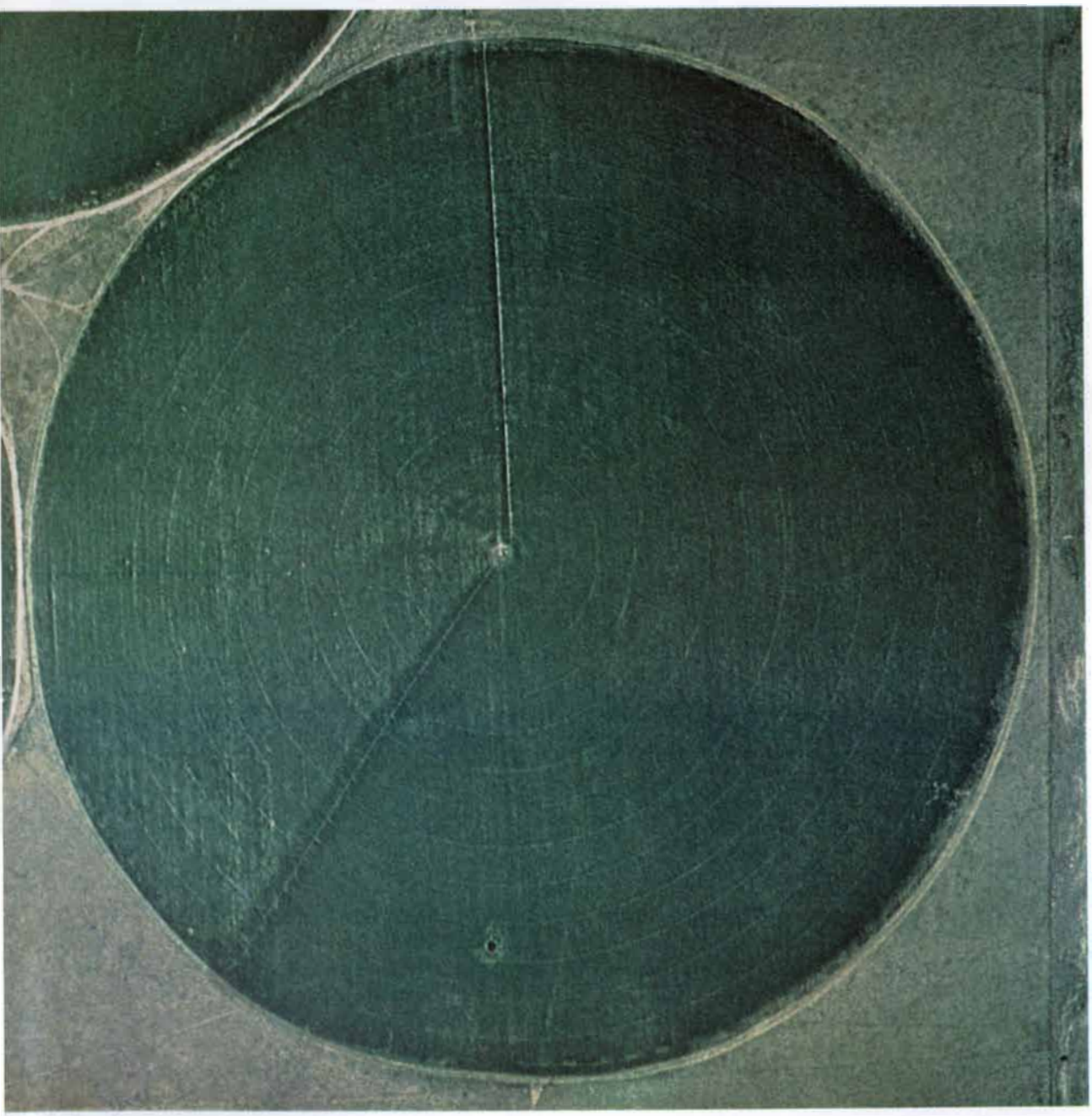
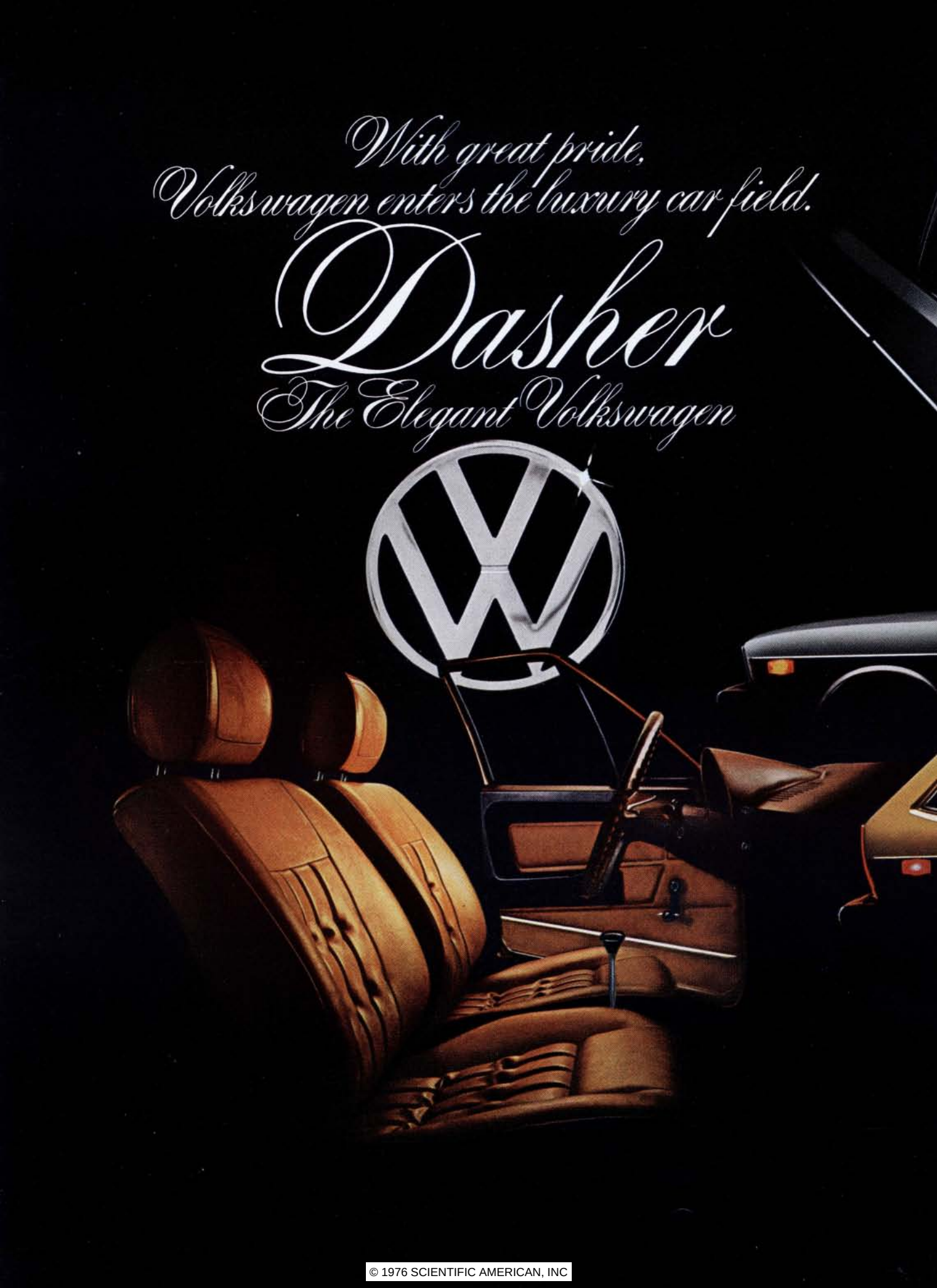
SCIENTIFIC AMERICAN

