find $x$ using binary search (Algorithm B), the mathematics is only a little more complicated. In this case the answer is $k-[(2^k-k-1)/n]$, where $k$, as before, is the number of comparisons required in the worst case. For large values of $n$ this answer is approximately equal to $k-1$; therefore the average case of Algorithm B is only about one comparison less than its worst case. By carefully extending the argument made earlier it is possible to show that binary




**BINARY-SEARCH ALGORITHM (Algorithm B in the text) is a substantial improvement over the sequential-search algorithm when the table to be searched is large. The entries in the table must first be arranged in order. Here the 25 names are listed in alphabetical order. Again the input word $x$ sought is "Grant". The algorithm compares "Grant" first with the key in the middle location, $j$, of the table. It calculates the initial value of $j$ by setting the left boundary $l$ of the search at 0 and the right boundary $r$ at $n + 1$. In this case $r$ is 26. Then $l$ and $r$ are added together and divided by 2, rounding down to the nearest integer if the answer is not already an integer. The midpoint $j$ of the table is 26/2, or 13, which is the location of "Lincoln" (*top*). Since the name "Grant" is alphabetically less than "Lincoln", the algorithm discards the entire right half of the table, containing all names alphabetically greater than or equal to "Lincoln". For the remaining half of the table the algorithm calculates a new midpoint, first setting $r$ equal to the location $j$ just examined, which is 13 (*second from top*). The new midpoint $j$ is (0 + 13)/2, which must be rounded down to 6, location of "Garfield". "Grant" is alphabetically greater than "Garfield", so that the left quarter of the table is discarded and the left boundary $l$ is set equal to 6 (*second from bottom*). When procedure is repeated once more, "Grant" is found in position 7 (*bottom*). If input word $x$ had been "Gibbs", Algorithm B would have executed one more step, with $l$ still equal to 6 and $r$ set at 7. Midpoint $j$ would have been 6, which is left boundary of search, meaning that "Gibbs" is not in table.**

*The legend continues...*

# Introducing the world's most sophisticated Diesel passenger car. The new Mercedes-Benz 300D.


*The new, 5-passenger Mercedes-Benz 300D—the state of the Diesel passenger car art.*

**Here is a most ingenious alternative to the conventional automobile. A truly remarkable new Mercedes-Benz. With a contemporary new look. With ample room for five people, an astonishing 5-cylinder engine and an unusually complete array of luxurious appointments and safety systems. The new 300D. The most sophisticated Diesel passenger car the world has ever seen.**

For years, you've heard about exotic and promising alternatives to the conventional automobile engine. To date, only one alternative has kept its promise: the Diesel engine—for over 60 years, the most efficient combustion power plant in use.

Now Mercedes-Benz has synthesized its proven, 5-cylinder Diesel engine with new, technologically advanced body design, suspension, steering and safety systems to produce the most ingenious alternative to the conventional automobile.

### A matter of taste

Though only a trim 190.9 inches from bumper to bumper, the new 300D is an honest 5-passenger sedan. The secret of its spaciousness lies in new Mercedes-Benz technology that puts the room in the car *in* the car—without adding bulk or sacrificing safety.

Enter a new 300D and you're surrounded by a complete array of security and convenience features. All are standard equipment. Such things as cruise control, bi-level climate control, electric windows, AM/FM radio, central locking system, 3-speed windshield wipers.

The new 300D is not an exercise in opulence. But it does exhibit meticulous taste. And as your senses will tell you, there's quite a difference between the two ideas.

### Sports car handling

The new 300D is one of the most sparkling road cars Mercedes-Benz has ever engineered. Its sophisticated power train, suspension and steering are those of a sports car. And that is why the new 300D handles like one.

The new 300D's unique, 5-cylinder engine is the most powerful, the smoothest Diesel yet engineered into a passenger car. But you pay no penalty for this performance bonus. The EPA estimates that the new 300D should deliver up to 28 mpg on the highway, 23 mpg in town. (Your mileage will depend on how and where you drive and the condition and equipment of your car.)

### The state of the art

For over 40 years, Mercedes-Benz has pioneered many of the major advances in Diesel passenger car engineering. The new 300D is the culmination of that experience. It is the state of the art Diesel passenger car art.

Test drive the new 300D. Experience the most ingenious alternative to the conventional automobile. The most sophisticated Diesel passenger car in the world.

**Mercedes-Benz**

**Engineered like no other car in the world.**


# The Small Computer

Twenty-five years ago a computer as powerful as the new Processor Technology Sol-20 priced out at a cool million.

Now for only $995 in kit form or $1495 fully assembled and tested you can have your own small computer with perhaps even more power. It comes in a package about the size of a typewriter. And there's nothing like it on the market today. Not from IBM, Burroughs, DEC, HP or anybody else!

## It fills a new role

If you're an engineer, scientist or businessman, the Sol-20 can help you solve many or all of your design problems, help you quantify research, and handle the books too. For not much more than the price of a good calculator, you can have high level computer power.

## Use it in the office, lab, plant or home

Sol-20 is a smart terminal for distributed processing. Sol-20 is a stand alone computer for data collection, handling and analysis. Sol-20 is a text editor. In fact, Sol-20 is the key element of a full fledged computer system including hardware, software and peripheral gear. It's a computer system with a keyboard, extra memory, I/O interfaces, factory backup, service notes, users group.

It's a computer you can take home after hours to play or create sophisticated games, do your personal books and taxes, and a whole host of other tasks.

Those of you who are familiar with small computers will recognize what an advance the Sol-20 is.

Sol-20 offers all these features as standard:

8080 microprocessor — 1024 character video display circuitry — control PROM memory — 1024 words of static low-power RAM — 1024 words of preprogrammed PROM — built-in cassette interface capable of controlling two recorders at 1200 bits per second — both parallel and serial standardized interface connectors — a complete power supply including ultra quiet fan — a beautiful case with solid walnut sides — software which includes a preprogrammed PROM personality module and a data cassette with BASIC-5 language plus two sophisticated computer video games — the ability to work with all S-100 bus products.

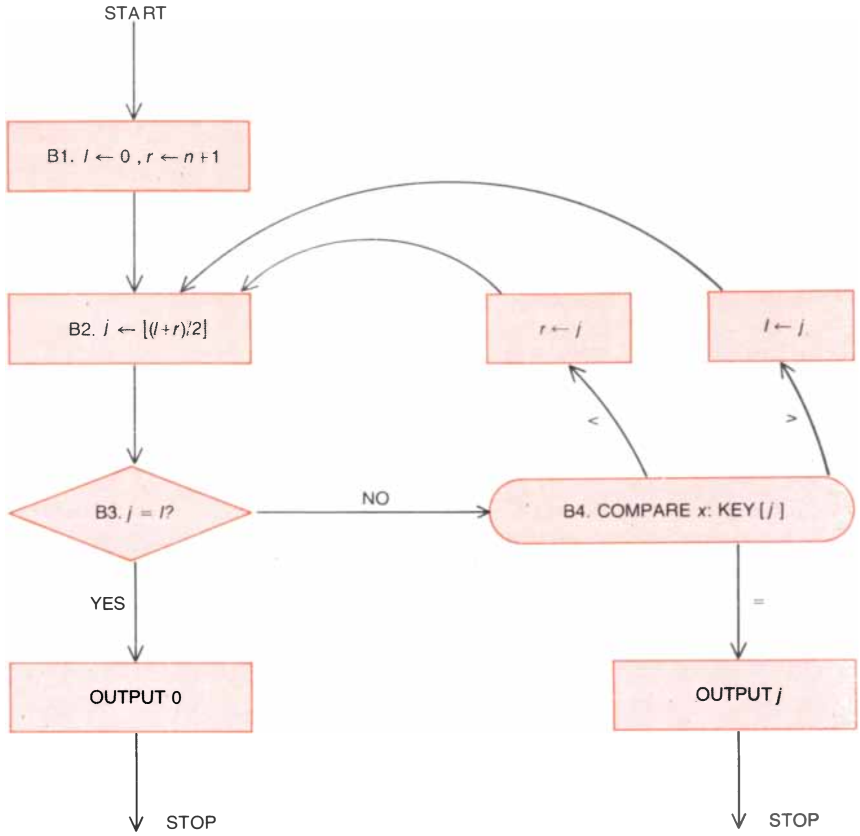
## Full expansion capability

Tailor the Sol-20 system to your applications with our complete line of peripheral products. These include the video monitor, audio cassette and digital tape systems, dual floppy disc system, expansion memories, and interfaces.

## Write for our new 22 page catalog. Get all the details.

Processor Technology, Box N, 6200 Hollis St., Emeryville, CA 94608. (415) 652-8080.

Processor Technology Corporation




**FLOW CHART OF ALGORITHM B illustrates the rules governing binary search. The algorithm searches for the input word $x$ in a table of $n$ keys that have previously been arranged in order. First $x$ is compared with the middle entry of the table. If $x$ is greater than ( > ) the middle entry, it is compared with the midpoint of the right half of the table. If $x$ is less than ( < ) the middle entry, it is compared with the midpoint of the left half of the table. The process continues, with half of the remaining table being discarded each time, until either $x$ is found or the search reveals that $x$ is not in the table. The half brackets (⌊ ⌋) mean "Round down to the nearest integer." Like Algorithm A, Algorithm B is written out in detail in first column of text on page 65.**

search is also the best possible algorithm from the standpoint of the average case: every search algorithm must make at least $k - [(2^k - k - 1)/n]$ comparisons on the average, and many do worse.

## Better than the Best

As soon as something has been proved impossible, a lot of people immediately try to do it anyway. This seems to be an inherent component of human behavior. I have just proved that binary search is the best possible way to search a computer's memory, and so naturally I shall now look for a better way.

In the first place, when the number of words in a table is small, Algorithm A actually turns out to be better than Algorithm B. Why does this not contradict the proof that binary search is best? The reason is that in comparing Algorithm A and Algorithm B I have so far been contrasting only the number of comparisons each algorithm makes. Actually Algorithm A requires less bookkeeping activity, so that it takes less time for a machine to execute each comparison. On a typical computer Algorithm A can be made to take about $2n + 6$ units of time, on the average, for a table of size $n$. Algorithm B, on the other hand, will require an average of about $12 \log_2 n - 11 + 12(k + 1)/n$ units of time, under the same assumptions. Thus unless there are 20 or more keys to be searched, Algorithm A will be better than Algorithm B. These numbers will vary slightly from computer to computer, but they show that the efficiency of an algorithm cannot be determined by counting only the comparisons made.

There is another reason Algorithm B can be beaten. When we look up someone's name in a telephone directory and compare the desired name $x$ with the names on a page, our subsequent action is not influenced solely by whether the comparison shows that the desired name is alphabetically less than or greater than the names on the page; we also observe how much less than or how much greater than, and we turn over a larger chunk of pages if we think we are farther from the goal. The above proof that binary search is best does not apply to algorithms that make use of such things as the degree of difference between $x$ and a particular key. The proof for Twenty Questions can be attacked on similar grounds. For example, the players might notice the length of the secret word as it is being written down, or they might be able to gain information from the length of time the player being questioned hesitates before answering "Yes" or "No."

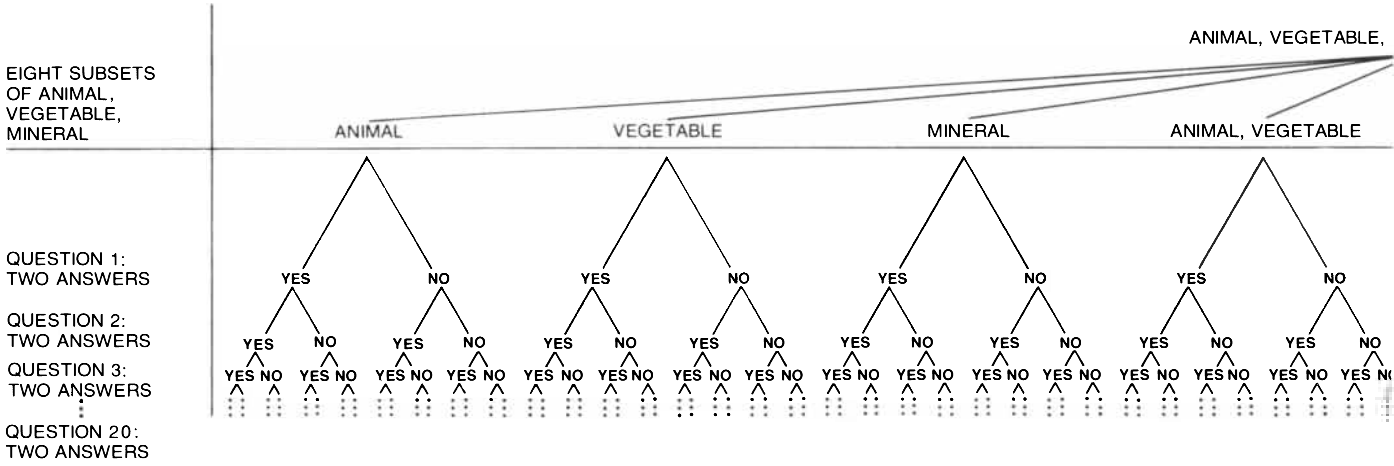
Therefore a human being concerned about efficiency need not begin searching a telephone directory by bisecting it as a computer would; the time-honored method of interpolation with the aid of alphabetical order probably works better in spite of the proof that the binary search is best. In fact, Andrew C. Yao of the Massachusetts Institute of Technology and F. Frances Yao of Brown University have recently shown that the average number of times an interpolation-search algorithm needs to access the table is only $\log_2 \log_2 n$ plus at most a small constant, provided that the table entries are independent and uniformly distributed random numbers. When $n$ is very large, $\log_2 \log_2 n$ is much smaller than $\log_2 n$, so that interpolation search will be significantly faster than binary search. The idea underlying the Yaos' proof is that each iteration of an interpolation search tends to reduce the uncertainty of the position of $x$ from $n$ to the square root of $n$. Furthermore, they have proved that interpolation search is nearly the best possible, in a very broad sense: any algorithm that searches such a random table by making appropriate comparisons must examine approximately $\log_2 \log_2 n$ entries, on the average.

These results are of great theoretical importance, although computational experience has shown that an interpolation search is usually not an improvement over binary search in practice. The reason is that the data stored in a table are typically not random enough to conform to the assumption of a uniform distribution; in addition $n$ is typically small enough so that the extra calculation per comparison required by each interpolation outweighs the amount of time saved by reducing the number of comparisons. The simplicity of binary search is one of its virtues, and it is important to maintain a proper balance between theory and practice.

## Binary Tree Search

The binary search can be improved, however, in another way: by dropping the assumption that every key in the table is equally likely to be sought. When some keys are known to be far more likely candidates than others, an efficient algorithm will examine the more likely ones first.

Before we explore this notion it will be helpful to look first at the binary search in a different way. Consider the 31 words that are used most frequently in the English language (according to Helen Fouché Gaines in her book




**THE GAME TWENTY QUESTIONS demonstrates a fundamental limitation on the power of any branching-search method. In the game one player thinks of an object, which he describes as being animal, vegetable or mineral, or any combination of those characteristics. The opposing players try to guess what the object is by asking as many as 20 questions, which must be answered "Yes" or "No." It can be proved that the players cannot identify more than $2^{23}$, or 8,388,608, objects correctly. The reason is that the set of characteristics animal, vegeta-**

*Cryptanalysis*). When these words are arranged alphabetically in the locations KEY[1], KEY[2], KEY[3], . . . ,KEY[31] of a table, Algorithm B first compares the desired word $x$ to the midpoint KEY[16], which is the word "I". If $x$ is alphabetically less than "I", the next comparison will be with KEY[8], which is the word "by"; if $x$ is greater than "I", the next comparison will be with KEY[24], which is "that". In other words, Algorithm B acts on the table of words by following a structure that looks like an upside-down tree, starting at the top and going down to the left when $x$ is less and down to the right when $x$ is greater [*see top illustration on page 72*]. It is not hard to see that any algorithm designed to search an ordered table purely by making comparisons can be described by a similar binary tree.

The tree for binary search is defined implicitly in Algorithm B by arithmetic operations on $l$, $r$ and $j$. It can also be defined explicitly by storing the tree information in the table of words itself. For this purpose let LEFT[$j$] be the position in the table at which we are to look if word $x$ is less than KEY[$j$], and let RIGHT[$j$] be the position at which we are to look if $x$ is greater than KEY[$j$]. For example, binary search in a table of 31 words would have LEFT[16] equal to 8 and RIGHT[16] equal to 24, since the search starts at KEY[16] and then proceeds to either KEY[8] or KEY[24]. If the search is to terminate unsuccessfully after determining that the desired word $x$ is less than KEY[$j$] or greater than KEY[$j$], we respectively let LEFT[$j$] equal 0 or RIGHT[$j$] equal 0. In the illustrations on page 72 those 0's are represented by little square nodes at the bottom of the tree.

The location of the first key to be examined in a binary tree is traditionally known as the root; in the 31-word example the root is 16. It is possible to construct search algorithms that do not start by looking at KEY[16], and these may well be more efficient than Algorithm B if some words are looked up much more often than others. A generalized tree-search procedure follows:

Algorithm C; tree search.
C1. [Initialize.] Set $j$ equal to the location of the root of the binary search tree.
C2. [Unsuccessful?] If $j = 0$, output $j$ and terminate the algorithm.
C3. [Compare.] If $x$ = KEY[$j$], output $j$ and terminate the algorithm. If $x <$ KEY[$j$], set $j \leftarrow$ LEFT[$j$] and go back to step C2. If $x >$ KEY[$j$], set $j \leftarrow$ RIGHT[$j$] and go back to step C2.

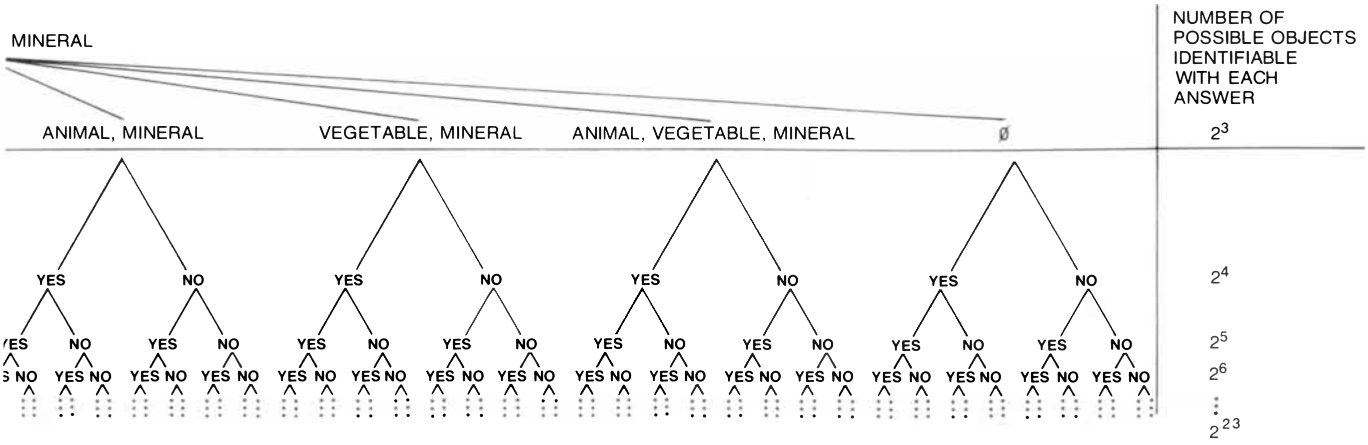
Algorithm C is analogous to a programmed textbook in which, depending on the answer to a certain question, each page tells the reader what page to turn to next. It works on any binary tree where all keys accessible from LEFT[$j$] are less than KEY[$j$] and all keys accessible from RIGHT[$j$] are greater than KEY[$j$], for all locations $j$ in the tree. Such a tree is called a binary search tree.

One of the advantages of Algorithm C over Algorithm B is that no arithmetic calculation is necessary, so that the search goes slightly faster on a computer. The main advantage of Algorithm C, however, is that the tree structure provides extra flexibility because the entries in the table can now be rearranged into any order. It is no longer necessary that KEY[1] be less than KEY[2] and so on up to KEY[$n$]. As long as the pointers LEFT and RIGHT define a valid binary search tree, the actual locations of the keys in the table are irrelevant. This means that one can add new entries to the table without moving all the other entries. For example, the word "has" could be added to the 31-word binary search tree simply by setting KEY[32] $\leftarrow$ "has" and changing RIGHT[$j$] from 0 to 32, where $j$ is the location of the key "had". One might think that such additions at the bottom of the tree would upset the balanced structure, but it can be shown mathematically that if new entries are added in random order, the result will almost surely be a reasonably well-balanced tree.

## Optimum Binary Search Trees

Since Algorithm C applies to any binary search tree, one can hand-tailor the tree so that the most frequently examined keys are examined first. Such tailoring reduces the average time required for a computer to carry out the search, although it cannot reduce the worst-case time. The bottom illustration on page 72 shows the best possible binary search tree for the 31 commonest English words, based on Gaines's estimates of their frequency. The average number of comparisons needed to search for $x$ in this optimum binary search tree is only 3.437, whereas the average number of comparisons needed in the balanced binary search tree is 4.393. It is worth noting that the optimum tree, which is based on the frequencies of the words, does not start by comparing $x$ with the word "the". Even though "the" is by far the commonest English word, it comes so late in alphabetical order that it is too far from the middle of the list to serve as the optimum root.

From the standpoint of conventional mathematics it is trivial to find the optimum binary tree for any particular set of $n$ words and frequencies because there are only finitely many search trees. In principle one merely has to list all the trees and choose the one that works best. In practice, however, this observation is useless because the number of possible binary trees with $n$ elements is equal to $(2n)!/n!(n+1)!$, where $n!$ stands for the product $1 \times 2 \times 3 \times \cdots \times n$. This for-




**ble and mineral has only eight, or $2^3$, possible subsets (including the null set Ø for an object with none of those characteristics), and these eight possibilities combine with only $2^{20}$ possible outcomes to the 20 yes-no questions. A similar argument can be used to show that a search algorithm asking at most 20 "less-equal-greater" questions cannot distinguish more than $2^{20} - 1$ different key values, since $1 + 2 + 4 + 8 + \dots + 2^{19} = 2^{20} - 1$. Binary search is able to attain this maximum limit, thus it is the most efficient search algorithm of its kind.**

mula shows that there are very many binary trees indeed, approximately $4^n/\sqrt{(\pi n^3)}$ of them, where $\pi$ is the familiar 3.14159. For example, when $n$ is 31, the total number of possible binary trees is 14,544,636,039,226,909, and each of these 14 quadrillion trees will be optimum for some set of assumed frequencies for the 31 words. How, then, is it possible to show that the particular tree I have chosen is the best one for Gaines's frequencies? The fastest modern computer is far from fast enough to examine 14 quadrillion individual possibilities; if one tree were considered per microsecond, the task would take 460 years.

There is, however, an important principle that does make the computation feasible: Every subtree of an optimum tree must also be optimum. In the optimum binary search tree for the 31 commonest English words the subtree to the left of the word "of" must represent the best possible way to search for the 20 words "a", "and" and so on over to "not". If there were a better way, it would lead to a better overall tree, and the given tree would therefore not be optimum. Similarly, in that subtree the even smaller subtree to the right of "for" must represent the best possible way to search for the 11 words "from", "had" and so on over to "not". Each subtree corresponds to a set of consecutive words KEY[$i$], KEY[$i + 1$], . . . ,KEY[$j$], where $1 \leq i < j \leq n$. It is possible to determine all the optimum subtrees by finding the small ones first and doing the computation in order of increasing values of $j - i$. For each choice of indices $i$ and $j$ there are $j - i + 1$ possible roots of the subtree. As one proceeds up the tree with the computation and examines each possible subtree root the optimum subtrees to the left and right will have already been calculated.

By this procedure the best possible binary search tree for $n$ keys and frequencies can actually be found by doing about $n^3$ operations. In fact, I have been able to improve the method even further, so that the number of operations required can be reduced to $n^2$. In the case of the 31 commonest words this means that the optimum binary search tree can be discovered after only 961 steps instead of 14 quadrillion.

I should point out that the preceding paragraphs discuss several algorithms whose sole purpose is to determine the best binary search tree. In other words, the output of those algorithms is itself an algorithm for solving another problem! This example helps to explain why computer science has been developing so rapidly as an independent discipline. In the study of how to use computers properly, problems arise that are interesting in their own right, and many of these problems require both a new and interrelated set of concepts and techniques.

It is amusing and instructive to consider the worst possible binary search tree for the 31 commonest English words in order to see how bad things could possibly become with Algorithm C. As in the case of the optimal trees, there is a way to determine such "pessimal" trees in about $n^2$ operations. For the 31 words with Gaines's frequencies the pessimal binary search tree requires Algorithm C to make an average of 19.158 comparisons per search. By way of comparison the worst arrangement of the keys for a sequential search requires Algorithm A to make an average of 22.907 comparisons per search. Hence even the worst case for Algorithm C can never be quite as bad as the worst case for Algorithm A.

## Hashing

The above algorithms for searching are closely related to the way people look for words in a dictionary. There is actually a much better way to search through a large collection of words by computer. It is called hashing, and it is a completely different approach that is quite unsuitable for human use because it is based on a machine's ability to do arithmetic at high speeds. The idea is to treat the letters of words as if they were numbers ($a = 1$, $b = 2$, $c = 3$ and so on through $z = 26$) and then to hash, or scramble, the numbers in some way in order to get a single number for each word. The number is the "hash address" of the word; it tells the computer where to look for the word in the table.

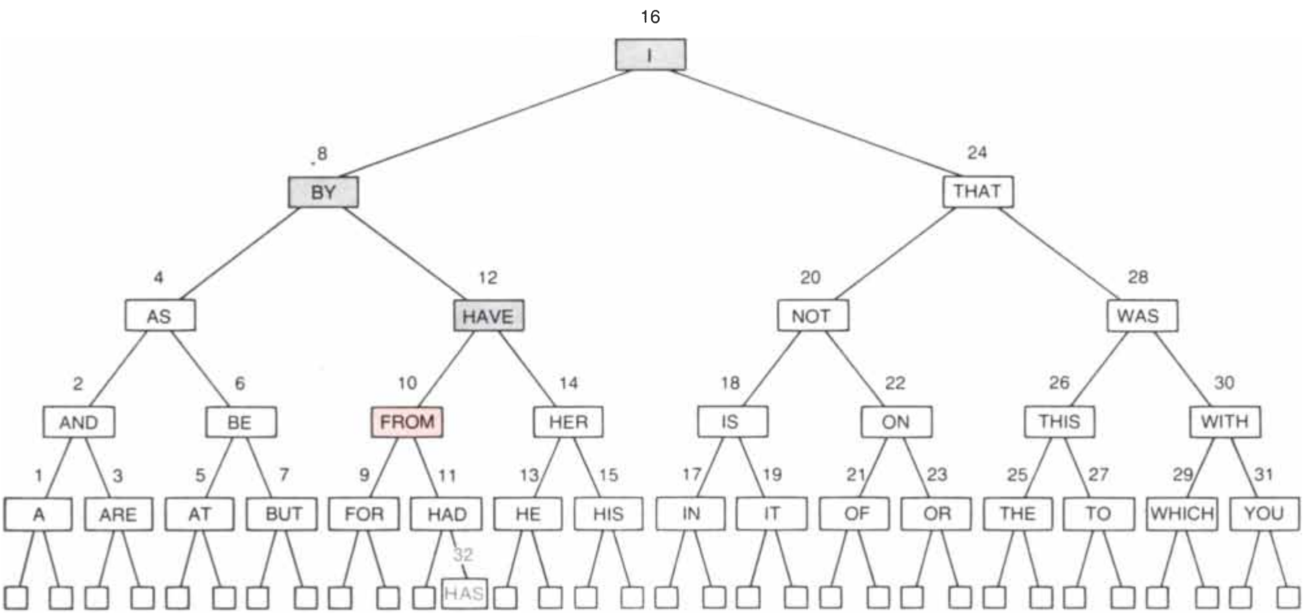
In the case of the 31 commonest English words we could convert each key into a number between 1 and 32 by adding up the numerical values of its letters and throwing away excess multiples of 32. For example, the hash address of "the" would be $20 + 8 + 5 - 32 = 1$, the hash address of "of" would be $15 + 6 = 21$, and so on for the rest of the list. If one is lucky, each word will lead to a different hash address and any search will be very fast.

In general, suppose there are $m$ locations in the computer's memory, and suppose we want to store $n$ keys, where $m$ is greater than $n$. Since $n$ is equal to 31, let us say $m$ is equal to 32. Suppose also there is a hash function $h(x)$ that converts each possible word $x$ into a number between 1 and $m$. A good hash function will have the property that $h(x)$ is unlikely to be equal to $h(y)$, if $x$ and $y$ are different words to be put into the table.

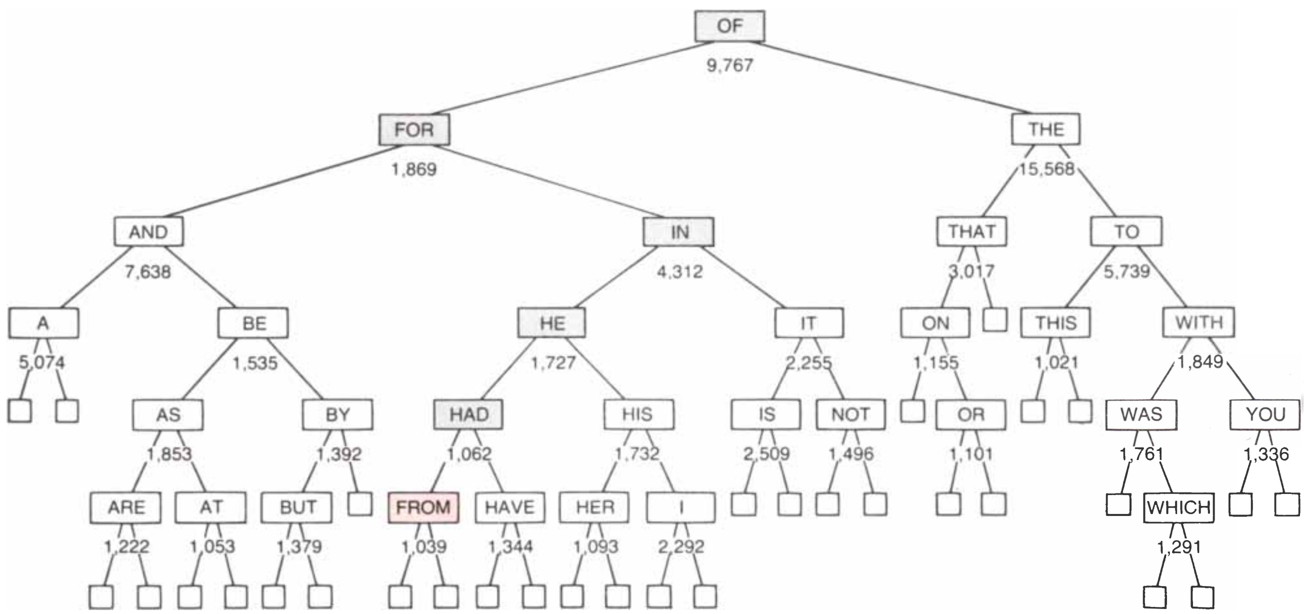
Unless $m$ is much larger than $n$, however, nearly every hash function will lead to at least a few "collisions" between the values $h(x)$ and $h(y)$. It is extremely improbable that $n$ independent random numbers between 1 and $m$ will all be different. Consider a common example: It is well known that when 23 or more people are present in the same room there is a better than even chance that two of them will have the same birthday. Moreover, in a group of 88 people it is likely that there will be three individuals with the same birthday. Although this phenomenon seems paradoxical to many people, the mathematics can be easily checked, and many seemingly impossible coincidences can be explained in the same way.

Another way to state the birthday paradox is to say that a hash function with $m$ equal to 365 and $n$ equal to 23 will have at least one collision, more often than not. Thus any search procedure based on a hash function must be able to deal with the problem of collisions.

Suppose we want to search a table for $x$ but the hash address $h(x)$ already contains word $y$. The simplest way to handle the collision is to search through locations $h(x)$, $h(x) - 1$, $h(x) - 2$ and so on until we either find $x$ or come to an empty position. If the search runs off one end




**BINARY SEARCH TREE is implicit in Algorithm B. Here a tree graphically illustrates the order in which Algorithm B would probe an alphabetical table of the 31 commonest words in English. Starting at the "root," or top, of the tree, the input word $x$ is first compared with the midpoint of the table, the word "I". If $x$ is alphabetically smaller than "I", the search proceeds down the left branch of the tree; if $x$ is greater, the search proceeds down the right branch. For example, if $x$ is the word "from", the search first finds that $x$ is less than "I", then that $x$ is greater than "by", then that $x$ is less than "have", finally that $x$ is equal to "from". If $x$ were not in the table, the search would stop at one of the 32 zeros (*square nodes*) at bottom of the tree. When branches of tree are represented explicitly in computer's memory, rather than implicitly as in Algorithm B (which requires calculation of midpoints), search goes slightly faster. It also becomes easier to insert new information: if one wants to add "has" (*word in gray*) to tree, one inserts it in alphabetical order in place of one of zeros.**




**OPTIMUM BINARY SEARCH TREE shows the best order of the 31 words in the tree, based on the relative frequency of each word estimated by Helen Fouché Gaines. The frequency of each word is represented by the number below it. This tree is not as well balanced as the tree implicitly defined by the standard binary-search algorithm and shown in illustration above, and the search will therefore take longer in some cases. For example, to find the word "from" in this tree takes six steps instead of four (*path in gray and color*). On the average the optimum tree is faster for a computer to search, however, because the commoner words are tested sooner. Note that although the word "the" is by far the most frequently used word in English, it is not placed at root of the tree because it is too far from center of alphabet.**

*inal prints. Furthermore, since the Polaroid prints and negatives are produced in only 30 seconds, the original photographs can lie, untouched, under the camera until the archivist knows he has satisfactory copies. The irreplaceable originals can then be removed to safe storage under optimum conditions with no further handling risks. Thus, in a manner of speaking, some of the great pioneers in the history of British photography will owe their lasting reputations to a film made by* Polaroid Corporation.

REPRODUCED FROM A POLAROID TYPE 665 NEGATIVE (3¼ X 4¼)

IMPERS.

ATINUM PRINT)

*nry Peach Robinson*

# The new $2984* Colt. Isn't a Datsun. Isn't a Toyota. It's a lot of little Dodge.

**The new Dodge Colt is such a lot of car,** it's got Mr. D. and Mr. T. confused. Because Colt offers you the value you'd expect from this import, plus Dodge Colt sales and service coast to coast.

**47 MPG highway, 30 MPG city.** Colt will give you great mileage, according to EPA estimates.** Your mileage may vary, according to your car's condition, equipment, and your driving habits. And Dodge Colt runs on either regular or unleaded gas.

**Looking for a long list of standard features?** Well, you get it on all of our '77 Colt models. Even our lowest priced two-door coupe gives you whitewall tires, two reclining bucket seats, tinted glass in all windows, carpeting, adjustable steering column, simulated wood-grained instrument panel, four-speed manual transmission, quiet sound insulation, trip odometer, locking gas cap, and electric rear window defroster.

And we offer you an optional automatic transmission to go with the standard 1.6 liter engine.

**So if you're thinking "import," think about Dodge Colt.** It's not a Datsun. Not a Toyota. For $2984, it's a lot of little Dodge. See it at your Dodge Colt Dealer's.

*Manufacturer's suggested retail price, not including destination charge, taxes, title, and options. California prices higher.

**Equipped with standard 1.6 liter engine, four-speed manual transmission, and 3.31 rear axle ratio. California mileage lower.


Dodge

A PRODUCT OF CHRYSLER CORPORATION

**Dodge Colt. It's a lot of little Dodge.**

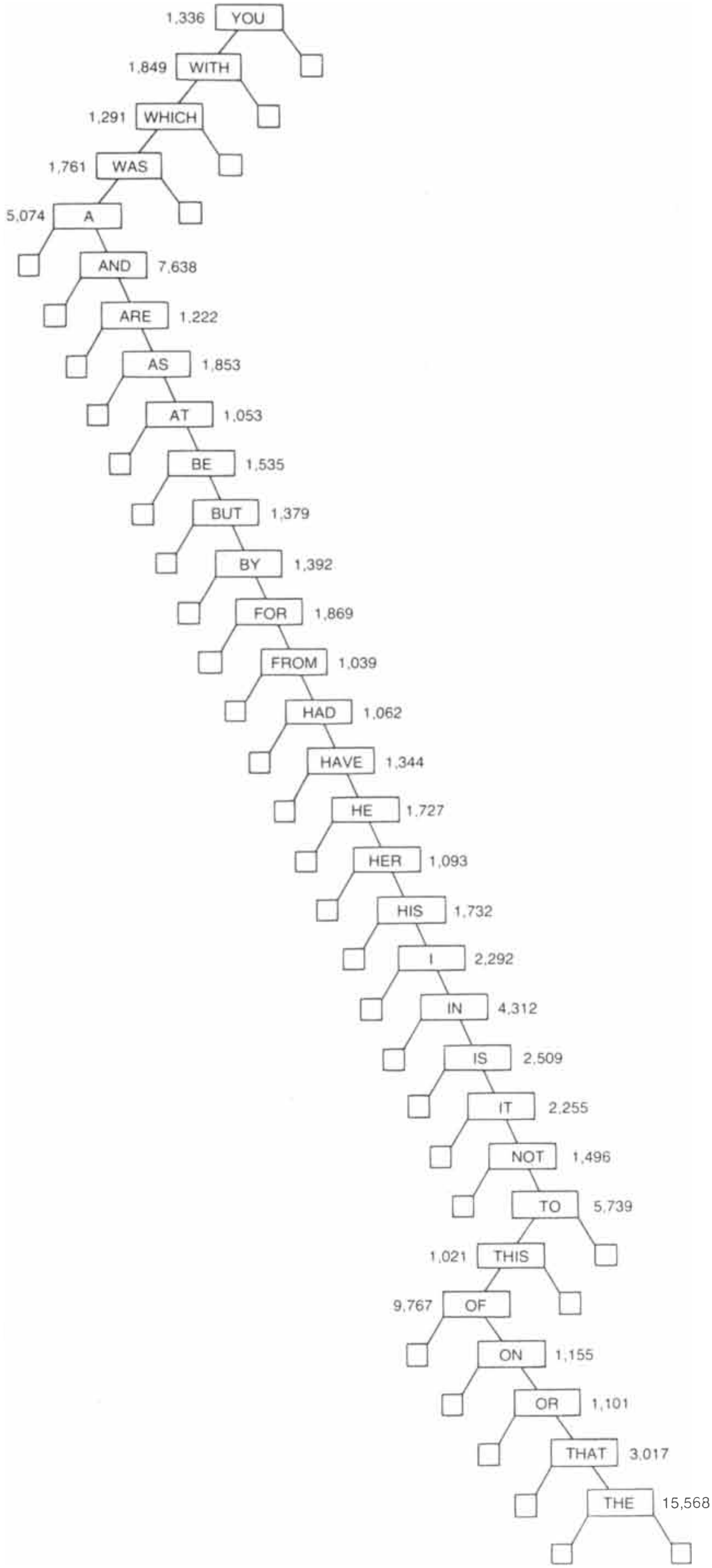
of the table before it is completed, it resumes at the other end. This procedure, which is known as linear probing, can be spelled out in an algorithm:

Algorithm D; hashing with linear probing.
D1. [Initialize.] Set $j \leftarrow h(x)$.
D2. [Unsuccessful?] If table entry $j$ is empty, output 0 and terminate the algorithm.
D3. [Successful?] If $x = \mathrm{KEY}[j]$, output $j$ and terminate the algorithm.
D4. [Move to next location.] Set $j \leftarrow j - 1$; then if $j = 0$, set $j \leftarrow m$. (Location $m$ is considered to be next to location 1.) Return to step D2.

If $x$ is not in the table, and if the algorithm terminates unsuccessfully in step D2 because table entry $j$ is empty, we could set $\mathrm{KEY}[j] \leftarrow x$ using the current value of $j$. This would insert $x$ into the table so that the algorithm could retrieve it later. A subsequent search for $x$ will follow the same path as it did the first time, starting at position $h(x)$, moving to $h(x) - 1$ and so on, finding $x$ in position $j$. Thus the search will proceed properly even when collisions occur.

Returning to the example of the 31 commonest English words, suppose the words are inserted one by one into an initially empty table in decreasing order of their frequency ("the" is inserted first, "of" is inserted second and so on). The result is the hash table shown in the illustration on the next page. Most of the words appear at or near their hash addresses except for the ones that are inserted into the table last; the least frequent word, "this", has been placed in position 8 although its hash address is 24, because positions 9 through 24 were already filled by the time it was inserted. In spite of such anomalies the average number of times the table must be probed by Algorithm D in order to find a word turns out to be only 1.666—less than half the average number of comparisons required to find the word with the optimum binary search tree. Of course, the time needed to compute $h(x)$ in step D1 must be added to the time for probing the table. For large collections of data, however, the hashing method will significantly outperform any binary-comparison algorithm.

In practice one would almost never let a hash table get as full as it does in the example. The number $m$ of table positions available is usually chosen to be so large that the table will never become more than 80 or 90 percent full. It can be shown that the average number of




**"PESSIMAL" TREE shows the worst possible binary search tree for searching for the 31 commonest English words. This tree has lost advantage of tree structure because one branch of each comparison is always "dead."**

| | | |
|---|---|---|
| 1 | THE | (1) |
| 2 | HAVE | (4) |
| 3 | TO | (3) |
| 4 | HIS | (4) |
| 5 | | |
| 6 | BE | (7) |
| 7 | FOR | (7) |
| 8 | THIS | (24) |
| 9 | I | (9) |
| 10 | BUT | (11) |
| 11 | WAS | (11) |
| 12 | HAD | (13) |
| 13 | HE | (13) |
| 14 | FROM | (20) |
| 15 | AT | (21) |
| 16 | NOT | (17) |
| 17 | THAT | (17) |
| 18 | WHICH | (19) |
| 19 | AND | (19) |
| 20 | AS | (20) |
| 21 | OF | (21) |
| 22 | ON | (29) |
| 23 | IN | (23) |
| 24 | ARE | (24) |
| 25 | YOU | (29) |
| 26 | BY | (27) |
| 27 | WITH | (28) |
| 28 | IS | (28) |
| 29 | IT | (29) |
| 30 | HER | (31) |
| 31 | OR | (1) |
| 32 | A | (1) |

**"HASH" TABLE provides a better way for a computer to search through large files of data. For each word $x$ one uses a computer's ability to do high-speed arithmetic by computing a hash address $h(x)$, where the search for $x$ is to start. The hash address for each of the 31 commonest words is shown in parentheses after each word; in this example each hash address was obtained by adding the numerical value of each letter ($a = 1$, $b = 2$ and so on up to $z = 26$) and discarding excess multiples of 32. Sometimes two different words $x$ and $y$ have the same hash address $h(x)$, so that they "collide." If $x$ is not stored in position $h(x)$, the search continues upward through positions $h(x) - 1$, $h(x) - 2$ and so on. For example, the hash address of "his" is $h + i + s$, or $8 + 9 + 19 - 32 = 4$. The hash address of "have" is also 4. To search for "have" the algorithm looks first in position 4 (*light gray*), then in position 3 (*dark gray*) and finally in position 2 (*color*), where "have" is located. If word $x$ is not in table, search for it will stop at empty position 5.**

probes needed to find one word out of $n$ equally likely words that have been randomly inserted into a table of size $m$ is $1 + [(n-1)/m + (n-1)(n-2)/m^2 + (n-1)(n-2)(n-3)/m^3 + \cdots]/2$. Let the symbol $\alpha$ stand for $n/m$, the fullness ratio or "load factor" of the table. As $n$ approaches infinity it can be shown that the average number of probes required to find any word $x$ in a table approaches the value $1 + (\alpha + \alpha^2 + \alpha^3 + \cdots)/2$, which is equal to $[1 + 1/(1-\alpha)]/2$. Furthermore, the true average number of probes will always be less than this limiting value. Therefore when the table being searched is 80 percent full, Algorithm D makes fewer than three probes per successful search on the average.

It is important to note that the stated upper limit on the average number of probes per successful search holds for all tables that are equally full, no matter how large the table is. The same cannot be said about binary-comparison algorithms, because their average running time per successful search will grow arbitrarily large as the number $n$ of words to be searched increases.

## Improving Unsuccessful Searches

My statements in the preceding paragraphs about the small number of probes required with Algorithm D apply only to cases where $x$ is actually found in the table. If $x$ is not present, the average number of probes needed to ascertain that fact will be larger, namely $1 + [2n/m + 3n(n-1)/m^2 + 4n(n-1)(n-2)/m^3 + \cdots]/2$; when $n$ is large, this number is approximately equal to $[1 + 1/(1-\alpha)^2]/2$. In other words, an average unsuccessful search in a large hash table that is 80 percent full requires nearly 13 probes. Moreover, in my example of the 31 words in 32 spaces, note that all unsuccessful searches must terminate at the single empty position 5 regardless of the location of the starting address $h(x)$. A precisely analogous situation occurred with the sequential search Algorithm A, where all unsuccessful searches end at position 0.

In 1973 O. Amble of the University of Oslo noted that the problem of unsuccessful search could be alleviated by combining the concept of hashing with the concept of alphabetical order. Suppose the 31 commonest English words are inserted into the table in decreasing alphabetical order instead of in decreasing order of frequency. Since the table is probed by starting at the address $h(x)$ and moving to $h(x) - 1$ and so on, all words lying between the address $h(x)$ and the actual location of $x$ must be alphabetically greater than $x$ lest there be a collision. A search for $x$ can therefore be terminated unsuccessfully whenever a word alphabetically less than $x$ is encountered. In other words, the following algorithm can be used:

Algorithm E; linear probing in an ordered hash table. This algorithm assumes that KEY[$j$] is 0 when entry $j$ is empty, and that all words $x$ have a numerical value that is greater than 0.

E1. [Initialize.] Set $j \leftarrow h(x)$.

E2. [Unsuccessful?] If KEY[$j$] $< x$, output 0 and terminate the algorithm.

E3. [Successful?] If KEY[$j$] $= x$, output $j$ and terminate the algorithm.

E4. [Move to next.] Set $j \leftarrow j - 1$; then if $j = 0$, set $j \leftarrow m$. Return to step E2.

The advantage of Algorithm E is illustrated in the ordered hash table on page 80. Suppose one wants to determine if "has" is one of the 31 commonest English words. Its hash address is $8 + 1 + 19 = 28$. With Algorithm E the search terminates in six steps when it reaches $j = 22$ ("by") instead of continuing through the table until it encounters the empty table entry at $j = 5$.

In an ordered hash table the average number of probes per unsuccessful search is reduced to $1 + [n/m + n(n-1)/m^2 + n(n-1)(n-2)/m^3 + \cdots]/2$, and this number is always less than $[1 + 1/(1-\alpha)]/2$. Thus the limit for a successful search and the limit for an unsuccessful one are identical. On the average, when an ordered hash table is 80 percent full, Algorithm E will make less than three probes regardless of the size of $n$.

This is all very well if the ordered hash table has been set up by inserting the keys in decreasing alphabetical order as I have described it. In practice, however, one cannot always assume that the words of a table have been entered in such a manner. Tables often grow dynamically with use, and new words enter in random order. Although the structure of a binary tree (Algorithm C) and of an unordered hash table (Algorithm D) will handle dynamic growth with ease, the structure of an ordered hash table (Algorithm E) is not so obviously adaptable. Fortunately there is a very simple algorithm for inserting a new word into an ordered hash table:

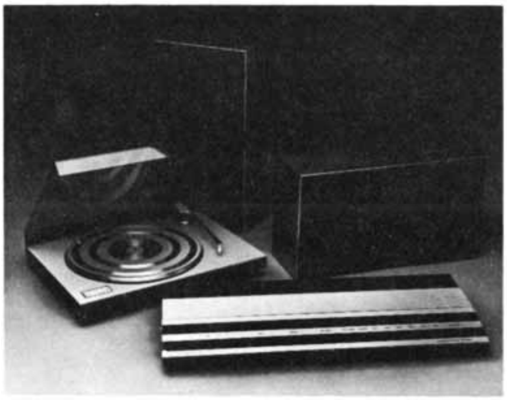
Algorithm F; insertion into an ordered hash table. This algorithm puts a new word $x$ into an ordered hash table and appropriately rearranges the other entries so that searching with Algorithm E remains valid.