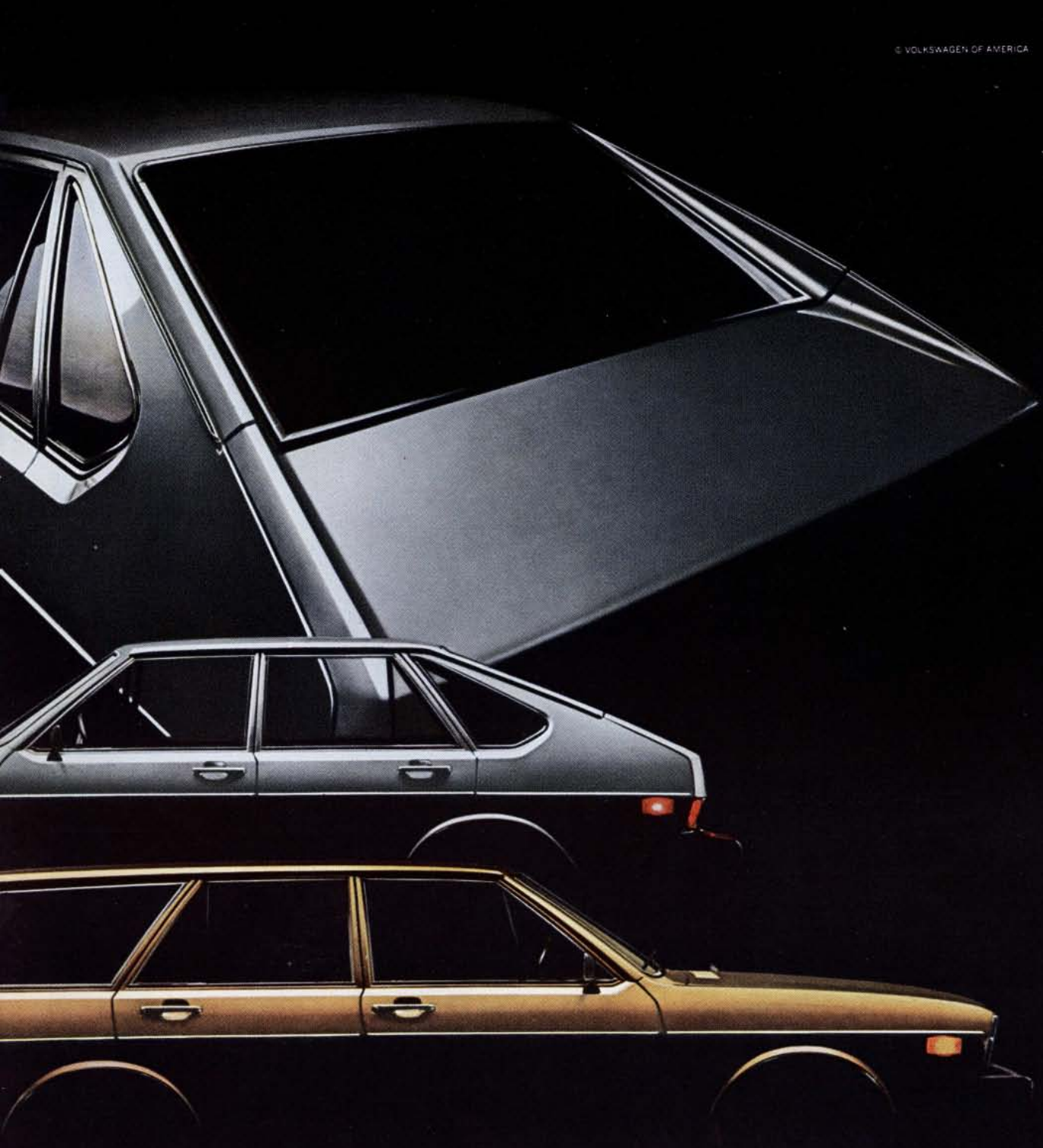


CENTER-PIVOT IRRIGATION

$1.25

June 1976





The elegant Volkswagen. New for us, but not strange for us. All Volkswagens have been elegant in their simplicity. Elegant in design. In concept. In function. Dasher is all of these. With the added elegance of timeless styling. A striking interior. Rich appointments. Dasher is a cultivated car. In size. (Extravagant inside, conservative outside.) In performance. (0-50 mph in 8.0 seconds.) In economy: 37 mpg highway, 24 mpg city. (1976 EPA estimates with standard transmission. Actual mileage may vary with your type of driving, driving habits, car's condition and optional equipment.) In serviceability. In features like steel-belted radial tires, fuel injection and front-wheel drive. You may drive the graceful sedan or the gracious station wagon. We offer these cars with great pride, to be owned with great pride. Volkswagens before Dasher have been elegantly simple. Dasher is simply elegant.

DEWAR'S® PROFILES

(Pronounced Do-ers "White Label")

BLENDED SCOTCH WHISKY • 86.8 PROOF • ©SCHENLEY IMPORTS CO., N. Y., N. Y.

PAUL SZEP

HOME: Scituate, Massachusetts

AGE: 33

PROFESSION: Nationally syndicated editorial cartoonist

HOBBIES: Lithography, golf, classical music

MOST MEMORABLE BOOK: Low's Autobiography

LAST ACCOMPLISHMENT: Won the 1974 Pulitzer Prize for editorial cartoons.

QUOTE: "The satirist is an extremely important person in a society that takes itself far too seriously and is incapable of laughing at its own frailties and insignificance."

PROFILE: Relaxed and easygoing, but very sensitive to social injustice. In his work, he reduces complex events to what concerns him most, their effect on ordinary people.

SCOTCH: Dewar's "White Label"®

Authentic. There are more than a thousand ways to blend whiskies in Scotland, but few are authentic enough for Dewar's "White Label." The quality standards established in 1846 have never varied. Into each drop go only the finest whiskies from the Highlands, the Lowlands, the Hebrides.

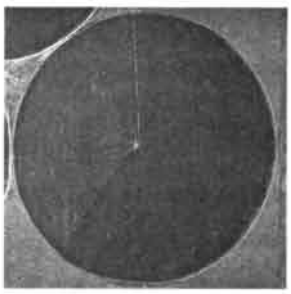
Dewar's never varies.

Established 1845 **SCIENTIFIC AMERICAN** June 1976 Volume 234 Number 6

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $15 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $18 PER YEAR, ALL OTHER COUNTRIES.



THE COVER

The photograph on the cover shows a circular cornfield half a mile in diameter near the town of Sterling in northeastern Colorado. The photograph was made from an altitude of 6,500 feet above the center point of the field, where a well draws water from an underground aquifer of the neighboring South Platte River. Water from the well is pumped directly into a quarter-mile pipe extending radially to the perimeter of the circle. Sprinklers mounted along the pipe apply water to the cornfield automatically as the pipe is advanced slowly around the center pivot by means of a row of mobile towers. Automatic irrigation machines operating on this principle are transforming the pattern of agriculture in many arid and semiarid regions of the world (see "Center-Pivot Irrigation," page 90). In this case the sprinkler system (*lower left*) is running clockwise. Line at top is an access road. Dark spot at bottom is a water hole.

THE ILLUSTRATIONS

Cover photograph by Harold Duke and Dale F. Heermann

Page	Source
22–27	Total Communications Industries, Ltd.
28–29	George V. Kelvin
30	J. P. Wolfe, W. L. Hansen, E. E. Haller, R. S. Markiewicz, C. Kittel and C. D. Jeffries, University of California at Berkeley
31–37	George V. Kelvin
38–39	Walther Stoeckenius, University of California at San Francisco
40–45	Allen Beechel
46	Walther Stoeckenius, University of California at San Francisco
58–74	Dan Todd
77	Thomas W. Jacobsen, Indiana University
78–83	Alan D. Iselin
84	Thomas W. Jacobsen, Indiana University
85–87	Alan D. Iselin
91	Harold Duke and Dale F. Heermann, Agricultural Research Service, U.S. Department of Agriculture
92	National Aeronautics and Space Administration
93	Andrew Christie
95	Valmont Industries, Inc.
97	Rain Bird Sprinkler Manufacturing Corporation (*top*), Valmont Industries, Inc.(*bottom*)
98	EROS Data Center, U.S. Geological Survey (*top*); Derek Bayes, *Fortune*, © Time Inc. (*bottom*)
101–102	Total Communications Industries, Ltd.
103–107	Lorelle M. Raboni
109	Bulova Watch Company, Inc.
110–125	Andrew Christie
127	*The Healing Hand: Man and Wound in the Ancient World.* © 1975, Harvard University Press
128	*Mites of Moths and Butterflies.* © 1975, Cornell University Press
129	*Muybridge: Man in Motion.* © 1976, University of California Press
130	*Wings over Hellas: Ancient Greece from the Air.* © 1974, Oxford University Press

Discover

the one organized (and inexpensive) way to keep up with what's new in computer techniques and management—

The Library of Computer and Information Sciences

and start with
a whale of a value—

Take any 3 books for only $1 each

(values to $57.95)

if you willjoin now for a trial period
and take three more books—at member
discount prices—within the next 12 months.