F1. [Initialize.] Set $j \leftarrow h(x)$.

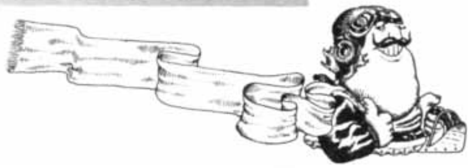
F2. [Compare.] If KEY[$j$] $< x$, interchange the values of KEY[$j$] and $x$. (That is, set $x$ to the former value of KEY[$j$] and set KEY[$j$] to the former value of $x$.)

F3. [Done?] If $x = 0$, terminate the algorithm.

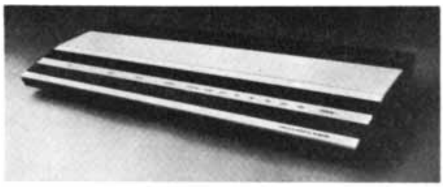
F4. [Move to next.] Set $j \leftarrow j - 1$; then



# The Bang & Olufsen Beosystem 1900. It's so simple, most people don't get it.

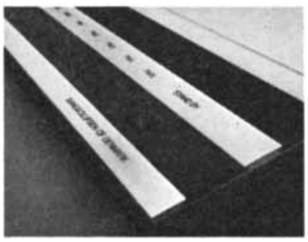


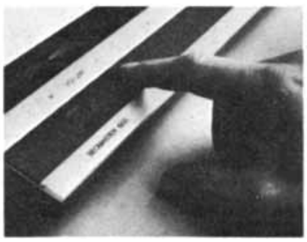
**An alternative to the airplane cockpit school of audio design.** It's not knobs and dials that make superb sound, it's superb engineering. In the Beosystem 1900, there is almost none of the former, and a great deal of the latter.



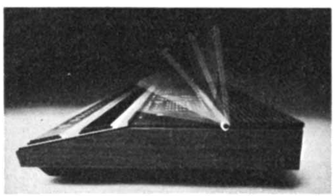
A few cases in point:

**Your finger, the component.** With the Beomaster® 1900, you become part of the system. All major controls are electronically activated by a light touch of your finger on the front control panel. The instant you touch it, an illuminated indicator appears for each function, you always know the operational status of the 1900, even in the dark.

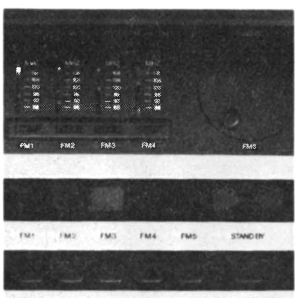




**For details, look inside.** Secondary controls, for bass, treble, and FM tuning, are out of sight, literally—concealed behind an aluminum door that opens and closes in a manner reminiscent of the Starship "Enterprise."



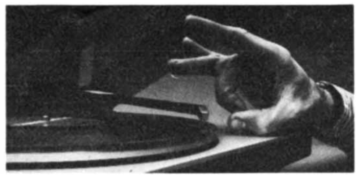
**Thanks for the memory.** The Beomaster 1900 also allows you one unforgettable convenience. You may pre-set the volume level and pre-tune up to five FM stations. Then, at the instant you want it, you have the station you want, at the level you want. Why clutter your memory when the system has one?

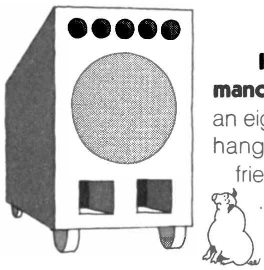



**The turntable, taken to its logical conclusion.** The Beogram® 1900 turntable's very low mass tone arm and MMC 4000 cartridge work magnificently with each other, because they are *made* to work with each other, by engineers who talk to each other, listen to each other, and design for each other. If that strikes you as overwhelmingly logical, you'd be surprised how other turntables are put together.

**A scratched record is forever. (How to protect your investment.)** No matter how little you've spent on your record collection, chances are some of it is irreplaceable, which makes it priceless. It makes sense to protect it—the way our MMC 4000 cartridge does with an effective tip mass of only 0.4 milligrams. (A tiny square of this page, this big □ weighs 1.0 milligram.) This results in a touch so delicate that it's almost impossible to scratch your records while playing them. It also reduces wear considerably enabling your records to continue working well past normal retirement age.

We don't recommend this, but with the MMC 4000 cartridge, it won't hurt.





**It's not size that counts. It's performance.** Can a speaker small enough to fit on an eight-inch shelf (or unobtrusive enough to hang on a wall) impress your audiophile friends? Yes, if they keep their eyes closed ...and their ears open.

**The missing link: Our 100% solution.** Phase distortion—a principal villain in speaker performance—was identified in 1973 by Bang & Olufsen engineers. The first practical solution to the problem was presented in London in 1975 to the international organization, the Audio Engineering Society, by Bang & Olufsen engineers.

Today that solution is an integral part of our Beovox® Phase-Link® Loudspeakers (Pat. Pend.).

**If a child can operate it, will an adult buy it?** Because *usability* is at the heart of the Beosystem 1900's design, it is true that a child can operate it. But only a very sophisticated adult can truly appreciate it. Welcome. Write to us at: Bang & Olufsen of America, 515 Busse Road, Dept. 13K Elk Grove Village, Illinois 60007, we'll be happy to send you our brochure and dealer list.

**Bang & Olufsen**

if $j = 0$, again set $j \leftarrow m$. Go back to step F2.

If we choose to insert the word "has" into the ordered hash table of the 31 commonest English words by means of Algorithm F, the procedure would place "has" in position 22, making room for it by moving "by" from position 22 to position 18, moving "at" from position 18 to position 15, moving "as" from position 15 to position 14, moving "are" from position 14 to position 8 and finally moving "and" from position 8 to the empty position 5. That may seem like a lot of work, but it takes only slightly longer than the task of inserting "has" into an unordered hash table using Algorithm D. In general the insertion of a word into an ordered hash table takes the same number of iterations as the insertion of the same word into an unordered hash table. Furthermore, the average number of words in the table that must be interchanged by way of step F2 to accommodate the new word is $(n - 1) / 2m + 2(n - 1)(n - 2) / 3m^2 + 3(n - 1)(n - 2)/4m^3 + \cdots$, which is approximately equal to $1/(1 - \alpha) + [\log_e(1 - \alpha)]/\alpha$, where $e$ is the familiar 2.71828. Thus inserting words by Algorithm F is quite a reasonable task.

In this specific case we actually should not have inserted "has" into the table because hash tables ought to have at least one empty position. By coincidence the smallest possible word in alphabetical order, "a", is present in this completely full table. Hence linear probing with Algorithm E will still work correctly in all cases. If "a" were not in the table, however, an empty position would be needed in order to avoid endless searching when the input word $x$ was equal to "a".

One of the most surprising properties of ordered hash tables is that each one is unique. If we use Algorithm F to build an ordered hash table from any set of words, the same table will be obtained regardless of the order in which the words are inserted. The reader may find it entertaining to prove this fact for himself.

| | | |
|---|---|---|
| 1 | THE | (1) |
| 2 | HAVE | (4) |
| 3 | TO | (3) |
| 4 | HIS | (4) |
| 5 | | |
| 6 | BE | (7) |
| 7 | FOR | (7) |
| 8 | AND | (19) |
| 9 | I | (9) |
| 10 | BUT | (11) |
| 11 | WAS | (11) |
| 12 | HAD | (13) |
| 13 | HE | (13) |
| 14 | ARE | (24) |
| 15 | AS | (20) |
| 16 | NOT | (17) |
| 17 | THAT | (17) |
| 18 | AT | (21) |
| 19 | WHICH | (19) |
| 20 | FROM | (20) |
| 21 | OF | (21) |
| 22 | BY | (27) |
| 23 | IN | (23) |
| 24 | THIS | (24) |
| 25 | IS | (28) |
| 26 | IT | (29) |
| 27 | ON | (29) |
| 28 | WITH | (28) |
| 29 | YOU | (29) |
| 30 | A | (1) |
| 31 | HER | (31) |
| 32 | OR | (1) |

**ORDERED HASH TABLE, which combines the concept of hashing with the advantage of alphabetical order, reveals more quickly when the input word is not present. Here all the words between position $h(x)$ and the actual location of $x$ are alphabetically greater than $x$. Thus an unsuccessful search need not stop only at the empty position 5; it will also stop as soon as a word alphabetically less than $x$ is encountered. If desired word $x$ is "has", with hash address 28 (*light gray*), search will stop when it reaches "by" at position 22 (*dark gray*).**

## Conclusion

My discussion of ways to search for information stored in a computer's memory is intended to illustrate several important points about algorithms in general. As we have seen, an algorithm must be stated precisely, and it is not as easy to do that as one might think. When one tries to solve a problem by computer, the first algorithm that comes to mind can usually be greatly improved. Data structures such as binary trees are important tools for the construction of efficient algorithms. When one starts to investigate how fast an algorithm is, or when one attempts to find the best possible algorithm for a specific application, interesting issues arise and one often finds that the questions have subtle answers. Even the "best possible" algorithm can sometimes be improved if we change the ground rules. Since computers "think" differently from people, methods that work well for the human mind are not necessarily the most efficient when they are transferred to a machine.

# 200-SX


## Suddenly from Datsun: A sporty car with everything but a sports car price.

Exit dull, sluggish economy cars.
Enter Datsun's spicy 200-SX. Sweet-handling. Tasty appointments. And no bitter price to swallow. Enjoy.

**Fun and frugal 5-speed.**
Sporty 5-speed transmission works like overdrive. Thus, saving gas.

According to EPA estimates, 200-SX squeezes 34 MPG on the Highway, 23 City. Naturally, your actual mileage depends on driving habits, optional equipment and condition of car. California mileage lower.

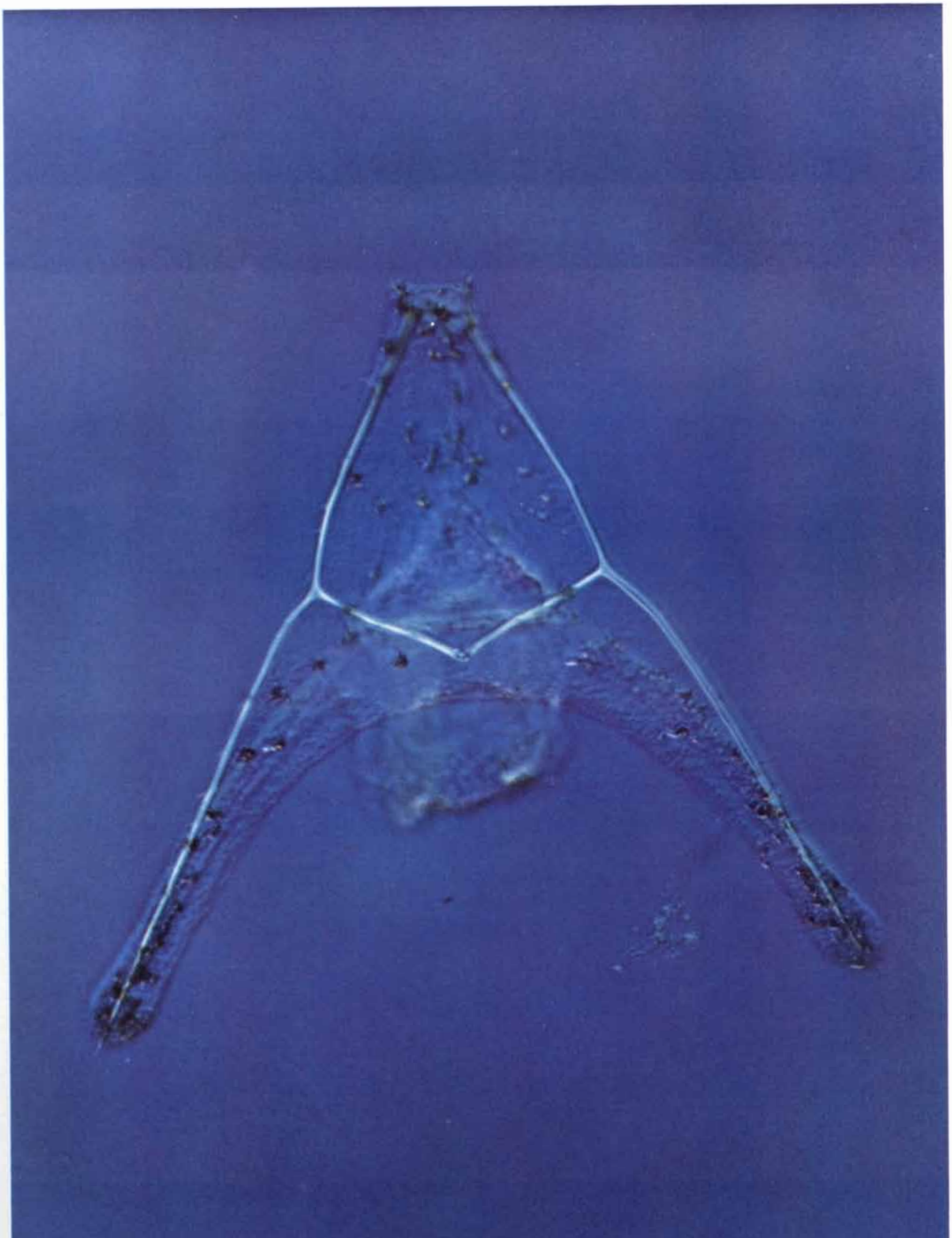
**Extras, yes. Extra cost, no.**
- AM/FM multiplex stereo radio
- Steel belted radial tires
- Tachometer
- Fully reclining bucket seats
- Cut-pile carpeting
- Electric rear window defogger
- Tinted glass
- Electric clock
- Sporty 5-speed gearbox
- Power-assist front disc brakes

All for $4399! (Manufacturer's Suggested Retail Price not including destination charges, taxes, license or title fees and optional tape stripe and mag type wheel cover package.)

**Tough sport.**
Solid, all-steel unibody is but one example of how the Datsun 200-SX is put together to stay together. Fact is, when we made this fun little car, we made sure of one thing.
The fun would last.

**Suddenly it's going to dawn on you.**

DATSUN SAVES



**CRYSTALLINE SKELETON** glows within a sea urchin larva in this micrograph made by one of the authors (Inoué) using a combination of differential-interference and polarized-light microscopy. Two of the larva's four horns are visible. Magnification is 640 diameters.

# Biocrystals

*The development of an elaborate crystalline skeleton within the transparent embryo of the sea urchin provides a model for the shaping of hard mineralized tissues such as shell and bone*

by Shinya Inoué and Kayo Okazaki

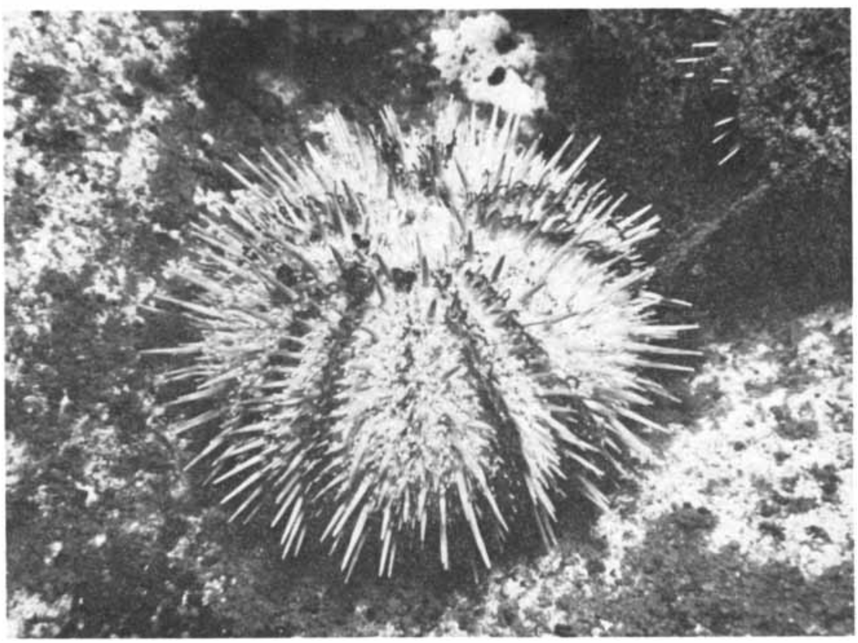
Precisely how embryonic development unfolds through the successive division, rearrangement and differentiation of cells arising from a single fertilized egg is a central question of biology. A part of that question is how the developing embryo molds inorganic substances into hard crystalline structures such as bone, tooth and shell. These materials, which can be characterized as biocrystals, consist of regular three-dimensional arrays of calcium, silicon, phosphate, carbonate or some combination of these substances, and they are chemically indistinguishable from the crystals found in the inanimate world. They are remarkable, however, in that they exist not in the simple geometric forms of nonbiological crystals but in intricately sculptured forms that vary with the living organisms that make them. Biocrystals are found throughout the animal kingdom, ranging from the silica skeletons of sponges to the layered crystals of calcium phosphate that form the mineral component of human bones and teeth.

The sea urchin, the spine-covered marine organism that lives in the shallow water of coral reefs or tidal pools, is an excellent organism for the investigation of biocrystals. In its embryonic form it grows two spicules, or starlike biocrystals, that eventually fuse to create a supporting skeleton for the four-horned larval form. Moreover, the embryo is transparent, so that it affords a clear view of living cells in the act of building mineral elements into crystalline structures.

The growth of the spicules within the developing sea urchin embryo has been described in considerable detail over the past half century. In season the adult sea urchin nurtures millions of ripe germ cells (eggs or sperm) within its spiny shell. Either naturally or on artificial provocation (such as electrical stimulation or the injection of potassium chloride) the organism releases the germ cells through five microscopic pores. In the surrounding seawater the sperm and eggs released by the male and female sea urchins unite and the fertilized eggs begin to divide. First the egg cleaves symmetrically three times, giving rise to two tiers of four equal-sized cells stacked on top of each other. Then at the fourth division the cells divide unequally, giving rise to eight intermediate-sized cells, four large cells and four small cells called micromeres. Most of the cells continue to divide until they form the blastula: a hollow sphere consisting of a single layer of cells. Some of these cells push into the hollow interior of the blastula while still maintaining contact with their neighbors, forming the primitive gut and mouth of the embryo, which is now called the gastrula.

At this point descendants of the micromeres, the four small cells arising at the fourth division, have migrated into the cavity of the gastrula. They are now designated primary mesenchyme cells, and with the aid of pseudopods—long contractile extensions—they associate in two groups, one group on each side of the primitive gut. There the pseudopods fuse, and within them appears a tiny calcareous (limy) granule. Next the primary mesenchyme cells, still connected by their pseudopods, organize themselves into three radiating cables 120 degrees apart. As the embryo continues to develop, the pseudopodial cables guide the growth of the calcareous granule into a triradiate spicule, a crystal shaped like a three-pointed star.

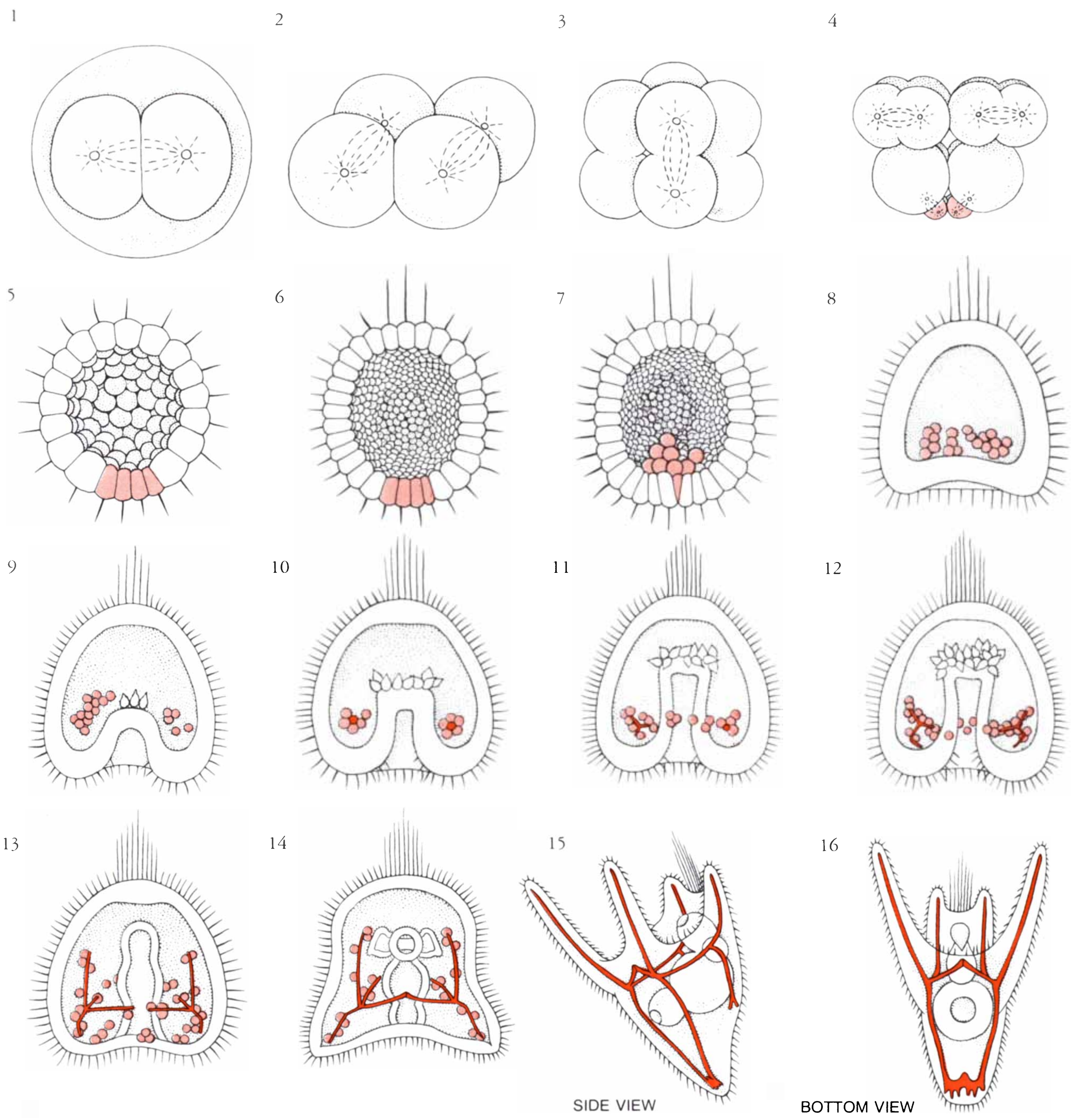
Later the chains of mesenchyme cells assume other configurations, with each chain migrating in a different direction. The flat triradiate spicule is then sculptured by the pseudopodial cables into a three-dimensional form termed the skel-



**ADULT SEA URCHIN is a common inhabitant of marine tidal pools and the shallow waters of coral reefs. In season it nurtures millions of eggs or sperm within its spine-covered shell.**

etal spicule. The process begins with a slight right-angle bend in one of the arms of the triradiate spicule. The bent portion elongates and establishes the long axis of the skeletal spicule, which depending on the species has one rod or more projecting from it. Eventually a joined pair of skeletal spicules shaped somewhat like a folding chair provides an internal skeleton for the four-horned sea urchin larva. The shape of the larval spicule in each species of sea urchin is unique to that species.

Chemically the larval spicule is calcium carbonate ($CaCO_3$) containing a small amount of magnesium (5 percent) and organic matrix (1 percent). Also a major constituent of coral, chalk and marble, calcium carbonate crystallizes in different forms, all of which exhibit the optical property of birefringence: they split light waves into two components that oscillate in planes perpendicular to each other. When such a crystal is placed in a beam of plane-polarized light (that is, light in which the




**DEVELOPMENT OF THE SEA URCHIN from the fertilized egg to the larval form is diagrammed in this illustration. After fertilization the egg cleaves three times (*1–3*). The resulting eight cells then divide unequally (*4*), giving rise to eight intermediate-sized cells, four large cells and four small cells called micromeres (*light color*). The dividing cells next form a hollow sphere, or blastula (*5, 6*). Descendants of the micromeres, the primary mesenchyme cells, migrate into the blastular cavity (*7, 8*). An inward folding of the blastular wall forms the primitive gut of the embryo, now called a gastrula, and the mesenchyme cells associate on each side of the fold (*9*). There the pseudopods, or contractile extensions, of the mesenchyme cells fuse to form radiating chains, or cables. At the vertex of the cables a calcareous granule (*dark color*) appears (*10*). The two granules in each embryo are gradually sculptured into flat triradiate spicules (*11, 12*) and then into complementary three-dimensional skeletal spicules. Finally the skeletal spicules fuse to form the larval skeleton (*13–16*).**

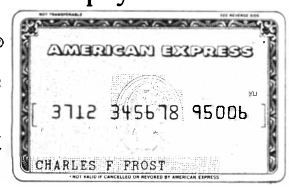
# Who can give you everything from Henry's Fort to Louise's Lake?

This year why not come to Canada for your vacation? We've got so much to go for. And with Air Canada's special tours you can go for so much less.

Take Air Canada's Canadapass. It's a unique one week fly/drive vacation that starts as low as $178* per person, double occupancy, plus air fare. Canadapass gives you the total freedom of a car with unlimited mileage, and a guaranteed room every night, wherever you are, without reservations. Dine on fine French food in Québec. Explore historic Fort Henry in Kingston, Ontario. Or relax on the beautiful beaches of Atlantic Canada.

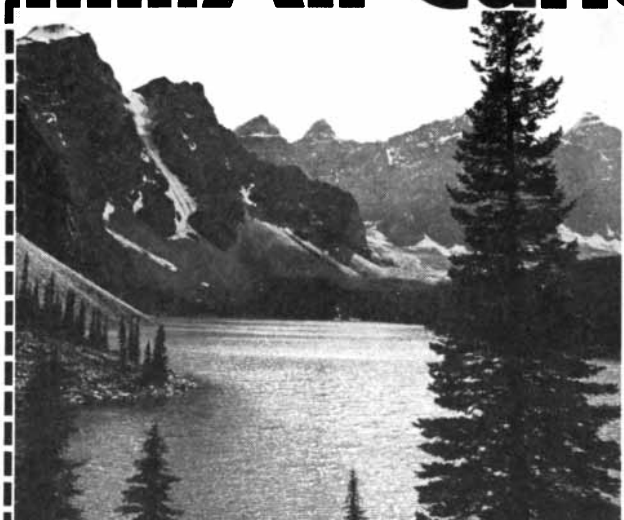
You don't go for fly/drive tours? Take our Rockies and Vancouver by Motorcoach tour this spring or fall. For $169† per person, double occupancy, plus air fare, you get six nights in really nice hotels, all transfers, provincial taxes, and sightseeing. And what sights you'll see! The rugged beauty of Fraser Canyon. The emerald green of Lake Louise. And, of course, Vancouver, one of the world's most spectacularly situated cities.

So come to Canada. You can pay for your vacation—and even extend payments—with the American Express Card® "Sign & Travel"® Plan. If you don't have the card, pick up an application wherever it is welcomed.




The American Express Card. Don't leave home without it.℠

# Air Canada can!



SA4·77

Air Canada, P.O. Box P,
Staten Island, New York 10305.

Please send me your Yes, Canada book. I understand it's got tours that can give me everything from Henry's Fort to Louise's Lake to Peggy's Cove to Edward's Island to Victoria's City.

Canada. So much to go for.

Name____________________

Street____________________

City____________________

State____________ Zip________

My travel agent is____________________

AIR CANADA

*ITAC—FRM07 †ITAC UTL WF6

waves are oscillating in only one plane) and viewed through a polarizing filter, a characteristic phenomenon is seen. When the plane of polarization is oriented in a direction that is neither parallel nor perpendicular to the planes of atoms in the crystal, the crystal dramatically lights up.

There are few sights to be seen through the microscope that are as lovely as sea urchin larvae viewed in a microscope equipped with polarizing filters above and below the specimen. When the axes of the filters are crossed so that the field is darkened, the two skeletal spicules within each swimming larva alternately light up and go dark like twinkling stars. With every 360-degree rotation of a larva each spicule shines four times at orientations exactly 90 degrees apart.

The birefringence of the skeletal spicule of the sea urchin larva is strong compared with that of other birefringent materials: the refractive index of the spicule is some 12 percent greater when the plane of polarization of polarized light is perpendicular to the long axis of the spicule than when the plane is parallel to that axis. This behavior, together with information obtained by X-ray diffraction, suggests that the calcium carbonate in the spicule is crystallized in the calcite form. The structure of calcite consists of planes of calcium ions, which have two positive charges, alternating with planes of carbonate groups ($CO_3$), which have two negative charges. The calcium and carbonate planes are perpendicular to the crystal's crystallographic axis and thus define that axis. The bonding forces in calcite are weak across planes parallel to the faces of a rhombohedron, with the result that nonbiological calcite crystals have a nearly perfect rhombohedral form.

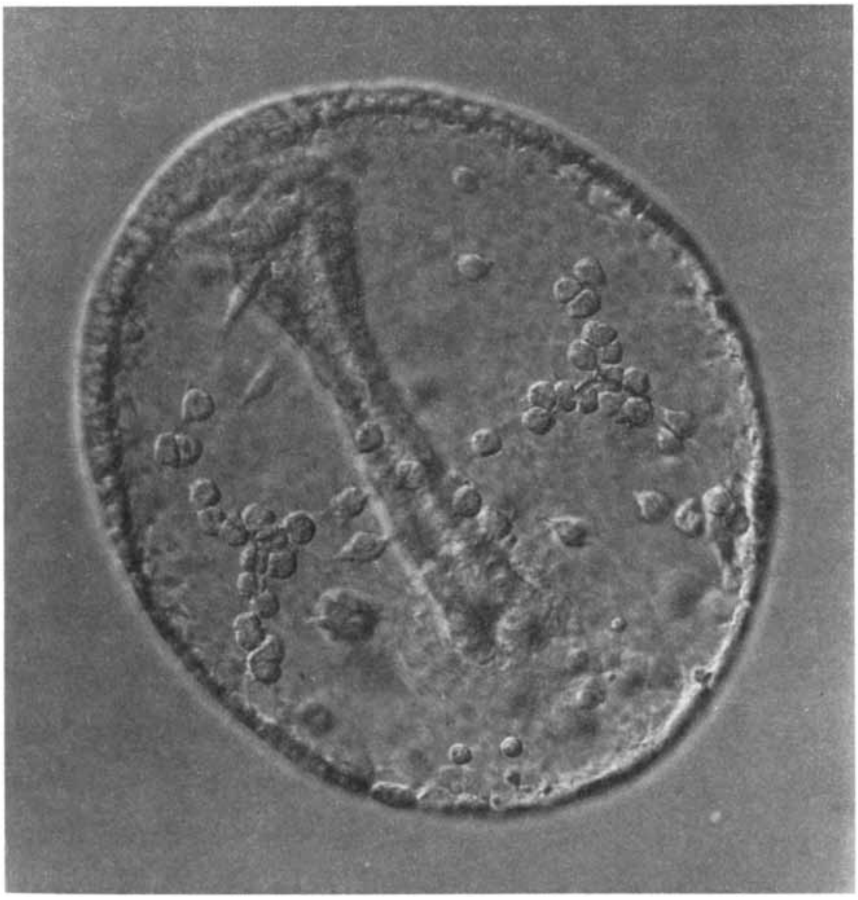
The birefringence exhibited by the skeletal spicule of the sea urchin larva tells us that each spicule, in spite of its complex external form, behaves optically as though it were carved out of a larger crystal of nonbiological calcite. This larger crystal would be oriented so that the crystallographic axis runs along the length of the spicule. Hence the skeletal spicule of a sea urchin larva apparently combines the internal form of a single calcite crystal with the sculptured external form characteristic of the species.

We first became fascinated by these structures 30 years ago while we were working with a hand-built polarizing microscope in the laboratory of Katsuma Dan at the Misaki Marine Biological Station south of Tokyo. Two summers ago we resumed our collaboration at the Marine Biological Laboratory in Woods Hole, Mass., where one of us (Okazaki) held a Lillie Fellowship. Kent McDonald, a graduate student from the University of Colorado, was operating the laboratory's scanning electron microscope. We all felt that here, perhaps, was an ideal tool for looking at the detailed structure of the spicule of the sea urchin larva in order to extend the knowledge we had gained with the polarizing microscope. Accordingly we isolated both triradiate spicules and skeletal spicules by placing sea urchin embryos in fresh water, which causes the cells of the embryo to osmotically take up water and burst. The spicules were then washed clean with dilute sodium hydroxide or Clorox (the same agents that are used to clean the adult sea urchin shells one finds in souvenir shops).

The scanning electron microscope displayed the isolated spicules splendidly. The simple elegance of the triradiate spicules and the sculptured beauty of the skeletal spicules were wonderful to behold. But there was something strange. No matter what magnification we used—and at the highest magnification we could resolve structures only 100 angstroms ($10^{-5}$ millimeter) across—there were no signs of crystalline texture on the surface of the isolated spicule.

We tried etching the spicules with weak acids and with a calcium-binding agent. The surface became eroded and rough but again we found no trace of the geometric crystalline forms that are commonly found in nonbiological calcite crystals exposed to weak acids. Moreover, nonbiological calcite tends to cleave along certain crystallographic planes and readily expose crystal faces that make characteristic rhombohedral angles with one another. When we fractured skeletal spicules, however, we could find no cleavage faces in the scanning electron microscope. Instead the spicules seemed to fracture along smooth, irregular faces, the way a noncrystalline material such as glass does. In short, although the skeletal spicules behaved optically as though they were single crystals, the scanning electron microscope failed to reveal any sign of crystallinity.

In an attempt to resolve this paradox we decided to see whether we could use isolated spicules as seeds on which to grow calcite crystals. If this could be done, the calcium carbonate molecules in the growing crystal would line up with similar molecules in the biocrystalline spicule, revealing the underlying crystal structure. We therefore exposed isolated spicules to a saturated solution of calcium carbonate by first immers-

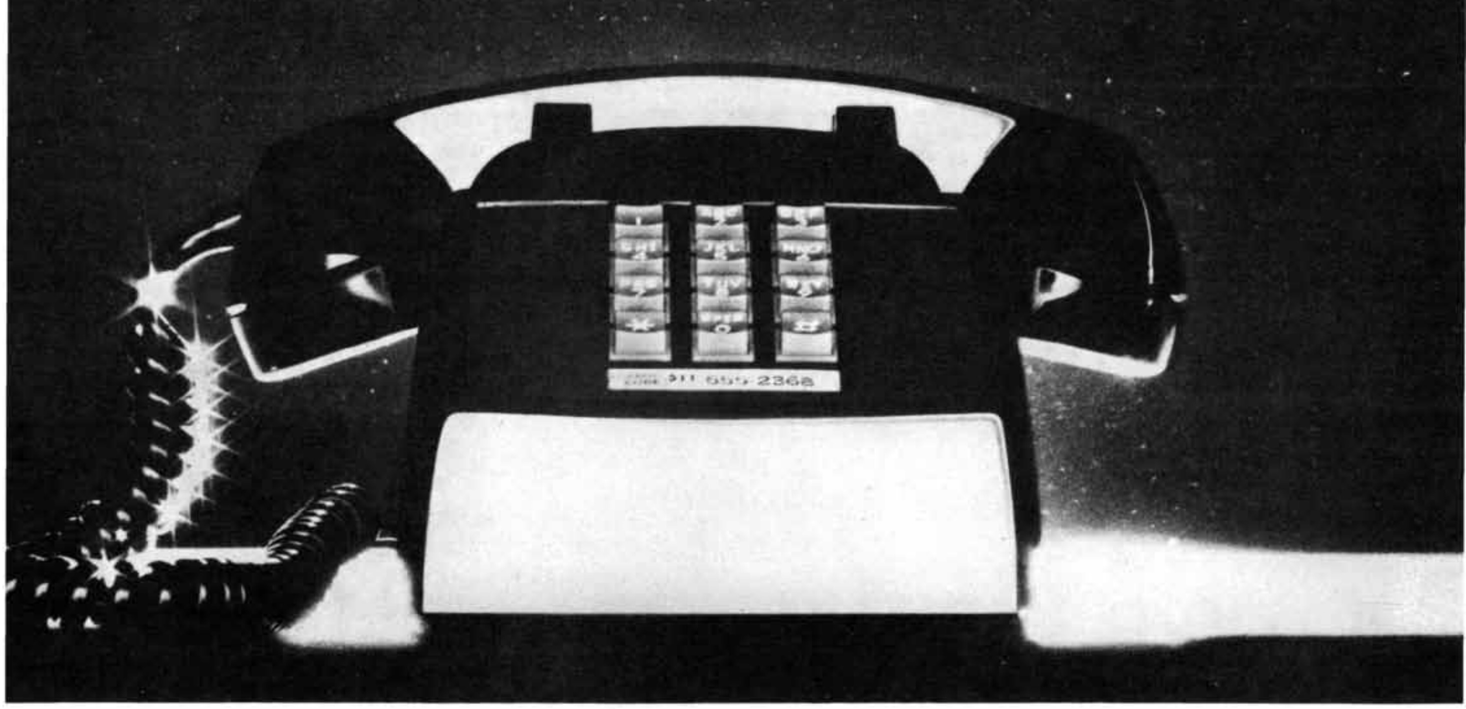


**CHAINS OF MESENCHYME CELLS guide the sculpturing of the two triradiate spicules within the cavity of the sea urchin embryo in this micrograph. Magnification is 500 diameters.**

## Willkommen, Bienvenue, Welcome...

We've fifty different ways of saying "Pleased to meet you, make yourself at home", and easily as many places where home will seem another world away – with Her Majesty's Regiments on Parliament Hill, or the poignant skirl of the Highland Piper, amid banter and barter at open-air markets, or the magic and pageantry of our Caribbean cultures, or the oom-pah-pah of Bavarian beerfests, and at any and all, por favor, pro szę bardzo, you're more than welcome.
For a thousand ways to have fun this summer with a friendly foreign flavour, call COLLECT (416) 965-4008, or write: Ontario Travel, Dept. F., Queen's Park, Toronto M7A 2E5.
We'll send you our Traveller's Encyclopaedia.

**Ontario Canada**



# There's at least one thing you can afford that's the very best on earth.

American telephone service is the best in the world. Yet it's well within the reach of virtually every American.

This didn't happen by accident.

Both the quality and economy of your telephone service benefit greatly from technological innovations Bell Laboratories and Western Electric produce for the Bell System.

## Money Well Spent...

Genuine innovation doesn't come cheap.

But the money Western Electric spends on research and development at Bell Labs produces a lot of innovation.

About 40% of what Western Electric makes today didn't exist in 1972.

That's triple the amount of innovation of industry in general.

Though such innovation costs a lot, it saves you money.

## ...Saves You Money.

Take long distance calling, for example. It's one of the few things that costs little more today than 25 years ago. That's because technological improvements have helped hold costs down.

For example, we have increased the capacity of coaxial cable systems from 600 simultaneous conversations in 1941 to 132,000 today.

We've increased sevenfold the capacity of our major microwave transmission system.

And our largest electronic switching system can handle four times the capacity of the previous system.

Innovations like these involve answering engineering and manufacturing questions at the forefront of technology.

Without the intimate flow of information that exists between Bell Labs and Western Electric, many of our innovations would have been much longer in coming.

Some might not have come at all.

## A Complex Future.

Many future innovations will be even more complex, requiring even closer interaction.

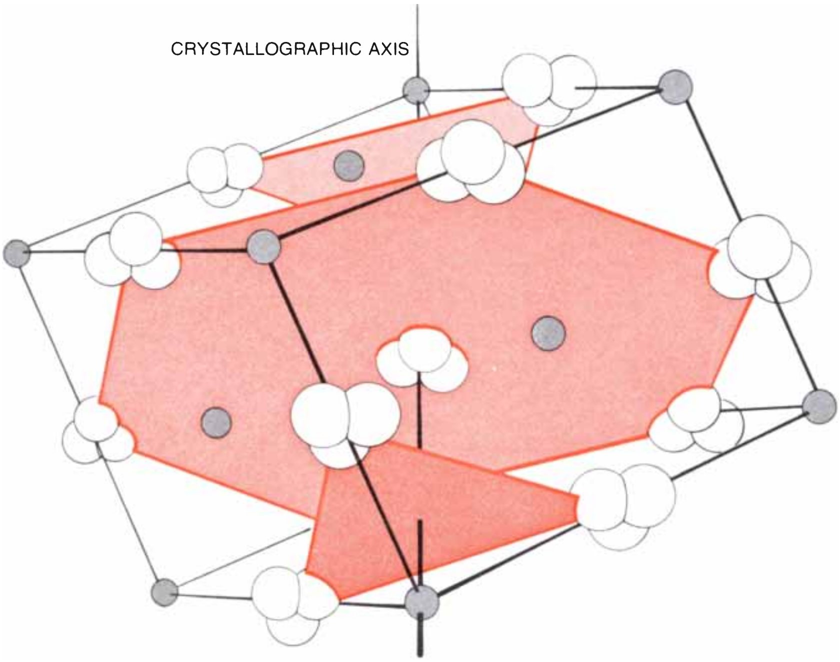
As part of the Bell System, Western Electric and Bell Labs will continue to put technology to work for you.

So even if you can't afford the best of everything, you'll still be able to afford the best of something: The best telephone service on earth.

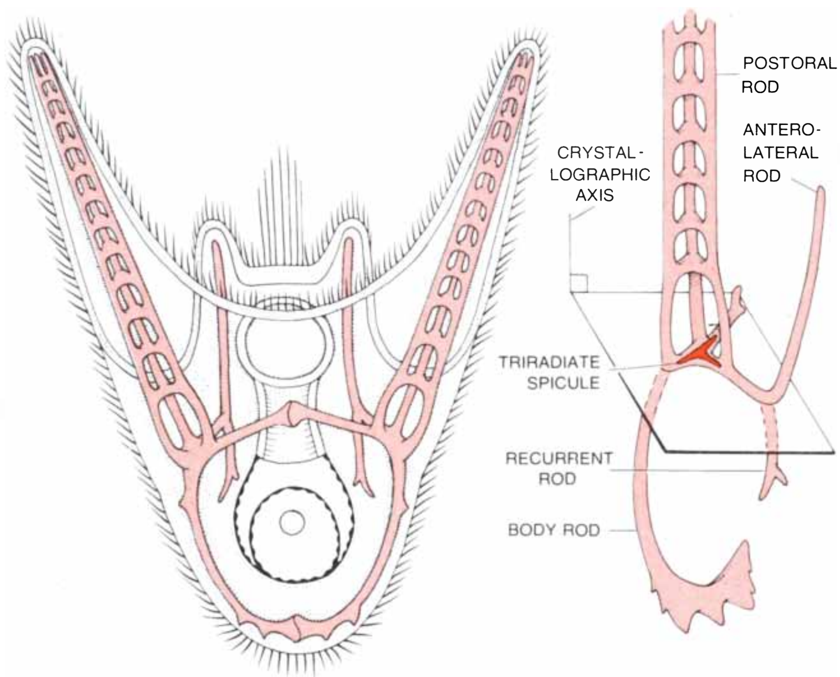
**Bell Laboratories/Western Electric**

ing them in a water solution of sodium bicarbonate and then adding calcium chloride in solution.

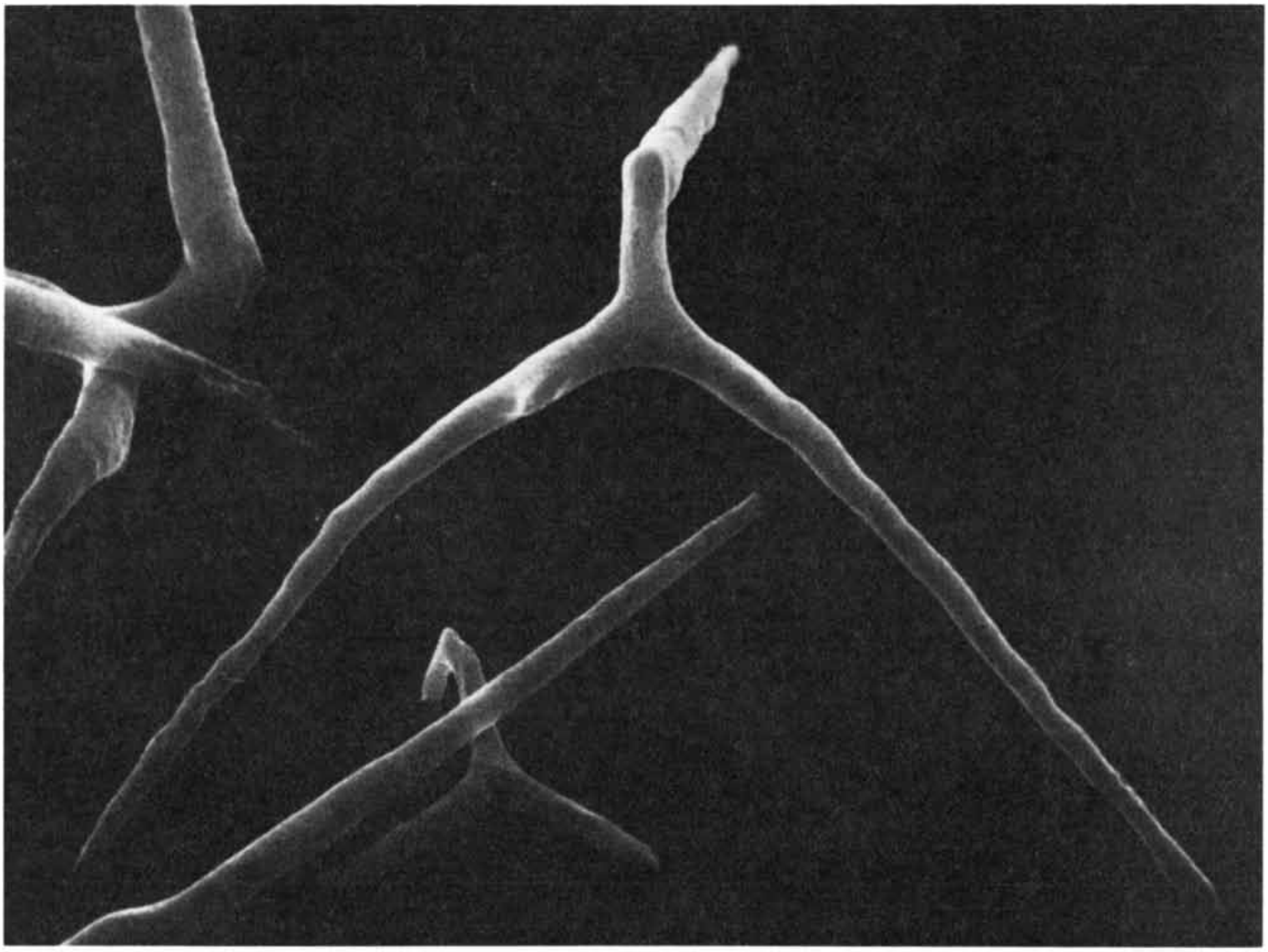
When we observed the seed spicules under the polarizing microscope, they appeared to get fatter and shine brighter in a few minutes. The edges of each spicule became jagged, yet the enlarged spicule still went dark as soon as its orientation was the same as when it went dark before treatment. It appeared that the calcium carbonate was depositing on the surface of the spicule with the molecules lined up in the same repeating



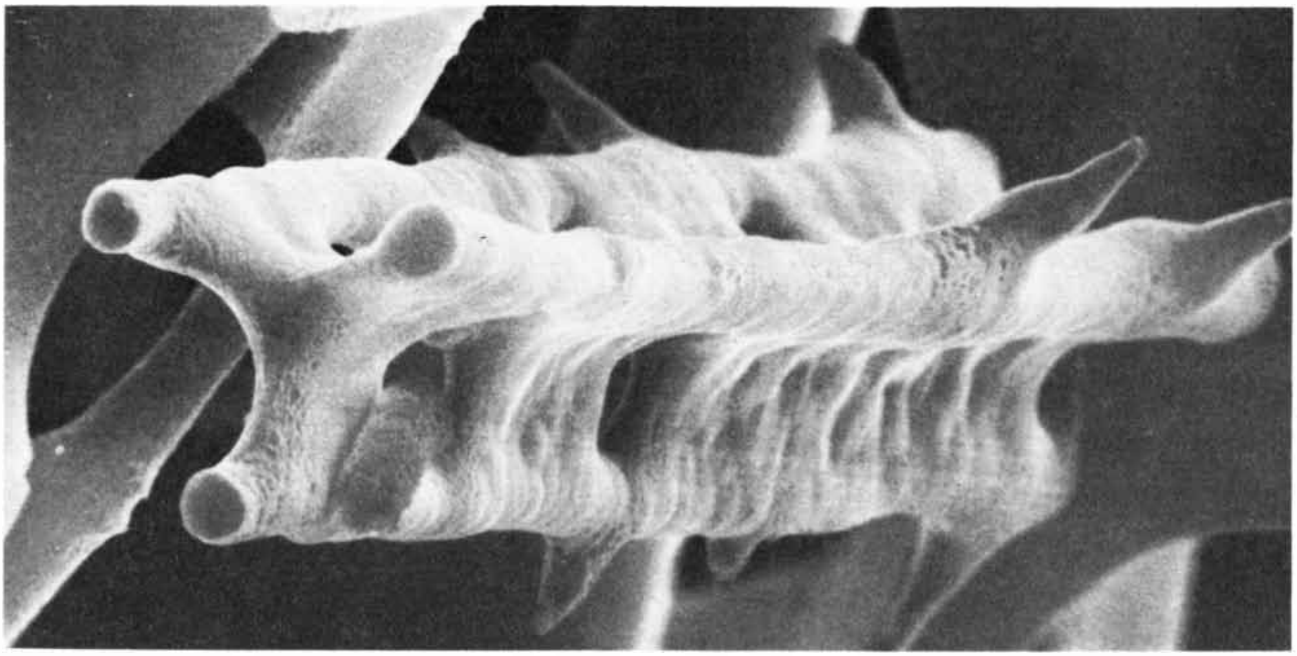

**STRUCTURE OF CALCITE consists of alternating planes of calcium ions (*gray*) and carbonate groups (*white*) perpendicular to a vertical crystallographic axis. Binding forces within the crystal are weak along the faces of a rhombohedron, giving rise to its characteristic shape.**




**FUSED PAIR OF SKELETAL SPICULES (*light color*) supports four-horned larva of the sea urchin *Arbacia punctulata*. Schematic diagram at the right shows how the elaborate rods of each skeletal spicule grow out from the arms of a flat triradiate spicule (*dark color*). Each skeletal spicule, in spite of its complex shape, behaves optically as if it were carved out of a single calcite crystal in which the crystallographic axis runs along spicule's major axis of elongation.**



**TRIRADIATE SPICULES** extracted from sea urchin embryos are magnified 4,200 diameters in this scanning electron micrograph. One arm of each triradiate spicule has begun the right-angled bend that will eventually grow to form the main body rod of the skeletal spicule.