(Publishers' Prices shown)

39950. COMPUTER-AIDED EXPERIMENTATION: Interfacing to Minicomputers. *Jules Finkel.* Lays out the basic concepts and techniques needed to specify, design, and use mini-computers in a real-time environment. *Counts as 2 of your 3 books.* **$24.95**

84385. TOP-DOWN STRUCTURED PROGRAMMING TECHNIQUES. *Clement McGowan and John R. Kelly.* Lays out the basic theory and spells out the applications of this powerful new technique. For experienced programmers. **$14.95**

55331. INFORMATION SYSTEMS. *John G. Burch, Jr. and Felix R. Strater, Jr.* All the background needed to successfully develop information systems for just about any organization. **$12.95**

56420. INTRODUCTION TO SEQUENCING & SCHEDULING. *Kenneth R. Baker.* **$16.95**

86540. THE WAY THINGS WORK BOOK OF THE COMPUTER. Encyclopedia of information science, cybernetics, and data processing. **$8.95**

42045. DESIGN OF MAN-COMPUTER DIALOGUES. *James Martin.* The elements used in interactive systems, the basic process involved, and system application. *Counts as 2 of your 3 books.* **$22.50**

41651. DATA MANAGEMENT SYSTEMS. *Carl Cagan.* How they operate, how they can best be implemented. **$14.75**

76920. SCIENTIFIC ANALYSIS ON THE POCKET CALCULATOR. *Jon M. Smith.* Shows how any inexpensive pocket calculator can easily be used to handle complex computations. **$12.95**

56370. INTRODUCTION TO PROGRAMMING LANGUAGES. *Harry Katzan, Jr.* Comprehensive introduction to BASIC, COBOL, and FORTRAN. **$13.75**

65915. OPERATING SYSTEMS PRINCIPLES. *S. Kurzban, T. S. Heines, and A. P. Sayers.* The many levels of hardware and software that make up a modern system identified with their complex interactions explained. **$16.95**

84430. TOWARD CREATIVE SYSTEMS DESIGN. *Henry C. Lucas, Jr.* Demonstrates how users (rather than systems analysts) can design their own systems. **$12.00**

70190. PRINCIPLES OF OPERATIONS RESEARCH. Second Edition. *Harvey M. Wagner.* A complete overview of all the fundamental concepts involved in operations research. 1039 pages. **$19.95**

70250. PRINCIPLES OF SYSTEMS PROGRAMMING. *Robert M. Graham.* Examines basic problems in the design, implementation, and maintenance of major software components—assemblers, loaders, macros, time sharing systems, and more. **$16.95**

76910. THE SCIENTIFIC PROCESS AND THE COMPUTER. *Donald N. Streeter.* Detailed description of the entire range of computer-based systems and services currently available in support of industry, technology, and government. **$18.50**

62760. MINICOMPUTER SYSTEMS: Structure, Implementation, and Application. *Cay Weitzman.* The latest minicomputer hardward and software technology, tools, procedures, and approaches. **$19.50**

How the Club Operates • The Book Club **News**, describing the coming Main Selection and Alternate Selections, will be sent to you 15 times a year at 3-4 week intervals. • If you wish to purchase the Main Selection, do nothing and it will be shipped to you automatically. • If you prefer one of the Alternates, or no books at all, simply indicate your decision on the reply form always enclosed with the **News** and mail it so that we receive it by the date specified. • The **News** is mailed in time to allow you at least 10 days to decide if you want the coming Main Selection. If, because of late mail delivery of the **News**, you should ever receive a Main Selection without having had the 10 day consideration period, that Selection may be returned at Club expense. • Continued membership past the trial period brings you the added benefits of our Bonus Book Plan—an important way to save even more, at least 70% off publishers' prices.

41665. DATA PROCESSING PROJECT MANAGEMENT. *Thomas R. Gildersleeve.* A step-by-step guide, stressing broadly applicable concepts. **$14.95**

42035. THE DESIGN OF REAL TIME APPLICATIONS. *Maurice Blackman.* How to design and implement data processing projects using real time modes of processing. *Counts as 2 of your 3 books.* **$22.00**

81880. SYSTEM ANALYSIS TECHNIQUES. *Edited by J. Daniel Couger and Robert W. Knapp.* 25 carefully selected papers tell how to properly evaluate an existing system, determine managerial information requirements, and assess the benefits. **$15.95**

78335. SIMULATION USING GPSS. *Thomas J. Schriber.* Includes 27 GPSS case studies and over 300 problems. **$16.95**

81860. SYSTEMS: Analysis, Administration, and Architecture. *John W. Sutherland.* Contains the most comprehensive inventory of system concepts ever assembled. **$13.95**

33440. THE ANATOMY OF A COMPILER. Second Edition. *John A. N. Lee.* Advice and information on syntactic analysis, lexical analysis and SYMTAB, storage allocation and management, etc. **$17.95**

55410. INFORMATION TECHNOLOGY. The Human Uses of Computers. *Harry Katzan, Jr.* From how computers work to the future of cybernetics. **$11.95**

The Library of Computer and Information Sciences 7-8AD
Riverside, New Jersey 08075

Please accept my application for membership and send me the three volumes indicated, billing me only $1 each. I agree to purchase at least three additional Selections or Alternates during the first 12 months I am a member, under the club plan described in this ad. Savings range up to 30% and occasionally even more. My membership is cancelable any time after I buy these three books. A shipping and handling charge is added to all shipments. **Send no money. Members are billed when books arrive.**

3 books for $1 each.
Indicate by number the 3 books you want.

A few expensive books (noted in book descriptions) count as 2 choices.

Name _____

Address _____

City _____ State _____ Zip _____

(Offer good in Continental U.S. and Canada only. Prices slightly higher in Canada.) Book selections purchased for professional purposes may be a tax-deductible expense.

What do Isuzus do?

You have no idea, do you? Probably because you have no idea what an Isuzu is. Let alone what one does.

Well, that's what we're here for. To tell you. Isuzu is a new kind of Opel, built by Isuzu. We call it, for lack of imagination, the Opel Isuzu. You may do the same, if you wish.

What an Opel Isuzu does, basically, is offer a lot of miles per gallon while whirring happily along providing grins for its occupants. Thanks to the fact that it comes standard with a short throw 4-speed transmission. And an exotic little 1817 cc overhead cam hemi engine.

Equipped with crossflow heads and 5 main bearings, even.

With that uncanny combination, according to EPA tests, an Opel Isuzu delivered an estimated 23 mpg in the city test. And an estimated 36 in the highway test.

Of course, your mileage may vary, depending on the road, the condition of the car, the equipment and how you drive. Which is as you might expect.

But, gee whiz. What are we standing around with our heads stuck under the hood for? Let's go inside. Where the fun is.

WELCOME TO THE ISUZU.

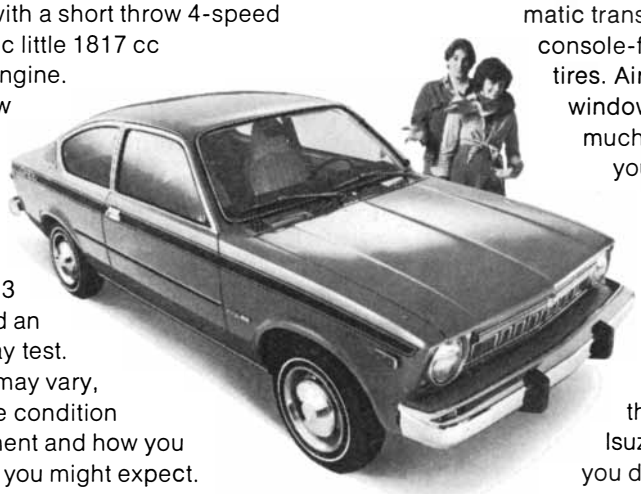
Like the reclining bucket seats? The flow through ventilation? The tinted glass? The locking glove box? The column-mounted dimmer switch? The rack-and-pinion steering? The front disc brakes? They're all standard.

If you want more, you can order more. An automatic transmission. Carpeting. A console-full of gauges. Radial tires. Air conditioning. Rear window defroster. And so much simulated wood grain, you'll think you're driving a tree.

Opel Isuzu.

All in all, it's one solid little investment. And a heck of a lot of fun.

At this point, only one question remains. Now that you know what an Isuzu will do, what are you doing without an Isuzu?



OPEL ISUZU

SOLD BY BUICK

In California see your Buick-Opel Dealer for EPA information.

GM
MARK OF EXCELLENCE

LETTERS

Sirs:

We found "Forage Crops," by Harlow J. Hodgson [SCIENTIFIC AMERICAN, February], extremely informative. We nonetheless feel that several rather glaring distortions permeate his conclusions. He seriously considers only the two extremes: grain-fed livestock and no livestock. To make the most of our food-producing potential it is clearly necessary to utilize the unique abilities of ruminants to convert cellulose into protein. The article never assesses, however, the quantitative effects of supporting our ruminant population solely on land unsuitable for food production. More research on forage crops could promote such a shift.

Hodgson's bias is revealed by his lack of seriousness in considering a replacement for the current high-meat diet in the U.S. He implies that a vegetarian diet must be high-calorie and low-protein by his choice of a rather ludicrous example: the consumption of a kilogram of corn per person per day. Anyone who has read and sampled the recipes in Frances Moore Lappé's book *Diet for a Small Planet,* for example, knows that a vegetarian diet can be both highly nutritious and appealing.

The shift we propose may reduce the meat supply, but this need not have an adverse effect on either the protein content or the palatability of the American diet. Indeed, some evidence suggests that a lower consumption of meat would have a beneficial effect on health.

Of course, it would be unfortunate if prices rose to the extent that only the wealthy could afford meat. Even now 12 million Americans have an inadequate diet because they do not have enough money to spend on food. The problem is one of distribution; it is not created by the level of food production.

A large fraction of the world's population is underfed. This is partly due to the use of large areas of arable land for animal feed and cash crops. Wealthy nations such as the U.S. and the U.S.S.R. exacerbate this situation: their demand for these crops makes profitable a shift in production away from direct food crops. Worldwide inequities in the *control and distribution* of resources are the major causes of malnutrition.

ERIC ENTEMANN

ROBERT SHAPIRO

Science for the People
Somerville, Mass.

Sirs:

I am pleased that Messrs. Entemann and Shapiro found my article informative. Evidently I failed to make some points with adequate clarity. I do not believe I considered only the extremes of grain-fed livestock and no livestock. On page 74 of the article I mentioned that if, as seems likely, the export demand for grain is high and prices rise, grain will be less available and attractive as a livestock feed. Ruminant-livestock production would become still more dependent on forages and rangelands. I indicated that ruminants could be raised entirely on forages and a variety of by-products and be totally noncompetitive with human beings. Increases in forage production that are entirely possible would enable us to achieve this end without decreasing the supply of meat and dairy products.

My reason for using the corn v. meat illustration is that corn is the principal grain produced in the U.S. and it is now fed to cattle. If meat were not produced, we would have to eat whatever else was available. Corn is biologically the most efficient crop that farmers grow, and it gives them the best economic return on their investment. It is easy to say that farmers could grow soybeans or vegetables instead of corn. In view of the comparative yields and production costs, however, the prices for these crops would have to be significantly higher than those for corn.