**DETAIL OF SKELETAL SPICULE** from *Arbacia*, showing the elaborately sculptured surface of the postoral rod, is magnified 8,600 diameters in this scanning electron micrograph. Note the absence of any crystalline faces on the smooth outer surface of the spicule or on the three fractured surfaces at left. Instead fracture faces are smooth and concave, like those in a noncrystalline material such as glass.

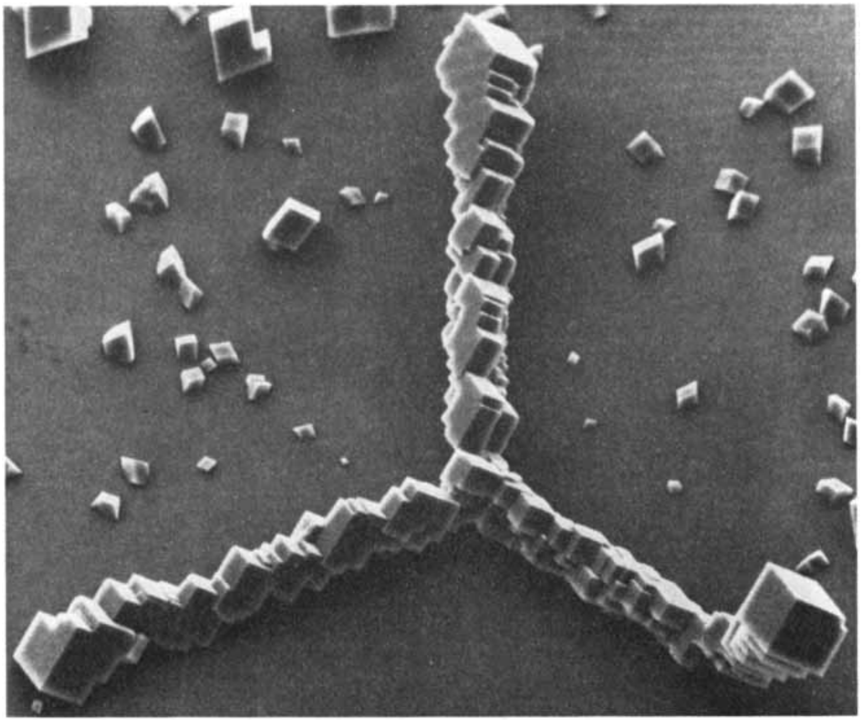
pattern as that of the molecules in the spicule itself.

The scanning electron microscope immediately confirmed this interpretation. Both the triradiate spicules and the skeletal spicules were "decorated" with miniature rhombohedral crystals resembling a stack of crystalline tiles. No matter what part of a spicule we viewed, the corresponding crystal faces all lay parallel to one another, regardless of the branching, sculpturing, bending or elongation of the spicule. The fact that the angle between the adjacent faces of the decorating crystals was characteristic of calcite was ascertained by viewing a stereoscopic pair of electron micrographs of a decorating crystal side by side with a rhombohedron of nonbiological calcite.

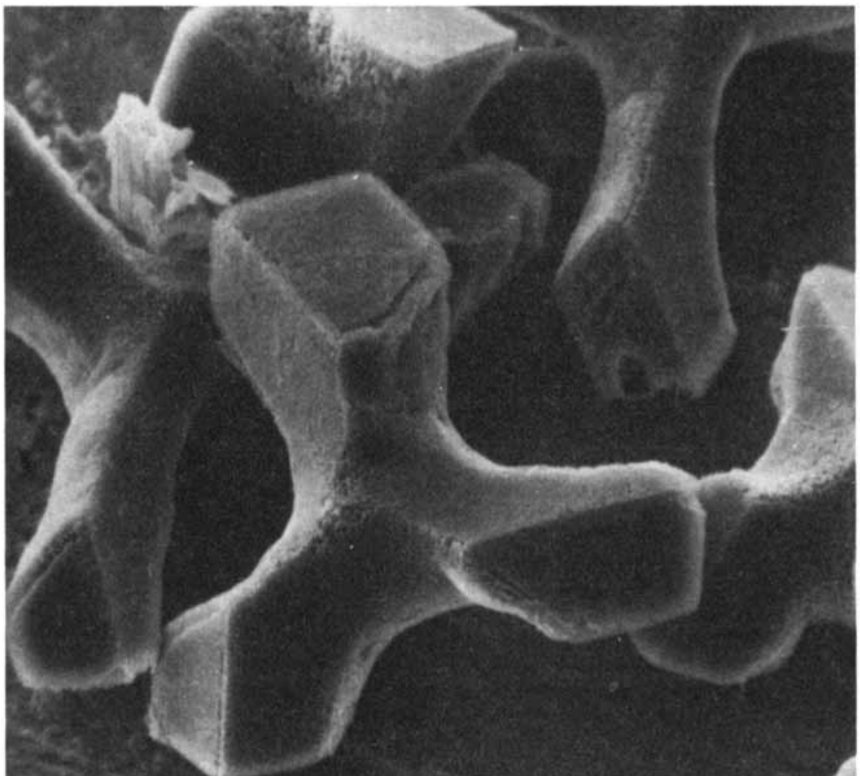
It appeared that we had successfully amplified the crystallographic arrangement of the calcium carbonate molecules in the spicule. Since all the corresponding surfaces of the decorating crystals were parallel to one another, all the calcium carbonate molecules that had acted as templates in the spicule also had to be oriented exactly parallel to one another. In fact, the crystallographic axes of crystals decorating a single spicule were always oriented parallel to the crystallographic axis of the spicule itself. The same held true independently for both the left and the right spicule in a fused pair, so that at the junction between the two spicules the two sets of crystals were joined with their crystallographic axes inclined to each other by the same angle as that at which the left and the right spicule were.

The calcium carbonate molecules in each spicule are thus arranged as regularly as they are in a single crystal of nonbiological calcite. Yet the biocrystal, instead of conforming to nonbiological calcite in external shape, assumes the elaborate forms characteristic of each species of sea urchin. How does the sea urchin embryo accomplish the feat?

As we have mentioned, the formation of each spicule in the developing embryo is guided by three radiating pseudopodial cables, or chains of mesenchyme cells. It turns out that these cells suppress the growth of even the smallest crystal faces in the spicule, thereby preventing the formation of mechanical weak spots and of abrasive surfaces that could irritate the surrounding tissue cells. In a series of elegant experiments Sven Hörstadius of the University of Stockholm showed that it is the nucleus of the micromere cell (the precursor of the mesenchyme cell) that carries inherited messages instructing the cells where and when to grow the biocrystal so as to generate the pattern unique to the species. Just how the mesenchyme cells decipher these messages and know their individual roles in collectively building up the biocrystal re-



**TILE OF CALCITE CRYSTALS completely "decorates" a triradiate spicule that was isolated and used as a seed for crystal growth. The spicule was incubated in a solution of calcium chloride and sodium bicarbonate. The axes of the decorating crystals are all exactly parallel, indicating that the arrangement of the template calcium carbonate molecules in the biocrystal is identical with that of a single crystal of nonbiological calcite. Magnification is 1,850 diameters.**



**ABNORMAL TRIRADIATE SPICULES were produced experimentally by growing sea urchin embryos in seawater with a diminished content of magnesium and calcium. These conditions weakened the mesenchyme cells so that they were unable to form normal pseudopodial cables. As a result the inorganic forces inherent in the calcite crystal overcame the biological forces exerted by the cells, and crystal faces appeared on the arms of the triradiate spicules.**

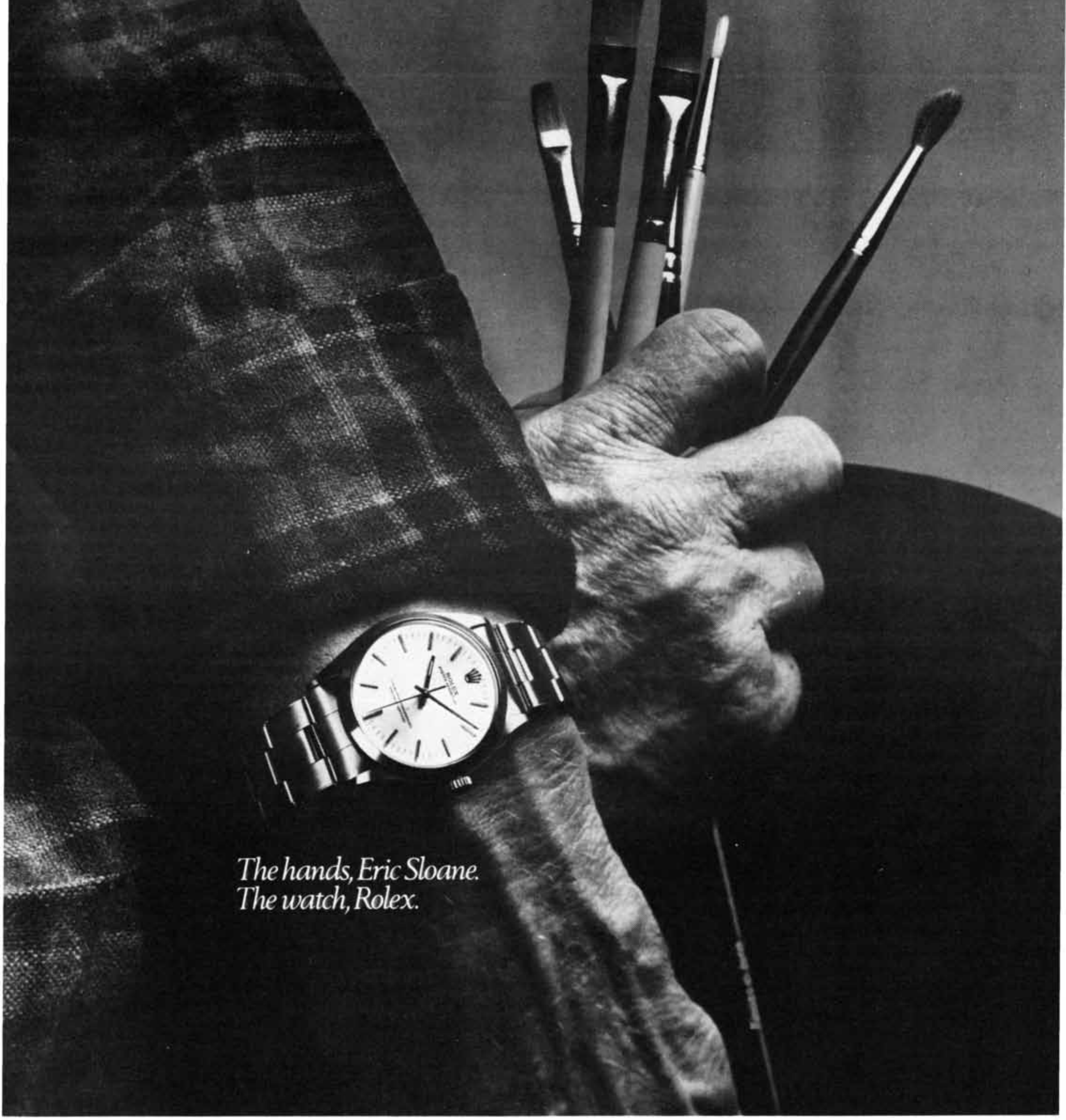
mains to date an intriguing mystery.

Recently one of us (Okazaki) succeeded in isolating micromeres from sea urchin embryos and growing them in tissue culture. The 20 or so mesenchyme cells descended from each micromere could make spicules with shapes characteristic of the species. This confirms that the micromeres alone can express sufficient genetic information to direct the growth of the skeletal spicules.

What would happen if the embryo was not capable of arranging its pseudopodial cables normally? Several years ago one of us (Okazaki) transferred embryos containing triradiate spicules from normal seawater to seawater containing a reduced concentration of calcium ions. The mesenchyme cells then formed less extended pseudopodial cables, and the normally slender tips of the triradiate spicules became bulbous. If the concentrations of both calcium and magnesium in the seawater were reduced to a tenth of normal, the bulbous tips became angular, as though calcite crystals were growing there. Observing these spicules in three dimensions side by side with a similarly oriented rhombohedron of nonbiological calcite (with stereoscopic pairs of scanning electron micrographs), we found that every face at all three triradiate-spicule tips lined up exactly parallel to the faces of a single calcite rhombohedron.

In an embryo deprived of adequate calcium and magnesium, then, crystal faces of calcite appear in the growing spicule. The forces that direct nonbiological calcite growth have overpowered the cellular control mechanisms and have created crystal faces that are inherent in the molecular bonding forces in calcite rather than the pattern normally seen in the biocrystal. To put it another way, the cells can no longer express the pattern encoded in their genes, and the tendency of the nonbiological crystal to form its own characteristic faces dominates.

In sum, when living organisms are building the mineral components of hard structures, they appear to take advantage of the crystallizing tendency of substances such as calcium carbonate and phosphate, utilizing the strong forces that hold the regularly arranged molecules together in a single crystal and yet guiding the growth of the crystals to the size and shape, and at the time and place, appropriate to the organism. Biocrystals dramatize the interplay between inanimate molecular forces, which tend toward a minimum-energy configuration, and the organizational capacity of living cells. Life of necessity follows inanimate principles, but instead of doing so by brute force it seems to channel the flow of energy, guiding the arrangement of matter into increasingly complex and thermodynamically improbable forms.


The hands, Eric Sloane.
The watch, Rolex.

The HP-25

The HP-67

The HP-97

# The HP programmables; or how to make a one-time investment in a machine you won't outgrow.

Users of preprogrammed hand-helds who seek to extend their problem-solving skills can find themselves stalled by their machines' built-in limitations. For users whose needs are evolving, the practical entry-level choice for personal data processing may well be the programmable machine.

Until about four and a half years ago, the slide rule was the personal number processor "of choice"– by necessity. Then, in 1972, Hewlett-Packard introduced an alternative: the first of a series of preprogrammed hand-held calculators that virtually revolutionized numerical data processing.

Since then, we have pioneered another alternative: hand-held *programmable* machines that expand their utility beyond the power resident in the key functions, and thus offer the capacity to deal realistically with a user's evolving needs:

• **The keystroke programmable.** The HP-25 ($145*) can be programmed by a series of keystrokes to solve automatically the repetitive problems every scientist and engineer faces. Its capabilities include branching, conditional tests, full editing, and eight addressable memories. It also has 72 built-in functions and operations, which can be incorporated into your programs to save program steps, or used "as is."

• **The programmable with continuous memory.** The HP-25C ($200*) adds a continuous memory to the HP-25, so that it can retain programs or data even when it is turned off.

• **The fully programmable.** The HP-67 ($450*) is the most powerful hand-held programmable HP has ever built. Programmed by prerecorded magnetic cards, or by keystroke for programs *you* develop (which can also be stored on magnetic cards for re-use), the HP-67 handles routines of up to 224 steps. But of even greater importance is the HP-67's capacity for automatic merging of keycodes (squeezing as many as three keystrokes into one step of program memory). There are program packages for math, stat, EE, ME, clinical lab and nuclear medicine, and surveying, available at $35* per pac.

• **The fully programmable with thermal printer.** The HP-97 ($750*) has all the capability of the HP-67, plus a built-in printer. Besides generating a permanent record of your results, the printer can aid significantly in programming, listing all or part of your programs complete with key memories, or tracing operations and results of each step of a running program to facilitate editing or debugging. While larger than pocket-sized, the HP-97 is small enough (8 by 9 by 2½ inches) to fit very portably into your briefcase.

**A few words about machine "language."**

All our hand-held and portable programmables incorporate HP's special RPN logic system. It's simple, direct, and trustworthy. You can work involved equations in the order you find natural, without having to enter or keep track of parentheses and brackets. And you get answers you can trust! Programs using HP's logic system are generally shorter than those using the logic systems of most other machines, making the most efficient use of program memory space. This, plus fully merged keystrokes, makes each programming step a stride, not a hobble.

**And added value.**

If you are a professional, or aspire to be one, there's a programmable with the range of power and features you need. Plus the engineering quality that assures reliable performance. Plus worldwide service. Plus manuals and aids to make programming a snap. If the notion of having a personal computer in your pocket or briefcase seems attractive, the next step is up to you.

## This HP computer advance automates financial reporting—and doubles as a scientific problem solver too.

For technical enterprises big enough to computerize their financial accounting, and small enough to put the same computer to work for scientific problem solving, the new HP 9896 desk-top system can fill both needs admirably and affordably.

The HP 9896 system puts it all together in a functional desk enclosure. Controlling the system is the HP 9831A desk-top computer, with integral keyboard and display. It has a standard read-write memory of 7,162 bytes, and is programmable in easy-to-use BASIC language. The system also includes two flexible disk drives for rapid access to stored programs and data, and a 9871A impact printer whose 96-character interchangeable print disks provide typewriter-quality printouts.

You also get a complete package of programs for general ledger, payables, receivables, payroll, and inventory. For larger organizations, the system can handle selected tasks in areas of production control, forecasting, financial analysis, or personnel records management.

The 9896 also has a library of scientific programs that include medical data analysis, structural and civil engineering data analysis, and general statistics for scientific problem solving.

The 9896 system is expandable. Memory can be added by 8 Kbytes up to 32 Kbytes. Two additional disk drives can be added, each providing 500 Kbytes of storage. Other available peripherals include a thermal printer, paper tape reader or punch, card reader, and data cartridge cassette memories.

Basic system price is $21,700*. If you'd rather buy programs independently, complete 9896 system hardware without business information management programs costs $18,700*. Leases start at $500* per month. As a Hewlett-Packard customer, you are assured of one responsible source for computer equipment, programs, training, support, and after-sale service.

HEWLETT hp PACKARD

1503 Page Mill Road, Palo Alto, California 94304

For assistance call: Washington (301) 948-6370, Chicago (312) 255-9800, Atlanta (404) 955-1500, Los Angeles (213) 877-1282.

Mail to: Hewlett-Packard, 1503 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) HP hand-held and portable programmables
( ) HP 9896 computer system

Name ____________
Company ____________
Address ____________
City ________ State ______ Zip ______

*Domestic U.S. prices only.

00743

# The Companions of Sunlike Stars

*Many stars have other stars as companions; do any of them have planets? An intensive spectrographic search for companions of 123 comparatively nearby stars provides a clue to the frequency of such planetary systems*

by Helmut A. Abt

The sun is accompanied on its journey through space by a retinue of nine planets, 32 natural satellites revolving around the planets, hundreds of comets and thousands of asteroids. Do other stars have similar arrays of companions? At present there is no direct way of telling, because if such a collection of bodies were associated with a nearby star, the most powerful telescopes on the earth could not detect them. The feeble light reflected from the companions would be lost in the brilliant glare of the central star.

Someday, perhaps, when large telescopes in space enable astronomers to view the stars unhindered by the distortions of the earth's atmosphere, it may be possible to resolve the brightest of the planetary companions around another star. With more sensitive auxiliary instruments ground-based telescopes may even be able to determine whether a star is being perturbed by a large planet revolving around it. There is some evidence that certain nearby stars, notably Barnard's star, may be attended by such unseen "dark companions," but the available data are inconclusive.

On the other hand, there are many stars that are known to have other stars as companions. In some cases the members of a stellar pair are identical in brightness; in other cases the secondary star is quite faint. The sun does not appear to have a companion star, although the possibility has not been completely eliminated that a faint companion star exists far outside the edge of the solar system. Is it possible that multiple-star systems are formed in a variety of combinations, ranging from pairs of nearly equal brightness through pairs of quite different brightness to solitary stars with planets?

In 1960 Frank D. Drake, who was then working at the National Radio Astronomy Observatory in Green Bank, W.Va., devised an experiment called Project Ozma, which was designed to detect other civilizations by listening for intelligible radio signals from certain nearby stars. For his first attempts Drake selected stars not known to have companion stars, since it seemed to him they were the ones most likely to have planets. Was that a reasonable approach?

In an effort to answer questions such as these my colleague Saul Levy and I recently undertook a different kind of investigation at the Kitt Peak National Observatory in Arizona. We studied the spectra of a large number of comparatively bright stars for clues to the possible existence of unseen companions. In so doing we obtained some important new information about the frequency of occurrence of various multiple-star combinations and about the likelihood of eventually finding other planetary systems.

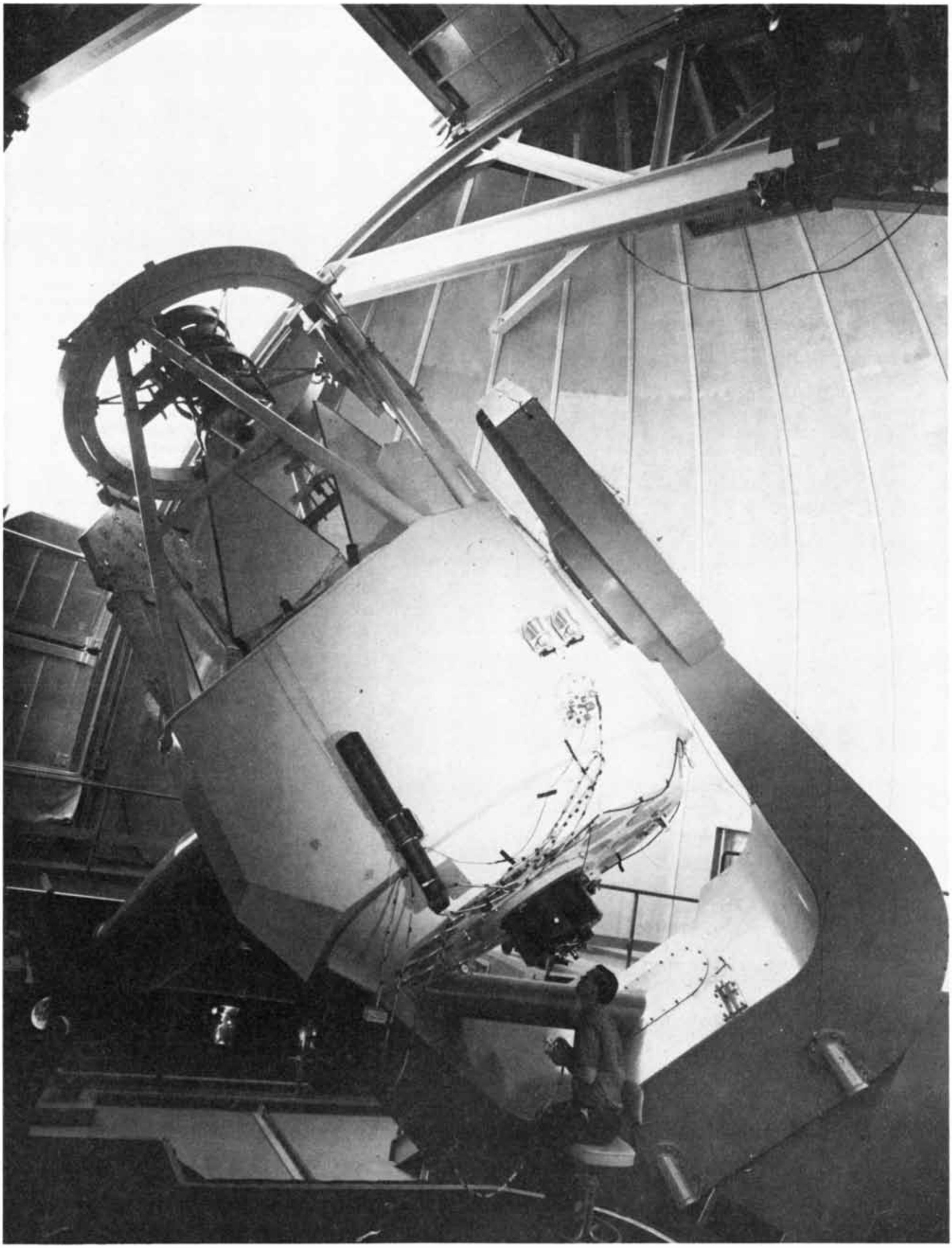
As a star the sun is probably not much different from many other stars in the sky. It is true that one can see more features—sunspots and prominences—on the sun than on any other star, but that is because the sun is the only star whose surface can easily be resolved from the earth. Other stars would appear as pinpoints of light through even the largest telescopes if it were not for the fact that the earth's turbulent atmosphere smears out those pinpoints into apparent disks. There are several sophisticated methods that show gross surface features in other stars, for example by studying certain stars as they are gradually eclipsed by other stars. In general, however, it seems reasonable to suppose that millions of stars are essentially identical with the sun. Stars that have roughly the same surface temperature (6,000 degrees Kelvin), the same diameter (1.4 million kilometers) and the same evolutionary history as the sun has are accordingly called sunlike stars.

There are basically four ways to determine whether a star is a member of a multiple-star system. The most obvious way is to watch (usually through a telescope) two stars that are close together and to record their relative positions in order to determine whether they gradually revolve around a common center of gravity. A complete revolution may take decades or centuries, but precise measurements can often show during a tenth of a revolution that the stars are following elliptical paths rather than the straight paths that stars accidentally passing in space would follow. Stellar pairs identified in this way are called visual doubles.

Sometimes the stars of a pair are so widely separated that for them to revolve once around each other would take thousands or millions of years. It may not be possible to observe any orbital motion in a human lifetime, or even in several human generations. Such stars may appear to have the same relative positions for decades, whereas another two stars, one near and one far away, that happen to be seen projected close together on the sky would appear to drift gradually apart, because each star typically has its own motion through space. Pairs of stars that are observed to stay together as they move across the sky are called common-proper-motion pairs.

In some cases the companion star may be too faint or too close to the brighter star to be observed directly, but many accurate measurements of the brighter star's position may reveal that its path across the sky is a wavy line. Stars tend to move in straight lines unless they are perturbed by a companion, so that a star showing this kind of motion may be a member of a pair. A stellar system of this type is called an astrometric double. Few such systems are known because it generally takes hundreds of measurements of the position of an individual star to detect a significant perturbation.

Finally, there are some stars that appear to be single in a telescope but are revealed by their spectrograms to be doubles. If the light from a star is spread out by color (or wavelength) with a prism or a diffraction grating, a pattern of absorption lines, or gaps in the array of colors, appears. The pattern of lines is characteristic of the star's temperature and composition. Each line is caused by a certain chemical element that absorbs



**2.1-METER REFLECTING TELESCOPE at the Kitt Peak National Observatory in Arizona was used by the author and Saul Levy to determine the frequency of multiple-star systems among the 123 sunlike stars visible to the unaided eye in the Northern Hemisphere. A total of 20 spectrograms of each star were obtained with the aid of the large spectrograph at the coudé focus (*see diagram on next two pages*). The 2.1-meter reflector is unusually compact for such a large telescope, owing to small focal ratio (*f*/2.63) of its primary mirror.**

that specific wavelength of light. The relative positions of the lines are known to a high accuracy (six or seven decimal places), but for most stars the entire pattern is displaced to shorter or longer wavelengths because of the star's motion with respect to the observer (the Doppler effect). If the star is approaching, the lines are shifted toward the violet end of the visible spectrum; if the star is receding, the lines are shifted toward the red end. An overall red shift of this type, known as the cosmological red shift, is characteristic of distant galaxies and furnishes the basic evidence for the widely accepted theory that the universe is expanding.

For some stars the Doppler shifts vary with time. On some days the lines are shifted toward the violet end of the spectrum and on other days they are shifted toward the red end. In most such cases the variable Doppler shifts are attributable to a star that is revolving in an orbit around a common center of gravity with an invisible companion. The companion is invisible simply because it happens to be much fainter than the primary star; hence the spectrum of the system is dominated by the primary. Such a system is called a spectroscopic double, since its duplicity is apparent only with the aid of the spectrograph.

If the two components of a spectroscopic double are nearly equal in brightness, the spectra of both stars are superposed on the spectrogram. The lines are double when one star is approaching and the other is receding; the lines are single when the orbital motions are at right angles to the line of sight. Stellar pairs of this type are called double-lined spectroscopic doubles.

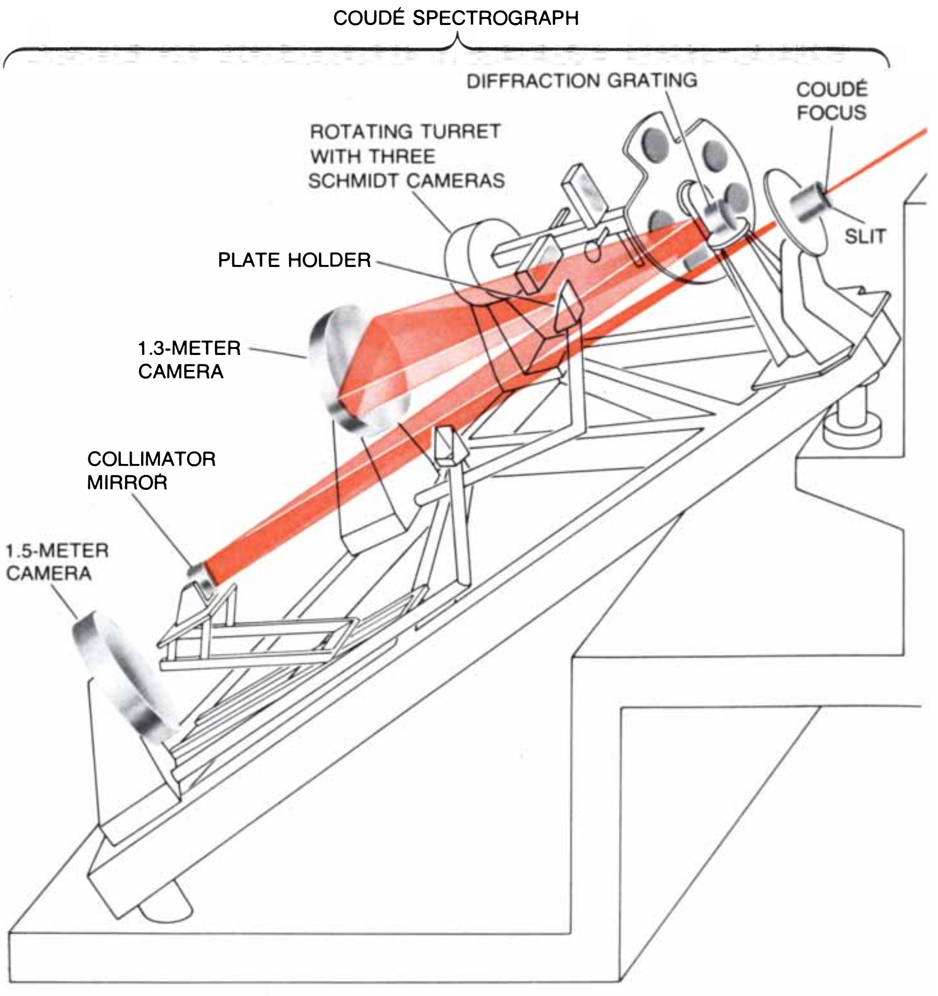
Astronomers have been collecting data about various double-star systems among the brighter sunlike stars for nearly a century. Attempts were made in the 1930's and the 1960's to determine what fraction of all the stars visible to the unaided eye are double stars, but the estimates differed drastically. At one extreme the estimate was that only 18 percent of the sunlike stars are double; at the other extreme the estimate was 54 percent. The difference arose from different interpretations of whether the Doppler shifts of many of the stars are really variable or only reflect measuring inaccuracies.

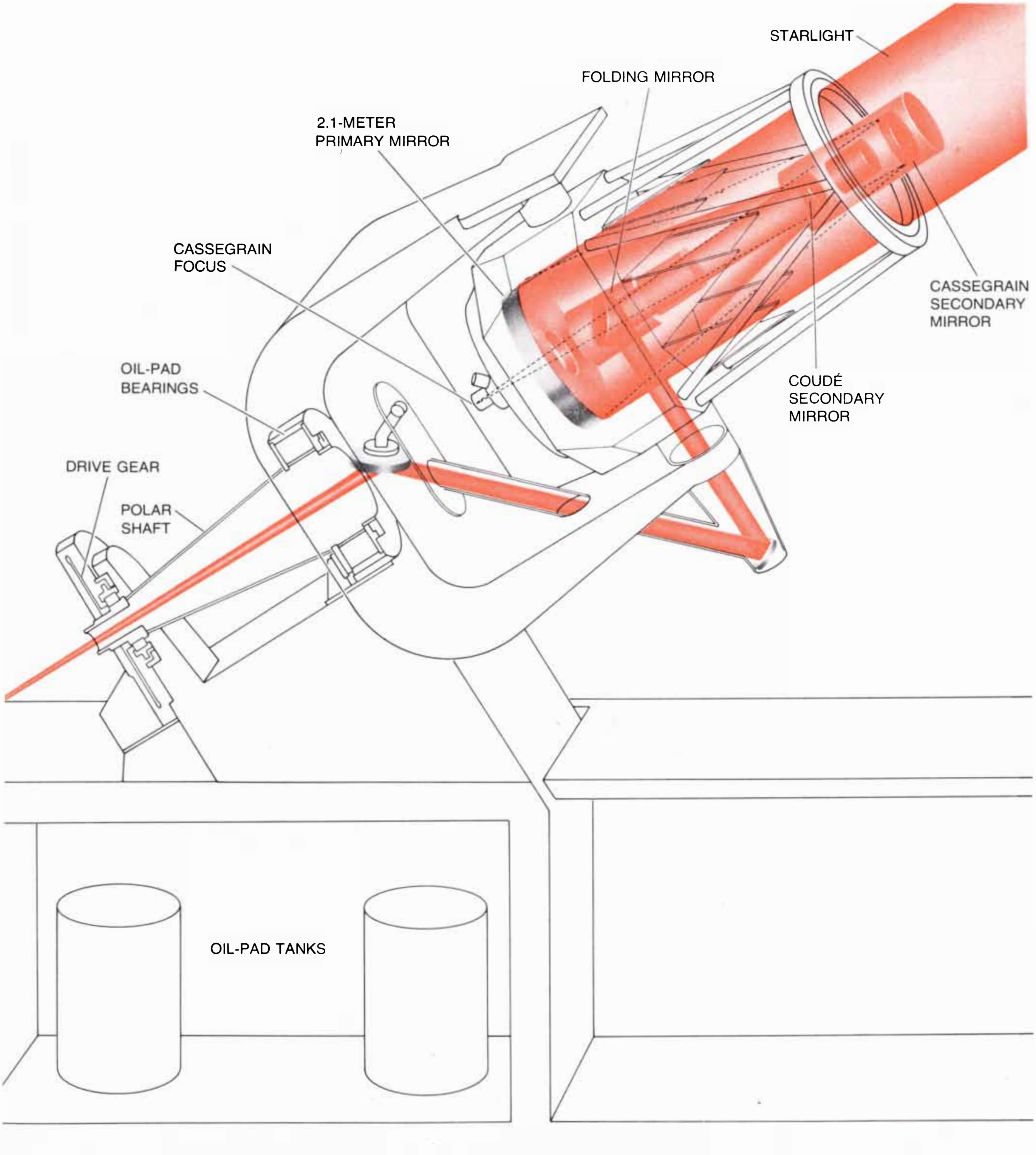
A few years ago Levy and I decided to determine more carefully the frequency of doubles and other multiple stars among sunlike stars. The reasons for the project were several. One was to find out whether in the process of star formation the tendency for single stars to form was greater than that for multiple stars to form. Another was to learn, if we could, something about the probable incidence of planets outside the solar system. There were additional reasons for undertaking the project, but as happens so often in science many of the more interesting results come as a surprise.

Levy and I observed all 123 sunlike stars visible to the unaided eye in the Northern Hemisphere. Each of these stars is within 85 light-years of the sun, and it seemed likely that among them essentially all the visual doubles that could be detected with current equipment had already been detected. For many of these stars, however, there was no evidence for variable Doppler shifts. Therefore we obtained about 20 spectrograms of each star with the large spectrograph installed at the coudé focus of the 2.1-meter (84-inch) reflecting telescope on Kitt Peak. The spectrograms were made at random times because we had no knowledge in advance whether individual stars were double, and if they were, whether they would have orbital periods measured in days, months or years. By measuring the Doppler shifts at random times for each star, however, we could search for and usually find the period that causes the Doppler shifts to repeat periodically. In our sample of 123 stars we discovered 25 new spectroscopic doubles and improved the orbital elements for most of the 21 spectroscopic doubles previously known. In addition there were 23 visual doubles with known orbital elements in the sample and a total of 25 common-proper-motion pairs. Some pairs were detected in more than one way, so that the net total was 88 pairs.

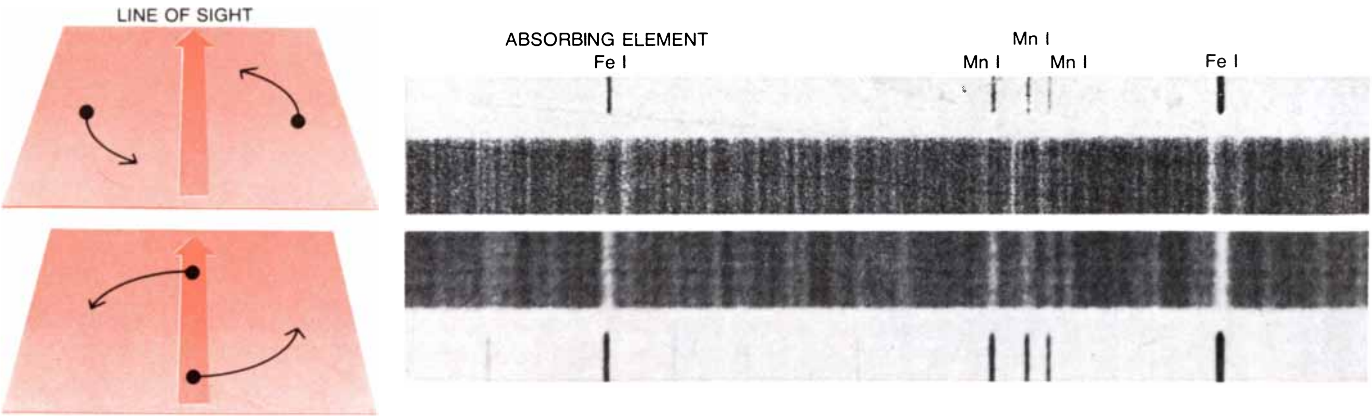
Some of the systems we observed are triple or even quadruple. A typical example would be a visual pair in which one component is a spectroscopic double, making a total of three stars, even though the spectroscopic companion is not actually seen. Among the 123 primary stars 57 were found to have one companion, 11 to have two companions and three to have three companions. In percentages the observed frequencies of single to double to triple to quadruple stars are 42 to 46 to 9 to 2 percent. Thus more than half (57 percent) of the stars have at least one stellar companion.

The preceding inventory is only part of the picture, because we are unable to detect all the companions that are likely to be present. For example, if a pair of stars were so close together that the components could not be detected as a visual pair and it happened that the plane of the orbits was almost exactly perpendicular to the line of sight, then as






**ALTERNATE LIGHT PATHS** available with the Kitt Peak 2.1-meter reflector are shown in this cutaway schematic diagram. In the Cassegrain arrangement a convex mirror at the upper end of the telescope tube reflects the light from the 2.1-meter primary mirror through a hole in the center of the primary mirror to a focus just below the lower end of the tube. Access to the Cassegrain focus is provided by means of a mobile steel platform. In the coudé arrangement an inside ring at the upper end of the telescope tube is rotated to substitute the convex coudé secondary mirror for the Cassegrain mirror. Three additional flat mirrors then reflect the starlight to a focus south of the telescope mounting, where the coudé spectrograph is permanently mounted. The circuitous light path was chosen to avoid the necessity of attaching the 13.5-ton spectrograph directly to the telescope. The spectrograph consists of five large cameras, ranging in diameter from .5 meter to 1.5 meters, that photograph spectra produced as a result of the dispersal of starlight by the diffraction grating at the coudé focus.




**DOUBLE-LINED SPECTROSCOPIC BINARY, designated 64 Piscium, is represented by this pair of spectrograms, made on different nights. Stellar spectrograms of this type are continuous streaks of color (*dark horizontal bands on black-and-white negative prints*) with narrow absorption gaps (*light vertical lines*) attributable to the presence of certain atoms in the atmospheres of the stars. In this case the dark (positive) reference lines for the stellar absorption lines were produced by an arc lamp and correspond to various emission lines**

the stars moved in their orbits they would be neither approaching nor receding by significant amounts. The Doppler shifts would therefore be virtually constant regardless of how fast the stars were traveling in their orbits. We would not be able to detect such a "face on" pair.

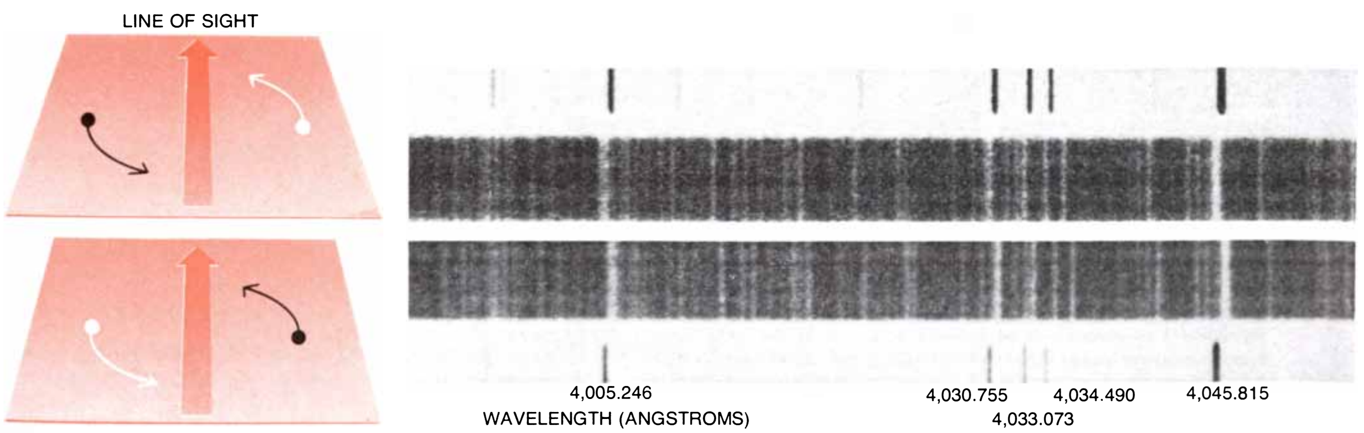
It is possible, however, to tell how many such face-on doubles were missed. There is good evidence that the orbital planes of spectroscopic doubles are oriented at random. The best evidence comes from determining the fraction of stars in various parts of the sky that are eclipsing doubles. These doubles are at the opposite extreme: they are ones whose orbital planes are in the line of sight. Su-Shu Huang and Clarence Wade, Jr., of Northwestern University found that the fraction of eclipsing doubles is the same in all directions, so that orbital planes are randomly oriented. We can therefore calculate statistically how many doubles we failed to detect because their orbital planes are perpendicular to the lines of sight. It turns out that we probably missed only one pair.

More important, there are many doubles that cannot be detected because the mass of the companion is too small to cause detectably variable Doppler shifts of the primary star. The observations themselves show that if the orbital speed of a primary is less than about two kilometers per second, it is very difficult to distinguish between a real variation and an effect caused by measuring inaccuracies. Typical orbital speeds for the spectroscopic doubles previously known are anywhere from 10 to 75 kilometers per second; for the newly discovered doubles the orbital speeds are between two and 10 kilometers per second.

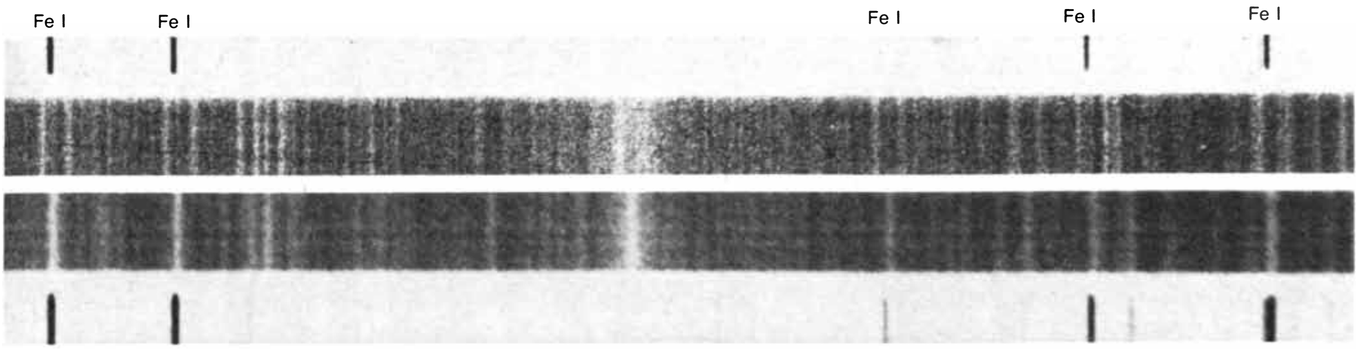
Again it turned out that after consideration of the doubles we did detect we could estimate how many we missed because the companions have a small mass. For systems in which the companion has half the mass of the primary star, we probably missed less than one spectroscopic pair for the 12 systems we detected; for systems with companions that are an eighth of the mass of the primary star, we probably missed two spectroscopic pairs for the six spectroscopic pairs we discovered.