Soybeans are grown principally for oil and for the meal used in livestock feed. Without the livestock market, prices would be much lower and much smaller acreages would be planted to soybeans. The American market for soybeans, corn and other feed grains for human food is small, probably less than 5 percent of the total production, even though the technology is availa-

Scientific American, June, 1976; Vol. 234, No. 6. 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 30 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

Pioneer has conquered the one big problem of high-priced turntables.

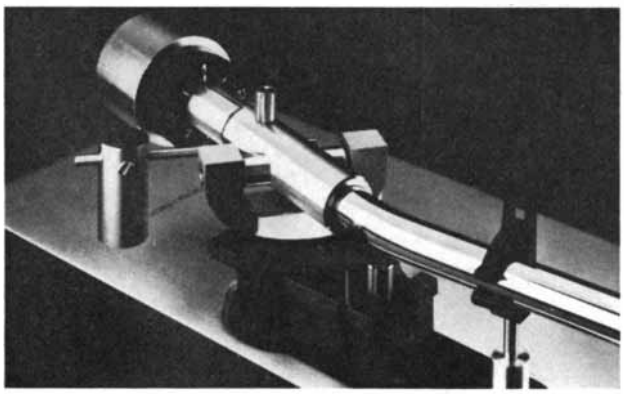
The high price.

For under $200,* you can now own the direct-drive PL-510.

The best way to judge the new Pioneer PL-510 turntable is to pretend it costs about $100 more. Then see for yourself if it's worth that kind of money.

First, note the precision-machined look and feel of the PL-510.

The massive, die-cast, aluminum-alloy platter gives an immediate impression of quality. The strobe marks on the rim tell you that you don't have to worry about perfect accuracy of speed at either 33⅓ or 45 RPM.



The S-shaped tone arm is made like a scientific instrument and seems to have practically no mass when you lift it off the arm rest. The controls are a sensuous delight to touch and are functionally grouped for one-handed operation.

But the most expensive feature of the PL-510 is hidden under the platter. Direct drive. With a brushless DC servo-controlled motor. The same as in the costliest turntables.

That's why the rumble level is down to -60 dB by the super-stringent JIS standard. And that's why the wow and flutter remain below 0.03%. You can't get performance like that with idler drive or even belt drive. The PL-510 is truly the inaudible component a turntable should be.

Vibrations are damped out by the PL-510's double-floating suspension. The base floats on rubber insulators inside the four feet. And the turntable chassis floats on springs suspended from the top panel of the base. Stylus hopping and tone arm skittering become virtually impossible.

But if all this won't persuade you to buy a high-priced turntable, even without the high price. Pioneer has three other new models for even less.

The PL-117D for under $175.*

The PL-115D for under $125.* And the amazing PL-112D for under $100.*

None of these has a rumble level above -50 dB (JIS). None of them has more wow and flutter than 0.07%.

So it seems that Pioneer has also conquered the one big problem of low-priced turntables.

The low performance.

U.S. Pioneer Electronics Corp., 75 Oxford Drive, Moonachie, New Jersey 07074.

PIONEER®

Anyone can hear the difference.

*For informational purposes only. The actual resale prices will be set by the individual Pioneer dealer at his option.

ble to process them into human food. They are, however, cheaper than meat, and an increased human demand would certainly divert them from livestock feed to human food. Because they are so little used as human food, I conclude that they are less appetizing or appealing to the great majority of Americans than foods of animal origin. And since the countries where 75 to 80 percent of the world's people live are seeking to add more animal products to their diet, I conclude that this preference is a nearly worldwide phenomenon.

The word "distribution" is often used by critics of livestock agriculture to explain inequities in the availability of food both in the U.S. and around the world. This usage does not tell us very much. It is a matter of economics and personal preferences in dietary matters. The implication is that in the U.S. we usurp for animal production large quantities of grain that should be made available as human food. There is no escaping the fact, however, that the demand for more grain as human food is not large. More than 85 percent of the coarse grains exported by the U.S. are fed to livestock in other countries.

Those who advocate that these grains should not be fed to livestock seem to imply that farmers should make them available to hungry people at their own expense. Perhaps these people should raise private funds and go into the marketplace, purchase the grains and distribute them to hungry people. Alternatively they could arouse public opinion to persuade Congress to appropriate more funds for food aid. Like the members of any other sector of the nation's economy, farmers will not embark on an activity unless it yields an economic return.

It seems to be often assumed that if the American people consumed a little less meat, grains would flow automatically to hungry people around the world. They most assuredly would not. Those who wish to eat less meat or no meat at all have every right to do so. The right of choice also extends, however, to those who want to have animal products in larger amounts.

HARLOW J. HODGSON

Winrock Conference Center
Morrilton, Ark.

Sirs:

The proposition that Habakkuk was the smallest man in the Bible because he "stood on his watch," so ably refuted by John A. Church ["Letters," SCIENTIFIC AMERICAN, February], is interesting.

However, I have always understood that the honor in question should go to Bildad the Shuhite.

Incidentally, the most precocious person in the Bible was certainly Job, who "cursed the day of his birth."

CALEB HATHAWAY

Bethesda, Md.

CHOOSE ANY ONE OF THESE VALUABLE SETS

as a trial member of the BOOK-OF-THE-MONTH CLUB®

The Suggested Trial: You simply agree to buy four Club choices during the next twelve months at substantial savings on most books you choose.

CARL SANDBURG's *Pulitzer Prize-Winning*
ABRAHAM LINCOLN

The entire six-volume set

More than a portrait of a great American, this monumental 6-volume biography is a tapestry of America in crisis and resolution. Winner of a Pulitzer Prize for history... a work of timeless beauty and humanity.

for only $18.50

PUBLISHER'S LIST PRICE: $120

THE STORY OF CIVILIZATION

by WILL and ARIEL DURANT

Ten matched volumes

This multi-volume history of all mankind—ancient and modern, Occidental and Oriental, is one of the great literary achievements of the twentieth century.

for only $22.50

PUBLISHER'S LIST PRICE: $224.50

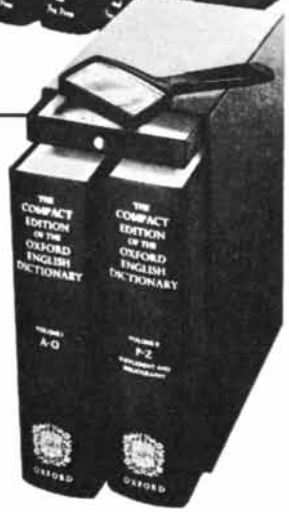
THE ENCYCLOPEDIA OF PHILOSOPHY

The complete contents of the original, eight-volume, $219.50 edition

The only major philosophical reference work published in English since 1901. The collaboration of 500 philosophers cover every aspect of the subject—ancient, modern, Eastern and Western.

for only $17.50

PUBLISHER'S LIST PRICE: $99.50



THE COMPACT EDITION of the OXFORD ENGLISH DICTIONARY

Contains the complete contents of the 13-volume *Oxford English Dictionary*, generally considered the final arbiter of the meaning, origin, history and usage of words in the English language. Bausch & Lomb magnifying glass included.

for only $17.50

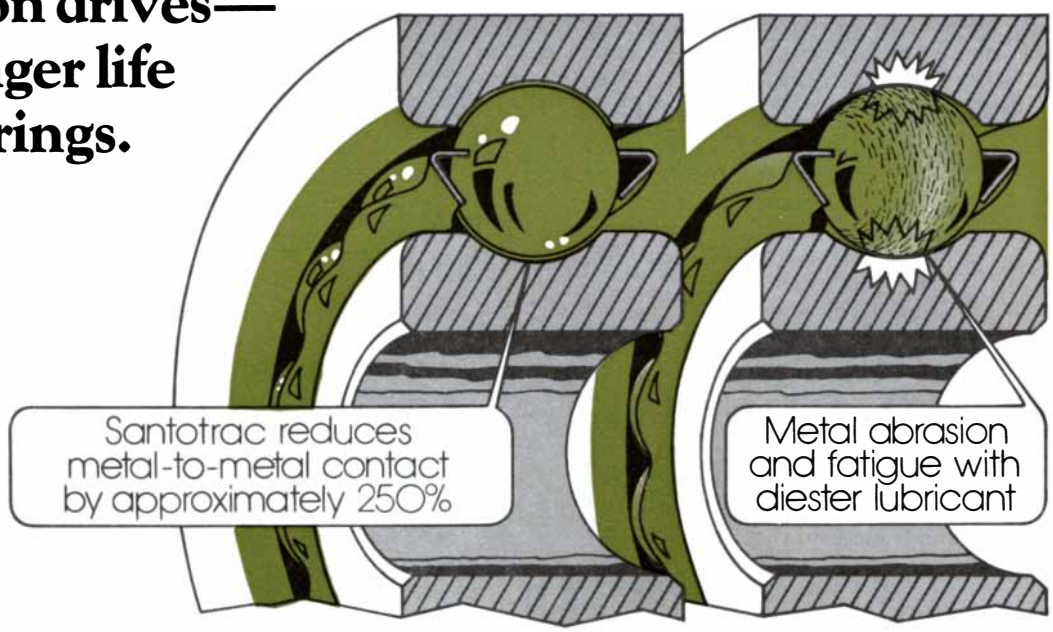
PUBLISHER'S LIST PRICE: $90

TO SIGNALIZE its 50th Anniversary, the Book-of-the-Month Club extends this special invitation. Until now, high retail prices have put the magnificent works shown here beyond the reach of many families. However, as a trial member, you can acquire one of these sets at extraordinary savings—as much as $202. This is not a one-time bounty for booklovers but a demonstration of a continuing benefit. As a member, not only will you be able to keep abreast of the important new books, but if you continue you will enjoy similar savings on other expensive library sets.

Save as much as $202

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

Santotrac® Lubricants: after improved traction drives— Now longer life for bearings.



Does it work? Let's try it and see:

Among the scientific disciplines, none of their practitioners is more practical and down to earth than mechanical engineers. This pragmatism is having a profound effect on the commercial development of Santotrac traction lubricants.

In the December '74 SCIENTIFIC AMERICAN, Monsanto reported on the intriguing prospects that were now opening ever wider for traction devices. Six years of prior testing had proved that the new synthetic cyloaliphatic molecule made a top-drawer lubricant – but a lubricant *with grip!* Trademarked Santotrac, the formulated product offered a traction coefficient of 0.1 – half again as high as any previous lubricant. That meant it should practically double the power capacity of any existing traction drives. But more exciting still, by making it possible to increase traction coefficient between smooth rolling surfaces by 35-50%, Santotrac gave these toothless variable-speed drives their big chance to bite into dozens of applications once the exclusive province of gear boxes, belt pulleys and electric motors. The door was opened for more efficient traction drives to control speed and feed power to machine tool spindles, passenger car drive shafts, lathes, pumps and knitting machines.

In short, with power transmission efficiency upped – traction drives can take on lots more jobs. So development work concentrated on demonstrating to their makers how the film integrity of a Santotrac lubricant is not disrupted by pressures as high as 750,000 psi and 300°F temperatures; that between rolling metal-to-metal contact points it appears to turn solid for a microsecond; that, notwithstanding, it was a *non pareil* lubricant. Manufacturers of traction drives were impressed. In fact, some baker's dozen of new traction drives now still in their development stage are engineered to operate with Santotrac lubricants. However, a diversion entered the picture.