There are undoubtedly other pairs of stars that escaped detection, for example a widely spaced pair that can easily be resolved with a telescope but in which the companion is so faint that it is lost among the many faint background stars that surround any bright star. Instead of going into further details about what types of missing pairs can or cannot be predicted and allowed for in our attempt to count all the pairs in a sample of stars, however, I shall proceed to our results.

In our sample of 123 primary stars we counted the number of companions, observed or inferred from the incompleteness calculations, with various masses. We combined the pairs in groups: first those with approximately equal masses, then those with secondary stars of approximately half the mass of the primary stars, and so on. We also separated




**SINGLE-LINED SPECTROSCOPIC BINARY, named ι Pegasi, can be detected in this pair of spectrograms, even though the spectrum of the secondary star does not appear at all. In the upper spectrogram the stellar absorption lines are shifted to the left of the reference lines, indicating that the visible primary star was approaching on that night. In the lower spectrogram the stellar lines are shifted to the right, in-**




**associated with singly ionized iron atoms (Fe I) and singly ionized manganese atoms (Mn I). In the upper spectrogram the absorption lines are double because one component of the system is approaching while the other is receding. (According to the Doppler effect, the spectral lines of an approaching object are shifted toward the violet end of the spectrum, whereas the lines of a receding object are shifted toward the red end.) In lower spectrogram both stars are moving at right angles to line of sight, so that two sets of lines are coincident.**

the pairs by their orbital period. The results were then arranged in a graph [*see illustration on next page*].

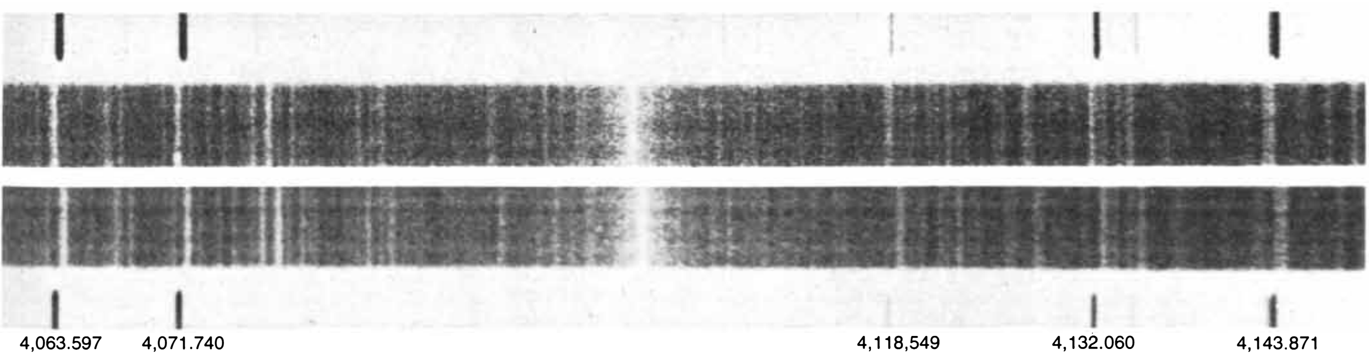
In the graph the results for pairs with an orbital period of less than 100 years are shown in color; the results for pairs with longer periods are in black. There are some missing points because we were unable to compute very well the numbers of undetected pairs in certain circumstances.

It is evident from the graph that for the pairs with an orbital period of less than 100 years, as one progresses from pairs with nearly equal masses to pairs with small secondary masses one finds fewer and fewer systems. The results for the pairs with a period of more than 100 years are the opposite: as one progresses to lower masses one finds more pairs, at least down to secondaries with a fourth of the primary mass.

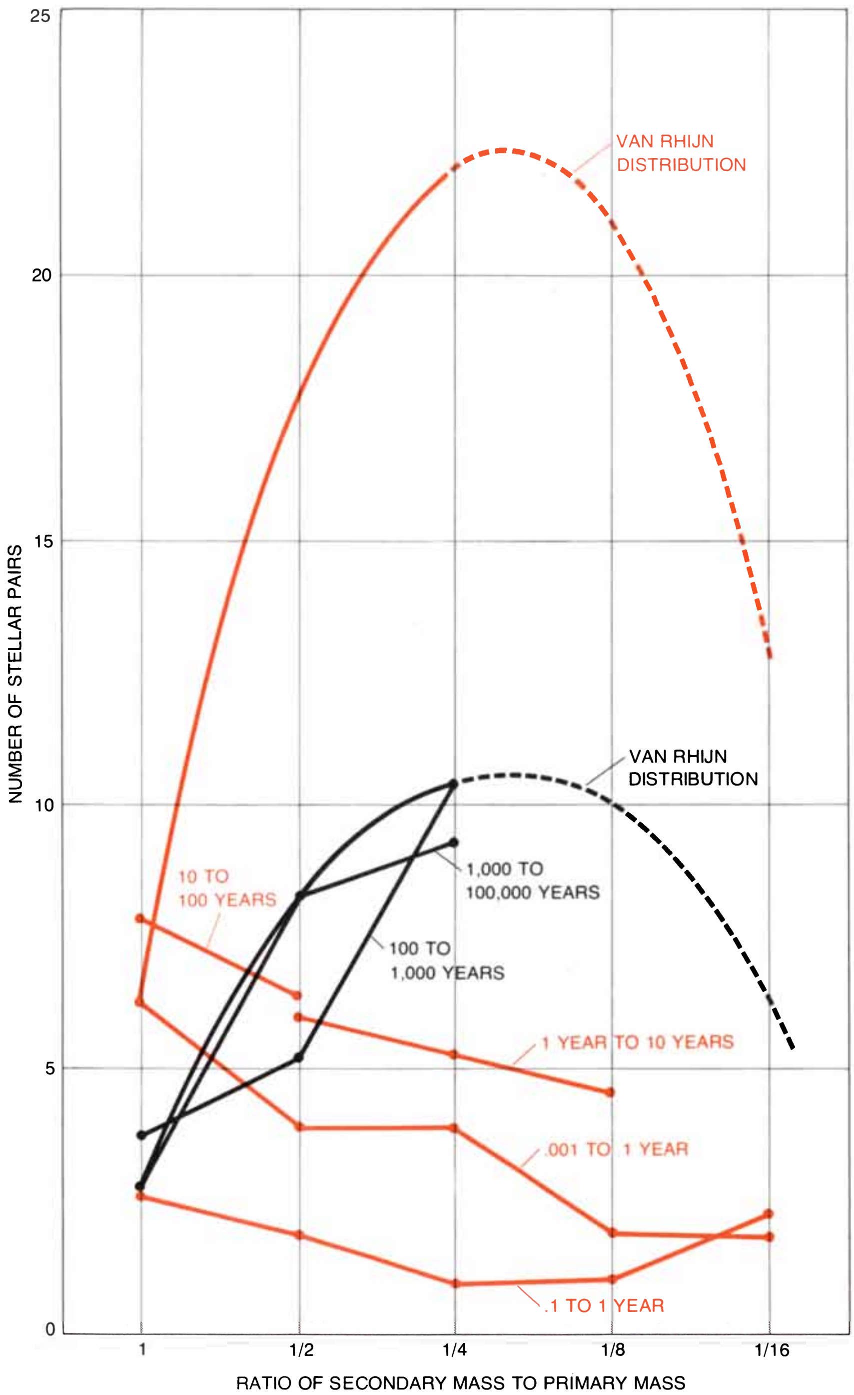
The primary stars themselves follow a smooth curve called the van Rhijn distribution, after the late Dutch astronomer Pieter J. van Rhijn, who counted stars in sample areas of the sky and with known individual distances. He found, for instance, that if within a given volume of space near the sun there are six stars of approximately one solar mass, there will be 18 stars with half that mass, 22 stars with a fourth as much mass and then perhaps decreasing numbers with still smaller masses. Van Rhijn's data were probably not very good for stars with a mass that is less than a fourth of a solar mass; hence beyond that point the curves are broken. For the pairs with a period of more than 100 years the curve is adjusted arbitrarily to start with 2.7 stars of approximately one solar mass. Apparently in such pairs both the primaries and the secondaries follow the van Rhijn distribution, but they do not do so for the pairs with a period of less than 100 years.

These data suggest that there are two types of double-star system. Before discussing the significance of this finding, however, it will be helpful to review something of what is known about the formation of stars.

Very young stars can be recognized in several ways. One of the easiest is to identify any star that is burning its nuclear fuel so rapidly that its entire lifetime as a dwarf star will be less than a million years. Such stars are almost always found in clouds of gas, such as the Great Nebula in Orion. In rare cases they are also found traveling rapidly through space away from their points of formation; such stars are known as runaway stars. Both from the observed incidence of young stars in clouds of dust and gas and from theoretical calculations it is apparent that stars originate in dense clouds. The mechanism of formation is only partly understood. It is necessary for parts of the gas clouds to reach a critical density such that the inward force of self-gravitation overcomes the outward forces of gas pressure and turbulence and also the forces exerted by magnetic fields, which affect the motions of charged particles. Parts of a cloud may reach such a density because of accidental jostling or a shock wave from a nearby supernova explo-




**dicating that the primary star was receding on that night. The shifts measure about .04 millimeter in the original spectrograms and correspond to an orbital motion of approximately 40 kilometers per second. All four spectrograms on these two pages, made with the coudé spectrograph of the 2.1-meter telescope on Kitt Peak, show a small portion of violet region of spectrum; longer wavelengths are to right.**




**RESULTS OF THE AUTHOR'S SPECTROGRAPHIC SURVEY of the frequency of multiple-star systems among the 123 sunlike stars visible to the unaided eye in the Northern Hemisphere are plotted in this graph. First the observed (and inferred) stellar companions were classified according to mass. Pairs with approximately equal masses were combined in one group; pairs with secondary stars of approximately half the mass of the primary stars were combined in another group, and so on. The pairs were then segregated by their orbital period. Pairs with a period of less than 100 years are represented in color; pairs with longer periods are shown in black. It is evident that for the pairs with an orbital period of less than 100 years, as one progresses from pairs with nearly equal masses to pairs with small secondary masses one finds fewer and fewer systems, whereas for the pairs with an orbital period of more than 100 years, as one progresses to lower masses one finds more pairs, at least down to secondaries with a fourth of the primary mass. For the pairs shown in black both the primary stars and the secondary stars appear (within the accuracy of the measurements) to follow a smooth curve called the van Rhijn distribution, but for the pairs shown in color they do not do so, indicating that there are two types of double-star systems: stars that were formed together as a pair and stars that were formed separately. Extrapolation of the data presented here for both types of double-star systems suggests that all of the 123 primary stars in the sample have companions: 67 percent of the secondaries are normal stars, 15 percent are black dwarfs and 20 percent are planets.**

sion. When the conditions are right, however, an entire cluster of stars is generally formed at one time.

The dense parts of gas clouds have much internal motion or turbulence that is not easily eliminated. Normally as a protostar, or part of a contracting cloud, contracts it rotates faster, eventually reaching the point where it becomes unstable. The protostar may partly avoid that fate by throwing off the particles that have the most angular momentum, but it cannot succeed entirely in eliminating the rapid rotation. If a cloud with a mean density of 100 atoms per cubic centimeter produces stars with a mean density of $10^{23}$ atoms per cubic centimeter, the mean internal gas motion of the cloud, which is about one kilometer per second, would increase to 10 million kilometers per second. Such a speed is 20,000 times faster than the highest rotational speed possible for a stable star. Again, separating out and throwing off the atoms with the most angular momentum will not solve the problem entirely. How then can a single stable star be formed? Perhaps it cannot.

When a protostar contracts and spins faster, its shape deviates from the roughly spherical form. First it becomes an oblate, or flattened, spheroid and then at still higher speeds it becomes an ellipsoid with three unequal radii. Finally the object divides into two separate ellipsoids, a process of bifurcation that leads to a double star with a small separation. In the process roughly 99 percent of the angular motion goes into the orbital motion and only about 1 percent goes into the rotation of the two stars. Hence by splitting in half and by losing some of its faster atoms a rapidly rotating protostar can develop into a pair of stable stars.

The mechanism described above seems to be the one that applies to the stars in the first category: those pairs with periods of less than 100 years. The decreasing frequency of double stars that have a secondary star with a smaller mass indicates that in the bifurcation process the dominant tendency is for the protostar to divide into a pair of stars with equal masses. Pairs with very unequal masses are less frequent.

How many such pairs of bifurcation doubles are there in our sample of 123 primaries? All the jagged colored lines in the graph have roughly the same slope, within our measuring accuracy for such a small sample of stars. (At this stage we wish we had started with a larger sample, but the observations of the 123 stars took 60 nights of observing time with a large telescope.) If one extrapolates those lines and counts up all the secondaries with masses ranging from the same mass as the primary to a sixteenth of the mass of the primary, one finds that there are 83 close secondaries among the 123 primaries; that is, two-thirds of the primaries have nearby secondaries.

A mass equal to a sixteenth of the mass of the sun is the smallest amount of material that can give rise to a star generating light from nuclear reactions in its interior. This important fact was discovered in 1963 by Shiv S. Kumar of the University of Virginia. If a gaseous sphere has less mass, the weight of that mass pushing on the central core will not produce a temperature sufficiently high for nuclear reactions to occur. In the sun the weight of the material causes the gas temperature to be about 14 million degrees Kelvin at the center, at which point the atomic nuclei (mostly protons) have an average speed of 100 kilometers per second. The fastest of the protons have enough speed so that when two of them are on a collision course they will collide, in spite of the repulsion due to their like electric charges. After the collision they are likely to form a deuterium nucleus and after additional collisions a helium nucleus. One helium nucleus weighs slightly less than the four hydrogen nuclei that formed it, and the material lost turns up as energy in the amount $E = mc^2$ (the energy generated is equal to the mass lost times the speed of light squared). In a gaseous sphere with a mass equal to less than a sixteenth of a solar mass, however, the central temperature is less than a million degrees K. and the protons are traveling with an average speed of only 25 kilometers per second. Virtually none of those protons have a speed great enough so that two protons on a collision course can overcome their electrical repulsion. The protons will approach each other, slow down and then back away without ever having the chance to collide and form a heavier nucleus.

What happens to those objects that start out to become stars but never do so because they have too little material in them? Kumar traced their history and found that the ones with masses of between a sixteenth and a hundredth of a solar mass will glow for nearly a billion years while they are contracting because they are converting gravitational energy into electromagnetic energy. After that time they will continue to exist indefinitely as nonluminous black dwarfs (not to be confused with black holes, which are something entirely different).

A gas sphere with a mass of less than a hundredth of a solar mass will hardly glow at all (and then only in the infrared wavelengths). Such an object—a sphere of gas that has too little mass to become a star or even a black dwarf—would be called a planet. If a body of this kind had a very small mass, it would have a solid surface.

It is unlikely that the jagged colored curves in the graph drop to zero at masses of less than a sixteenth of a solar mass for the secondary bodies. The lines probably continue on, indicating that some of the secondaries will be black dwarfs or planets.

Originally we thought that the average slope of the colored lines in the graph was such that the number of double systems was proportional to the cube root of the mass of the secondary. For example, at an eighth of a solar mass there would be half as many double systems as there would be at one solar mass: $(1/8)^{1/3} = 1/2$. David R. Branch of the University of Oklahoma pointed out two effects that were overlooked, with the result that the cube-root relation may actually be another relation, possibly a square-root one. In other words, an extrapolation from the present data is risky. Physically that means we have not actually observed doubles with black dwarfs or planets as secondaries, and our data for the more massive secondaries are not good enough to extrapolate with confidence.

Nevertheless, if one extrapolates on the basis of the cube-root relation, the results make good sense. If one continues the colored curves first to a hundredth of a solar mass and then to zero mass, one finds that there are about 20 secondaries that are black dwarfs and 25 that are planets. The sum of these plus the 83 stellar secondaries is almost exactly 123, which means that all the primaries have secondaries! In 67 percent of the cases the secondaries are normal stars, in 15 percent they are black dwarfs and in 20 percent they are planets. The reason such a result makes sense, in spite of the weak extrapolation on which it is based, is that it means all primaries are bifurcation doubles; the problem of forming stable stars in spite of the excess angular motion is roughly solved by most of the motion's going into orbital motion.

The sun is among the 20 percent of the stars in the sample that have planets for companions. The fact that in the solar system the secondary developed into a set of planets rather than into one giant planet does not alter the model significantly. Thus we conclude, as others have done on a more intuitive basis, that in searching for planets one should concentrate on those stars that are not known to have stars as companions.

What can one say about the widely spaced pairs: the pairs represented by the black curves in the graph of our results? A pair with an orbital period of more than 100 years is clearly not the result of a bifurcation of a rapidly rotating protostar. Such a wide pair is probably the result of two protostars that contracted separately but that are bound to each other by gravity. Since they represent two parts of a gas cloud, they can start with any amount of gas. The frequency distribution of their masses approximates the van Rhijn distribution for both the primaries and their distant companions. Evidently the way a turbulent gas cloud under the forces of self-gravitation, internal pressure and interstellar shock waves breaks up into stars of various masses results in the van Rhijn distribution, even though astronomers are currently unable to predict that distribution from theory. The important fact here is that the tendency to form roughly equal primary and secondary masses that arises from bifurcation is not evident among the wide doubles.

If one extrapolates the data represented in black in the graph with the use of the van Rhijn distribution, one finds that nearly three-fourths (72 percent) of the primaries have distant stellar companions (primaries with masses as small as a sixteenth of the sun's mass). If one knew the van Rhijn distribution beyond a sixteenth of the solar mass, the curve might not drop to zero but might include the formation of black dwarfs and planets. It could turn out that most or all primaries have distant companions as well as nearby ones.

Would it not be true, then, that the distant protostars will have the same problem of having too much angular motion to form a stable star, a problem that can be solved only by bifurcation? It would appear so. I would guess that the distant companions are really close bifurcation pairs, just as the primaries are. In other words, the most likely situation is a close bifurcation pair gravitationally held to another distant bifurcation pair, in short, a quadruple system. In either pair the less massive companion can be either a star (in roughly two-thirds of the cases), a black dwarf (a sixth of the cases) or a planet (roughly a sixth of the cases).

The foregoing analysis conjures up a tantalizing picture where most stars form as members of multiple systems. Such an outcome is not completely unexpected, since it has been thought for some time that the individual stars in the sky are all escapees from clusters that have gradually disintegrated. This suspicion comes from the theoretical difficulty of trying to explain how a single cloud of say one solar mass contracts to form a star. It is also known that most open clusters are gravitationally unstable and will lose stars one by one during the millions of years of their existence.

The picture is only tantalizing, because the data presented are not numerous enough or sensitive enough for one to be sure about the extrapolation to black dwarfs and planets. That conclusion must await the development of space telescopes or better equipment for ground-based telescopes to provide much more reliable information on the frequency of planets. Meanwhile one may be allowed to wonder about the conditions on those planets and to speculate about whether radio signals from the earth are likely to raise a listener and perhaps even a speaker at the other end.

# New Sinclair Cambridge Programmable. An astonishing $29.95!

**How pocket calculators grew up**

A couple of years ago, calculators took a step forward. Programmability transformed the slick slide-rule calculator into an advanced scientific machine.

Sadly, it also transformed a cheap little calculating aid into a piece of capital investment.

*Now the all-new Sinclair Cambridge Programmable puts programmability where it belongs: in the palm of your hand, for less than $30.*

**The features of the Sinclair Cambridge Programmable**

The Cambridge Programmable is genuinely pocketable. A mere 4½" x 2", it weighs about 2oz.

Yet there is absolutely no compromise in the package of functions it offers.

Because the Cambridge Programmable is both *a scientific calculator* with memory, algebraic logic and brackets (which means you enter a calculation exactly as you write it), and a *programmable calculator* which offers simple, flexible through-the-keyboard program entry and operation.

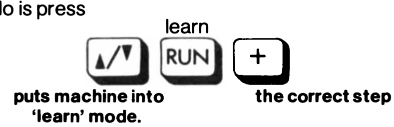
The Cambridge Programmable has a 36-step program memory, and features conditional and unconditional branch instructions (*go to* and *go if negative*).

There is also a step facility, which allows you to step through the program to check that it has been entered correctly. If there is any programming error, the learn key allows you to correct single steps without destroying any of the remainder of the program.

To achieve this, each program key-stroke has an identifying code, or 'check symbol'. (The symbols for the digit keys are the digits themselves, while the symbols for the operator keys are letters printed beside the keys.)

The check symbol for ⊟, for example, is F. So if, as you step through the program, the display shows


it means that ⊟ is programmed as step 26. If step 26 should have been ⊞, all you have to do is press




It's as simple as that!

These facilities make the Cambridge Programmable exceptionally powerful, whether it's running programs you devise for yourself or the programs in the Program Library.

**Use the 294-program library to tailor the machine to your own specialty**

Like a full-size computer – and *unlike* far more expensive specialist calculators – the Sinclair Cambridge Programmable can be programmed to handle calculations concerned with *any* specialty.

And of course, whatever it's doing the Programmable is error-free – in fact, once it's programmed, it can even be given to an operator who doesn't understand the program!

To save you time, and to help inexperienced programmers, Sinclair have produced a library of 294 programs ready to be entered straight into the calculator.


Using these standard programs, the Cambridge Programmable solves problems from quadratic equations (where the program gives both real and imaginary roots) to twin-T filter design, and from linear regression to bond yields. It even plays a lunar landing game! To realise the full power of the Cambridge Programmable, the Program Library is a must.

(The calculator is supplied with 12 sample programs, and full instructions for entering your own program. The four books in the program library are available at $4 each, or $10 for the complete set.)

**Why the Cambridge Programmable costs so little**

The Sinclair Cambridge Programmable uses the Sinclair talent for miniaturisation to the full – as you'd expect from the company that pioneered the truly pocketable pocket calculator, and recently introduced the world's first pocket TV.

Chip and circuitry design are unique to Sinclair, and the Cambridge Programmable is assembled by Sinclair's own staff at their headquarters plant. Shipped direct, and sold to you direct, the Cambridge Programmable accumulates no middleman's profits on the way.

The result is a pocket programmable calculator of advanced design, sold by the manufacturer with the manufacturer's own 1-year comprehensive guarantee, at a price unmatched by any comparable calculator.

## 10-day no-obligation offer

There's a lot more to this remarkable calculator than a brief written description can cover.

You need to see it and handle it . . . to program it yourself in a few seconds to save you hours . . . to check its performance against tables and graphs . . . to test the full range of programs available . . . to evaluate, perhaps, its use as an educational aid in developing a student's computer understanding.

So we're offering a 10-day trial. Send your check or money order with the order form below, and you'll receive a calculator direct. Use it for 10 days, and if you don't feel it's the finest $29.95 you've ever invested, send it back. We'll refund your money without question.

There's nothing to lose, and so much calculating power to gain.

Post your order today.

sinclair

**Sinclair Radionics Inc., Galleria,**
**115 East 57th Street, New York, N.Y. 10022 USA.**
**Tel: (212) 355 5005.**

**To: Sinclair Radionics Inc., Galleria, 115 East 57th Street, New York, N.Y. 10022, USA.**

Please send me______(qty) Sinclair Cambridge Programmable(s) at $29.95 each, including full instructions, and sample programs. $______

______ (qty) AC line adaptor(s) at $4.95 $______

______ (qty) Program Library **Book 1** at $4 **General/Finance/Statistics** (percentage, metric system, memory, games, dates, finance, mortgages, statistics) $______

______ (qty) Program Library **Book 2** at $4 **Mathematics** (algebra, calculus, geometry, trigonometry, number theory, transcendental functions) $______

______ (qty) Program Library **Book 3** at $4 **Physics and Engineering** (astronomy, statics and dynamics, relativity, mechanics, properties of matter, fluids, structures, thermodynamics) $______

______ (qty) Program Library **Book 4** at $4 **Electronics** (networks, circuits, filters, electrodynamics, radiation and propagation) $______

______ **(qty) complete set(s) of 4 books at $10 per set** $______

Sales tax (NY residents) $______

Post and packing $ 2.50

**Total** $______

Enclosed is check/MO payable to Sinclair Radionics Inc.

Name ______

Address ______

City ______

State ______

Zip ______

(PLEASE PRINT)

I understand that you will refund purchase price in full if I return calculator(s) and accessories in saleable condition within 10 days of receipt.

Signature ______

# The Spread of the Bantu Language

*Bantu is spoken by 130 million people in southern Africa. Linguistic and archaeological evidence suggests that the original Bantu-speakers began their migration from a region in the north some 2,300 years ago*

by D. W. Phillipson

The 19th-century European explorers of Africa who outfitted their expeditions at Zanzibar or the adjacent mainland were surprised and pleased by a bonus they received: their porters and escorts spoke a language that was understood through wide areas of subequatorial Africa and could understand many of the tribal dialects they encountered in the eastern half of the region. The reason was that those concerned were speaking one or another form of Bantu, a family of languages common to more than 130 million black Africans, comprising several hundred separate tribal groups, who still inhabit much of Africa south of the Equator. The Zanzibar version of the language, Ki-Swahili, is enriched with loan words from several foreign tongues; its very name comes from the Arabic word *sahil,* or "coast," so that it can be rendered as "coast dialect." The basic form of Swahili, however, is pure Bantu.

The word Bantu is an exclusively linguistic label and has no other primary implications, either of race or of culture. The family of dialects was named by a 19th-century German philologist, Wilhelm H. I. Bleek. The words *ki-ntu* and *bi-ntu* in these African dialects respectively mean "a thing" and "things"; the words *mu-ntu* and *ba-ntu,* "a man" and "men." Thus *bantu* can be construed as "people" or "the people," and this seemed to Bleek a name eminently suitable for a language held in common by so many.

Since late in the 19th century it has generally been recognized by students of linguistics that the various Bantu tongues must have come from a common ancestral language within the relatively recent past. Indeed, it has been a prime goal for students of African history to elucidate the processes responsible for this remarkably widespread linguistic distribution. Unfortunately the oral traditions of the Bantu-speakers themselves generally extend only a short distance into the past. Furthermore, no Bantu language was reduced to writing until after the time of contact with Arabs or Europeans, so that both tradition and written records are silent on ultimate Bantu origins. To trace Bantu ancestry backward in time therefore becomes the task of archaeology and of historically oriented linguistic studies. Here I shall present the latest evidence concerning the dispersal of Bantu-speakers in Africa south of the Equator that has been developed by scholars in both disciplines.

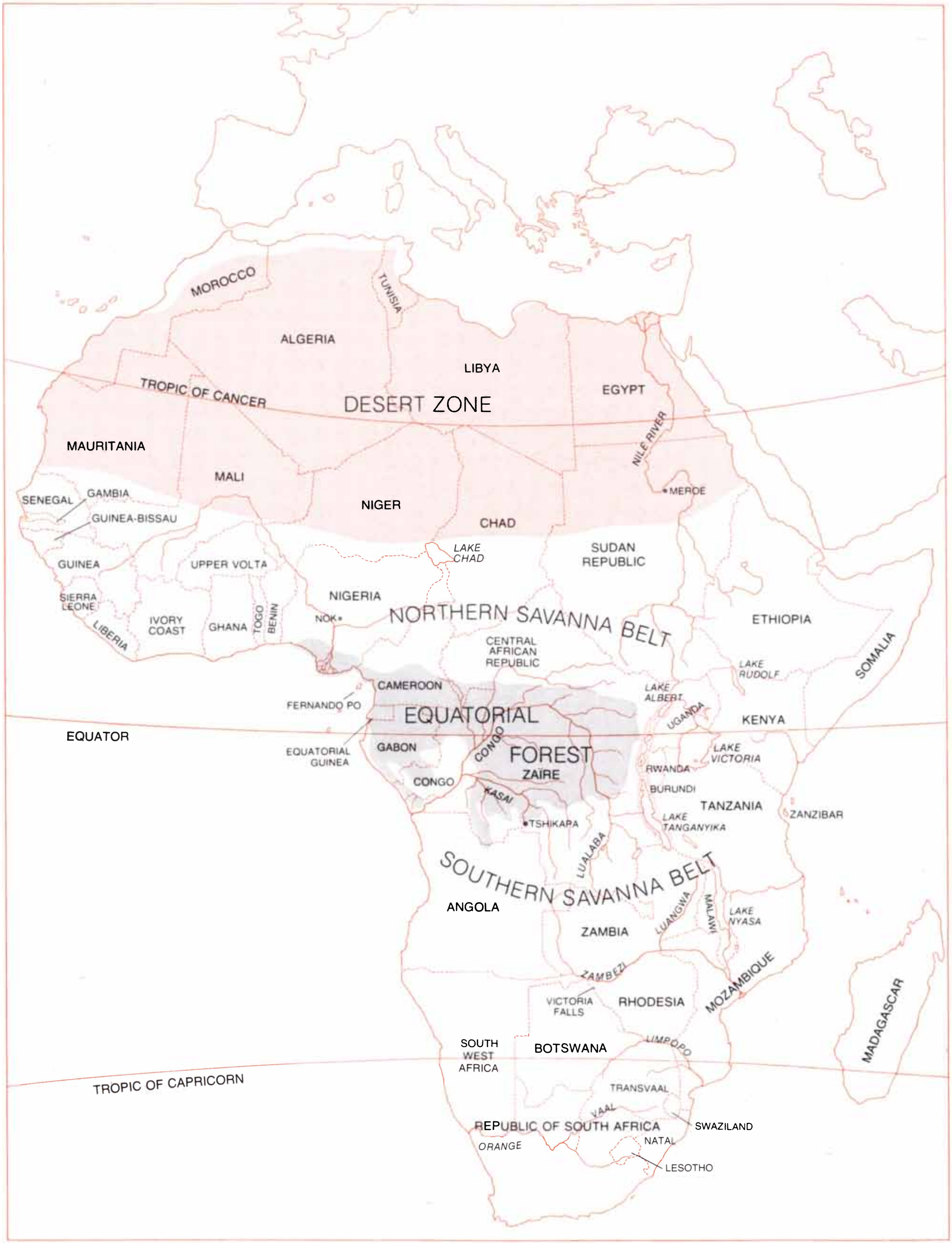
To deal first with the archaeological evidence, from about 300 B.C. to A.D. 600 a major change took place in the greater part of Africa lying between the Equator and the Vaal River. The change was marked by the appearance of a characteristic type of pottery that clearly belongs to a single stylistic tradition even though regional variations are readily apparent. The pottery is found in association with evidence for the working of metal: iron and in some areas copper. Because this is the earliest-known evidence of metallurgy in the entire subequatorial region the cultural complex to which the artifacts belong has been named the Early Iron Age.

The often elaborate decoration and the varied forms of Early Iron Age pottery allow detailed stylistic analyses of the complex, and these analyses in turn provide a framework for studies of its subdivisions and interregional relations. Although the extreme geographical limits of the complex are still imperfectly known, more than 350 Early Iron Age sites have now been identified. Virtually all are in areas that in more recent times are known to have been inhabited by Bantu-speaking people.

In most instances the Early Iron Age population appears to have consisted of settled village dwellers who practiced an economy of mixed farming, that is, both raising crops and herding animals. In the regions to the south of central Tanzania they are the earliest followers of such a life-style to appear in the archaeological record. Indeed, the contrast between the Early Iron Age population and its predecessors is much greater in southern subequatorial Africa than it is farther to the north. There, notably in the Rift Valley and the adjacent highlands of northern Tanzania and southern Kenya, settled communities practicing a pastoral economy and possibly cultivating cereal crops were established at least as early as 1000 B.C. To the south, however, it seems that Early Iron Age techniques of food production were introduced to a region where such practices had been virtually unknown and the indigenous population lived by hunting and gathering. Here, with the possible exception of the extreme south, pottery of the Early Iron Age is the oldest pottery known.

Many aspects of Early Iron Age culture other than pottery making and mixed farming evidently represented major innovations in most of the southern areas. The culture was introduced in full-fledged form almost everywhere, not only with respect to its various components but also as a viable socioeconomic system. Pottery that might be regarded as being in some way ancestral to Early Iron Age wares has yet to be found in any part of the region. Metallurgy too seems to have been introduced as an already efficient and fully developed technology into a region where previously even a rudimentary knowledge of metalworking did not exist. Again, the domesticated animals and several of the domesticated plants were species unknown in subequatorial Africa even in wild forms. It thus seems that the hypothesis viewing Early Iron Age culture as being alien to the region is amply supported. Evidently the culture was introduced by means of a rapid and coherent movement of people who brought with them a full-fledged culture that had undergone its formative processes elsewhere.

The scale of human migration involved in the dispersal of Early Iron Age culture in these parts of Africa is not easy to assess. Clearly the people moved in numbers sufficient to support a technology that was largely independent of the stone-tool-using technology of the peoples whose lands they were entering. Indeed, in many of these new territories the indigenous population continued to




**MAJOR GEOGRAPHICAL BARRIERS that inhibit north-south movement in Africa are the great desert zone along the Tropic of Cancer and the belt of equatorial forest that extends eastward almost to the shores of Lake Albert and Lake Tanganyika. Between the forest and the desert is a zone of savanna: the Sudanic belt. At its southern edge the equatorial forest gives way to a similar zone of savanna. The two grasslands meet in the vicinity of Lake Victoria. It was from this general region that the Early Iron Age culture of Africa began its subequatorial spread about 2,000 years ago. Political subdivisions appear on the map only where the author makes reference to them.**

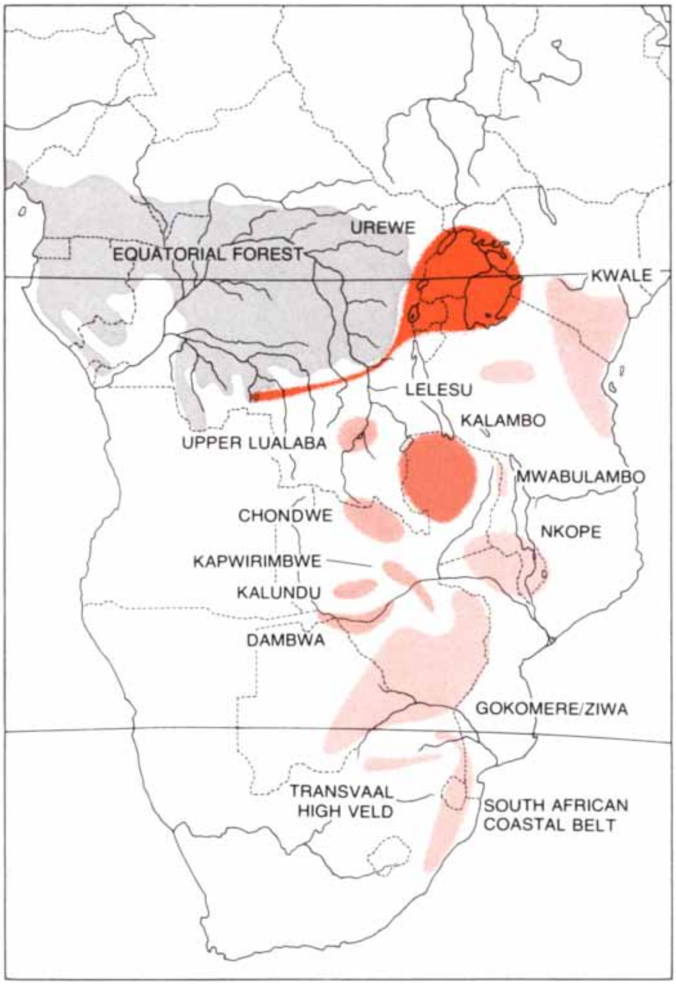
practice the traditional occupations of hunting and gathering for centuries after the arrival of the Early Iron Age settlers.

Archaeologists now recognize more than a dozen Early Iron Age regional industries, differentiated mainly on the basis of pottery styles. At least in the eastern part of subequatorial Africa these are remarkably distinct from the pottery styles that became popular in later times. Up to now the most detailed comparison of the wares of various Early Iron Age groups is one made by Robert Soper of the University of Ibadan in Nigeria. On the basis of his work and my own it is possible to discern two major ceramic subdivisions. These I have provisionally termed the eastern stream and the western stream.

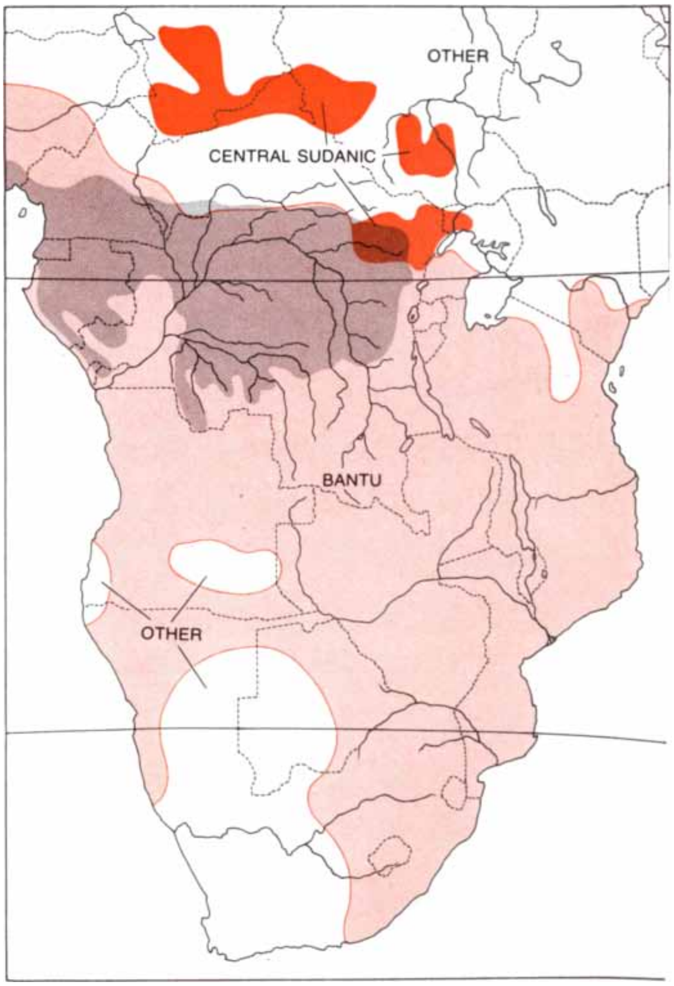
Sites attributable to the eastern stream are found mainly in the coastal hinterland of East Africa, in Malawi, in southern and eastern Zambia, in most of Rhodesia and in the Transvaal, Natal and Swaziland. The western stream is represented in territories that are archaeologically less well known, but the stream is apparent in central and western Zambia and in adjacent parts of Zaïre and Angola. Although the two streams were first distinguished on the basis of pottery style, it is also evident that they are geographically distinct. Furthermore, there are significant differences between the two in terms of relative chronology and economic practices.

The earliest-known manifestation of Early Iron Age culture is found in the Great Lakes region of East Africa and is recognized by a characteristic pottery known as Urewe ware. The makers of Urewe ware were probably settled in the country around the western and southwestern shores of Lake Victoria by 500 B.C. or soon thereafter. Later they spread to the eastern shore of the lake to settle in what is now southwestern Kenya. The Urewe people were certainly workers in iron; smelting furnaces of this period have been found in both Rwanda and eastern Zaïre. So far there is only indirect evidence that the Urewe people practiced agriculture. Analysis of the plant pollens in bottom sediments from Lake Victoria off southern Uganda indicates a substantial reduction in the forest cover in about the first half of the first millennium B.C. One can view this as evidence of some form of land clearing, possibly for agriculture. Those responsible for the clearing could have been Early Iron Age pioneers.

Far to the southwest of the Great Lakes, near the modern border between Angola and Zaïre on the southern fringe of the tropical forest, typical Urewe pot-




**EARLY IRON AGE SITES** in subequatorial Africa fall into three general classes based on pottery characteristics. The first and earliest is known as Urewe; most Urewe sites are clustered around Lake Victoria, but one western outlier, Tshikapa, is near the border between Angola and Zaïre. Sites in the class known as the eastern stream include five groups north of the Zambezi River and four to the south. Sites in the class known as the western stream include four groups; most are north of the Zambezi. Sites of the Kalambo group, near Lake Nyasa, show features of both streams.




**BANTU LANGUAGES** spoken by more than 130 million black Africans are distributed across Africa in a pattern strikingly similar to the distribution of Early Iron Age archaeological sites (*see map at left*). Speakers of Bantu languages are also found in the equatorial forest (*gray*) as far north as where forest gives way to savanna. There, near Lake Edward and Lake Albert, Bantu-speakers are in contact with speakers of Central Sudanic languages (*dark color*). Blanks indicate other non-Bantu areas.

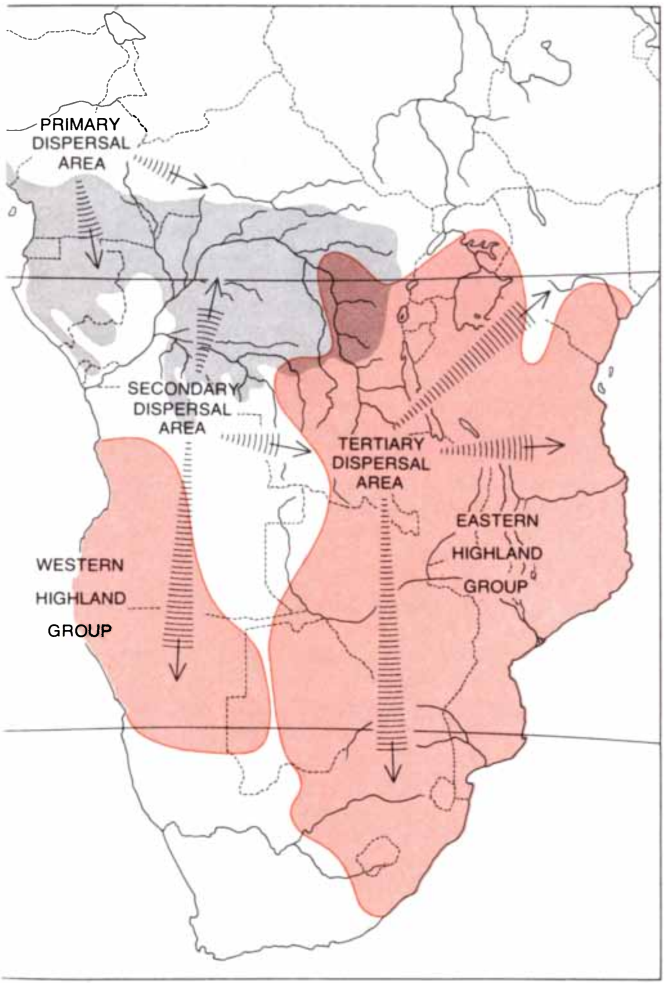
tery has been found at the site of Tshikapa. Jacques Nenquin of the State University of Ghent has been able to demonstrate that the vessels were made of local clay, rather than having been imported from the east. The significance of this extension of Urewe potmakers toward the west will become evident in what follows.

The two-stream phenomenon is most clearly recognizable in the spread of Early Iron Age culture into the south-central and southern parts of subequatorial Africa, a series of movements that began only after the initial settlement of the Great Lakes region. The eastern stream is clearly derived from the Great Lakes Urewe settlements. The western stream may either be derived from the same settlements or share a common ancestry with them. In any event the eastern stream can be shown to have reached the coastal hinterland of southern Kenya, adjacent parts of Tanzania and perhaps Somalia in about the second century of the Christian Era. In the fourth century there was a rapid movement of Early Iron Age culture southward through Malawi, eastern Zambia and Rhodesia and on into the Transvaal and Swaziland. This was a further thrust of the eastern stream.

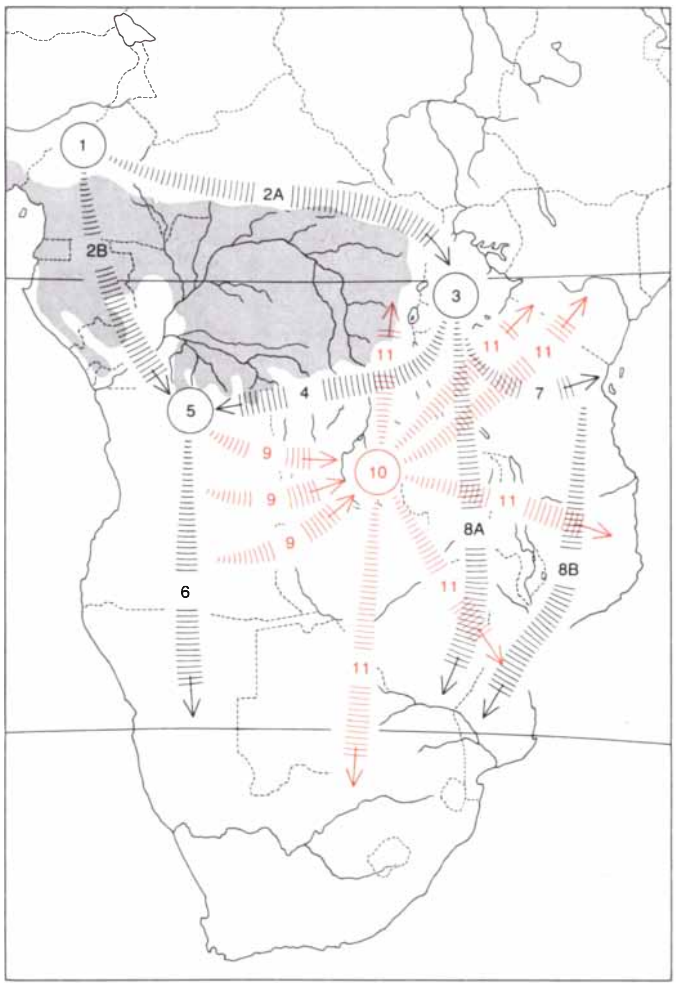
The process, which seems to have taken place with amazing speed, may have affected Rhodesia slightly later than it did Malawi and eastern Zambia. In any event a date of about A.D. 400 is indicated for the start of the Early Iron Age south of the Limpopo River, in what is now the Republic of South Africa. Not enough carbon-14 dates are yet available to establish the timing of the event with any greater precision. In the Victoria Falls region of southern Zambia, where the settling of Early Iron Age people can be shown to have stemmed from Rhodesia, the imported culture did not reach full florescence until the sixth century.

The Early Iron Age western stream is well represented in the archaeological record only in central Zambia, west of the Luangwa River, and in an adjacent part of Rhodesia, across the Zambezi River. Related sites are also known in the province of Shaba (formerly Katanga) in Zaïre, and in a few places farther to the west, all the way to the Atlantic coast. It is known that in the coastal area pottery (not necessarily representative of the Early Iron Age complex) was in use as early as 200 B.C.

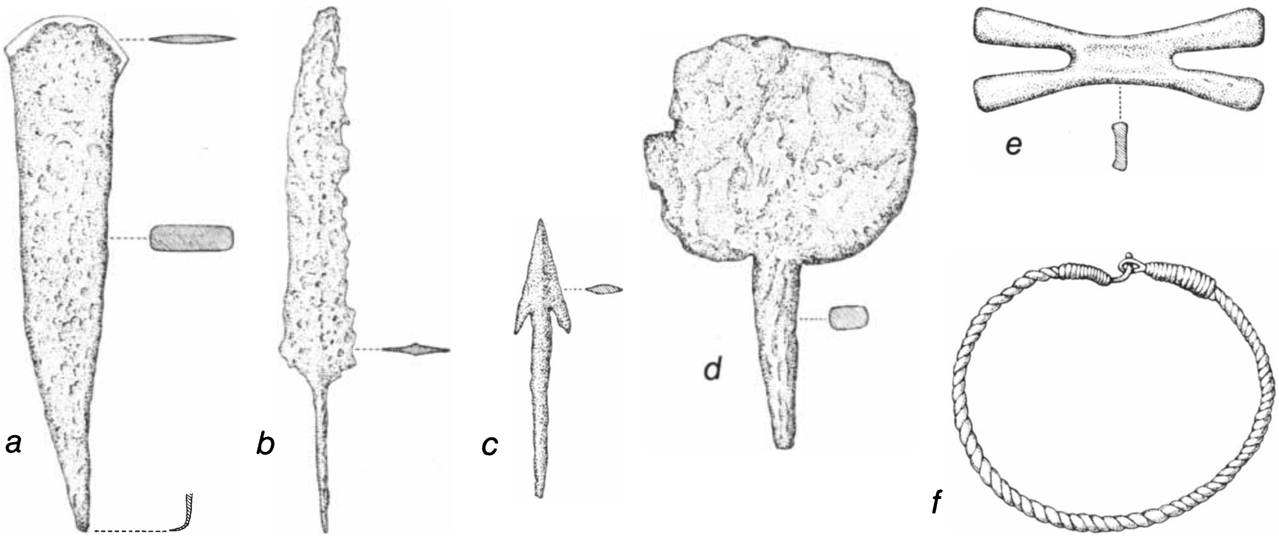
In the 11th century A.D. there was a rapid eclipse of Early Iron Age culture over almost all the eastern half of subequatorial Africa. The traditions affected included some that were typical of the western-stream settlements as well as the eastern-stream ones. The successor




**DISPERSAL OF BANTU LANGUAGES** is shown schematically; the data are drawn principally from the research of Bernd Heine of the universities of Cologne and Nairobi and of David Dalby of the University of London. The homeland is near the forest margin in Cameroon; this is where modern Bantu dialects show the greatest diversity. Two further dispersals followed; the Western Highland Group of Bantu languages became established significantly earlier than the Eastern Group.




**HYPOTHETICAL PROGRESSION** of Bantu-speakers over some 2,000 years sees the language arising among a Neolithic people before 1000 B.C. (*1*). A dual movement then seems to have brought Bantu-speakers, some using iron, beyond the forest (*3*, *5*). An east-to-west movement (*4*) reinforced southward expansion of languages ancestral to the Western Group (*6*). Early in the first millennium A.D. the eastern stream expanded to the south (*7*, *8*). The western stream (*9*) gave rise to a center (*10*) that sent forth languages ancestral to the Eastern Group (*11*) in the 11th century A.D.




**EARLY IRON AGE METALWORK found at sites in subequatorial Africa includes these characteristic artifacts. The iron axe (*a*) has a curved tang, and the iron spearhead (*b*) has a strengthening ridge on both faces. The barbed iron arrowhead (*c*) is tapered in cross section to provide cutting edges, and the iron hoe blade (*d*) has a sturdy tang. The Early Iron Age smiths were familiar with copper; the ingot of copper (*e*) may have been used as a form of currency, but the necklet of twisted copper wire (*f*) was probably only a personal ornament.**

cultures, which are known collectively as later Iron Age, show considerable diversity. Nevertheless, many factors—cultural, technological and, as we shall see, linguistic—point to connections between the later Iron Age cultures and the more westerly settlements of the Early Iron Age people, particularly those in southeastern Zaïre. A significant degree of continuity between the Early Iron Age and the later Iron Age is also evident in several eastern regions, but it is by no means as marked there as it is farther to the west.

Evidence in favor of the hypothesis that the spread of Early Iron Age culture into subequatorial Africa can be correlated with the dispersal of populations speaking Bantu languages is largely circumstantial. For example, arguing backward from the present and the recent past, proponents of the hypothesis point out that the parts of Africa occupied by Bantu-speaking peoples are broadly those where evidence of Early Iron Age settlement is found. Supporting circumstantial evidence is also provided by strong indications that most of the non-Bantu-speaking inhabitants of these same areas had adopted an Iron Age culture only in part or not at all at the time of the first foreign contacts, and that some non-Bantu-speakers remain only partly acculturated to this day.

Two other items of evidence, although they too are circumstantial, are persuasive. One is the close linguistic similarity between the Bantu dialects that are now spread across an enormous range of subequatorial Africa; the similarity suggests that all the variants were probably derived from a common ancestral tongue within relatively recent times. (As we have seen, a similarly brief history and common ancestry is attributed to Early Iron Age culture on the basis of archaeological evidence.) The other item of circumstantial evidence comes from linguistic reconstructions of proto-Bantu languages. Among the reconstructed words are a number that describe certain key elements of Early Iron Age culture.

How are such reconstructions accomplished? Where and by whom was proto-Bantu spoken? With respect to the second question most students of the problem now support, at least in its basic outlines, the view of Joseph H. Greenberg of Stanford University. Greenberg places the Bantu linguistic homeland close to the northwestern limit of the present Bantu range. Not only is this the area where the greatest diversity exists among modern Bantu dialects but also it is where a strong degree of similarity is found, both in vocabulary and in grammatical form, between the Bantu dialects and the local non-Bantu languages.

With respect to how a reconstruction is accomplished, the one that is presented below is based on the work of Bernd Heine of the universities of Cologne and Nairobi and that of David Dalby of the University of London. The complex processes involved can be summarized in a map [*see illustration at left on preceding page*]. An initial center of proto-Bantu appears to have been in the far northwest of the modern Bantu area, most probably in what is now central Cameroon. From this center dialects ancestral to those Bantu tongues found today in parts of northeastern Zaïre dispersed eastward along the northern margin of the equatorial forest.

Perhaps at roughly the same time a more direct southward expansion from the Cameroon center brought Bantu speech along a coastal or river route to the southwestern margin of the equatorial forest near the mouth of the Congo River. From this second center Bantu dialects appear to have been introduced southward into Angola and South West Africa (Namibia). Subsequent developments apparently gave rise to a third Bantu speech center, in the area of the upper Lualaba and Kasai rivers in southeastern Zaïre. From this center Heine and Dalby derive the modern Bantu dialects of the eastern half of subequatorial Africa, all of which show strong similarities.

The study of words that have entered Bantu languages from non-Bantu sources can be used to illustrate the sequence of Bantu linguistic evolution. The most intensive current research along these lines is that being done by Christopher Ehret of the University of California at Los Angeles. His work is of particular interest for what it suggests about the transmission of the cultural traits the loan words describe. For example, Ehret suggests that the names applied to cattle and sheep by many modern Bantu-speakers were probably derived from the non-Bantu languages known collectively as Central Sudanic. These tongues are still spoken in those areas of what are now the southern Sudan, Uganda, Zaïre and the Central African Republic that lie to the northwest of Lake Albert.

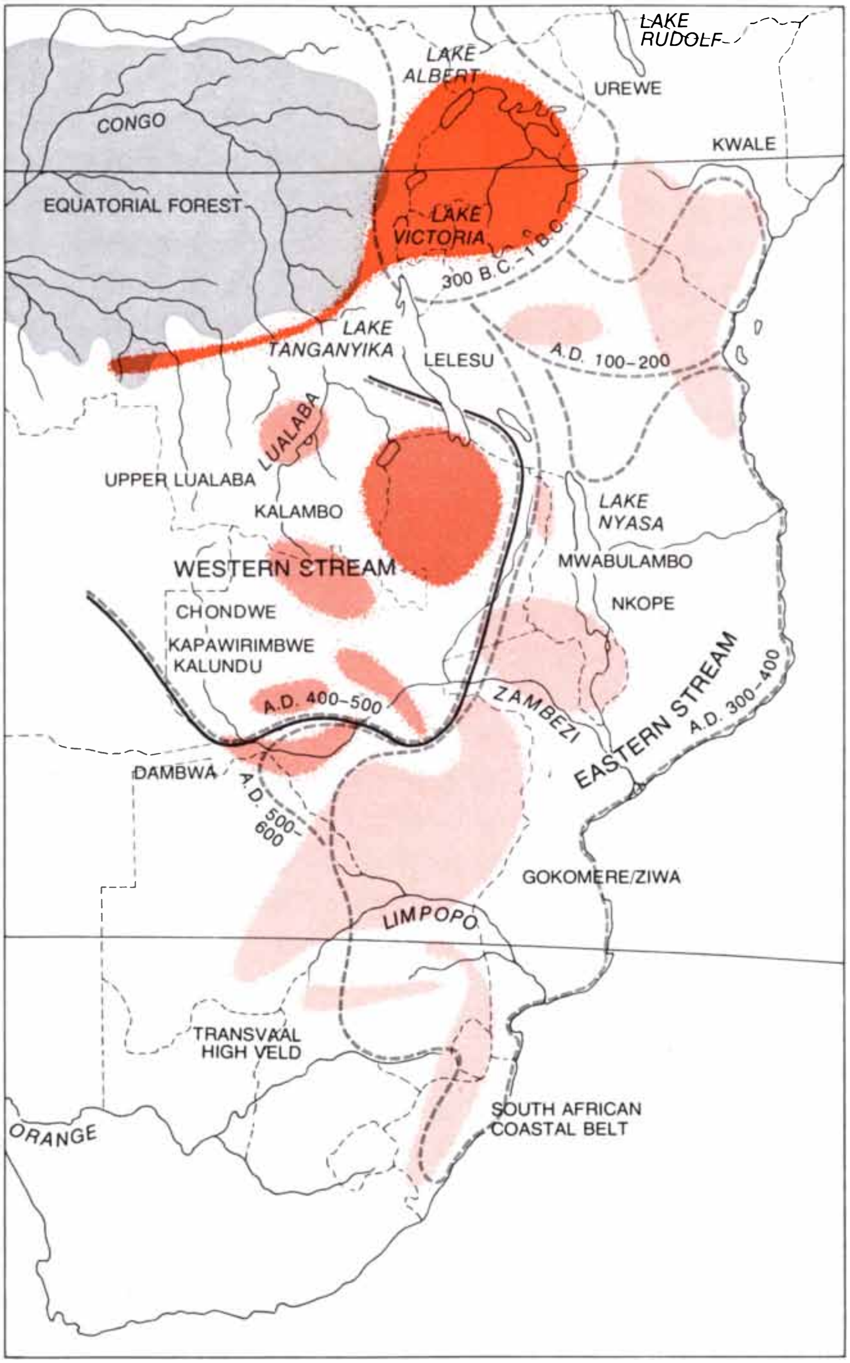
To return from linguistics to archaeology, by the second millennium B.C. or even earlier a food-producing economy had become established throughout most of the Sudanic belt, the broad region that stretches across Africa between the southern fringes of the Sahara and the northern limits of the equatorial forest. Most of the cereals cultivated by Early Iron Age farmers farther to the south were species known and perhaps originally domesticated in the Sudanic belt. The animals that accompanied the Early Iron Age farmers—goats, sheep and cattle—although not originally domesticated in the region, were in all probability herded there 2,000 years before the Christian Era.

The evidence suggesting a Sudanic origin for the Early Iron Age culture is not confined to the agricultural economy. The two best-known early centers of ironworking in sub-Saharan Africa are Nok in Nigeria and Meroë in Nubia. Both are adjacent to the Sudanic belt, and the knowledge of iron metallurgy doubtless diffused rapidly through the region, although the direction of the diffusion remains to be ascertained. Still another item of evidence is found in the Sudanic pottery of the period. Although little is known about these wares, they seem to reflect the same tradition that is evident in Early Iron Age pottery. The relation is most apparent at the time, presumably in about the middle of the first millennium B.C., when knowledge of iron was spreading among the stone-tool-using farmers of the belt. Archaeologically the central part of the belt remains almost unknown, but evidence is accumulating to suggest that the populations ultimately giving rise to the Early Iron Age complex were somewhere within this general region.

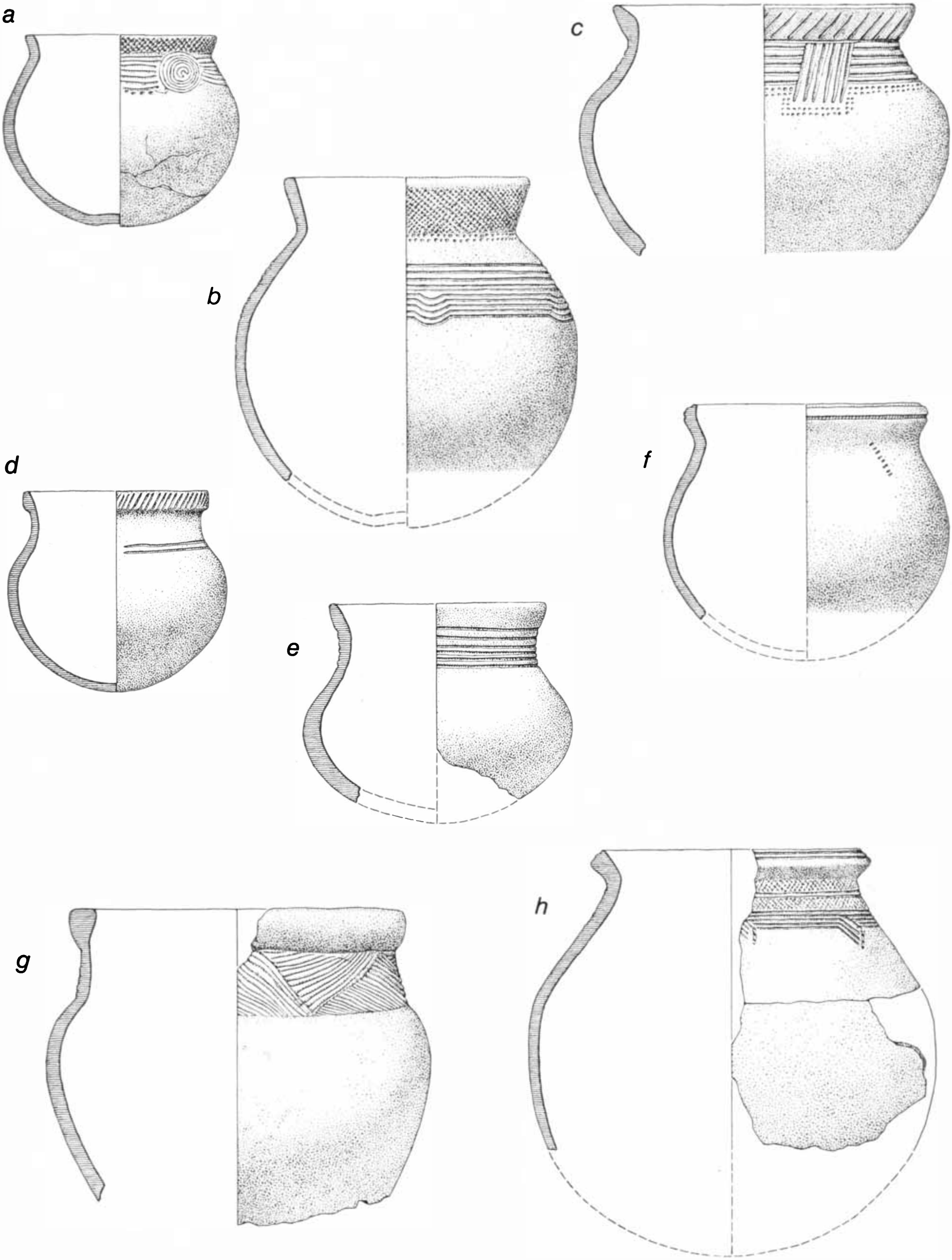
Our circumstantial linguistic hypothesis linking Bantu speech with Early Iron Age culture must surmount one major obstacle. The best prehistoric evidence related to Early Iron Age culture comes almost exclusively from the territory south of the great equatorial forest, that is, the side exactly opposite the one where the culture's formative processes seem to have centered. How did the culture cross the barrier? The great forest is itself inimical to farming practices of any kind; that a farming people could make even a rapid journey through hundreds of kilometers of dense equatorial growth is inherently improbable. By what other route could the carriers of the culture have reached the subequatorial savanna?

The alternate route requires an initial movement in an easterly direction, followed by a turn to the south. Along the first part of this route we find an area east of Lake Chad where the pottery that was made in the first millennium B.C. strongly resembles Early Iron Age wares. The area is adjacent to the territory where Central Sudanic languages are spoken today. The linguistic evidence indicates that some early Bantu-speaking people acquired domestic cattle and sheep, together with the words for both animals, from speakers of Central Sudanic. Did they acquire them during the eastward stage of such a detour around the equatorial forest? There is further evidence favoring this interpretation: around Lake Victoria, where the southward turn must have been made, we find the sites that have yielded Urewe ware.

It is appropriate at this point to emphasize the significance of the Urewe sites in the Great Lakes region. First, these Early Iron Age sites are geographically closest to the region northwest of Lake Albert that is proposed here as the homeland of the culture. Second, the carbon-14 evidence shows that Urewe pottery is the oldest of all Early Iron Age ceramics. Third, it is in the Urewe




**PROGRESS OF EASTERN STREAM, from initial Urewe settlements around Lake Victoria to the final Early Iron Age coastal thrust south of the Limpopo River, is divisible into two successive phases. The first is a movement southward and eastward in the second and third centuries A.D. that produced the Lelesu and Kwale groups of sites. The second and far more extensive movement was southward in the fourth and fifth centuries and produced the Mwabulambo, Nkope, Gokomere/Ziwa, Transvaal high-veld and South African coastal-belt groups. Starting in the fifth century and continuing in the sixth, an eastern movement by western-stream groups produced the Upper Lualaba, Chondwe, Kapwirimbwe and Kalundu groups, and an eastern-stream movement in the sixth and seventh centuries produced the Dambwa group.**




**EARLY IRON AGE POTTERY** from eight regions of subequatorial Africa shows strikingly similar forms, although details of decoration, lip finish and constriction of neck differ. The first example (*a*) is a pot from the western Urewe site, Tshikapa. Beside it (*b*) is a pot from one of the Lake Victoria Urewe sites. The four eastern-stream pots represent sites in the Nkope group (*c*), the Dambwa group (*d*), the Transvaal high-veld group (*e*) and the South African coastal-belt group (*f*). One western-stream example is shown (*g*); it is from a site in the Kapwirimbwe group. The largest of the eight (*h*) is from a site in the Kalambo group, whose stream relations remain uncertain.

wares that one finds preserved the greatest number of features that may be called ancestral.

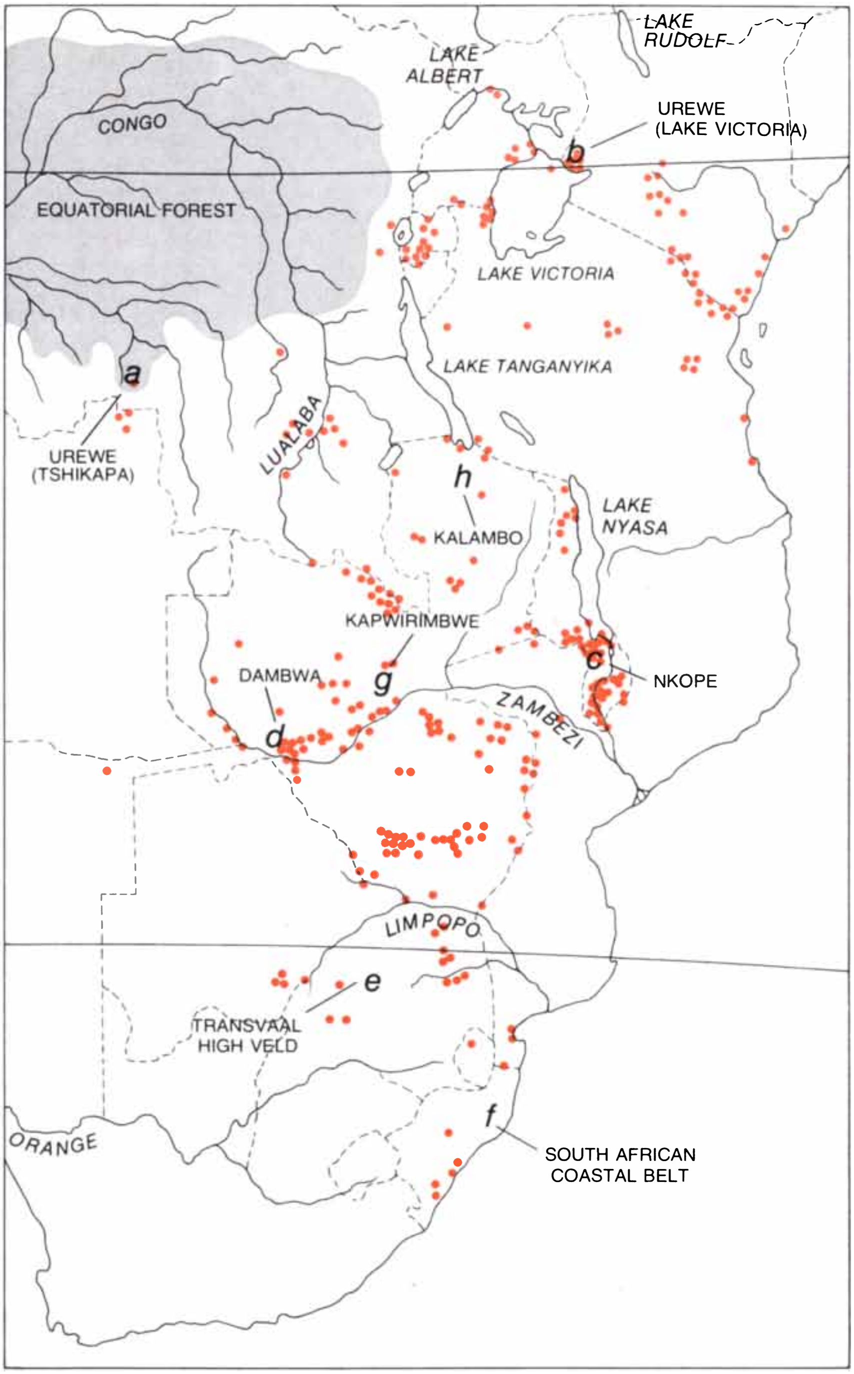
Thus there is a striking similarity between the archaeological evidence that gives the Urewe sites a position early in the prehistory of the Early Iron Age culture and the linguistic evidence cited by Heine in support of an early eastward spread of the Bantu languages. I think it is a plausible further step to suggest that the language of the Urewe potters and their descendants was a Bantu dialect derived from those that spread along the northern margin of the equatorial forest to the vicinity of Lake Albert at a relatively early date. The direct modern descendants of these early dialects are Nyali and Mbuti, spoken today in a part of extreme northeastern Zaïre, where pottery is still made in an Early Iron Age style and where neighboring peoples still speak Central Sudanic languages. Both archaeology and linguistics therefore point to the area northwest of Lake Albert as the place where Early Iron Age culture underwent its formative processes. It is unfortunate that the archaeology of the region remains virtually unknown.

Some Bantu-speaking people appear to have reached the savanna south of the equatorial forest by a more westerly route leading directly south from the Cameroon homeland along or close to the Atlantic coast. It is therefore open to question whether the Early Iron Age Bantu dialect hypothesized above was spoken only by those who eventually made up the eastern stream or by all the Early Iron Age migrants. Specifically, could the people of the western stream have acquired a Bantu dialect independently? Such an independent tongue could have been derived from the southward spread of Bantu speech by a westerly route. That such was the case seems highly probable, at least in some areas. If my dating of the easterly spread of Bantu across the northern fringe of the forest barrier is even approximately correct, then at about the same time when an Early Iron Age presence is demonstrable in the southwestern savanna the carriers of the culture must have spoken western Bantu dialects. Loan words support this contention.

Loan words from Central Sudanic languages have been detected in the modern Bantu dialects spoken in areas initially colonized by migrants of the Early Iron Age western stream. Indeed, both domestic cattle and the Central Sudanic loan words related to them can be shown to have entered much of southern Africa with the western stream. Similarly, the ceramic tradition of the western stream shows strong relations with Urewe pottery. As an example, consider the Tshikapa site. The finds here may well represent an early spread of the Urewe pottery tradition (perhaps accompanied by a full Early Iron Age smithing-and-farming economy) to the territory in northwestern Angola where the western stream subsequently arose. There contact could have been established with those speakers of Bantu dialects who had spread southward from the Cameroon homeland.

Archaeological evidence of this early southward spread in the extreme west is found in two areas. First, pottery that resembles Early Iron Age wares is found on the island of Fernando Po in association with a Neolithic industry characterized by finely polished stone tools that may have been hafted for use as either axes or hoes. Second, similar polished stone tools have been found in the region around the mouth of the Congo River, in association with pottery that is less ornamented than most Early Iron Age wares; the Congo sites are dated at about 200 B.C.

We do not know whether the Neolithic people of Fernando Po and the lower Congo had any domestic animals. The




**EARLY IRON AGE SITES, now numbering more than 350, are shown in relation to the modern political boundaries of subequatorial Africa. The eight groups of sites that yielded the examples of Early Iron Age pottery illustrated on the opposite page are named and located.**

difficulties of moving herds through the equatorial forest suggest that the Congo people, at least, were not herdsmen. A possible exception, however, is the herding of goats. Whereas the Bantu words for cattle and sheep are probably Central Sudanic in origin, the word for goat is common to almost all Bantu languages, which suggests that goats were already a part of the agricultural economy in the Bantu homeland in Cameroon. In any event these coastal users of Neolithic tools may have practiced the variety of tropical and subtropical agriculture that concentrates on root crops such as the yam rather than on cereals. Whether or not they were Bantu-speakers pushing south from Cameroon, they were certainly a people having only a premetallurgical technology. Whether and to what extent they were farmers and herders has not yet been conclusively demonstrated.

These premetallurgical people reached the southern fringe of the equatorial forest belt along the lower reaches of the Congo. They did not, however, penetrate the opener savanna of northern Angola. Indeed, the Angolan grasslands may already have been occupied by full-fledged Early Iron Age herdsmen and grain growers who had bypassed the forest barrier by following the Great Lakes route, bringing with them a Urewe-related pottery tradition such as the one found at Tshikapa. One may envisage a meeting and fusion of the two Bantu-speaking populations: the Great Lakes group now practicing metallurgy, herding flocks and raising cereal crops, and the coast travelers still using stone tools and planting tubers. Perhaps it was this fusion that gave rise, at about the beginning of the Christian Era, to the Bantu culture of the Kongo area in northern Angola, a culture that soon afterward spread to the more southerly highlands of Angola and South West Africa.

A final expansion of this western stream of Early Iron Age culture led migrants eastward from the highlands into west-central Zambia and southeastern Zaïre, where their presence is attested in the archaeological record beginning in about A.D. 500. It was at this point that the people of the western stream made contact with those of the eastern stream along a common boundary in central Zambia and northern Rhodesia.

In describing the progression of the western stream I have skipped the time when the people of the eastern stream were expanding rapidly southward from their initial bases in southern Kenya and northern Tanzania. Let us turn back and consider the routes the eastern stream followed in reaching the regions farther to the south. Here the archaeological evidence is more comprehensive, and one can map the general trends, notably the thrust through Malawi and eastern Zambia to Rhodesia and the Transvaal between A.D. 300 and 400, followed about 200 years later by a further spread into southwestern Zambia.

Carbon-14 dates indicate that the eastern stream was well established in Zambia and Rhodesia for a significant period before the eastward movement of the western stream established contact between the two. In its early stages the eastern stream appears to have lacked certain elements of the full-fledged Early Iron Age culture. Among these were possession of cattle and of some sophisticated metallurgical techniques at a time when both were in the hands of the people of the western stream. The greater western-stream cultural inventory may be attributable, at least in part, to an innovative period in the Kongo area following the fusion of peoples that gave rise to the western stream.

The form of Bantu spoken by the people of the eastern stream was evidently a dialect derived from the early eastward spread of proto-Bantu from Cameroon to the Great Lakes region, enriched along the way by a number of Central Sudanic loan words relating particularly to agricultural and metallurgical technology. Dalby, Heine and others, however, have demonstrated conclusively that virtually all the modern Bantu dialects spoken in the eastern half of the Bantu linguistic area are in fact western in derivation. The distribution of these dialects, which Heine has named the Eastern Highland Group, does not coincide with the distribution of the archaeological remains identifiable with the Early Iron Age eastern stream. Instead it coincides to a remarkable degree with the archaeological distribution of later Iron Age industries.

As we have seen, these industries first appear in the archaeological record of eastern Africa during the 11th century A.D. and show connections with the region farther to the west. Indeed, the area Heine postulates as being the source of the Eastern Highland Group of dialects is almost exactly identical with the one I consider to be the homeland of many of the cultural innovations that distinguish the later Iron Age of subequatorial Africa from its predecessors. The modern distribution of the Eastern Highland languages is thus due to a second spread of Bantu speech overlying that of the Early Iron Age.

We can therefore see what a marked degree of similarity there is between the archaeological sequence of the Iron Age in subequatorial Africa and the linguistic evidence for the spread and development of the Bantu languages and their speakers. In this discussion it has been necessary to view the Iron Age Bantu-speakers alone, with little reference to the other peoples with whom they interacted. One should not, however, forget that the processes of population movement and settlement I have been describing unfolded in territory that was already occupied by a diverse population of earlier peoples, whose own distinctive life-styles were well adapted to the environment. It is not possible to take up here the substantial contributions—genetic, cultural, economic and technological—made by these earlier groups to the Iron Age peoples who eventually displaced them and who today are the major element in the population of subequatorial Africa.

**DAMBWA POT, uncovered during the excavation of an Early Iron Age house at the type site, Dambwa, near Livingstone in southern Zambia, was partially buried by mud that had fallen from the house wall. Scale in foreground measures inches; beyond it at each end are the charred remains of two fallen wood poles that had supported the wall. Floor is made of puddled earth.**

# Unique full-function 8-digit wrist calculator... available <u>only</u> as a kit.

A wrist calculator is the ultimate in common-sense portable calculating power. Even a pocket calculator goes where your pocket goes—take your jacket off, and you're lost!

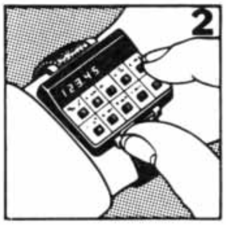
But a wrist-calculator is only worth having if it offers a genuinely comprehensive range of functions, with a full-size 8-digit display.

This one does. What's more, because it *is* a kit, supplied *direct* from the manufacturer, it costs only a very reasonable $19.95. And for that, you get not only a high-calibre calculator, but the fascination of building it yourself.

**How to make 10 keys dothe work of 27**

The Sinclair Instrument wrist calculator offers the full range of arithmetic functions. It uses normal algebraic logic ('enter it as you write it'). But in addition, it offers a % key; plus the convenience functions $\sqrt{x}$, 1/x, $x^2$; plus a full *5-function* memory.

All this, from just 10 keys! The secret? An ingenious, simple three-position switch. It works like this.



**1.** The switch in its normal, central position. With the switch centered, numbers—which make up the vast majority of key-strokes—are tapped in the normal way.

**2.** Hold the switch to the left to use the functions to the left above the keys...

**3.** and hold it to the right to use the functions to the right above the keys.

The display uses 8 full-size red LED digits, and the calculator runs on readily-available hearing-aid batteries to give weeks of normal use.

**Assembling the Sinclair Instrument wrist calculator**

The wrist calculator kit comes to you complete and ready for assembly. All you need is a reasonable degree of skill with a fine-point soldering iron. It takes about three hours to assemble. If anything goes wrong, Sinclair Instrument will replace any damaged components *free:* we want you to enjoy assembling the kit, and to end up with a valuable and useful calculator.


**Contents**

Case and display window.
Strap.
Printed circuit board.
Switches.
Special direct-drive chip (no interface chip needed).
Display.
Batteries.

Everything is packaged in a neat plastic box, and accompanied by full instructions. The only thing you need is a fine-point soldering iron.

All components are fully guaranteed, and any which are damaged during assembly will be replaced free.

**The wrist-calculator kit is available only direct from Sinclair Instrument. Take advantage of this 10 day money-back offer.**

**Send the coupon today.**

**Sinclair Instruments Inc.**
**6 Commercial Street**
**Hicksville, N.Y. 11801**

To: **Sinclair Instruments Inc.**
**6 Commercial Street**
**Hicksville, NY 11801**

Please send me

_____ Sinclair wrist calculator kit(s) at $19.95$_____

Sales tax (NY residents) _____

Shipping & handling ($2.50 per unit) _____

Enclosed is check/MO (payable to Sinclair Instruments Inc.) TOTAL $_____

Name _____

Address _____

City _____

State _____ Zip _____

I understand that you will refund the purchase price of the kit in full if I return it in saleable conditiion within 10 days of receipt.

SA4

# The Theory of the Rainbow

*When sunlight is scattered by raindrops, why is it that colorful arcs appear in certain regions of the sky? Answering this subtle question has required all the resources of mathematical physics*

by H. Moysés Nussenzveig

The rainbow is a bridge between the two cultures: poets and scientists alike have long been challenged to describe it. The scientific description is often supposed to be a simple problem in geometrical optics, a problem that was solved long ago and that holds interest today only as a historical exercise. This is not so; a satisfactory quantitative theory of the rainbow has been developed only in the past few years. Moreover, that theory involves much more than geometrical optics; it draws on all we know of the nature of light. Allowance must be made for wavelike properties, such as interference, diffraction and polarization, and for particlelike properties, such as the momentum carried by a beam of light.

Some of the most powerful tools of mathematical physics were devised explicitly to deal with the problem of the rainbow and with closely related problems. Indeed, the rainbow has served as a touchstone for testing theories of optics. With the more successful of those theories it is now possible to describe the rainbow mathematically, that is, to predict the distribution of light in the sky. The same methods can also be applied to related phenomena, such as the bright ring of color called the glory, and even to other kinds of rainbows, such as atomic and nuclear ones.

Scientific insight has not always been welcomed without reservations. Goethe wrote that Newton's analysis of the rainbow's colors would "cripple Nature's heart." A similar sentiment was expressed by Charles Lamb and John Keats; at a dinner party in 1817 they proposed a toast: "Newton's health, and confusion to mathematics." Yet the scientists who have contributed to the theory of the rainbow are by no means insensitive to the rainbow's beauty. In the words of Descartes: "The rainbow is such a remarkable marvel of Nature ... that I could hardly choose a more suitable example for the application of my method."

The single bright arc seen after a rain shower or in the spray of a waterfall is the primary rainbow. Certainly its most conspicuous feature is its splash of colors. These vary a good deal in brightness and distinctness, but they always follow the same sequence: violet is innermost, blending gradually with various shades of blue, green, yellow and orange, with red outermost.

Other features of the rainbow are fainter and indeed are not always present. Higher in the sky than the primary bow is the secondary one, in which the colors appear in reverse order, with red innermost and violet outermost. Careful observation reveals that the region between the two bows is considerably darker than the surrounding sky. Even when the secondary bow is not discernible, the primary bow can be seen to have a "lighted side" and a "dark side." The dark region has been given the name Alexander's dark band, after the Greek philosopher Alexander of Aphrodisias, who first described it in about A.D. 200.

Another feature that is only sometimes seen is a series of faint bands, usually pink and green alternately, on the inner side of the primary bow. (Even more rarely they may appear on the outer side of the secondary bow.) These "supernumerary arcs" are usually seen most clearly near the top of the bow. They are anything but conspicuous, but they have had a major influence on the development of theories of the rainbow.

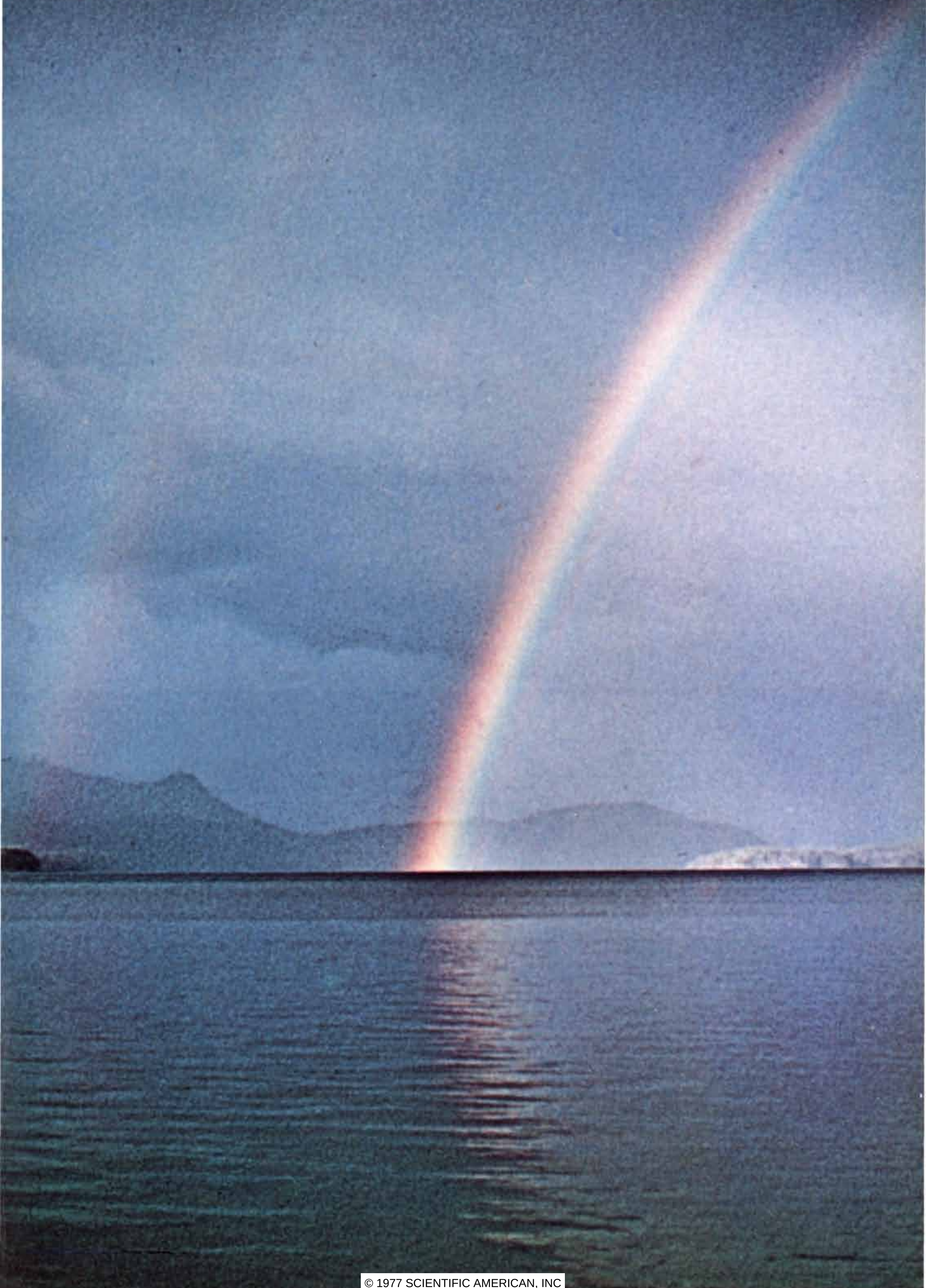
The first attempt to rationally explain the appearance of the rainbow was probably that of Aristotle. He proposed that the rainbow is actually an unusual kind of reflection of sunlight from clouds. The light is reflected at a fixed angle, giving rise to a circular cone of "rainbow rays." Aristotle thus explained correctly the circular shape of the bow and perceived that it is not a material object with a definite location in the sky but rather a set of directions along which light is strongly scattered into the eyes of the observer.

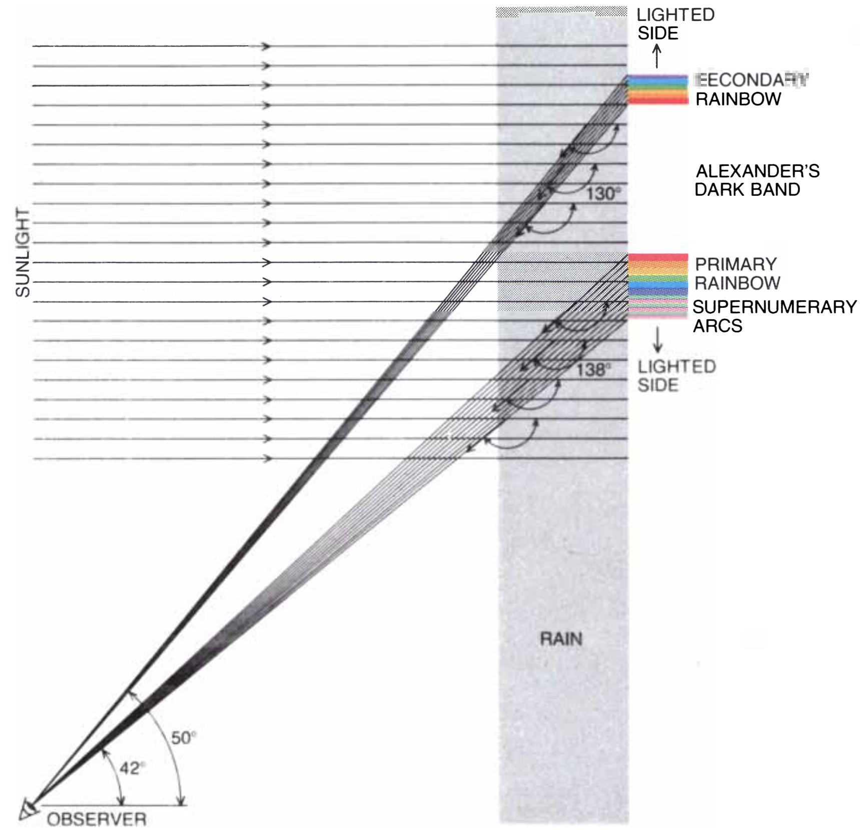
The angle formed by the rainbow rays and the incident sunlight was first measured in 1266 by Roger Bacon. He measured an angle of about 42 degrees; the secondary bow is about eight degrees higher in the sky. Today these angles are customarily measured from the opposite direction, so that we measure the total change in the direction of the sun's rays. The angle of the primary bow is therefore 180 minus 42, or 138, degrees; this is called the rainbow angle. The angle of the secondary bow is 130 degrees.

After Aristotle's conjecture some 17 centuries passed before further significant progress was made in the theory of the rainbow. In 1304 the German monk Theodoric of Freiberg rejected Aristotle's hypothesis that the rainbow results from collective reflection by the raindrops in a cloud. He suggested instead that each drop is individually capable of producing a rainbow. Moreover, he tested this conjecture in experiments with a magnified raindrop: a spherical flask filled with water. He was able to trace the path followed by the light rays that make up the rainbow.

Theodoric's findings remained largely unknown for three centuries, until they were independently rediscovered by Descartes, who employed the same method. Both Theodoric and Descartes showed that the rainbow is made up of rays that enter a droplet and are reflected once from the inner surface. The secondary bow consists of rays that have undergone two internal reflections. With each reflection some light is lost, which is the main reason the secondary bow is fainter than the primary one. Theodoric and Descartes also noted that along each direction within the angular

**DOUBLE RAINBOW was photographed at Johnstone Strait in British Columbia. The bright, inner band is the primary bow; it is separated from the fainter secondary bow by a region, called Alexander's dark band, that is noticeably darker than the surrounding sky. Below the primary bow are a few faint stripes of pink and green; they are supernumerary arcs. The task of theory is to give a quantitative explanation for each of these features.**






**GEOMETRY OF THE RAINBOW is determined by the scattering angle: the total angle through which a ray of sunlight is bent by its passage through a raindrop. Rays are strongly scattered at angles of 138 degrees and 130 degrees, giving rise respectively to the primary and the secondary rainbows. Between those angles very little light is deflected; that is the region of Alexander's dark band. The optimum angles are slightly different for each wavelength of light, with the result that the colors are dispersed; note that the sequence of colors in the secondary bow is the reverse of that in the primary bow. There is no single plane in which the rainbow lies; the rainbow is merely the set of directions along which light is scattered toward the observer.**

range corresponding to the rainbow only one color at a time could be seen in the light scattered by the globe. When the eye was moved to a new position so as to explore other scattering angles, the other spectral colors appeared, one by one. Theodoric and Descartes concluded that each of the colors in the rainbow comes to the eye from a different set of water droplets.

As Theodoric and Descartes realized, all the main features of the rainbow can be understood through a consideration of the light passing through a single droplet. The fundamental principles that determine the nature of the bow are those that govern the interaction of light with transparent media, namely reflection and refraction.

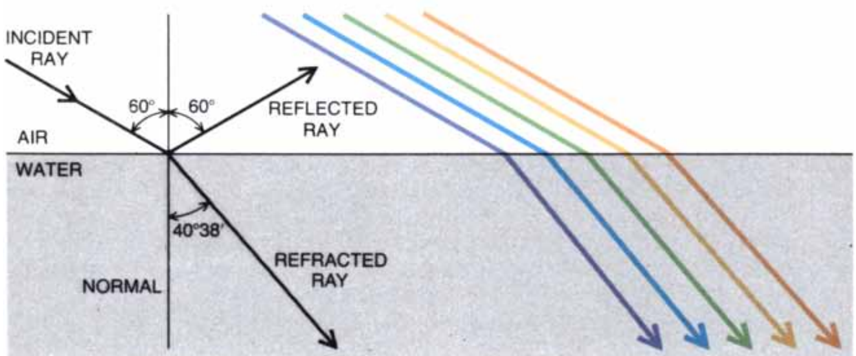
The law of reflection is the familiar and intuitively obvious principle that the angle of reflection must equal the angle of incidence. The law of refraction is somewhat more complicated. Whereas the path of a reflected ray is determined entirely by geometry, refraction also involves the properties of light and the properties of the medium.

The speed of light in a vacuum is invariant; indeed, it is one of the fundamental constants of nature. The speed of light in a material medium, on the other hand, is determined by the properties of the medium. The ratio of the speed of light in a vacuum to the speed in a substance is called the refractive index of that substance. For air the index is only slightly greater than 1; for water it is about 1.33.

A ray of light passing from air into water is retarded at the boundary; if it strikes the surface obliquely, the change in speed results in a change in direction. The sines of the angles of incidence and refraction are always in constant ratio to each other, and the ratio is equal to that between the refractive indexes for the two materials. This equality is called Snell's law, after Willebrord Snell, who formulated it in 1621.

A preliminary analysis of the rainbow can be obtained by applying the laws of reflection and refraction to the path of a ray through a droplet. Because the droplet is assumed to be spherical all directions are equivalent and there is only one significant variable: the displacement of the incident ray from an axis passing through the center of the droplet. That displacement is called the impact parameter. It ranges from zero, when the ray coincides with the central axis, to the radius of the droplet, when the ray is tangential.

At the surface of the droplet the incident ray is partially reflected, and this reflected light we shall identify as the scattered rays of Class 1. The remaining light is transmitted into the droplet (with a change in direction caused by refraction) and at the next surface is again partially transmitted (rays of Class 2) and partially reflected. At the next boundary the reflected ray is again split into reflected and transmitted components, and the process continues indefinitely. Thus the droplet gives rise to a series of scattered rays, usually with rapidly decreasing intensity. Rays of Class 1 represent direct reflection by the droplet and those of Class 2 are directly transmitted through it. Rays of Class 3 are those that escape the droplet after one internal reflection, and they make up the primary rainbow. The Class 4 rays, having undergone two internal re-




**REFLECTION AND REFRACTION of light at boundaries between air and water are the basic events in the creation of a rainbow. In reflection the angle of incidence is equal to the angle of reflection. In refraction the angle of the transmitted ray is determined by the properties of the medium, as characterized by its refractive index. Light entering a medium with a higher index is bent toward the normal. Light of different wavelengths is refracted through slightly different angles; this dependence of the refractive index on color is called dispersion. Theories of the rainbow often deal separately with each monochromatic component of incident light.**

flections, give rise to the secondary bow. Rainbows of higher order are formed by rays making more complicated passages, but they are not ordinarily visible.

For each class of scattered rays the scattering angle varies over a wide range of values as a function of the impact parameter. Since in sunlight the droplet is illuminated at all impact parameters simultaneously, light is scattered in virtually all directions. It is not difficult to find light paths through the droplet that contribute to the rainbow, but there are infinitely many other paths that direct the light elsewhere. Why, then, is the scattered intensity enhanced in the vicinity of the rainbow angle? It is a question Theodoric did not consider; an answer was first provided by Descartes.

By applying the laws of reflection and refraction at each point where a ray strikes an air-water boundary, Descartes painstakingly computed the paths of many rays incident at many impact parameters. The rays of Class 3 are of predominating importance. When the impact parameter is zero, these rays are scattered through an angle of 180 degrees, that is, they are backscattered toward the sun, having passed through the center of the droplet and been reflected from the far wall. As the impact parameter increases and the incident rays are displaced from the center of the droplet, the scattering angle decreases. Descartes found, however, that this trend does not continue as the impact parameter is increased to its maximum value, where the incident ray grazes the droplet at a tangent to its surface. Instead the scattering angle passes through a minimum when the impact parameter is about seven-eighths of the radius of the droplet, and thereafter it increases again. The scattering angle at the minimum is 138 degrees.

For rays of Class 4 the scattering angle is zero when the impact parameter is zero; in other words, the central ray is reflected twice, then continues in its original direction. As the impact parameter increases so does the scattering angle, but again the trend is eventually reversed, this time at 130 degrees. The Class 4 rays have a maximum scattering angle of 130 degrees, and as the impact parameter is further increased they bend back toward the forward scattering direction again.

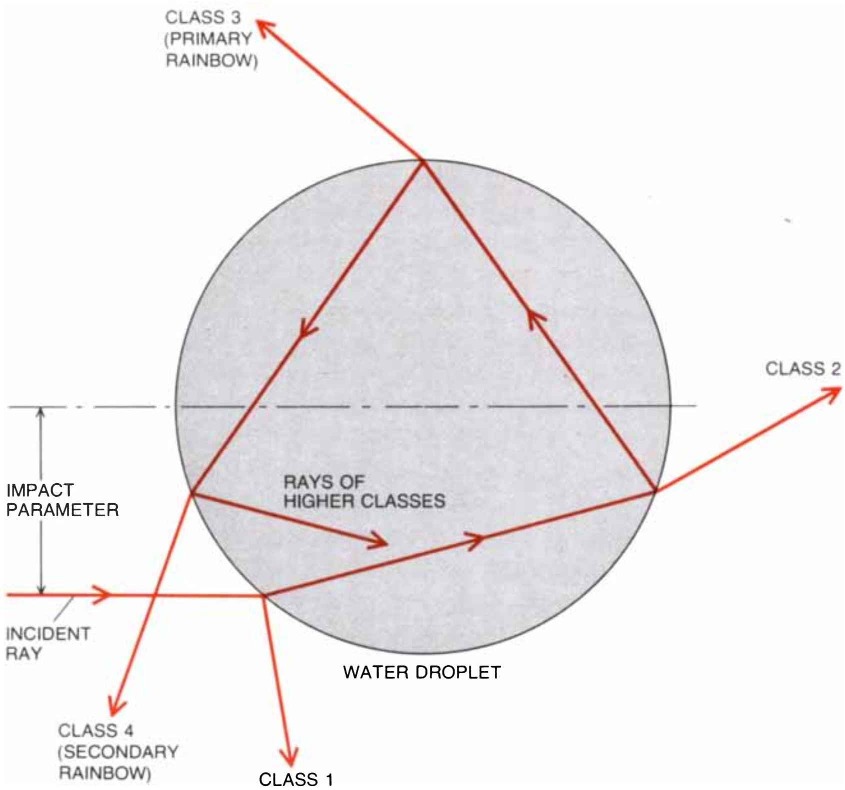
Because a droplet in sunlight is uniformly illuminated the impact parameters of the incident rays are uniformly distributed. The concentration of scattered light is therefore expected to be greatest where the scattering angle varies most slowly with changes in the impact parameter. In other words, the scattered light is brightest where it gathers together the incident rays from the largest range of impact parameters. The regions of minimum variation are those surrounding the maximum and minimum scattering angles, and so the special status of the primary and secondary rainbow angles is explained. Furthermore, since no rays of Class 3 or Class 4 are scattered into the angular region between 130 and 138 degrees, Alexander's dark band is also explained.

Descartes's theory can be seen more clearly by considering an imaginary population of droplets from which light is somehow scattered with uniform intensity in all directions. A sky filled with such droplets would be uniformly bright at all angles. In a sky filled with real water droplets the same total illumination is available, but it is redistributed. Most parts of the sky are dimmer than they would be with uniform scattering, but in the vicinity of the rainbow angle there is a bright arc, tapering off gradually on the lighted side and more sharply on the dark side. The secondary bow is a similar intensity highlight, except that it is narrower and all its features are dimmer. In the Cartesian theory the region between the bows is distinctly darker than the sky elsewhere; if only rays of Class 3 and Class 4 existed, it would be quite black.

The Cartesian rainbow is a remarkably simple phenomenon. Brightness is a function of the rate at which the scattering angle changes. That angle is itself determined by just two factors: the refractive index, which is assumed to be constant, and the impact parameter, which is assumed to be uniformly distributed. One factor that has no influence at all on the rainbow angle is size: the geometry of scattering is the same for small cloud droplets and for the large water-filled globes employed by Theodoric and Descartes.

So far we have ignored one of the most conspicuous features of the rainbow: its colors. They were explained, of course, by Newton, in his prism experiments of 1666. Those experiments demonstrated not only that white light is a mixture of colors but also that the refractive index is different for each color, the effect called dispersion. It follows that each color or wavelength of light must have its own rainbow angle; what we observe in nature is a collection of monochromatic rainbows, each one slightly displaced from the next.

From his measurements of the refractive index Newton calculated that the




**PATH OF LIGHT through a droplet can be determined by applying the laws of geometrical optics. Each time the beam strikes the surface part of the light is reflected and part is refracted. Rays reflected directly from the surface are labeled rays of Class 1; those transmitted directly through the droplet are designated Class 2. The Class 3 rays emerge after one internal reflection; it is these that give rise to the primary rainbow. The secondary bow is made up of Class 4 rays, which have undergone two internal reflections. For rays of each class only one factor determines the value of the scattering angle. That factor is the impact parameter: the displacement of the incident ray from an axis that passes through the center of the droplet.**

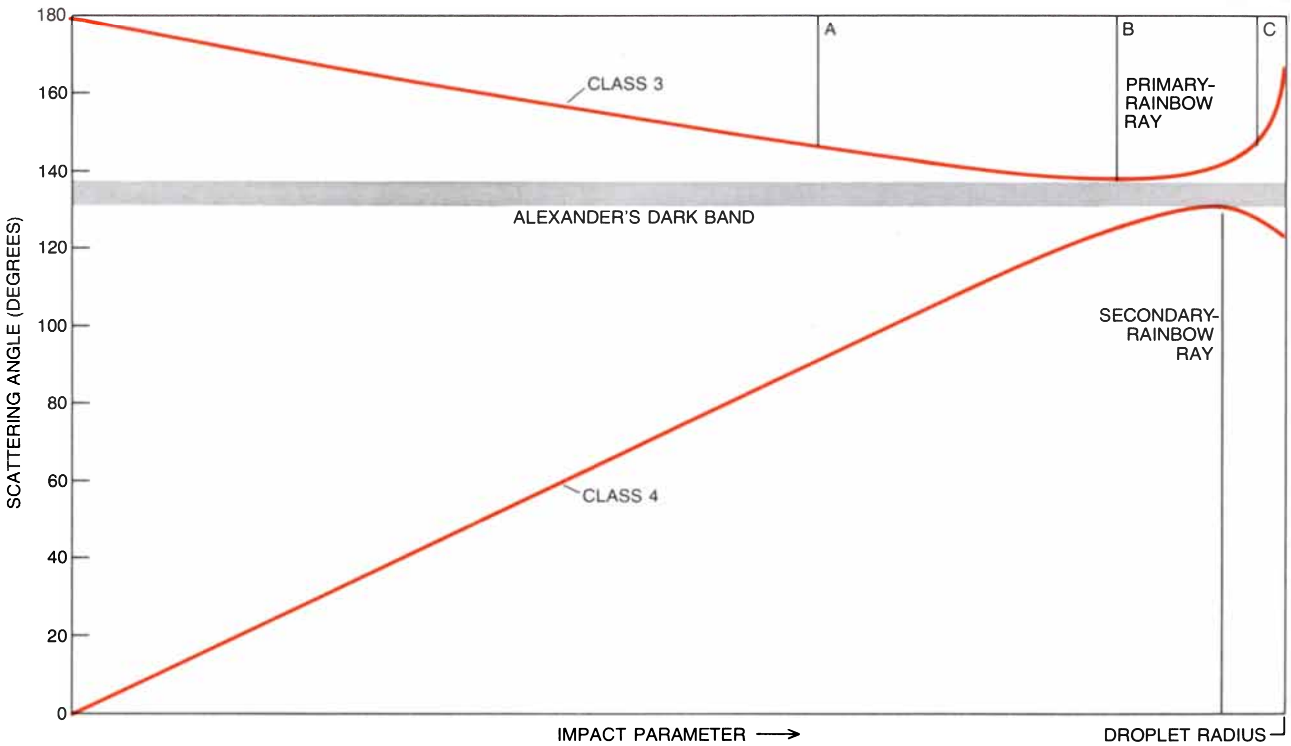
rainbow angle is 137 degrees 58 minutes for red light and 139 degrees 43 minutes for violet light. The difference between these angles is one degree 45 minutes, which would be the width of the rainbow if the rays of incident sunlight were exactly parallel. Allowing half a degree for the apparent diameter of the sun, Newton obtained a total width of two degrees 15 minutes for the primary bow. His own observations were in good agreement with this result.

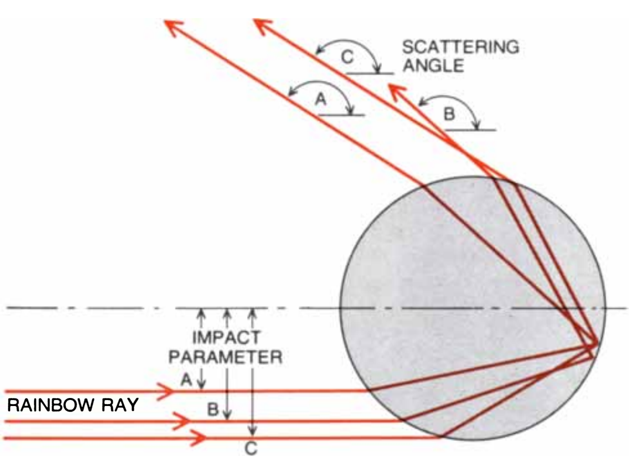
Descartes and Newton between them were able to account for all the more conspicuous features of the rainbow. They explained the existence of primary and secondary bows and of the dark band that separates them. They calculated the angular positions of these features and described the dispersion of the scattered light into a spectrum. All of this was accomplished with only geometrical optics. Their theory nevertheless had a major failing: it could not explain the supernumerary arcs. The understanding of these seemingly minor features requires a more sophisticated view of the nature of light.

The supernumerary arcs appear on the inner, or lighted, side of the primary bow. In this angular region two scattered rays of Class 3 emerge in the same direction; they arise from incident rays that have impact parameters on each side of the rainbow value. Thus at any given angle slightly greater than the rainbow angle the scattered light includes rays that have followed two different paths through the droplet. The rays emerge at different positions on the surface of the droplet, but they proceed in the same direction.

In the time of Descartes and Newton these two contributions to the scattered intensity could be handled only by simple addition. As a result the predicted intensity falls off smoothly with deviation from the rainbow angle, with no trace of supernumerary arcs. Actually the intensities of the two rays cannot be added because they are not independent sources of radiation.

The optical effect underlying the supernumerary arcs was discovered in 1803 by Thomas Young, who showed that light is capable of interference, a phenomenon that was already familiar from the study of water waves. In any medium the superposition of waves can lead either to reinforcement (crest on

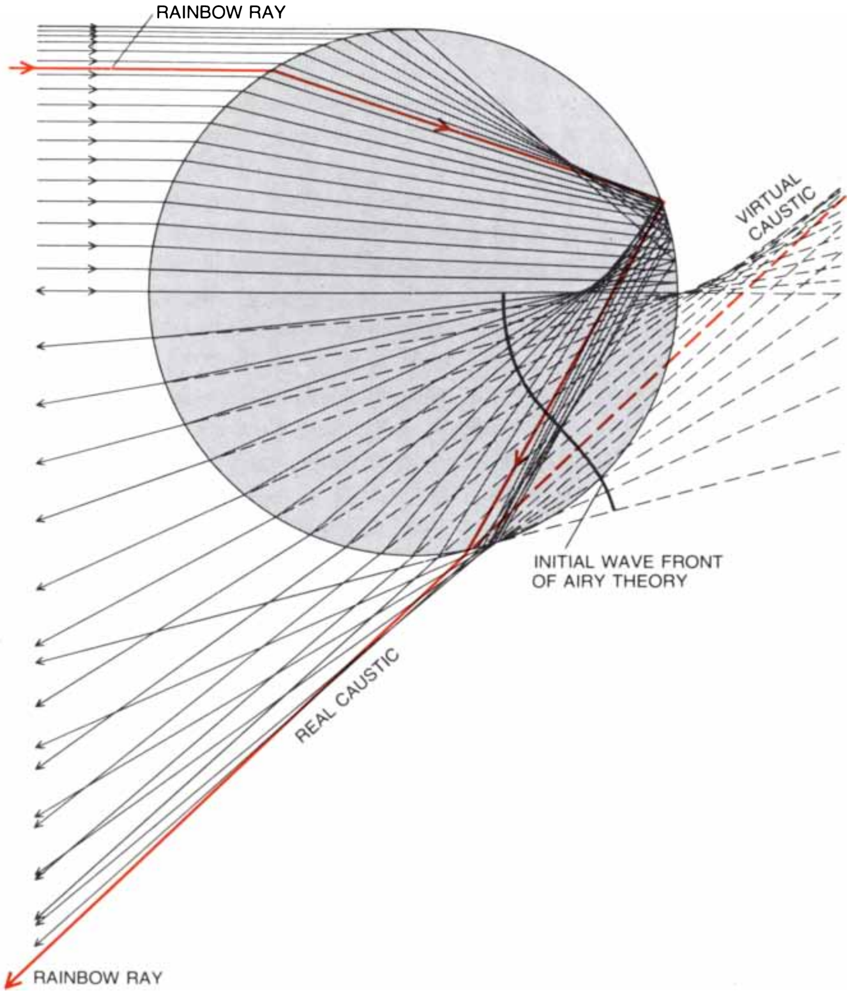



**RAINBOW ANGLE can be seen to have a special significance when the scattering angle is considered as a function of the impact parameter. When the impact parameter is zero, the scattering angle for a ray of Class 3 is 180 degrees; the ray passes through the center of the droplet and is reflected by the far surface straight back at the sun. As the impact parameter increases, the scattering angle decreases, but eventually a minimum angle is reached. This ray of minimum deflection is the rainbow ray in the diagram at the left; rays with impact parameters on each side of it are scattered through larger angles. The minimum deflection is about 138 degrees, and the greatest concentration of scattered rays is to be found in the vicinity of this angle. The resulting enhancement in the intensity of the scattered light is perceived as the primary rainbow. The secondary bow is formed in a similar way, except that the scattering angle for the Class 4 rays of which it is composed increases to a maximum instead of decreasing to a minimum. The maximum lies at about 130 degrees. No rays of Class 3 or Class 4 can reach angles between 130 degrees and 138 degrees, explaining the existence of Alexander's dark band. At the left two Class 3 rays, with impact parameters on each side of the rainbow value, emerge at the same scattering angle. It is interference between rays such as these two that gives rise to the supernumerary arcs.**

crest) or to cancellation (crest on trough). Young demonstrated the interference of light waves by passing a single beam of monochromatic light through two pinholes and observing the alternating bright and dark "fringes" produced. It was Young himself who pointed out the pertinence of his discovery to the supernumerary arcs of the rainbow. The two rays scattered in the same direction by a raindrop are strictly analogous to the light passing through the two pinholes in Young's experiment. At angles very close to the rainbow angle the two paths through the droplet differ only slightly, and so the two rays interfere constructively. As the angle increases, the two rays follow paths of substantially different length. When the difference equals half of the wavelength, the interference is completely destructive; at still greater angles the beams reinforce again. The result is a periodic variation in the intensity of the scattered light, a series of alternately bright and dark bands.

Because the scattering angles at which the interference happens to be constructive are determined by the difference between two path lengths, those angles are affected by the radius of the droplet. The pattern of the supernumerary arcs (in contrast to the rainbow angle) is therefore dependent on droplet size. In larger drops the difference in path length increases much more quickly with impact parameter than it does in small droplets. Hence the larger the droplets are, the narrower the angular separation between the supernumerary arcs is. The arcs can rarely be distinguished if the droplets are larger than about a millimeter in diameter. The overlapping of colors also tends to wash out the arcs. The size dependence of the supernumeraries explains why they are easier to see near the top of the bow: raindrops tend to grow larger as they fall.




**CONFLUENCE OF RAYS scattered by a droplet gives rise to caustics, or "burning curves." A caustic is the envelope of a ray system. Of special interest is the caustic of Class 3 rays, which has two branches, a real branch and a "virtual" one; the latter is formed when the rays are extended backward. When the rainbow ray is produced in both directions, it approaches the branches of this caustic. A theory of the rainbow based on the analysis of such a caustic was devised by George B. Airy. Having chosen an initial wave front—a surface perpendicular at all points to the rays of Class 3—Airy was able to determine the amplitude distribution in subsequent waves. A weakness of the theory is the need to guess the amplitudes of the initial waves.**

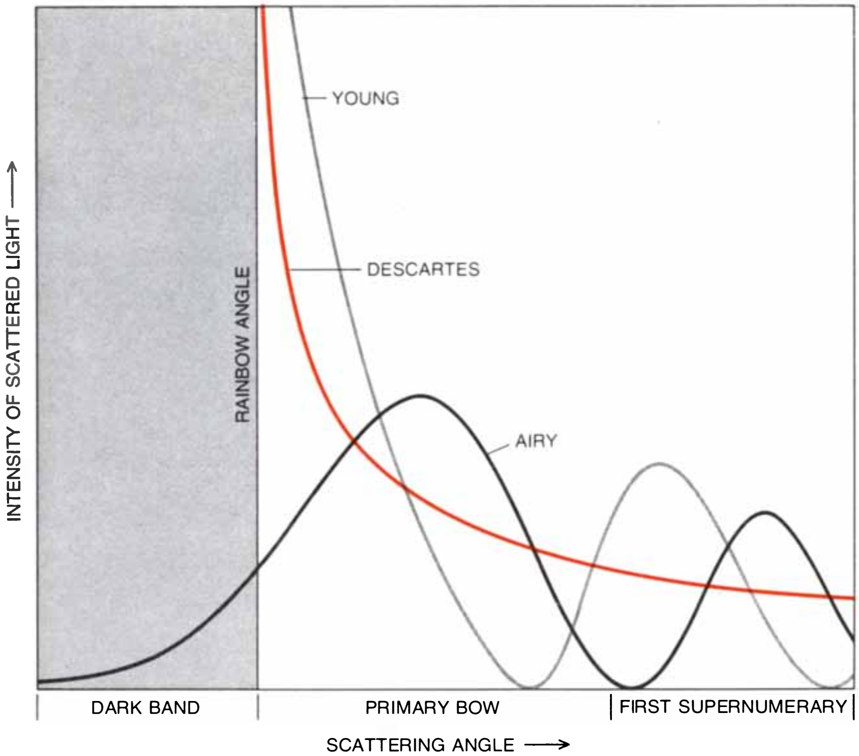
With Young's interference theory all the major features of the rainbow could be explained, at least in a qualitative and approximate way. What was lacking was a quantitative, mathematical theory capable of predicting the intensity of the scattered light as a function of droplet size and scattering angle.

Young's explanation of the supernumerary arcs was based on a wave theory of light. Paradoxically his predictions for the other side of the rainbow, for the region of Alexander's dark band, were inconsistent with such a theory. The interference theory, like the theories of Descartes and Newton, predicted complete darkness in this region, at least when only rays of Class 3 and Class 4 were considered. Such an abrupt transition, however, is not possible, because the wave theory of light requires that sharp boundaries between light and shadow be softened by diffraction. The most familiar manifestation of diffraction is the apparent bending of light or sound at the edge of an opaque obstacle. In the rainbow there is no real obstacle, but the boundary between the primary bow and the dark band should exhibit diffraction nonetheless. The treatment of diffraction is a subtle and difficult problem in mathematical physics, and the subsequent development of the theory of the rainbow was stimulated mainly by efforts to solve it.

In 1835 Richard Potter of the University of Cambridge pointed out that the crossing of various sets of light rays in a droplet gives rise to caustic curves. A caustic, or "burning curve," represents the envelope of a system of rays and is always associated with an intensity highlight. A familiar caustic is the bright cusp-shaped curve formed in a teacup when sunlight is reflected from its inner walls. Caustics, like the rainbow, generally have a lighted side and a dark side; intensity increases continuously up to the caustic, then drops abruptly.

Potter showed that the Descartes rainbow ray—the Class 3 ray of minimum scattering angle—can be regarded as a caustic. All the other transmitted rays of Class 3, when extended to infinity, approach the Descartes ray from the lighted side; there are no rays of this class on the dark side. Thus finding the intensity of the scattered light in a rainbow is similar to the problem of determining the intensity distribution in the neighborhood of a caustic.

In 1838 an attempt to determine that distribution was made by Potter's Cambridge colleague George B. Airy. His




**PREDICTED INTENSITY as a function of scattering angle is compared for three early theories of the rainbow. In the geometric analysis of Descartes, intensity is infinite at the rainbow angle; it declines smoothly (without supernumerary arcs) on the lighted side and falls off abruptly to zero on the dark side. The theory of Thomas Young, which is based on the interference of light waves, predicts supernumerary arcs but retains the sharp transition from infinite to zero intensity. Airy's theory relocates the peaks in the intensity curve and for the first time provides (through diffraction) an explanation for gradual fading of the rainbow into shadow.**

reasoning was based on a principle of wave propagation formulated in the 17th century by Christiaan Huygens and later elaborated by Augustin Jean Fresnel. This principle regards every point of a wave front as being a source of secondary spherical waves; the secondary waves define a new wave front and hence describe the propagation of the wave. It follows that if one knew the amplitudes of the waves over any one complete wave front, the amplitude distribution at any other point could be reconstructed. The entire rainbow could be described rigorously if we knew the amplitude distribution along a wave front in a single droplet. Unfortunately the amplitude distribution can seldom be determined; all one can usually do is make a reasonable guess for some chosen wave front in the hope that it will lead to a good approximation.

The starting wave front chosen by Airy is a surface inside the droplet, normal to all the rays of Class 3 and with an inflection point (a change in the sense of curvature) where it intersects the Descartes rainbow ray. The wave amplitudes along this wave front were estimated through standard assumptions in the theory of diffraction. Airy was then able to express the intensity of the scattered light in the rainbow region in terms of a new mathematical function, then known as the rainbow integral and today called the Airy function. The mathematical form of the Airy function will not concern us here; we shall concentrate instead on its physical meaning.

The intensity distribution predicted by the Airy function is analogous to the diffraction pattern appearing in the shadow of a straight edge. On the lighted side of the primary bow there are oscillations in intensity that correspond to the supernumerary arcs; the positions and widths of these peaks differ somewhat from those predicted by the Young interference theory. Another significant distinction of the Airy theory is that the maximum intensity of the rainbow falls at an angle somewhat greater than the Descartes minimum scattering angle. The Descartes and Young theories predict an infinite intensity at that angle (because of the caustic). The Airy theory does not reach an infinite intensity at any point, and at the Descartes rainbow ray the intensity predicted is less than half the maximum. Finally, diffraction effects appear on the dark side of the rainbow: instead of vanishing abruptly the intensity tapers away smoothly within Alexander's dark band.

Airy's calculations were for a monochromatic rainbow. In order to apply his method to a rainbow produced in sunlight one must superpose the Airy patterns generated by the various monochromatic components. To proceed further and describe the perceived image of the rainbow requires a theory of color vision.

The purity of the rainbow colors is determined by the extent to which the component monochromatic rainbows overlap; that in turn is determined by the droplet size. Uniformly large drops (with diameters on the order of a few millimeters) generally give bright rainbows with pure colors; with very small droplets (diameters of .01 millimeter or so) the overlap of colors is so great that the resulting light appears to be almost white.

An important property of light that we have so far ignored is its state of polarization. Light is a transverse wave, that is, one in which the oscillations are perpendicular to the direction of propagation. (Sound, on the other hand, is a longitudinal vibration.) The orientation of the transverse oscillation can be resolved into components along two mutually perpendicular axes. Any light ray can be described in terms of these two independent states of linear polarization. Sunlight is an incoherent mixture of the two in equal proportions; it is often said to be randomly polarized or simply unpolarized. Reflection can alter its state of polarization, and in that fact lies the importance of polarization to the analysis of the rainbow.

Let us consider the reflection of a light ray traveling inside a water droplet when it reaches the boundary of the droplet. The plane of reflection, the plane that contains both the incident and the reflected rays, provides a convenient geometric reference. The polarization states of the incident light can be defined as being parallel to that plane and perpendicular to it. For both polarizations the reflectivity of the surface is slight at angles of incidence near the perpendicular, and it rises very steeply near a critical angle whose value is determined by the index of refraction. Beyond that critical angle the ray is totally reflected, regardless of polarization. At intermediate angles, however, reflectivity depends on polarization. As the angle of incidence becomes shallower a steadily larger portion of the perpendicularly polarized component is reflected. For the parallel component, on the other hand, reflectivity falls before it begins to increase. At one angle in particular, reflectivity for the parallel-polarized wave vanishes entirely; that wave is totally transmitted. Hence for sunlight incident at that angle the internally reflected ray is completely polarized perpendicular to the plane of reflection. The angle is called Brewster's angle, after David Brewster, who discussed its significance in 1815.

Light from the rainbow is almost completely polarized, as can be seen by looking at a rainbow through Polaroid sunglasses and rotating the lenses around the line of sight. The strong polarization results from a remarkable coincidence: the internal angle of incidence for the rainbow ray is very close to Brewster's angle. Most of the parallel component escapes in the transmitted rays of Class 2, leaving a preponderance of perpendicular rays in the rainbow.

With the understanding that both matter and radiation can behave as waves, the theory of the rainbow has been enlarged in scope. It must now encompass new, invisible rainbows produced in atomic and nuclear scattering.

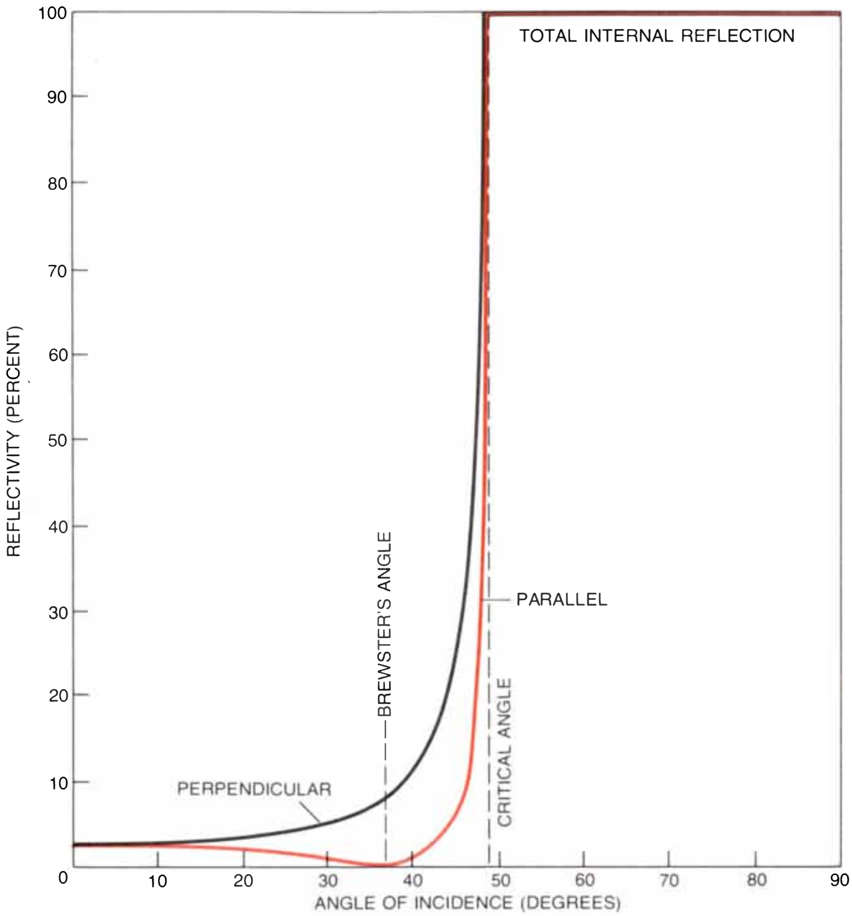
An analogy between geometrical optics and classical particle mechanics had already been perceived in 1831 by William Rowan Hamilton, the Irish mathematician. The analogues of rays in geometrical optics are particle trajectories, and the bending of a light ray on entering a medium with a different refractive index corresponds to the deflection of a moving particle under the action of a force. Particle-scattering analogues exist for many effects in optics, including the rainbow.

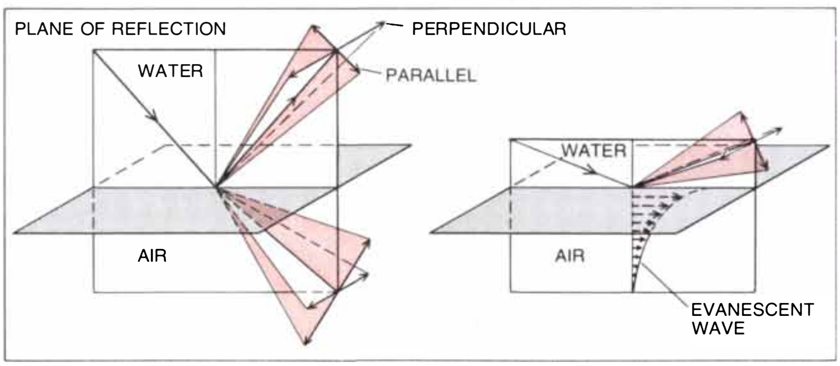
Consider a collision between two atoms in a gas. As the atoms approach from a large initial separation, they are at first subject to a steadily increasing attraction. At closer range, however, the electron shells of the atoms begin to interpenetrate and the attractive force diminishes. At very close range it becomes an increasingly strong repulsion.

As in the optical experiment, the atomic scattering can be analyzed by tracing the paths of the atoms as a function of the impact parameter. Because the forces vary gradually and continuously, the atoms follow curved trajectories instead of changing direction suddenly, as at the boundary between media of differing refractive index. Even though some of the trajectories are rather complicated, each impact parameter corresponds to a single deflection angle; moreover, there is one trajectory that represents a local maximum angular deflection. That trajectory turns out to be the one that makes the most effective use of the attractive interaction between atoms. A strong concentration of scattered particles is expected near this angle; it is the rainbow angle for the interacting atoms.

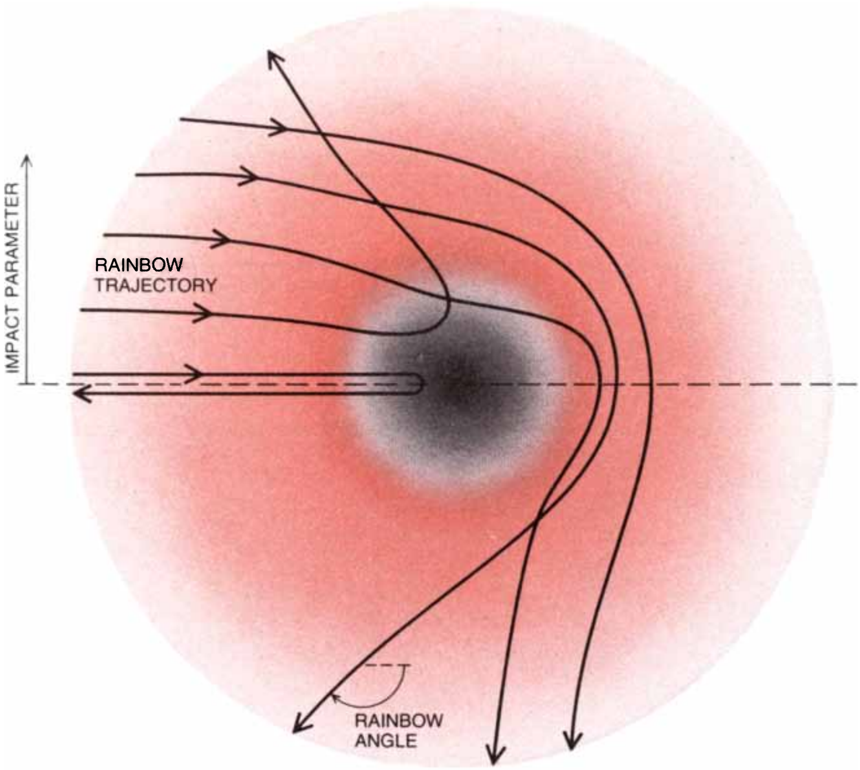
A wave-mechanical treatment of the atomic and nuclear rainbows was formulated in 1959 by Kenneth W. Ford of Brandeis University and John A. Wheeler of Princeton University. Interference between trajectories emerging in the same direction gives rise to supernumerary peaks in intensity. A particle-scattering analogue of Airy's theory has also been derived.

An atomic rainbow was first observed in 1964, by E. Hundhausen and H. Pauly of the University of Bonn, in the scattering of sodium atoms by mercury atoms. The main rainbow peak and two supernumeraries were detected; in more recent experiments oscillations on an even finer scale have been observed. The rainbows measured in these experiments carry information about the interatomic forces. Just as the optical rain-

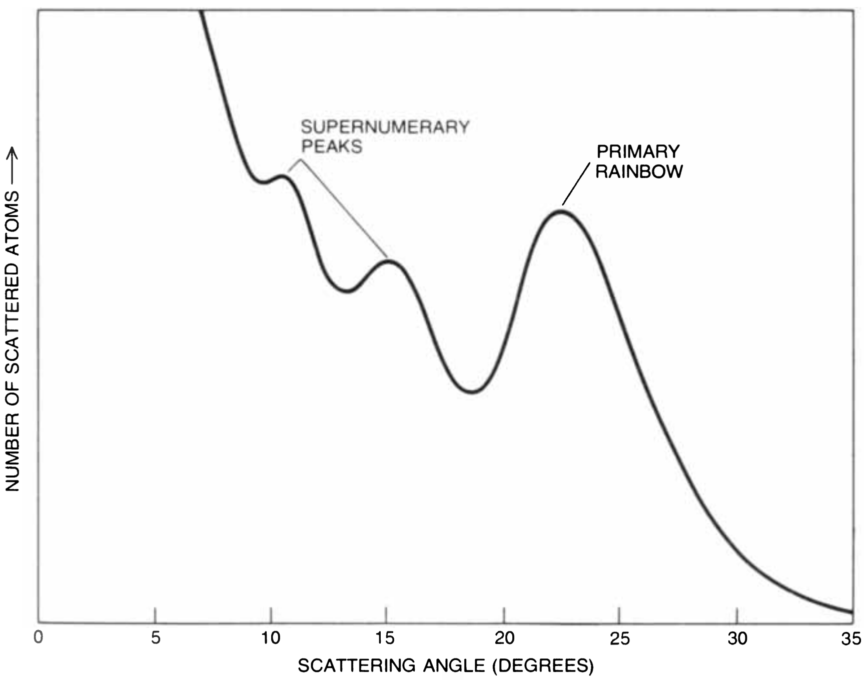



**POLARIZATION OF THE RAINBOW results from differential reflection. An incident ray can be resolved into two components polarized parallel to and perpendicular to the plane of reflection. For a ray approaching an air-water boundary from inside a droplet the reflectivity of the surface depends on the angle of incidence. Beyond a critical angle both parallel and perpendicular components are totally reflected, although some light travels parallel to the surface as an "evanescent wave." At lesser angles the perpendicular component is reflected more efficiently than the parallel one, and at one angle in particular, Brewster's angle, parallel-polarized light is completely transmitted. The angle of internal reflection for the rainbow ray falls near Brewster's angle. As a result light from the rainbow has a strong perpendicular polarization.**




**SCATTERING OF ATOMS BY ATOMS** creates a particulate rainbow. The role played in optical scattering by the refractive index is played here by interatomic forces. The principal difference is that the forces vary smoothly and continuously, so that the atoms follow curved trajectories. As one atom approaches another the force between them is initially a steadily growing attraction (*colored shading*), but at close range it becomes strongly repulsive (*gray shading*). A local maximum in the scattering angle corresponds to the optical rainbow angle. It is the angle made by the trajectory most effective in using the attractive part of the potential.




**ATOMIC RAINBOW** was detected by E. Hundhausen and H. Pauly of the University of Bonn in the scattering of sodium atoms by mercury atoms. The oscillations in the number of scattered atoms detected correspond to a primary rainbow and to two supernumerary peaks. A rainbow of this kind embodies information about the strength and range of the interatomic forces.

bow angle depends solely on the refractive index, so the atomic rainbow angle is determined by the strength of the attractive part of the interaction. Similarly, the positions of the supernumerary peaks are size-dependent, and they provide information about the range of the interaction. Observations of the same kind have now been made in the scattering of atomic nuclei.

The Airy theory of the rainbow has had many satisfying successes, but it contains one disturbing uncertainty: the need to guess the amplitude distribution along the chosen initial wave front. The assumptions employed in making that guess are plausible only for rather large raindrops. In this context size is best expressed in terms of a "size parameter," defined as the ratio of a droplet's circumference to the wavelength of the light. The size parameter varies from about 100 in fog or mist to several thousand for large raindrops. Airy's approximation is plausible only for drops with a size parameter greater than about 5,000.

It is ironic that a problem as intractable as the rainbow actually has an exact solution, and one that has been known for many years. As soon as the electromagnetic theory of light was proposed by James Clerk Maxwell about a century ago, it became possible to give a precise mathematical formulation of the optical rainbow problem. What is needed is a computation of the scattering of an electromagnetic plane wave by a homogeneous sphere. The solution to a similar but slightly easier problem, the scattering of sound waves by a sphere, was discussed by several investigators, notably Lord Rayleigh, in the 19th century. The solution they obtained consisted of an infinite series of terms, called partial waves. A solution of the same form was found for the electromagnetic problem in 1908 by Gustav Mie and Peter J. W. Debye.

Given the existence of an exact solution to the scattering problem, it might seem an easy matter to determine all its features, including the precise character of the rainbow. The problem, of course, is the need to sum the series of partial waves, each term of which is a rather complicated function. The series can be truncated to give an approximate solution, but this procedure is practical only in some cases. The number of terms that must be retained is of the same order of magnitude as the size parameter. The partial-wave series is therefore eminently suited to the treatment of Rayleigh scattering, which is responsible for the blue of the sky; in that case the scattering particles are molecules and are much smaller than the wavelength, so that one term of the series is enough. For the rainbow problem size parameters up to several thousand must be considered.

A good approximation to the solution by the partial-wave method would require evaluating the sum of several thousand complicated terms. Computers have been applied to the task, but the results are rapidly varying functions of the size parameter and the scattering angle, so that the labor and cost quickly become prohibitive. Besides, a computer can only calculate numerical solutions; it offers no insight into the physics of the rainbow. We are thus in the tantalizing situation of knowing a form of the exact solution and yet being unable to extract from it an understanding of the phenomena it describes.

The first steps toward the resolution of this paradox were taken in the early years of the 20th century by the mathematicians Henri Poincaré and G. N. Watson. They found a method for transforming the partial-wave series, which converges only very slowly onto a stable value, into a rapidly convergent expression. The technique has come to be known as the Watson transformation or as the complex-angular-momentum method.

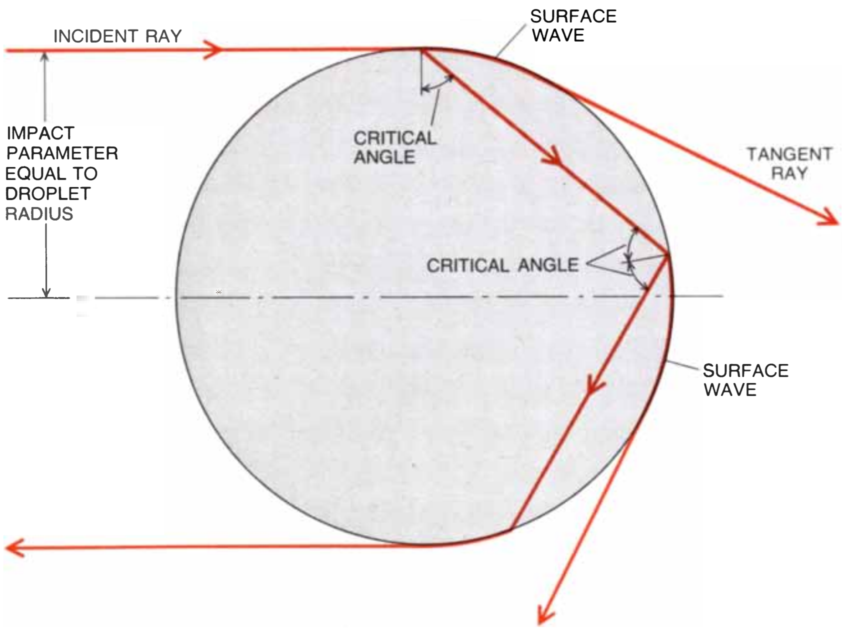
It is not particularly hard to see why angular momentum is involved in the rainbow problem, although it is less obvious why "complex" values of the angular momentum need to be considered. The explanation is simplest in a corpuscular theory of light, in which a beam of light is regarded as a stream of the particles called photons. Even though the photon has no mass, it does transport energy and momentum in inverse proportion to the wavelength of the corresponding light wave. When a photon strikes a water droplet with some impact parameter greater than zero, the photon carries an angular momentum equal to the product of its linear momentum and the impact parameter. As the photon undergoes a series of internal reflections, it is effectively orbiting the center of the droplet. Actually quantum mechanics places additional constraints on this process. On the one hand it requires that the angular momentum assume only certain discrete values; on the other it denies that the impact parameter can be precisely determined. Each discrete value of angular momentum corresponds to one term in the partial-wave series.

In order to perform the Watson transformation, values of the angular momentum that are conventionally regarded as being "unphysical" must be introduced. For one thing the angular momentum must be allowed to vary continuously, instead of in quantized units; more important, it must be allowed to range over the complex numbers: those that include both a real component and an imaginary one, containing some multiple of the square root of $-1$. The plane defined by these two components is referred to as the complex-angular-momentum plane.

Much is gained in return for the mathematical abstractions of the complex-angular-momentum method. In particular, after going over to the complex-angular-momentum plane through the Watson transformation, the contributions to the partial-wave series can be redistributed. Instead of a great many terms, one can work with just a few points called poles and saddle points in the complex-angular-momentum plane. In recent years the poles have attracted great theoretical interest in the physics of elementary particles. In that context they are usually called Regge poles, after the Italian physicist Tullio Regge.

Both poles and saddle points have physical interpretations in the rainbow problem. Contributions from real saddle points are associated with the ordinary, real light rays we have been considering throughout this article. What about complex saddle points? Imaginary or complex numbers are ordinarily regarded as being unphysical solutions to an equation, but they are not meaningless solutions. In descriptions of wave propagation imaginary components are usually associated with the damping of the wave amplitude. For example, in the total internal reflection of a light ray at a water-air boundary a light wave does go "through the looking glass." Its amplitude is rapidly damped, however, so that the intensity becomes negligible within a depth on the order of a single wavelength. Such a wave does not propagate into the air; instead it becomes attached to the interface between the water and the air, traveling along the surface; it is called an evanescent wave. The mathematical description of the evanescent wave involves the imaginary components of a solution. The effect called quantum-mechanical tunneling, in which a particle passes through a potential barrier without climbing over it, has a similar mathematical basis. "Complex rays" also appear on the shadow side of a caustic, where they describe the damped amplitude of the diffracted light waves.

Regge-pole contributions to the transformed partial-wave series are associated with surface waves of another kind. These waves are excited by incident rays that strike the sphere tangentially. Once such a wave is launched, it travels around the sphere, but it is continually damped because it sheds radiation tangentially, like a garden sprinkler. At each point along the wave's circumferential path it also penetrates the sphere at the critical angle for total internal reflection, reemerging as a surface wave after taking one or more such shortcuts. It is interesting to note that Johannes Kepler conjectured in 1584 that "pin-




**COMPLEX-ANGULAR-MOMENTUM theory of the rainbow begins with the observation that a photon, or quantum of light, incident on a droplet at some impact parameter (which cannot be exactly defined) carries angular momentum. In the theory, components of that angular momentum are extended to complex values, that is, values containing the square root of $-1$. The consequences of this procedure can be illustrated by the example of a ray striking a droplet tangentially. The ray stimulates surface waves, which travel around the droplet and continuously shed radiation. The ray can also penetrate the droplet at the critical angle for total internal reflection, emerging either to form another surface wave or to repeat the shortcut.**

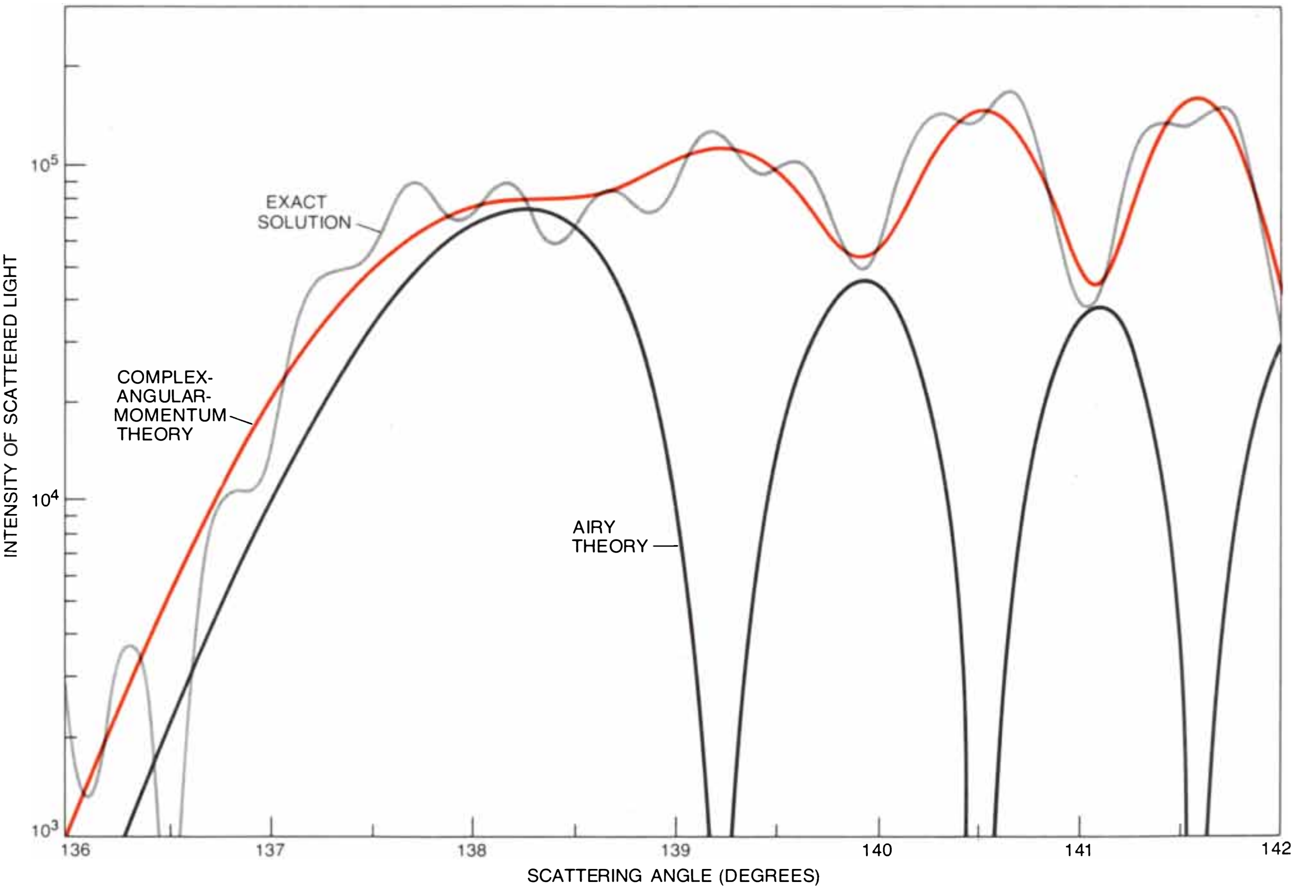
wheel" rays of this kind might be responsible for the rainbow, but he abandoned the idea because it did not lead to the correct rainbow angle.

In 1937 the Dutch physicists Balthus Van der Pol and H. Bremmer applied Watson's transformation to the rainbow problem, but they were able to show only that Airy's approximation could be obtained as a limiting case. In 1965 I developed an improved version of Watson's method, and I applied it to the rainbow problem in 1969 with somewhat greater success.

In the simple Cartesian analysis we saw that on the lighted side of the rainbow there are two rays emerging in the same direction; at the rainbow angle these coalesce into the single Descartes ray of minimum deflection and on the shadow side they vanish. In the complex-angular-momentum plane, as I have mentioned, each geometric ray corresponds to a real saddle point. Hence in mathematical terms a rainbow is merely the collision of two saddle points in the complex-angular-momentum plane. In the shadow region beyond the rainbow angle the saddle points do not simply disappear; they become complex, that is, they develop imaginary parts. The diffracted light in Alexander's dark band arises from a complex saddle point. It is an example of a "complex ray" on the shadow side of a caustic curve.

It should be noted that the adoption of the complex-angular-momentum method does not imply that earlier solutions to the rainbow problem were wrong. Descartes's explanation of the primary bow as the ray of minimum deflection is by no means invalid, and the supernumerary arcs can still be regarded as a product of interference, as Young proposed. The complex-angular-momentum method simply gives a more comprehensive accounting of the paths available to a photon in the rainbow region of the sky, and it thereby achieves more accurate results.

In 1975 Vijay Khare of the University of Rochester made a detailed comparison of three theories of the rainbow: the Airy approximation, the "exact" solution, obtained by a computer summation of the partial-wave series, and the rainbow terms in the complex-angular-momentum method, associated with the collision of two saddle points. For the dominant, perpendicular polarization the Airy theory requires only small corrections within the primary bow, and its errors become appreciable only in the region of the supernumerary arcs. For the scattered rays polarized parallel to the scattering plane, however, Airy's approximation fails badly. For the supernumerary arcs the exact solution shows minima where the Airy theory has maximum intensity, and vice versa. This serious failure is an indirect result of the near coincidence between the angle of internal reflection for the rainbow rays and Brewster's angle. At Brewster's angle the amplitude of the reflected ray changes sign, a change the Airy theory does not take into account. As a result of the change in sign the interference along directions corresponding to the peaks in




**QUANTITATIVE THEORIES of the rainbow predict the intensity of the scattered light as a function of the scattering angle and also with respect to droplet size and polarization. Here the predictions of three theories are presented for parallel-polarized light scattered by droplets with a circumference equal to 1,500 wavelengths of the light. One curve represents the "exact" solution to the rainbow problem, derived from James Clerk Maxwell's equations describing electromagnetic radiation. The exact solution is the sum of an infinite series of terms, approximated here by adding up more than 1,500 complicated terms for each point employed in plotting the curve. The Airy theory is clearly in disagreement with the exact solution, particularly in the angular region of the supernumerary arcs. There the exact solution shows troughs at the positions of Airy's peaks. The results obtained by the complex-angular-momentum method, on the other hand, correspond closely to the exact solution, failing only to reproduce small, high-frequency oscillations. These fluctuations are associated with another optical phenomenon in the atmosphere, the glory, which is also explained by the complex-angular-momentum theory.**

the Airy solutions is destructive instead of constructive.

In terms of large-scale features, such as the primary bow, the supernumerary arcs and the dark-side diffraction pattern, the complex-angular-momentum result agrees quite closely with the exact solution. Smaller-scale fluctuations in the exact intensity curve are not reproduced as well by the rainbow terms in the complex-angular-momentum method. On the other hand, the exact solution, for a typical size parameter of 1,500, requires the summation of more than 1,500 complicated terms; the complex-angular-momentum curve is obtained from only a few much simpler terms.

The small residual fluctuations in the exact intensity curve arise from higher-order internal reflections: rays belonging to classes higher than Class 3 or Class 4. They are of little importance for the primary bow, but at larger scattering angles their contribution increases and near the backward direction it becomes dominant. There these rays are responsible for another fascinating meteorological display: the glory [see "The Glory," by Howard C. Bryant and Nelson Jarmie; SCIENTIFIC AMERICAN, July, 1974].

The glory appears as a halo of spectral colors surrounding the shadow an observer casts on clouds or fog; it is most commonly seen from an airplane flying above clouds. It can also be explained through the complex-angular-momentum theory, but the explanation is more complicated than that for the rainbow. One set of contributions to the glory comes from the surface waves described by Regge poles that are associated with the tangential rays of Kepler's pinwheel type. Multiple internal reflections that happen to produce closed, star-shaped polygons play an important role, leading to resonances, or enhancements in intensity. Such geometric coincidences are very much in the spirit of Kepler's theories.

A second important set of contributions, demonstrated by Khare, is from the shadow side of higher-order rainbows that appear near the backward direction. These contributions represent the effect of complex rays. The 10th-order rainbow, formed only a few degrees away from the backward direction, is particularly effective.

For the higher-order rainbows Airy's theory would give incorrect results for both polarizations, and so the complex-angular-momentum theory must be employed. One might thus say the glory is formed in part from the shadow of a rainbow. It is gratifying to discover in the elegant but seemingly abstract theory of complex angular momentum an explanation for these two natural phenomena, and to find there an unexpected link between them.

# MATHEMATICAL GAMES

## *The pool-table triangle, a limerick paradox and divers other challenges*

by Martin Gardner

Here are eight miscellaneous problems, together with a note on the popular board game Othello and a discussion of last month's problem. I shall give the solutions to the eight miscellaneous problems next month.

1. Pool-Table Triangle. Colonel George Sicherman of Buffalo tells me that about 10 years ago, while he was watching a game of pool, the following problem occurred to him: Is it possible to form a "difference triangle" in arranging the 15 balls in the usual triangular configuration before the start of a game? In a difference triangle the numbers 1 through 15 are arranged in such a way that each number below a pair of numbers is the positive difference between that pair.

It is evident from the illustration below that the problem is trivial with the balls numbered 1, 2 and 3 and that two solutions can be obtained with them. The illustration also shows the four solutions for six balls and the four for 10 balls. To Sicherman's surprise the 15 pool balls have only one basic solution. (It can, of course, be reflected.) Can you find it?

Searching for the solution is considerably simplified by first exploring triangular patterns of even and odd to see which patterns have exactly eight odd and seven even spots. It does not take long to discover that there are only five arrangements for the triangle's top row: EEOEO, OEEEO, OOOEE, OOOOE and OOEEO. The 15 ball obviously must be in the top row and the 14 ball must be in the same row or below 15 and 1. Other dodges shorten an exhaustive analysis.

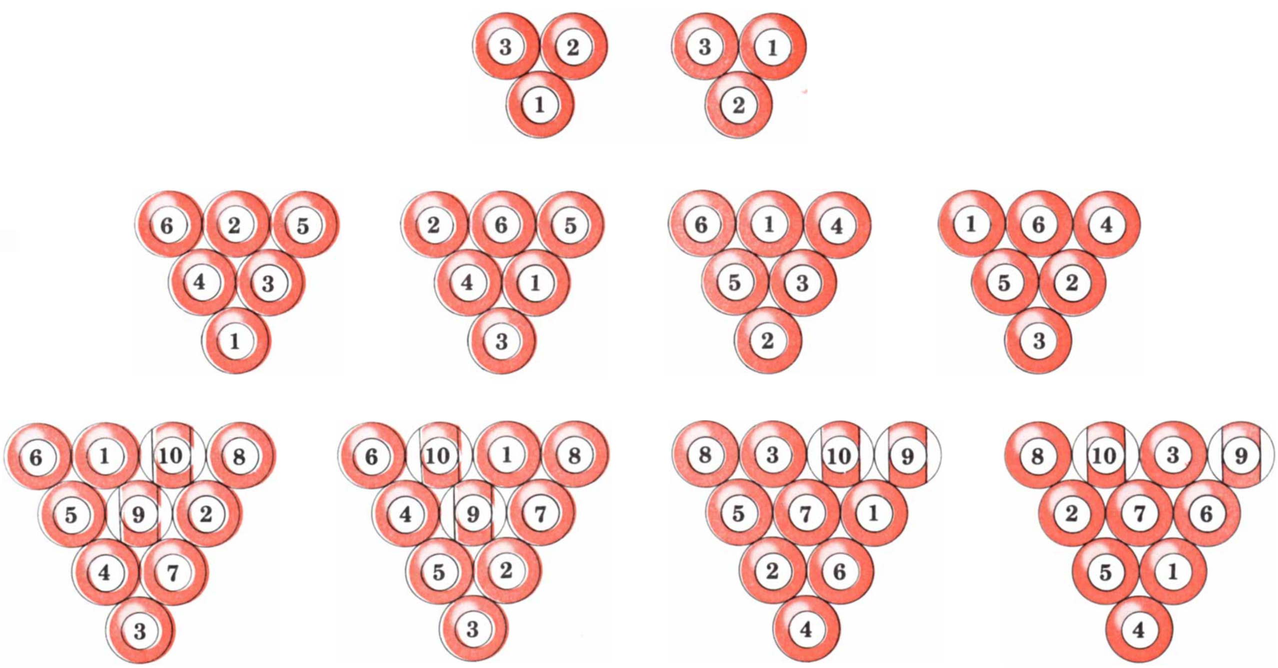
The problem is related to one posed by Hugo Steinhaus in his book *One Hundred Problems in Elementary Mathematics*. Given a triangular array with an even number of spots, is it always possible to form an even-odd difference pattern in which the number of even spots equals the number of odd ones? The problem remained unsolved for more than a decade until Heiko Harborth, in *Journal of Combinatorial Theory (A)*, Vol. 12, 1972, pages 253–259, proved that the answer is yes.

As far as I know no work has been done on what we might call the general pool-ball problem. Given any triangular number of balls, numbered consecutively from 1, is it always possible to form a difference triangle? If not, is there a largest triangle for which a solution is possible? If there is, what is it? We now know that odd-even patterns for such solutions exist for all triangles with an even number of balls. Do they also exist for all triangles with an odd number of balls?

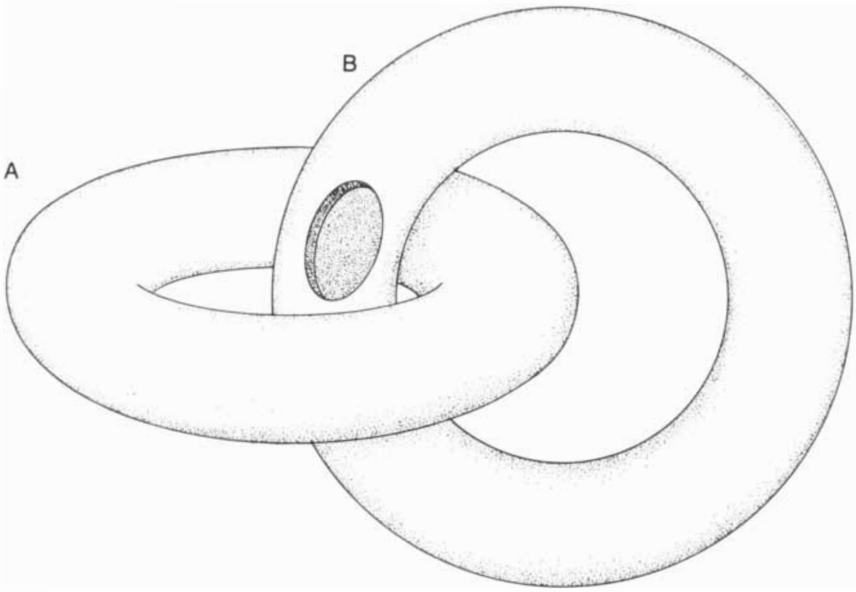
Since this is an April issue, let me add the following problem for readers who succeed in solving the 15-ball problem. Suppose the balls bear the 15 consecutive even numbers from 2 through 30. Is it possible to arrange the set in a difference triangle?

2. Toroidal Cannibalism. A torus is a surface shaped like a doughnut. Imagine a torus made of sheet rubber. It is well known that if there is a hole in such a torus, the torus can be turned inside out through the hole. (See this department for December, 1972.)

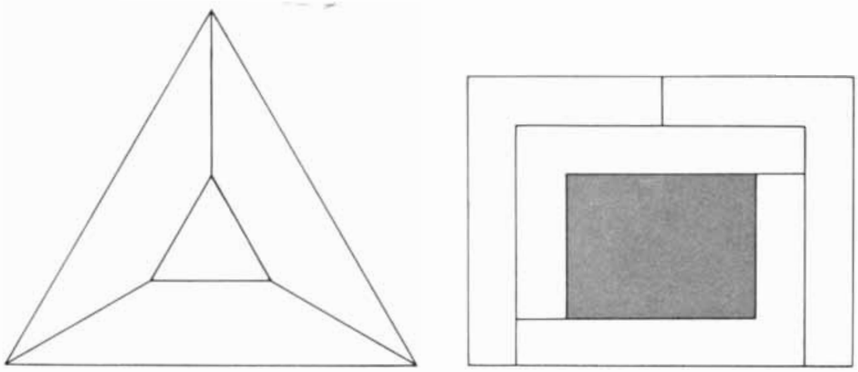
John Stillwell, a mathematician at Monash University in Australia, poses the following problem. Two toruses, *A*



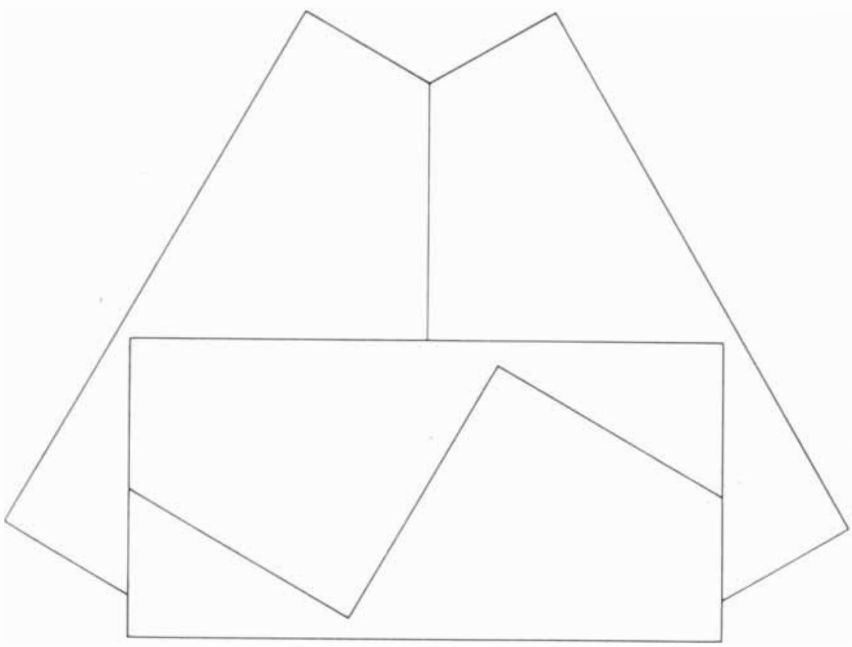
*Difference triangles for three, six and 10 pool balls*




*Can torus* **B** *swallow torus* **A?**



*A pair of tetrads*



*A tetrad of congruent hexagons*

and *B*, are linked as is shown in the top illustration on this page. There is a "mouth" (a hole) in *B*. We can stretch, compress and deform either torus as radically as we please, but of course no tearing is allowed. Can *B* swallow *A?* At the finish *B* must have its original shape, although it will be larger, and *A* must be entirely inside it.

3. Exploring Tetrads. The most sensational news last year in recreational mathematics was surely the announcement by two University of Illinois mathematicians that they had proved the four-color-map conjecture. This famous conjecture is often confused with a simpler theorem in topology, which is easily proved; it states that no more than four regions on the plane can have a mutual border. Michael R. W. Buckley, in *Journal of Recreational Mathematics*, Vol. 8 (1975), proposed the name tetrad for four simply connected planar regions, each pair of which shares a finite portion of a common boundary.

The tetrad at the left in the middle illustration on this page has no holes. Note that only three regions are congruent. The tetrad at the right has four congruent regions and a large hole. Is it possible, Buckley asked, to construct a tetrad with four congruent regions and no hole?

This question has been answered affirmatively by Scott Kim, a student at Stanford University. His results have not been published, and I am grateful to him for his permission to give some of them here.

The bottom illustration on this page shows a solution with four congruent hexagons. It is not known if a solution can be achieved with a polygon of fewer than six sides or if there is a solution with an outside border that is convex.

Part *a* of the illustration on page 132 shows a solution with congruent polyhexes of order 4. (A polyhex is a union of congruent regular hexagons.) It is easy to show that no solution with lower-order polyhexes is possible.

Part *b* of the illustration shows a solution with congruent polyiamonds of order 10. (A polyiamond is a union of congruent equilateral triangles.) It is not known whether there is a solution with lower-order polyiamonds.

Part *c* of the illustration shows a solution with congruent polyominoes of order 12. (A polyomino is a union of congruent squares.) It is not known whether there is a solution with lower-order polyominoes.

Part *d* of the illustration shows a solution with congruent pieces that have bilateral symmetry and a bilaterally symmetric border. Is there such a solution with polygons of fewer sides?

I shall be pleased to hear of any improvements that readers are able to make on the above results.

It has been known since the 1870's

that in three dimensions an infinite number of congruent solids can be put together so that every pair shares a common portion of surface. For readers who are unfamiliar with this result, here is a problem to be answered next month. Show how an infinite number of congruent polycubes can be put together, with no interior hole, so that every pair shares part of a surface. (A polycube is a union of identical cubes.) Such a structure proves that an unlimited number of colors are required for coloring any three-dimensional "map."

4. Knights and Knaves. Raymond Smullyan, a mathematician at the City University of New York whose remarkable chess problems are familiar to regular readers of this column, is responsible for the following four charming logic puzzles involving knights and knaves and perhaps some other people. In all four problems a knight always tells the truth, a knave always lies.

*A* says: "*B* is a knight."
*B* says: "*A* is not a knight."
Prove that one of them is telling the truth but is not a knight.

*A* says:"*B* is a knight."
*B* says: "*A* is a knave."
Prove either that one of them is telling the truth but is not a knight or that one is lying but is not a knave.

In the above problems we must consider the possibility that a speaker is neither a knight nor a knave. In the next two problems each of the three people involved is either a knight or a knave.

*C* says: "*B* is a knave."
*B* says: "*A* and *C* are of the same type [both knights or both knaves]."
What is *A?*

*A* says: "*B* and *C* are of the same type."
*C* is asked: "Are *A* and *B* of the same type?"
What does *C* answer?

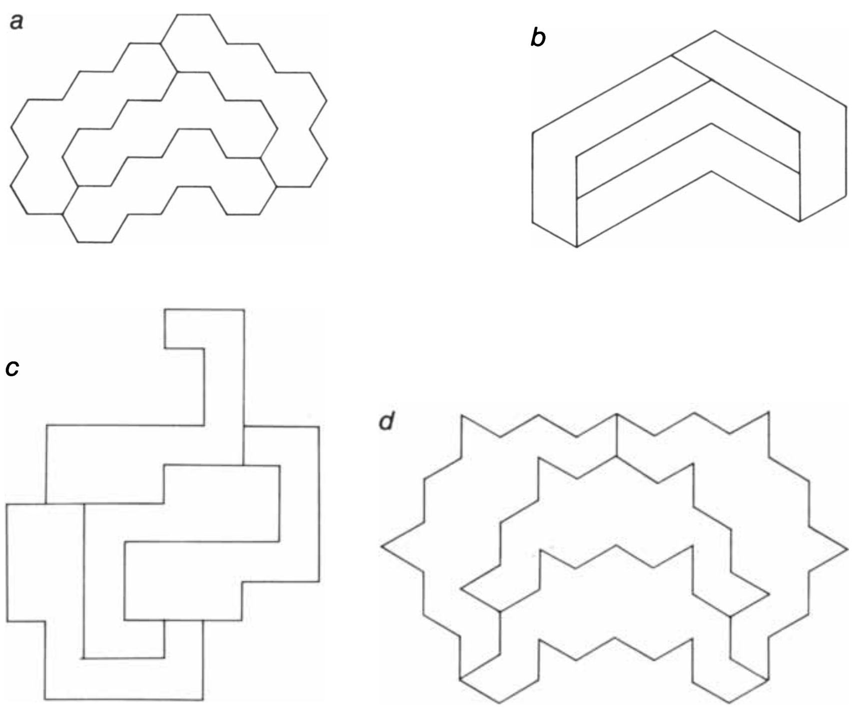
5. Lost-King Tours. Several years ago Scott Kim proposed what he calls the "Lost-King's Tour." This is a king's tour of a small chessboard subject to the following conditions:

First, the king must visit each cell once and only once.

Second, the king must change direction after each move, that is, it cannot move twice consecutively in the same direction.

Third, the number of spots where the king's path crosses itself must be minimized.

Part *a* of the top illustration on page 134 shows the only possible tour on a 3 × 3 board from cell *A* to cell *B*. It has one crossing and is unique, except of course for reflection by the main diagonal. A closed tour is impossible on this board. Closed tours with no crossings are easily found on the 4 × 4 board. On the 5 × 5 board a closed tour probably requires two crossings. The problem is less interesting on larger chessboards




*Solutions for a variety of tetrads*

because crossing-free closed tours and crossing-free open tours from any cell to any other cell are believed to be always possible.

Here are two beautiful tour problems devised by Kim:

On the order-4 board shown in part *b* of the illustration find a lost-king's tour from *A* to *B* with as few as three crossings. The solution is unique.

On the order-5 board shown in part *c* of the illustration find a lost-king's tour from *A* to *B* with as few as two crossings. This problem is unusually difficult. Kim does not know whether the solution he found is unique or whether the problem can be solved with only one crossing.

6. Steiner Ellipses. This oldie goes back to Jakob Steiner, a noted Swiss geometer of the 19th century. My excuse for reviving it is that it is one of the best examples I know of a problem that is difficult to solve by calculus or analytic geometry but that is ridiculously easy if it is approached with the right turn of mind and some knowledge of elementary plane and projective geometry.

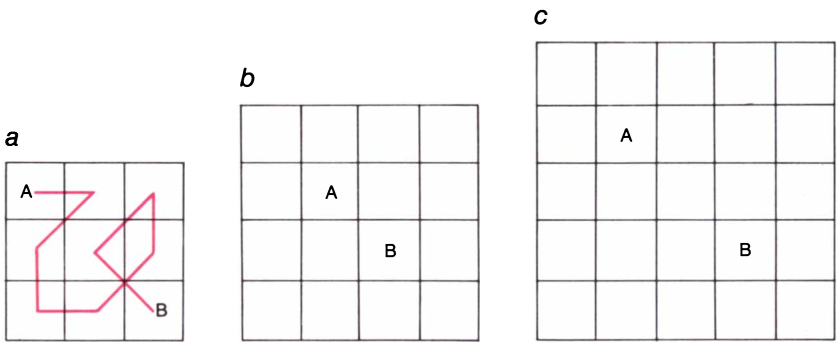
We are given a 3, 4, 5 triangle. Its area is six square units. We wish to calculate both the smallest area of an ellipse that can be circumscribed around it and the largest area of an ellipse that can be inscribed inside it.

7. Different Distances. It is easy to place three counters on the cells of a $3 \times 3$ checkerboard so that no two pairs of counters are the same distance apart. We assume that each counter marks the exact center of a cell and that distances are measured on a straight line joining the centers. Discounting rotations and reflections, there are three solutions, which appear in the top part of the bottom illustration on the next page.

It also is easy to put four counters on a $4 \times 4$ board so that all distances between pairs are different. There are 16 ways of doing it. On the $5 \times 5$ board the number jumps to 28.

In the January 1972 issue of *Journal of Recreational Mathematics* Sidney Kravitz asked for solutions to the order-5 and the order-6 squares. The solution for the order-6 square proved to be difficult because for the first time the 3, 4, 5 right triangle (the smallest Pythagorean triple) enters the picture. The fact that an integral distance of 5 is now possible both orthogonally and diagonally severely limits the patterns. As readers of the journal discovered, there are only two solutions. They appear in the bottom part of the bottom illustration on the next page.

For what squares of side *n* is it possible to place *n* counters so that all distances are different? As reported in the fall issue of the journal last year, John




*Lost-king tours on three boards of different sizes*

H. Muson proved (using a pigeonhole argument) that no solutions are possible on squares of order 16 and higher. Harry L. Nelson, the journal's new editor, lowered the limit to 15. Milton W. Green, with an exhaustive-search computer program, established impossibility for orders 8 and 9 and found a unique solution for order 7. David Babcock, an editor of *Popular Computing,* confirmed these results through order 8. The problem was finally laid to rest last November by Michael Beeler. His computer program confirmed the above results and proved impossibility for orders 10 through 14.

The order-7 board, therefore, is the largest for which there is a solution. The solution is unique and extremely hard to find without a computer. Readers may nonetheless enjoy searching for it.

*Journal of Recreational Mathematics,* now in its ninth year, continues to be crammed with new and exciting material. (Subscription rates in the U.S. for four issues are $10 for individual subscribers and $21 for institutions; subscriptions to the journal are available from the Baywood Publishing Company, 43 Central Drive, Farmingdale, N.Y. 11735.)

8. A Limerick Paradox. An amusing variant of the old liar paradox appeared a few years ago in the British monthly *Games and Puzzles.* It is presented here as the last of four "limericks."

There was a young girl in Japan
Whose limericks never would scan.
When someone asked why,
She said with a sigh,
"It's because I always attempt to
get as many words into the last
line as I possibly can."

Another young poet in China
Had a feeling for rhythm much fina.
His limericks tend
To come to an end
Suddenly.

There was a young lady of Crewe
Whose limericks stopped at line two.

There was a young man of Verdun.

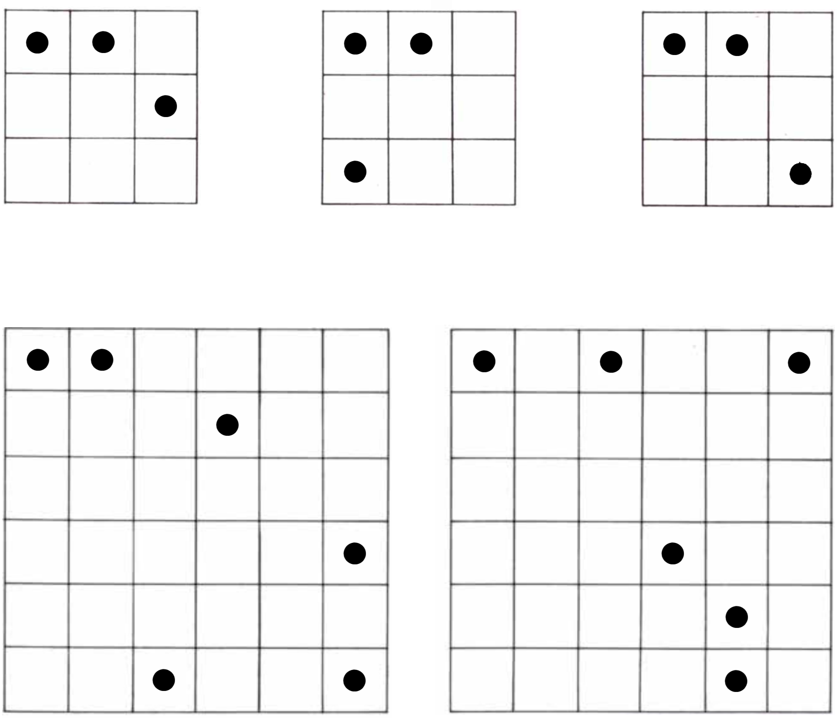
The popularity of the "new" board game Othello prompts this note. According to *Time* (November 22, 1976, page 97), the game was invented in 1971 by a Japanese salesman and named Othello by his father, a Shakespearean scholar, because it displayed such "dramatic reversals." After the game had swept Japan in 1973 (millions of sets were sold) Gabriel Industries bought the American rights, and the game is now a big seller in this country.

Othello .is the 19th-century English board game of reversi with nothing altered except the name. It is still sold in England under its original name. Readers who would like to avoid paying $9 for a small plastic board and its accompanying poker chips can play the game on any chessboard or checkerboard with a supply of chips in two colors. The history and rules of reversi were given in this department in April, 1960.

As I wrote in 1960, it is "a game that combines complexity of structure with rules of delightful simplicity, and a game that does not deserve oblivion." N. J. D. Jacobs, a mathematician in the data-handling group of CERN (European Organization for Nuclear Research, 1211 Geneva 23, Switzerland), has an excellent computer program for reversi and would like to hear from anyone who cares to play against it, with or without the player's own computer program.

Last month's task was to analyze a game (similar to nim) in which players may take from either of two piles or take one counter from one pile and two counters from the other. The last person to play wins. In the unbounded-chessboard model explained last month for W. A. Wythoff's nim the first rule is equivalent to the move of a rook west or south and the second rule is equivalent to a knight jumping southwest. The take-away game is therefore isomorphic with the game of cornering a chess piece that combines the powers of rook and knight.

If the piece moves only like a rook, the game on the chessboard is the same as standard nim with two piles. Safe pairs are any two equal positive integers. They correspond to cells on the board's main diagonal that passes through cor-



*A point-placement problem*

# How to make your fourth most important decision one you can live with.

Choosing a home is a decision you'll come home to every night.

You'll live with its look. You'll live with its feel. And you'll live with its costs.

As you search for a house you'll enjoy living with, you'll find that a lot of builders know just what your ideal home looks like. Theirs. But what we've found is that most home buyers have a very good idea of what they want in their homes. And what they need is someone with the inclination and the ability to help them put it together.

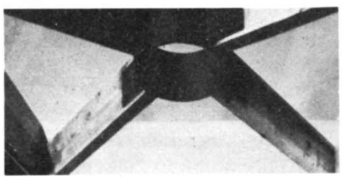
Few can do that like Acorn.

**The Design Options**

Acorn gives homebuyers a choice of 30 design approaches. Each is a contemporary, architect designed house. And each is a jumping off point for developing a personal home that combines interesting exterior shapes with attractive, functional interior spaces.

The key to building a home you'll enjoy living with is the ability to put yourself into it before you build. Our design department can help you do this by going over your ideas, giving advice and showing you how personal modifications will look and feel through drawings and models.

We encourage your participation in the process. After all, who are we to tell you your floor should be parquet if you want fieldstone?



**The Quality Factor**

There's a lot you should know about what's going on under the surface of your apparently beautiful home. The quality of the materials, the quality of workmanship and the way pieces fit together.

Acorn's technique of manufacturing homes has proven superior to total on-site construction in this respect. In our plant, where the materials are in stock, conditions are stable, proper equipment is on hand and quality is strictly controlled, the components of the homes are built to exact specifications. And attention is paid to each detail.

So when an Acorn home goes up, it goes up right.

**The Control Element.**

Unfortunately, cost overruns are commonplace in the home building industry. It's simply very difficult for most builders to account for the kind of modifications we do accurately, within an estimate.

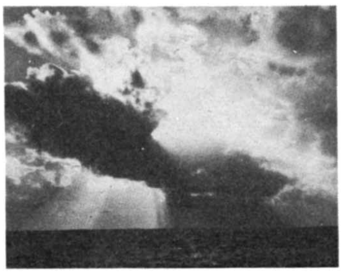
Acorn, however, enjoys a reputation for completing homes, with or without modifications, right on quote. We've achieved this by maintaining a controlling hand throughout the process. We purchase materials in quantity. We manufacture in our plant. Our own authorized builder/dealers carry out on-site construction. And since we've pretty much done it all in our years of building, we and our builder/dealers know precisely what effects particular modifications will have on the final price. Before we build.

Consequently, you do too.

**The Energy Considerations**

Energy consumption, both in terms of comfort and as a continuing financial factor has become a serious consideration in recent years. For obvious reasons.

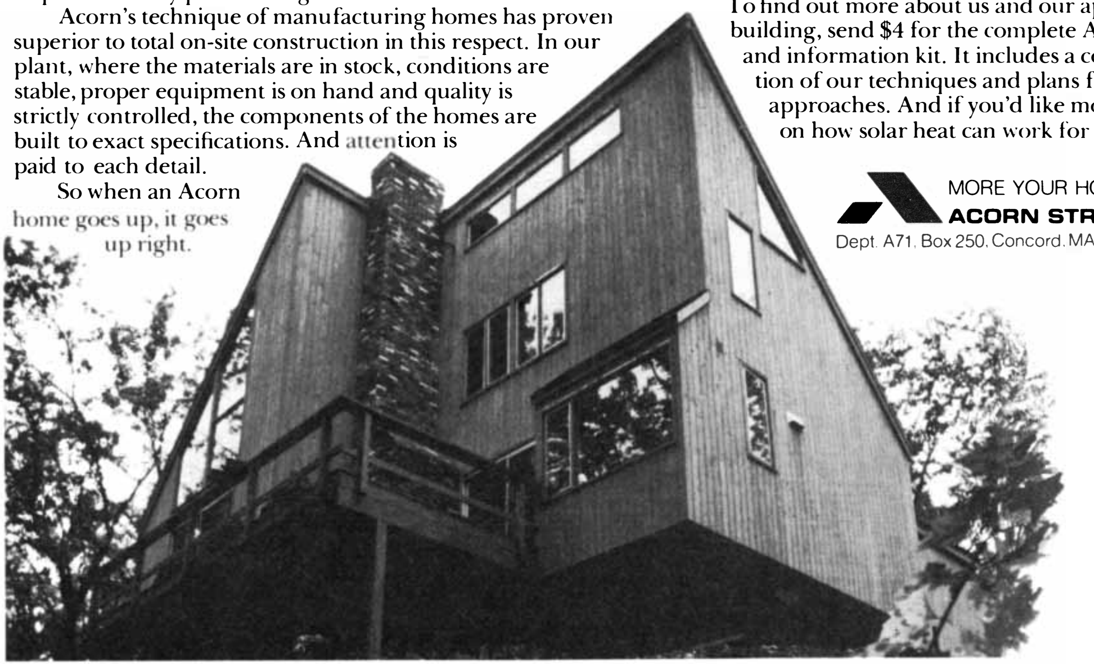
Acorn combines elements from the past and present to diminish the effects of rising energy prices. Many designs can minimize unfavorable exposures of wall and roof. Windows, purposefully arranged, and fireplaces, properly located, are used to maximize heat efficiency. Zone controlled heating, full insulation, weatherstripping and insulating glass are standard.

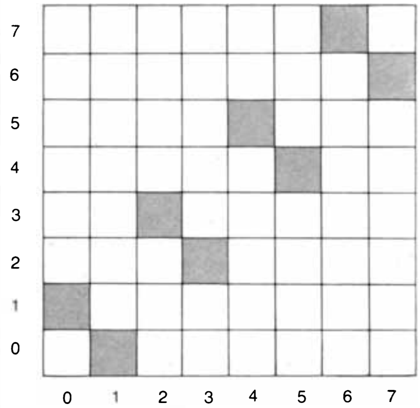


As one of the pioneers in the field of solar heat, having studied its potential in an experimental unit for a number of years, we can offer the alternative of providing much of your own energy. Independently. Adapting nicely to existing homes as well as fitting several of our own designs, there are many Acorn solar energy packages already in use today.

To find out more about us and our approach to building, send $4 for the complete Acorn planning and information kit. It includes a complete explanation of our techniques and plans for all 30 design approaches. And if you'd like more information on how solar heat can work for you, just ask.

MORE YOUR HOME THAN OURS.
**ACORN STRUCTURES, INC.**
Dept. A71, Box 250, Concord, MA 01742 (617) 369-4111






*Safe cells of rook-knight nim*

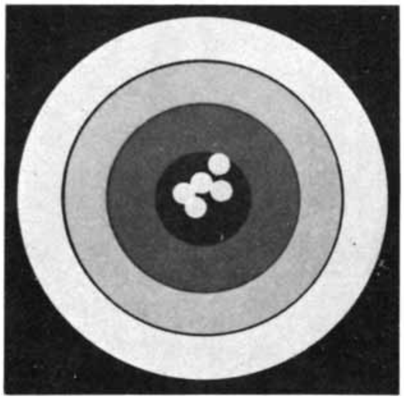
ner cells 0/0 and 7/7. The player who places the rook (on the top row or the column farthest to the right) wins only by putting it on 7/7. Thereafter his strategy is always to move to the diagonal. In the take-away game this means keeping the piles equal. The safe pairs are simply 1/1, 2/2, 3/3....

Surprisingly, giving the rook the additional power of a knight has no effect on this strategy. Applying the recursive technique explained last month, we find that the safe cells (or safe pairs) are exactly the same as in the rook game.

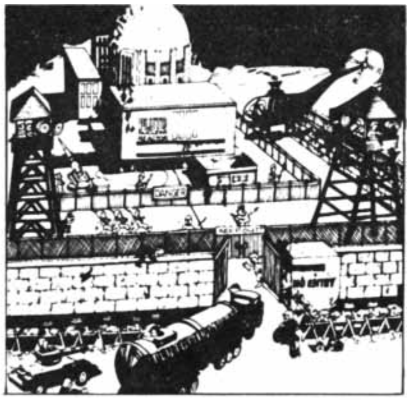
The misère form of rook-knight nim (the last person to play loses) is more interesting. The safe pairs are 0/1, 2/3, 4/5, 6/7.... On a chessboard these unordered pairs are the cells shown in gray in the illustration above. The "placer" has the win, but he must put the rook-knight on a cell adjacent to the cell in the top right corner. Thereafter he moves to occupy a safe cell. This procedure eventually brings him to 0/1 or 1/0, forcing his opponent to make the final move.

Readers might enjoy analyzing the game on a standard chessboard when the placed piece has other chess powers, in each case limiting moves to west, south and southwest. A superqueen (combining queen and knight) means a loss for the placer in standard and reverse play. A king loses for the placer in standard play but wins in misère. The same result emerges if the piece is a king-king or a king-rook. The placer wins in both types of play if the piece is a king-bishop.

A bishop or a bishop-knight provides a dull game because it cannot move from any cell in the bottom row or the column farthest to the left. Either piece obviously would result in a draw in both kinds of play if the object is to reach 0/0. If winning is defined as being the last one to play, the placer wins the standard game by putting either piece on 0/7 or 7/0; he wins the misère game by putting either piece on 1/7 or 7/1.



Megadeadly marksmanship.



How to live with Plutonium.


Inflation: Good or bad?

Will sex be out by 2001 A.D.?

Playing with fire.

Blackberry & apple – same branch?

# Our readership is international ...our coverage universal

*New Scientist* is unique, combining the immediacy of a newspaper with the authority of a learned journal. An international news magazine of science, technology, and medicine, we range from recombinant DNA to radio astronomy, from the politics of science to the psychology of scientists. A hybrid publication, renowned for cross fertilising information and ideas between the specialties, and between scientists and non-scientists. Twenty years after its first appearance, *New Scientist* has grown into a dependable and lively forum for news and opinion. An essential weekly experience for over half a million readers.

But little known outside of Britain. Now we plan to change that. Why not try *New Scientist,* for news and informed comment on the science scene worldwide, for gossip – and for sheer entertainment too? And only a dollar a week....

*Fill in this subscription order form today, and send to: New Science Publications, IPC Magazines Ltd., 205 East 42 Street, New York NY 10017.*

**SUBSCRIPTION ORDER FORM**

To: New Science Publications, IPC Magazines Ltd., 205 East 42 Street, New York NY 10017. Please enter my subscription for

☐ 1 year no index $49.50 ☐ 1 year with index $52.50

I enclose payment with order. Cheques/Postal Orders made payable to New Scientist

Name ______________________

(BLOCK LETTERS PLEASE)
Address ______________________

______________________

______________________

IPC Magazines Ltd. (Company registered in England.) Regd. No. 53626. Registered Office: King's Reach Tower, Stamford Street, London, SE1 9LS, England.

**new scientist**

# Physics...concrete math...astronomy and astrophysics...chemistry... probability...structural geology... undoubtedly the best books in every scientific area that interests you.

## Take any 3 books, values to $60.85, all for only $3.95 as a member of

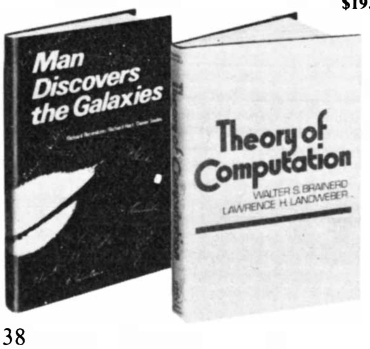
# The Library of Science

if you will join now for a trial period and agree to accept only 3 more books—at members' discount prices—over the next 12 months.

(Publishers' Prices shown)

**59800. MAN DISCOVERS THE GALAXIES.** *Richard E. Berendzen, Richard C. Hart & Daniel Seeley*. Topical, current look at modern astronomy dealing mostly with the discovery and arrangement of our Home Galaxy, the Milky Way System and the discovery of the basic properties of the universe of galaxies in which we find ourselves. **$15.95**

**83760. THE THEORY OF COMPUTATION.** *Walter S. Brainerd and Lawrence H. Landweber*. Demonstrates the equivalence of various algorithmic languages, treats unsolvable problems emphasizing methods for obtaining undecidability results, studies problems known to be effectively computable and provides a thorough treatment of combinatory calculus. **$19.95**





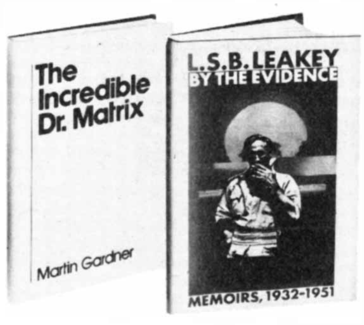
**36395. BLACK HOLES, QUASARS AND THE UNIVERSE.** *Henry L. Shipman*. Black holes, quasars, white dwarfs, neutron stars, supernovae, Seyfert Galaxies — explores the frontiers of astronomy. **$12.95**

**39865. COMPUTER DATE-BASE ORGANIZATION.** *James Martin*. How you can use the latest design and implementation techniques to improve your data system's real-time accessibility and improve flexibility — and do it at the lowest possible cost. *Counts as 2 of your 3 books*. **$26.50**

**34400. ARCHAEOLOGY BY EXPERIMENT.** *John Coles*. Scientific overview of an exciting new branch of archaeological research. Spotlights such archaeological experiments as the highly sophisticated simulated Iron Age settlements that have been established today. **$7.95**

**55190. THE INCREDIBLE DR. MATRIX.** *Martin Gardner*. The delightful incredibly brilliant Dr. Matrix returns with more amazing feats and adventures in number theory, sleight of word, numerological analysis and puzzles. **$8.95**

**37195. BY THE EVIDENCE.** *L.S.B. Leakey*. Fascinating reading of Leakey's early years in Africa and his development into one of the most respected scientists of the twentieth century. **$9.95**




**34140. APPLIED ANALYTICAL MATHEMATICS FOR PHYSICAL SCIENTISTS.** *James T. Cushing.* Deals with topics ranging from spectral analysis of linear operations to complex variables, to group theory. **$19.95**

**61530. MATHEMATICS DICTIONARY. Fourth Edition.** *Edited by Robert C. James and Edwin F. Beckenbach.* Revised and expanded edition of this leading mathematics dictionary offers clear, accurate definitions of more than 8,000 terms. **$17.95**

**70190. PRINCIPLES OF OPERATIONS RESEARCH. Second Edition.** *Harvey M. Wagner.* This 1039-page book offers a complete overview of all the fundamental concepts. **$19.95**

**49340. FUNDAMENTAL ASTRONOMY: SOLAR SYSTEM AND BEYOND.** *Franklyn W. Cole.* Traces our assault on the mysteries of the universe from the earliest sensory-oriented glimmerings to modern astronomical thought. **$16.50**

**81780. SYMMETRY DISCOVERED.** *Joe Rosen.* Fascinating look into one of the most beautiful and practical of the sciences. **$11.95**

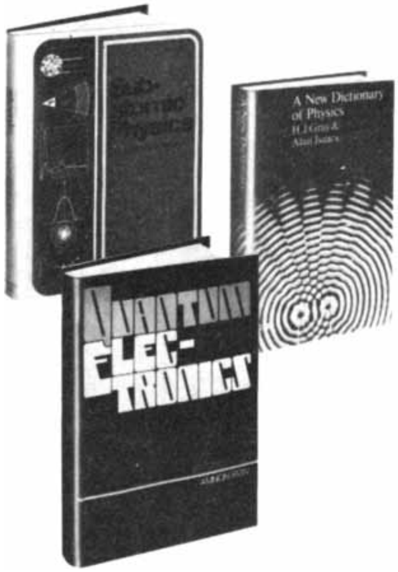

**44900. THE ENCYCLOPEDIA OF COMPUTER SCIENCE.** *Edited by Anthony Ralston and Chester L. Meek.* A monumental 1,550 page volume, offering condensed, accurate information on essential topics ranging from theory to applications, from program structures to the logical design of central processing units, from Boolean algebra to microprogramming. Bibliographies, cross references, complete indexing and 700 charts and tables. *Counts as 3 of your 3 books.* **$60.00**

**46500. THE EVOLUTION OF RADIO ASTRONOMY.** *J.S. Hey.* One of the pioneers of astronomy offers an absorbing history of this powerful technique's remarkable development."... an important work in the history of astronomy." — M.L. Meeks, in *Sky and Telescope.* **$10.00**

**73240. QUASARS, PULSARS, AND BLACK HOLES.** *Frederic Golden.* A thorough review of the spectacular discoveries in astronomy, as told by the science editor of *Time* magazine. Illustrated with twelve drawings and twenty-four color photographs. **$7.95**

**39745. COMPANION TO CONCRETE MATHEMATICS.** *Z.A. Melzak.* "A delightful collection of problems organized into over sixty sections, together with historical notes, solutions, observations, and remarks. Mathematicians at all levels will enjoy sampling this work." — *American Mathematical Monthly.* **$16.50**

**IMPORTANT GUARANTEE!** Examine your 3 introductory books for 10 days. If not delighted, simply return them and your membership will be canceled. You pay nothing, owe nothing!




**32300. ADVENTURES OF A MATHEMATICIAN.** *S.M. Ulam.* First-hand account of the birth and childhood of the Atomic Age. **$14.95**

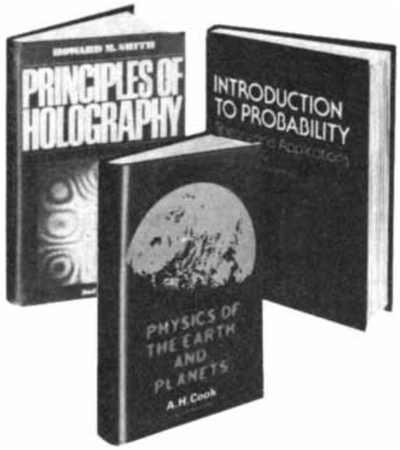
**81150. SUBATOMIC PHYSICS.** *Hans Frauenfelder and Ernest M. Henley.* Covers everything from how accelerators and detectors work to the structure of subatomic particles. **$17.50**

**64145. THE NEW DICTIONARY OF PHYSICS.** *Edited by H.J. Gay and Alan Isaacs.* Over 9,000 entries in 600 pages. A monumental reference that is intermediate between a strict dictionary and an encyclopedia. About 25% greater than the 1958 edition, it reflects all major changes characteristic of the field. *Counts as 2 of your 3 books.* **$35.00**

**82660. TECHNOLOGY MATHEMATICS HANDBOOK.** *Jan J. Tuma.* Instant access to concise summaries of all major definitions, formulas, graphs, tables of elementary and intermediate mathematics — from algebra to determinants and vectors and much, much more. **$15.95**

**73165. QUANTUM ELECTRONICS.** *Amnon Yariv.* Classic work incorporates all the latest advances in quantum electronics. Containing 70% new material, it deals with quantum mechanics, electromagnetism, laser oscillation, nonlinear optics, and much more. **$20.95**

**73530. THE REBIRTH OF COSMOLOGY.** *Jacques Merleau-Ponty and Bruno Morando.* A thorough account of the Greek, classical and astrophysical ages of the fascinating study of the universe as a whole physical system. **$12.50**

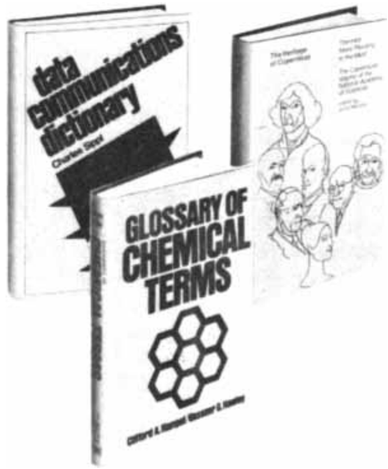

**70170. PRINCIPLES OF HOLOGRAPHY. Second Edition.** *Howard M. Smith.* From fundamentals to rigorous theory, from chemical formulas to applications. **$16.50**

**56380. INTRODUCTION TO PROBABILITY.** *Richard L. Scheaffer and William Mendenhall.* Covers everything from Markov chains to the arrival times for the Poisson process. **$13.50**

**68735. PHYSICS OF THE EARTH AND PLANETS.** *Alan H. Cook.* Brilliant investigation by one of the world's leading authorities on physical meteorology. *Counts as 2 of your 3 books.* **$24.75**

**49267. FRONTIERS OF ASTROPHYSICS.** *Edited by Eugene Avrett.* A contemporary source on astrophysical topics from recent solar research, to neutron stars, supernovae and intergalactic matter. *Counts as 2 of your 3 books.* **$20.00**

**50380. GLOSSARY OF CHEMICAL TERMS.** *Clifford A. Hampel and Gessner G. Hawley.* Over 2,000 concise definitions: all major chemical groups, important functional terms, basic phenomena and processes, all the chemical elements, and more. **$14.95**




**52650. THE HERITAGE OF COPERNICUS.** *Edited by Jerzy Neyman.* 25 highly readable essays on the quasi-Copernican revolutions that have occurred in the past five centuries. *Counts as 2 of your 3 books.* **$30.00**

**66570. AN OUTLINE OF STRUCTURAL GEOLOGY.** *Bruce E. Hobbs, Winthrop D. Means and Paul F. Williams.* Up-to-date account of every aspect of structural geology — from the nature of stress and strain to the latest information on plate tectonics. **$16.95**

**56590. INTRODUCTORY GROUP THEORY AND ITS APPLICATION TO MOLECULAR STRUCTURE.** *John R. Ferraro.* Includes derivation of selection rules, potential force fields, and added emphasis on space symmetry and point symmetry. **$19.95**

**41640. DATA COMMUNICATIONS DICTIONARY.** *Charles J. Sippl.* Over 14,500 definitions and explanations of data processing and communications terms, acronyms, concepts and abbreviations. **$19.95**

**55000. THE ILLUSTRATED ENCYCLOPEDIA OF ASTRONOMY AND SPACE.** *Ian Roxburgh.* Covers the full range of astronomical knowledge and space exploration, from ancient times to now. **$16.95**

---

If the reply card has been removed, please write to
**The Library of Science**
**Dept. 2-4AJ, Riverside, N.J. 08075**
to obtain membership information and an application.

# BOOKS

## *Galaxies, left and right, Isambard Kingdom Brunel and hunter-gatherers of the Kalahari*

by Philip Morrison

MAN DISCOVERS THE GALAXIES, by Richard Berendzen, Richard Hart and Daniel Seeley. Science History Publications ($15.95). SECOND REFERENCE CATALOGUE OF BRIGHT GALAXIES, by Gerard de Vaucouleurs, Antoinette de Vaucouleurs and Harold G. Corwin, Jr. University of Texas Press ($50). THE HUBBLE ATLAS OF GALAXIES, compiled by Allan Sandage. Academic Press ($14. 50). U.S. OBSERVATORIES: A DIRECTORY AND TRAVEL GUIDE, by H. T. Kirby-Smith. Van Nostrand Reinhold Company ($11.95). Full of unusual photographs, many of them published for the first time, *Man Discovers the Galaxies* tells in a simple, fresh way the history of how we came to know the universe of galaxies. It is a hybrid between an elementary textbook and an original historical monograph with biographical insets. Some pages are devoted to prologue, but the story is of our century, closing in about 1950. If you seek visual evidence of date and place, Edwin Hubble stands in one photograph next to a slight but romantic hero of the silent movies, Ramón Novarro. In a 1917 photograph the heavy base of the 100-inch telescope teeters precariously: the truck (an antique Mack?) that was hauling it up Mount Wilson had ditched. Here is the very plate from the 100-inch on which Hubble discovered in the great Andromeda spiral "three stars...2 of which were novae, and 1 proved to be a variable later identified as a Cepheid." (So reads the notebook entry for October 5, 1923, in Hubble's own hand.) Here too is the decisive diagram published by Robert Trumpler in *Lick Observatory Bulletin* for 1930, which shows the then best-supported model of our galaxy of stars as a little off-center disk ("the Kapteyn universe") at the edge of a ragged ball of globular clusters.

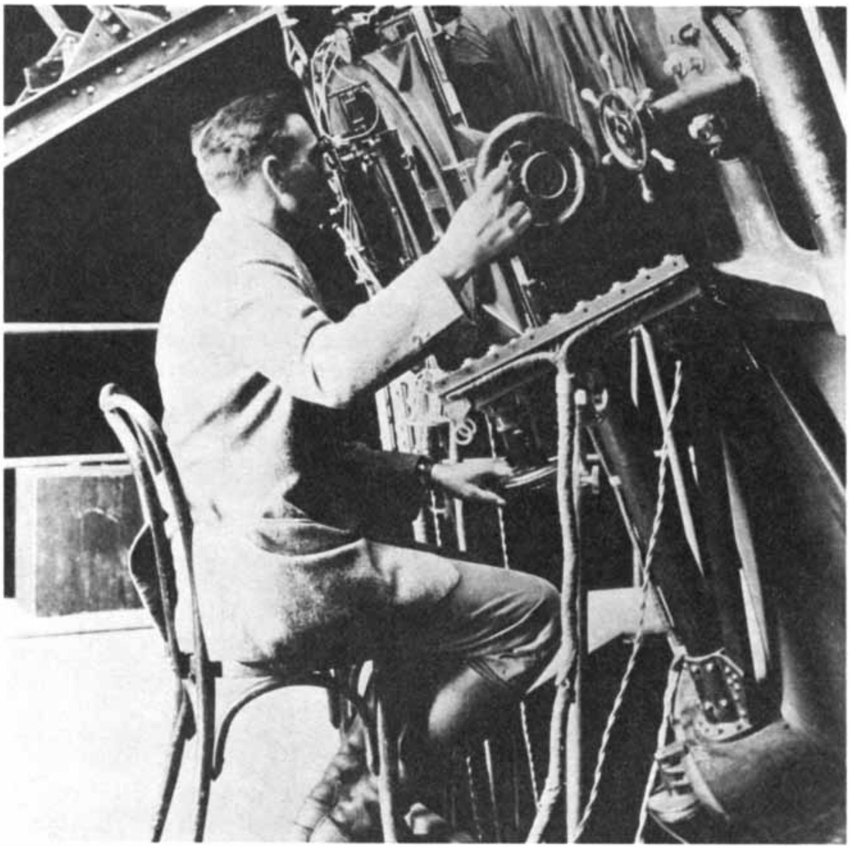
The issue and its outcome are clear. Do we live in one island galaxy, one bound system of stars, adrift on an ocean of space in an archipelago of similar galaxies? We are sure of that today, even if we are still in doubt about the true measures and wilder shores of the ocean. In *Man Discovers the Galaxies* the scholars are struggling confusedly toward the answer. In 1910 the possibility—first glimpsed in the 18th century—seemed unlikely, even bizarre to many, but by 1935 the question was resolved.

It is easy to list the results that were in contention. Jacobus Kapteyn of Leiden, a formidable figure, measured the star distribution in the sky as a function of brightness. From his long counts he could map our galaxy as a smallish disk along the plane of the Milky Way. Our sun lay rather near its center. In 1915 Kapteyn wrote the founder of the Mount Wilson Observatory, his frequent host George Ellery Hale: "One of the startling consequences is, that we have to admit that our solar system must be in or near the center of the universe.... Twenty years ago this would have made me very skeptical.... I have sometimes felt uneasy in my mind about this result, because... scattering of light in space has been neglected." Harlow Shapley, again neglecting space absorption, showed that the sun lay toward the edge of a big disk of globular clusters, a disk 10 times bigger than Kapteyn's galaxy. He thought the globulars avoided the plane of the Milky Way because they could not form there; a similar process must affect the small spirals dotted here and there. Had not a renowned and meticulous observer at Mount Wilson, working with the unmatched 100-inch, shown by repeated measurements that seven spirals were in rapid rotation, turning once in about a million years? Once seen to move so nimbly, the spirals had to be of subgalactic size. The machine with which the measurements were made was still on the mountain in 1972, bearing a typed label (shown here in a photograph made by the senior author) with the admonition "Do not use this Stereocomparator without consulting A. van Maanen." But Adriaan van Maanen had been dead since 1946; his results, first published in 1916 and reinforced in paper after paper over 20 years, had finally been shown to be wrong, even on the basis of van Maanen's own plates, by Hubble in 1935. We do not know exactly why van Maanen erred; the hundredfold error has been demonstrated to reside in some bias within the observations he made at the microscope—a small visual effect in the presence of a strong preconception?

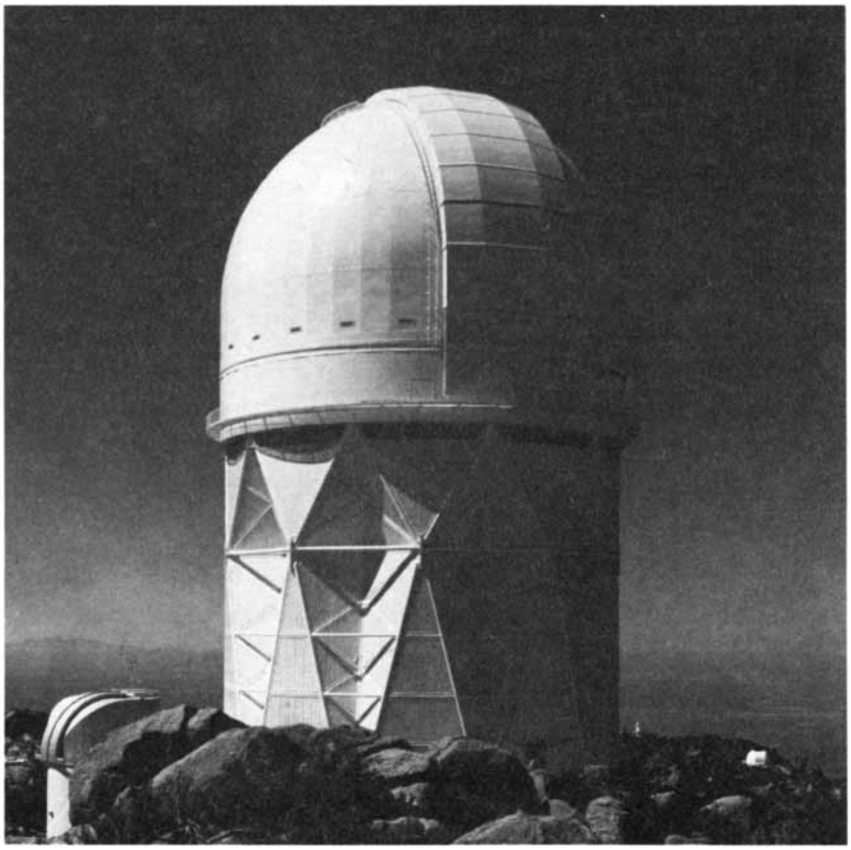
The other side also built its case haltingly. First came the novas in the spiral nebulas. Heber Curtis of Lick made the point. Six novas were observed in one little spiral patch in a few years; about the same number were displayed by our entire galaxy in the same span. Was that patch not itself an entire galaxy? The two bright stars seen in spirals, 100 times brighter than the novas, were to be regarded as a rare new class of objects, not as a confusion. They were to prove to be supernovas, but the ambiguity hurt. Then V. M. Slipher had shown that the spirals were receding at high velocities; by 1914 these, the first red shifts, had convinced no less a theorist than Ejnar Hertzsprung, who wrote Slipher: "With this discovery the great question, if the spirals belong to the system of the Milky Way or not, is answered with great certainty to the end, that they do not." Jan Oort had by 1927 gauged the direction and distance to the center of our galaxy on the basis of the relative motions of nearby stars assumed to be in Keplerian galactic orbits, with more or less the modern result. Trumpler showed in 1930 that open irregular clusters, matched for brightness on the basis of the star colors, seemed to be larger in diameter if they were farther away. Only one reason made sense: the region of the galactic plane was not transparent but dusty. At once this result turned Shapley's big disk of globular clusters—quite distinct in their size and distance from the open clusters—into a rough sphere and provided a simple explanation of their avoidance of the plane of the galaxy. Trumpler himself never dropped Kapteyn's model, even though his own work had doomed it. In the end Shapley's clusters were brought closer in, but the spirals were pushed out of the galaxy forever.

Curtis had been right all along. The famous debate of 1920 between Curtis and Shapley is here an introduction to the entire story. Curtis himself, who had been a college teacher of Latin before he went into astronomy, is not given biographical treatment (although he is engagingly shown in knee breeches and ruffled shirt, fancy dress for a party). He had dropped out of research before the issue was fully joined. The race is not always to the swift. Single results can be misleading; the theory of the color variation of absorption was as bad a guide to reality as the simple view of gravitational orbits was a good one. Perhaps we learn mainly that it takes time and a confirming diversity of argument to draw correct conclusions. One authoritative error in a theory or an experiment can delay any inference. The three historian-authors skimp the physics a bit more than is required for the level of the text.

In *Second Reference Catalogue of Bright Galaxies* dry pages of computer printout collect what we know today of galaxies. No book for the general reader, it will be



*Edwin P. Hubble at the camera of the 100-inch reflector, from* **Man Discovers the Galaxies**



*Dome of the 158-inch reflector on Kitt Peak, from* **U.S. Observatories**

invaluable to a pro, with its carefully prepared data on about 4,400 galaxies up to 1975. The list is not complete, of course. Galaxies come in a wide range of sizes; dwarf faint galaxies cannot be seen very far, even through the highly transparent space between the groups of galaxies. Out to a few tens of millions of light-years—looking back in time to the Miocene, say—the list probably contains roughly half of all the galaxies, excluding the dwarf galaxies with fewer than 100 million stars (our own holds 1,000 times that number). Beyond our near neighborhood the familiarity falls off rapidly. The list includes most galaxies for which at least one piece of significant informatiln is known. It includes all such galaxies that are not fainter, smaller or at a higher red shift than certain stipulated limits; the red-shift limit is a twentieth of the speed of light. Within these limits the attempt is to include all those unusual objects to which attention has recently been drawn: radio sources, ultraviolet sources, compact galaxies, galaxies that are host to a supernova, interacting galaxies and galaxies with bright lines. By the end of its reach, back in time in the Precambrian, half a billion or a billion light-years out, the catalogue can list only one in 10,000 of the galaxies we believe are there.

The list includes 2,700 galaxies with known red shifts. Add to those given here all the distant, faint but remarkable galaxies that have been studied because they were radio sources or were bright members of very distant clusters, include as well all the known quasars, and the total list of known red shifts would reach 3,000 or somewhat more. The references (note that the first version of this catalogue, issued in 1964, is relied on for all earlier references) make it plain that all the red-shift work, the foundation of cosmology today, is the contribution of only 50 or 60 investigators over the decades. They alone are our expeditionaries toward creation.

The main list covers some 170 pages; the notes on individual entries, including all references for the newer data, make up 80 more. The volume includes a good deal of apparatus, careful statements of the elaborate data treatment, lists of alternative names, galaxies with special features such as supernovas, and the like. Radio enters here as full partner; the radio power and spectra are cited in rather telegraphic form, and the 21-centimeter radio line results as well, now including many red shifts. The radio and optical velocities are "in excellent systematic agreement at all velocities," a result that would have delighted the founders of the subject. About 400 galaxies now have radio red shifts but not optical ones.

Not much is offered the eye here; only the mind is catered to. A few projections of the sky are shown dotted with galaxies in various categories; north and south are unevenly known and unevenly populous, at least close by. There is a list of compilations of galaxy photographs; some 800 photographs are identified at 18 sources, both accessible and esoteric. Many more photographs are cited in the list for individual galaxies. "This onerous undertaking" has cost the authors about 15,000 hours, and their students, keypunch operator and typists spent another 5,000. It will last; its authors are known for their care and judgment, and very few flaws can be spotted in the work. One misses a listing of galaxies by groups or clusters; the only mention of such associations is found in the notes for individual galaxies.

Sixteen years ago there appeared what is still the best collection of photographs of galaxies. It was lovingly compiled and dedicated to Hubble by his distinguished continuator at Mount Wilson and Palomar, Allan Sandage. It offers the peruser halftones of about 200 galaxies, all made with instruments of the Hale Observatories; many of the pictures spread their splendor over a full outsized page. The organizing principle is Hubble's scheme of classification of galaxies by their external form. That scheme is still standard—although it is much complemented by later workers, who draw more attention to overall luminosity, spectra and the relative importance of the galaxies' central regions, the nuclear domain whose concentration is frequently suppressed by the same photographic techniques that alone can show the delicate periphery.

There be monsters! Taxonomy and details apart, few reflective readers can look at these strange bright spindles and disks against the starry black sky without an overwhelming sense of awe and wonder. There are by now many books and journals that present some galaxies in color, but there is still no rival to the richness of *The Hubble Atlas of Galaxies* in its crisp elegance. Subsidized by the Carnegie Institution of Washington, it remains a bargain at preinflationary prices; still in print, it is any skyward library's best buy and a dreamer's delight.

It is a remarkable detail in the unfolding of modern cosmology that the theater of its main empirical advance was the U.S., and in particular the clear Southwest. Long before Los Angeles became a center of electromechanical industry and while Arizona still had few wealthy residents, astronomers flocked there from Uppsala and Leiden, from Göttingen and Basel. The dry mountain air was a natural resource, to be sure, but it had to be exploited by men who were ambitious for their science and who could persuade street-railway magnates and philanthropic foundations to put up the cash. The theater of this campaign is open to tourists, and a desirable journey it offers to amateur astronomers or to anyone with an interest in the look and feel of a historic explorers' base camp—an outpost where they are still laying plans for ever more daring routes to the summit.

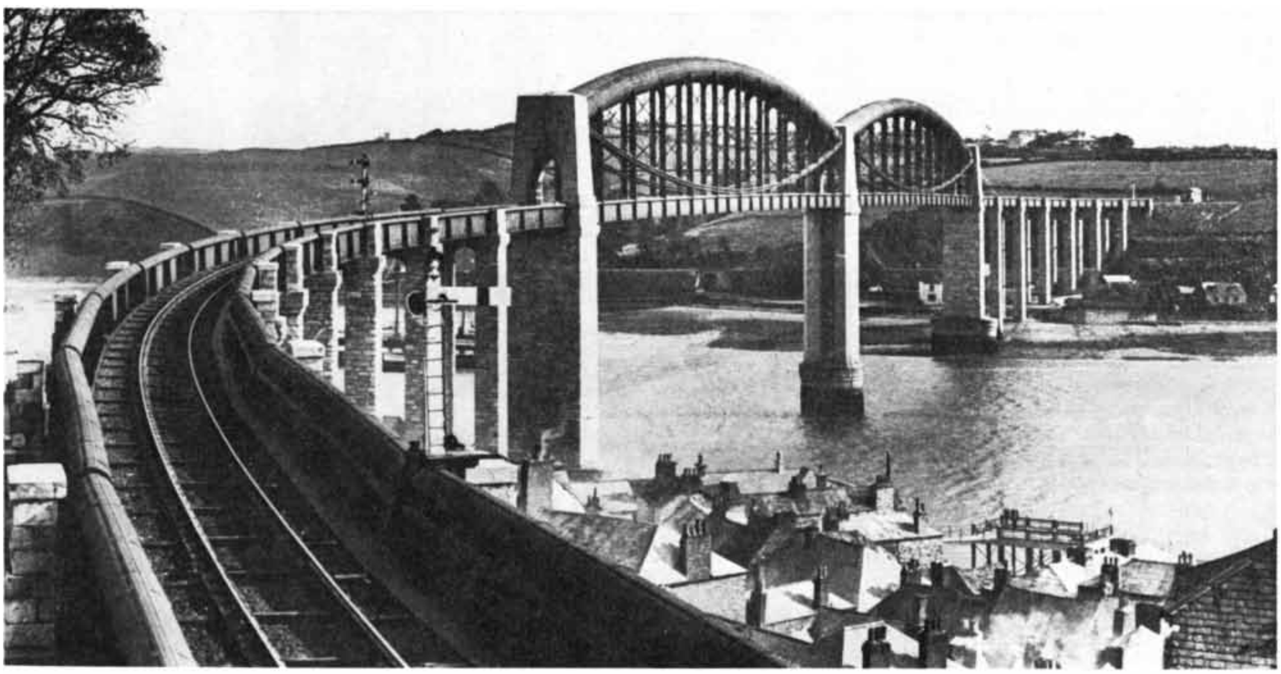
H. T. Kirby-Smith, professor of English at the University of North Carolina at Greensboro, is a devoted and knowing teacher and amateur of astronomy. He has gone to a number of the principal observatories across the land to see what he could see, and he has prepared a compendium of lesser but still exciting places to visit—one is surely near you—with excellent photographs and with comment that is practical, evocative and penetrating, befitting the poet he is. On the 100-inch at Mount Wilson: "The mounting of the telescope has a desperate look to it; all girders and rivets, and so ready to compromise with the requirements of rigidity that a large circle around the celestial pole is inaccessible." At Palomar he complains about the dim light and the glassed-in visitors' booths, which are required to keep the heat input low but which ensure that "the view of the giant telescopes is like that enjoyed by those who saw the Titanic founder: impressive but dim." He suggests another stance. He also gives valuable hints on motels and side trips—and on the problem of children running in circles along the reverberating circular walls! Fifteen centers are described in detail, mostly from on-the-spot experience, with a good serving of history and comment. Then there are 80 pages of briefer comments, state by state, based on a careful survey by mail of many possible places to view telescopes, astronomers and the sky. "It is a pleasure, though, to have discovered that many great research centers... encourage the interested and intelligent visitor and suffer patiently the ignorant and the inane. Good scientists do not forget that a sense of mystery, or just a healthy curiosity, animates the most valuable endeavors." The Very Large Array on New Mexico's Plains of San Augustin spans "perhaps the most beautiful remaining 'wide open space' in the country." He describes the mighty radio hammock at Arecibo in Puerto Rico, although he did not get there. That instrument, amidst its domed emerald hills, offers one of the most striking architectural views and settings in the hemisphere. This book is a model for informed travelers; one hopes that many another such guide to the local habitation of scientific interest will appear.

ANIMAL ASYMMETRY, by A. C. Neville. Edward Arnold, distributed in the U.S. by Crane, Russak & Company, Inc. ($3.75). THE PSYCHOLOGY OF LEFT AND RIGHT, by Michael C. Corballis and Ivan L. Beale. Lawrance Erlbaum Associates, Publishers, distributed by John Wiley & Sons ($14.95). The little crabs called fiddlers (actually they seem rather to carry a cello) are grotesquely asymmetrical. The big claw of the male accounts for about two-thirds of the animal's total weight, outweighing the smaller claw by a factor of 20. This handedness is unbiased: it is 50–50 for a fiddler male to be right- or left-fiddled. Chance determines it. The crab is adapted to let a claw go in times of need. Young males start with two small male-type claws, and if one is lost, it regenerates as a small female claw as the other claw grows, molt after molt, to be brandished proudly in maturity; if neither claw is lost, the crab will sport a pair of fiddles.

The wry-billed plover turns stones for a living on the shingle-bedded rivers of New Zealand. Its long beak is bent to the right side from the middle; it thus feeds on the bias, politely from the left. Owl ears tilt out of the plane, in one species the right ear up and the left one down, to help the birds' elegant acoustic scanning achieve three-dimensional performance. Muscles can be one-sided in insects or in man; a blindfolded person walks, swims and even drives in a progressing set of circles, but the direction is uncorrelated with handedness or with the stronger side of the body.

Structural and behavioral asymmetry is the theme of the little book by A. C. Neville, a zoologist at the University of Bristol, who covers a remarkable variety of animal forms (but no plants). He begins with molecular isomers and he gives a paragraph to bacterial flagella, whose left-handed helical fibers must rotate clockwise (viewed from the body) to prevent a terrible tangle. One of a series aimed at British undergraduates, this well-illustrated review is brief and lively, although it overuses the zoologist's technical volcabulary. Almost every page offers some small astonishment; what a lot is somehow built on the enantiomorphic bias of biological monomers!

*The Psychology of Left and Right* is no such directory of simple results. It is the work of two reflective experimental psychologists, a Canadian and a New Zealander. Their collaboration on mirror-image discrimination in pigeons set these hardheaded experimenters among



***The Royal Albert Bridge, from* The Works of Isambard Kingdom Brunel: An Engineering Appreciation**

philosophers and physicists in an effort to understand behavior in the looking glass.

Our cultural practices are plainly handed. The comparison of flipped pictures of a tree in the park with those of a busy street make the point, perhaps too sharply. Nature is normal enough in both views, whereas no adult could mistake the flipped gasoline sign. A Japanese street scene would be less obvious for most readers, however, and it is evident that nature is handed too, like culture, once one knows and examines it in specifics. A couple of chapters of careful logic establish tests for the discrimination of mirror-image stimuli, with responses that must not themselves be enantiomorphs, and of left-right response, demanding asymmetrical response to symmetrical cues. Not every experimenter has grasped the subtleties here; it is all too easy to feed small cues unknowingly to gain the one bit of a signal that discrimination requires.

Experimental results are here in plenty, all critically assessed. The subject is not structure, of course, but behavior; although animals claim some attention, people are the principal subjects, and as usual they are difficult to pin down. The octopus has a hard time with mirror-image discrimination, and at least one experimenter found it impossible to make dogs lift their right foreleg at the sound of a metronome and their left foreleg when they heard a buzzing sound coming from the same place. That same sound discrimination was easy for the dogs when left and right were not at issue; the left-right response could be learned when the two sounds were spatially separated. In children the task of mirror-image discrimination appears to be one that matures slowly between the ages of four and 10, with a real improvement at the age of school entry. Handedness, eye preferences and even ear preferences also develop in childhood, with credit going both to internal processes and to several kinds of learning experience. We mostly acquire a left-right sense, in a way.

All of this has its applied side. For a long time it has been argued that handedness lies deep below the special disabilities of reading that are such a trial to many children and their teachers. The circumstances are complex and the case is not proved, but these authors do see that the specific disability, unaccompanied by evidence of brain damage or plain disillusionment with school, might be the consequence of an unusual internal symmetry. Perhaps one child in 40 might fit that diagnosis. "If a child has a specific reading problem, but is otherwise intelligent and aware, the prognosis may be excellent; he may be another Leonardo da Vinci."

Mirror writing is a classical topic that plunges us immediately into the depths of the enigmatic architecture of the brain. It is true that mirror writing is often associated with damage to the left cerebral cortex. It is tempting to see in this the presence of a transformation that builds weakly in the right hemisphere a mirrored map of what is stored directly in the left hemisphere. When the right hemisphere is then called on to perform, or when it is in unusually active partnership, as it is for a Leonardo, it produces mirror writing. The entire thrust of the careful remarks of these authors, however, is to minimize the old easy judgments of the dominance of a single half of the brain, still more the glib new assignment of a particular cognitive style to one or another half of that subtle walnut. Some kind of mirrored transfer (which does not imply topographic mapping, only logical) is here preferred. But the brain and the mind are not to be summed up in a simple static image.

Finally, whence the human asymmetry? One can go back not to the failure of parity or even to molecules but to the first cellular divisions of the fertilized egg. The handed quality appears so early there that it can hardly be encoded genetically in the usual way; it may be stored in some cytoplasmic form. Assertions that monozygotic, identical twins have hair whorls rotating in opposite directions appear to be false. Handedness is not paired in identical twins either, although joined Siamese twins do usually show reversal in the position of the heart and other organs. Handedness is not to be teased easily out of the double helix!

Three thousand years ago the men of the tribe of Benjamin assembled, nearly 27,000 of them. "Among all this people were seven hundred chosen men lefthanded; every one could sling stones at an hair breadth and not miss." The handedness ratio is about the same today, and the word "dexterity" is no more apt than it was in the days of the Judges.

THE WORKS OF ISAMBARD KINGDOM BRUNEL: AN ENGINEERING APPRECIATION, edited by Sir Alfred Pugsley. Institution of Civil Engineers, 26–34

*A Kalahari hunter tests a hollow tree for water, from* **Kalahari Hunter-Gatherers**

Old Street, London ECIV 9AD ($18). Robert Stephenson and Isambard Kingdom Brunel were great contemporaries, friends if rivals, engineer sons of engineer fathers of even higher originality. It is Brunel the younger whose work is here assessed by a group of British engineers in a volume that looks at the man only through his manifold projects. In early Victorian times this man spanned a particular epoch. Before him engineers in Britain were, like the Stephensons father and son, singularly gifted workmen who made their way from mine shaft or coal-loading tramway. After Brunel engineers began to specialize, and few others could leave a legacy of tunnels, bridges, railways, ships, signals and engine design: path-breaking construction in stone, timber and wrought iron. Brunel and his fellows created a profession and a social group as much as a set of civil works; they became gentlemen, clubbable, cultivated, accepted by the arbiters of social class.

Brunel was no modern; he was through and through an eminent early Victorian. He was devoted to the idea of laissez-faire, spurning the honors offered by the state and even its patent protection. Of course he and his apprentices and staff were gentlemen without leisure. He wrote in a reference letter in 1849: "He is one of those who gets *up late,* go to their work at *gentlemanly hours*—and from whom it is difficult to get any real *work.*" He was meticulous, even picky, about expense accounts. He spent all his energies in engineering, and "found little time for recreation, holidays or spiritual activities." A stroke took him at 53 in the midst of his work, overworked and overstrained. He had a compelling dying urge to see his greatest bridge and "this he did, lying on a couch on a specially prepared wagon of the Cornwall Railway, which was drawn slowly over the spans of his last great masterpiece, the Royal Albert Bridge," across the River Tamar above Plymouth. That unadorned and beautiful bridge, lovingly and expertly maintained, serves perfectly today under heavy loads, its main members unchanged these 120 years. On the portal stands his "best valediction," I. K. BRUNEL: ENGINEER.

He was an audacious but painstaking designer and constructor. At 33 he came to Bristol (the city and university hold him high today) to build the Great Western Railway as a system, the first such railway. He designed the way stations and invented the signal scheme; he chose a wide gauge—nearly seven feet; he prepared the bridges and the right-of-way and found time to sketch a few ornamental lampposts in his exquisite style. The wide gauge was not a success and did not last, but his choice was reasoned: he wanted the load placed between the wheels for safety at high speeds, for he was an early proponent of speed in travel. By 1840 he had become somewhat disillusioned with the early locomotives, and he put into practice a bizarre "atmospheric system of traction," which powered the train by a piston working in an evacuated pipe, with the slot for the piston arm sealed before and after its passage by a continuous hinged leather flap. The scheme lasted only a year on the steepest grades of the new South Devon Railway. The assessment here looks away from the technical failure of the seal (the leather "sometimes froze, rats gnawed it, and the sea air completed the destruction") to the inflexibility of such a traction scheme, which made junctions tricky and required auxiliary traction for making up trains and switching, even if it could have been perfected technically.

Brunel built three famous iron ships, each one a pioneer. The first was the largest steam vessel launched up to that time and the first to demonstrate the feasibility of operating a transatlantic liner under steam. The second was a structural marvel that is still afloat as an exhibit and is without serious structural trouble after a century of indignities. It was Brunel who put into this hull the main requirement of large ships: longitudinal beamlike strength to resist bending under the weight and buoyancy forces in large waves. Smaller oceangoing vessels need mainly lateral strength, to resist local pressures of the sea. Wooden ships mostly lack such lengthwise resistance.

Brunel had predecessors, but he was the first to work in iron and on a large scale. His biggest ship, the much-described *Great Eastern,* scaled up his previous vessels by an astonishing factor of nearly eight in displacement. She was a structural success at sea but an economic disaster that proved to be useful only for laying cable. The calculations that led Brunel to fix the size of the great ship are cited in full. He argued that the fuel requirement increased only as the area of the hull but that the space available for coal could increase with the volume. He planned the ship for a round-the-world route, around both Capes, with perhaps one refueling stop, but there was no demand for such capacity. While she was being built he asked a visiting expert friend: "If she belonged to you, in what trade would you place her?" The reply was forthright: "Send her to Brighton, dig out a hole in the beach and bed her stern in it.... She would make a substantial pier... her hold magnificent salt water baths and her 'tween decks a grand hotel.... I do not know any other trade, at present, in which she would be likely to pay so well."

Even in that day of headlong expansion of industrial capital the dreams of engineers were not always sound. Brunel was a great pioneer who was concerned with systems concepts, skilled at mathematical design and demanding about careful tests of materials and methods. Error is easier to see looking backward, as indeed it remains today.

The book has excellent period drawings and photographs. One shows the raising of the last span of the Royal Albert Bridge, huge iron tubes bent into a bow and held bent by great chains. Those bony bridges have a remarkable unified beauty our framed steel structures lack. The rolling mill today makes economical the webbed members Brunel and his contemporaries could not get. But Brunel's big tubes (10 feet across and 450 feet long), so laboriously riveted together out of half-inch iron boiler plates only two feet by 10, catch the eye and the mind in a lost organic metaphor.

KALAHARI HUNTER-GATHERERS: STUDIES OF THE !KUNG SAN AND THEIR NEIGHBORS, edited by Richard B. Lee and Irven DeVore. Harvard University Press ($18.50). It is close to 15 years since the two men who brought this collection of related papers into a beautifully made and illustrated (if rather technical) volume carried out their own first field studies among the Northern Kalahari Bushmen called the !Kung. Around their work and that of their predecessors the Marshalls there has nucleated a diverse group of scientists who have extended the early ethnographic and ecological inquiries into a wide range of investigations: demographic, medical, genetic, psychological, folkloric and archaeological. The editors have brought a certain unity to this set of studies, which are beginning to exhibit the gleaming facets of an ancient way of life.

"Man the hunter and woman the gatherer" described these people until recently, as it has for a very long time. Deep in the past such societies amounted to a substantial sample of humankind, living close to our origins. One may estimate that a twentieth of human experience is held in the history of these handsome, gentle and loquacious people. Now about 95 percent of them are entering some degree of integration with the pastoral and agricultural economy of Botswana.

Certainly their way was remarkably adaptive to the specific nature of their world. A Japanese investigator has gone 400 kilometers south of the water holes of the !Kung at Dobe and has found people with a different language but with a mode of subsistence not unlike that of the !Kung. These isolated /Gwi have worked out life without the grand staple of the !Kung: the mongongo tree. Their domain is too dry for it, and they depend instead on 13 plant species over the different times of the year. There are no permanent water holes; they have access to standing water only in the days after heavy rain during the rainy season, a couple of months near the turn of the year. And so they depend for water mainly on the juicy tsama melon, which is like a small watermelon and which they eat except for the peel. In the driest months the trees and grasses wither and "every part of the Kalahari turns brown and barren." The blood and stomach contents of animals might provide water, but game is in short supply. The food and water of the /Gwi are won in those months chiefly by the digging stick. Hard work retrieves from the cool subsoil two species of tubers that provide food and a bitter tuber that yields water. The clever baboons that dwell in most of the desert cannot live hereabouts, because they are not masters of the digging stick; only the much cleverer human beings can survive. In contrast to the rest of !Kung land, the scene 100 kilometers north from Dobe "looked like paradise ... thick forests with great large trees, and tall lush grass." But the !Kung find it strange—the mongongo nut is absent.

Life is changing for the !Kung too. One poignant photograph shows a band of them moving under packs, spears at the port, along the border road paralleling the long jeep-patrolled wire fence now separating Botswana and Namibia. The people are squatters now on their own land; they try to raise goats and donkeys and even some horses and cows. A large minority live and work as laborers on the big farms nearby. They still tell of old Karã/'túma, their bumbling mythical forerunner: his bow shot the first cow, he spurned her milk and he set fire to the field of sorghum grain; therefore are they poor today. For a while it will be possible to use their living sites as models for the campsites of hunters of the Paleolithic, to examine their vital statistics (tabulated by Nancy Howell in a summary paper that ingeniously constructs quantitative life tables among a people who use a calendar based on memorable events), to map their annual wanderings or to find the mean spacing of the birthplaces of parents and of parent and child. For a while we can record their expert ethnoethology, inductive and empirical. Not for long.

We change too. These anthropologists are no longer detached witnesses. They are organizing to help the people we ought now to call San (no longer Bushmen, as they were dubbed in derogation by the Dutch settlers who waged exterminating war on them until 1850). Women now study women and (perhaps less happily) healing trance states are studied by a man who seems himself rather too entranced. The San may yet preserve their antique and supple roots. Hear the woman Kun/obe speak: "I refuse this guy [Karã/'túma] I do.... People should cry out for themselves.... People should protest. We who are Zhú/twãsi (San), let us cry out, so we will be lifted up. Unless we do, we are just going to ruin."

# BIBLIOGRAPHY

*Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.*


## THE IMPORTATION OF LIQUEFIED NATURAL GAS

SAFETY WITH CRYOGENIC FLUIDS. Michael G. Zabetakis. Plenum Press, Inc., 1967.

LIQUID NATURAL GAS. Noel de Nevers in *Scientific American,* Vol. 217, No. 4; October, 1967.

LNG INFORMATION BOOK, 1973. Prepared by the LNG Information Book Task Group of the Liquefied Natural Gas Committee. American Gas Association, 1973.

LIQUEFIED NATURAL GAS. W. L. Lom. John Wiley & Sons, 1974.

TECHNOLOGY AND CURRENT PRACTICES FOR PROCESSING, TRANSFERRING, AND STORING LIQUEFIED NATURAL GAS. Arthur D. Little, Inc., for the Office of Pipeline Safety, 1974.

LIQUEFIED NATURAL GASES: VIEWS AND PRACTICES, POLICY AND SAFETY. U.S. Coast Guard, 1976.

## THE COLLISION BETWEEN INDIA AND EURASIA

GEOLOGY OF THE HIMALAYAS. Augusto Gansser. Interscience Publishers, 1964.

TIBETAN, VARISCAN, AND PRECAMBRIAN BASEMENT REACTIVATION: PRODUCTS OF CONTINENTAL COLLISION. John F. Dewey and Kevin C. A. Burke in *The Journal of Geology,* Vol. 81, No. 6, pages 683–692; November, 1973.

CENOZOIC TECTONICS OF ASIA: EFFECTS OF A CONTINENTAL COLLISION. Peter Molnar and Paul Tapponnier in *Science,* Vol. 189, No. 4201, pages 419–426; August 8, 1975.

THE SUBDUCTION OF THE LITHOSPHERE. M. Nafi Toksoz in *Scientific American,* Vol. 233, No. 5; November, 1975.

SUR LE MÉCANISME DE FORMATION DE LA SCHISTOSITÉ DANS L'HIMALAYA. Maurice Mattauer in *Earth and Planetary Science Letters,* Vol. 28, No. 2, pages 144–154; December, 1975.

SLIP-LINE FIELD THEORY AND LARGE-SCALE CONTINENTAL TECTONICS. Paul Tapponnier and Peter Molnar in *Nature,* Vol. 264, No. 5584, pages 319–324; November 25, 1976.

## THE STATUS OF INTERFERON

INTERFERON. Alick Isaacs in *Scientific American,* Vol. 204, No. 5; May, 1961.

THE INDUCTION OF INTERFERON. Maurice R. Hilleman and Alfred A. Tytell in *Scientific American,* Vol. 225, No. 1; July, 1971.

INTERFERON. D. C. Burke in *Defence and Recognitions,* edited by R. R. Porter. University Park Press, 1973.

INTERFERONS AND INTERFERON INDUCERS. Edited by N. B. Finter. American Elsevier Publishing Co., Inc., 1973.

SELECTIVE INHIBITORS OF VIRAL FUNCTIONS. Edited by William A. Carter. CRC Press, 1973.

THE MECHANISM OF ACTION OF INTERFERON. David H. Metz in *Cell,* Vol. 6, No. 4, pages 429–439; December, 1975.

## ALGORITHMS

ANCIENT BABYLONIAN ALGORITHMS. Donald E. Knuth in *Communications of the ACM,* Vol. 15, No. 7, pages 671–677; July, 1972.

THE ART OF COMPUTER PROGRAMMING: VOLS. I–III. Donald E. Knuth. Addison-Wesley Publishing Company, 1973.

COMPUTER SCIENCE AND ITS RELATION TO MATHEMATICS. Donald E. Knuth in *The American Mathematical Monthly,* Vol. 81, No. 4, pages 323–343; April, 1974.

ORDERED HASH TABLES. O. Amble and D. E. Knuth in *The Computer Journal,* Vol. 17, pages 135–142; May, 1974.

BINARY SEARCH TREES AND FILE ORGANIZATION. J. Nievergelt in *Computing Surveys,* Vol. 6, No. 3, pages 195–207; September, 1974.

THE COMPLEXITY OF SEARCHING AN ORDERED RANDOM TABLE. Andrew C. Yao and F. Frances Yao in *Symposium on Foundations of Computer Science.* IEEE Computer Society's Technical Committee, IEEE Computer Society, 1976.

## BIOCRYSTALS

STUDIES ON THE CELLULAR BASIS OF MORPHOGENESIS OF THE SEA URCHIN EMBRYO: DEVELOPMENT OF THE SKELETAL PATTERN. L. Wolpert and T. Gustafson in *Experimental Cell Research,* Vol. 25, No. 2, pages 311–325; November, 1961.

SKELETON FORMATION OF SEA URCHIN LARVAE, IV: CORRELATION IN SHAPE OF SPICULE AND MATRIX. Kayo Okazaki in *Embryologia,* Vol. 17, No. 1, pages 21–38; January, 1962.

EXPERIMENTAL EMBRYOLOGY OF ECHINODERMS. Sven Hörstadius. Clarendon Press, 1973.

SPICULE FORMATION BY ISOLATED MICROMERES OF THE SEA URCHIN EMBRYO. Kayo Okazaki in *American Zoologist,* Vol. 15, No. 3, pages 567–581; Summer, 1975.

## THE COMPANIONS OF SUNLIKE STARS

PROBLEMS OF DOUBLE-STAR ASTRONOMY. G. P. Kuiper in *Publications of the Astronomical Society of the Pacific,* Vol. 47, No. 275, pages 15–42, February, 1935; Vol. 47, No. 277, pages 121–150, June, 1935.

DOUBLE STARS. Peter van de Kamp in *Publications of the Astronomical Society of the Pacific,* Vol. 73, No. 435, pages 389–409; December, 1961.

INTELLIGENT LIFE IN THE UNIVERSE. I. S. Shklovskii and Carl Sagan. Holden-Day, Inc., 1966.

BINARY AND MULTIPLE SYSTEMS OF STARS. Alan H. Batten. Pergamon Press, 1973.

## THE SPREAD OF THE BANTU LANGUAGE

ZUR GENETISCHEN GLIEDERUNG DER BANTU-SPRACHEN. Bernd Heine in *Afrika und Übersee,* Vol. 56, No. 4, pages 164–185; April, 1972.

THE CHRONOLOGY OF THE IRON AGE IN BANTU AFRICA. D. W. Phillipson in *The Journal of African History,* Vol. 16, No. 3, pages 321–342; 1975.

THE PREHISTORICAL IMPLICATIONS OF GUTHRIE'S COMPARATIVE BANTU, PART I: PROBLEMS OF INTERNAL RELATIONSHIP. David Dalby in *The Journal of African History,* Vol. 16, No. 4, pages 481–501; 1975.

LINGUISTIC EVIDENCE AND ITS CORRELATION WITH ARCHAEOLOGY. Christopher Ehret in *World Archaeology,* Vol. 8, No. 1, pages 5–18; June, 1976.

THE LATER PREHISTORY OF EASTERN AND SOUTHERN AFRICA. D. W. Phillipson. Heinemann Educational Books, 1977.

## THE THEORY OF THE RAINBOW

LIGHT SCATTERING BY SMALL PARTICLES. H. C. Van de Hulst. John Wiley & Sons, 1957.

LIGHT AND COLOUR IN THE OPEN AIR. M. Minnaert, translated by H. M. Kremer-Priest and revised by K. E. Brian Jay. G. Bell and Sons Ltd, 1959.

THE RAINBOW: FROM MYTH TO MATHEMATICS. Carl B. Boyer. Thomas Yoseloff, 1959.

INTRODUCTION TO METEOROLOGICAL OPTICS. R. A. R. Tricker. American Elsevier Publishing Co. Inc., 1970.

THEORY OF THE RAINBOW. V. Khare and H. M. Nussenzveig in *Physical Review Letters,* Vol. 33, No. 16, pages 976–980; October 14, 1974.

# Lancia

The Lancia HPE Estate Wagon. As elegant as it is versatile. With refinements that reflect Lancia's 70-year racing and car building heritage. A test drive will confirm its high standards of engineering, performance, styling, comfort. Lancia. The intelligent alternative to the average and the overpriced automobile.


LANCIA

**The intelligent alternative.**

Lancia Scorpion, HPE Estate Wagon, Coupe, Sedan. Now at your nearest Lancia dealer. Call toll free (800) 447-4700 or in Illinois (800) 322-4400.

# DEWAR'S® PROFILES

*(Pronounced Do-ers "White Label")*

BLENDED SCOTCH WHISKY • 86.8 PROOF • ©SCHENLEY IMPORTS CO., N. Y., N. Y.

# SHARON MILLER

HOME: Salmon, Idaho

AGE: 36

PROFESSION: White-water guide

HOBBIES: Skiing, kayaking, horseback riding, yoga.

MOST MEMORABLE BOOK: "Thomas Wolfe's Letters to His Mother"

LAST ACCOMPLISHMENT: Became a certified ski instructor.

QUOTE: "The challenge of white water presents ever-changing situations that demand instant, precise decisions. The implications reach far into everyday life. It's a very literal case of 'he who hesitates is lost'."

PROFILE: Vibrant in her love for life and its intrigue. She exudes this enthusiasm in her work, giving others a special awareness and a true appreciation of nature.

SCOTCH: Dewar's "White Label®"


*Ask for Dewar's or "White Label." Either way, you'll get the same great Scotch. Dewar's never varies.*