Something New, Something Tough

A handful of mechanical engineers far removed from power transmission took a hard look at that *film-forming phenomenon under immense pressure.* The initial small fraternity has now grown to the point where Monsanto's development staff needs more manpower to keep up with the questions, performance testing, and application evaluations. Their interest: *bearing fatigue life.*

Failures in rolling element bearings – on anything from freight cars to machine tools – most generally are due to two factors: 1) the metal attrition or gradual wear from abrasion at metal-to-metal contact points and/or 2) tiny cracks, pits or spalling caused by elastic deformation of the metal from repeated stress or metal fatigue. The first can be trimmed down to size by good elastohydrodynamic quality in the lubricant – one that maintains oil films thick enough to limit metal-to-metal "asperity" contact under great pressure. But except for degassing and surface treatment of bearing metals, nothing else offered a remedy for the second.

Until the advent of Santotrac – and its impact on the pragmatism of mechanical engineers. As a result, the myriad of test stands that have been contrived are somewhat mind-boggling – all devised with one objective: bearing life. Will it work? In such widely separated spots as London, England and Canton, Ohio, Santotrac

was put to the test.

At Britain's Imperial College Lubrication Laboratory, a sophisticated test machine was devised to study ball bearings. It measures fatigue life under controlled conditions of traction, spin and/or slip. How would Santotrac 50 stand up with standard ball bearings compared with a tough filming diester lubricant of equal viscosity? The first runs showed it did very well indeed: the fatigue life of ball bearings tested (under Hertz Stress 875,000 psi) was almost 4-fold greater when lubricated with Santotrac 50 under identical operating conditions.

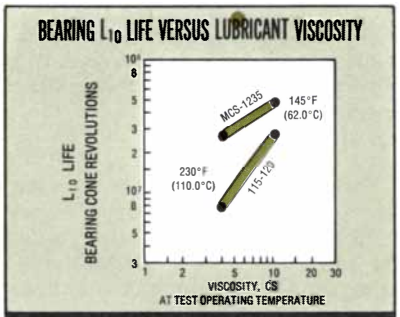
But this was not the most interesting finding. It was the machine's resolution of probably *why* the ball-bearings' fatigue life was so extended. What came out was the enormous difference in the *character* of the rolling contact. Controls were so adjusted to provide that both lubricants furnished a coefficient of traction equal to 0.054. But what showed up most dramatically was the fraction of total traction that was contributed by asperity contact (the metal-to-metal wearing rub) and the amount contributed by the elastohydrodynamic pull or (film strength) lubrication:

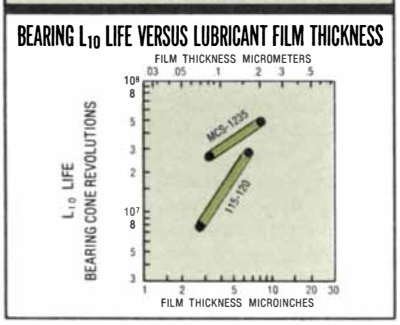
	ASPERITY CONTACT	EHD TRACTION	TOTAL TRACTION COEFFICIENT
DIESTER LUBRICANT	0.025	0.029	0.054
SANTOTRAC 50	0.010	0.044	0.054

Santotrac reduced the metal-to-metal contact by approximately 250%. The cumulative data has led the British engineers to theorize that sensitivity to fatigue under traction is highly dependent on the magnitude of asperity contact.

Tough Films Show Their Mettle

Two American engineers cut the cake another way with a different machine. They calculated film thickness under load, using two different temperatures and related this to fatigue life. Their findings were reported to the 30th annual ASLE meeting. Engineers D. V. Culp and J. D. Stover of The Timken Company, Canton, Ohio, tested tapered roller bearings, four per run in a test machine that maintained the bearing cone speed at 2,700 rpm under a maximum Hertz Stress load of 272,000 psi. The fatigue life vis-a-vis lubricant film thickness and viscosity was recorded for a 32-bearing test. A formulation of Santotrac 50 lubricant (MCS-1235) was compared to a high viscosity index paraffinic oil (115-120), with both compounded to have equal viscosity at the two test temperatures.





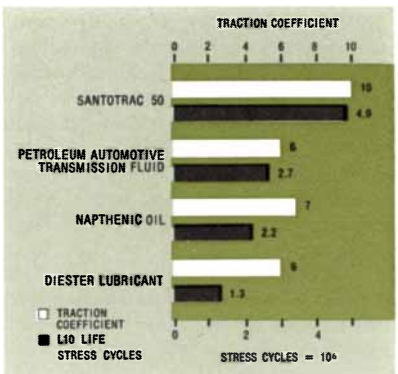
Their findings show up on these two graphs. The graphs show that at equal viscosities, the bearings had longer life lubricated with Santotrac: 1.7 times longer at 11 cs. (145°F) and 3.4 times longer at 4 cs. (145°F). In scanning their findings of the relation of bearing life to *film thickness,* however, these astute researchers caught the anomaly squarely by the tail. They observed:

> ". . . the life improvement at 230°F (110.0°C) is over and above that which is readily explicable by increased film thickness alone. This would indicate that yet another mechanism, or relationship between bearing fatigue life and the specific oils involved, accounts for a major portion of the life improvement at the 230°F (110.0°C) temperature."

There is indeed. *Part* is probably due to the London findings that show Santotrac lubricants minimize the asperity contact. Another is a theory that Monsanto researchers are trying hard to find a way to prove experimentally.

Do The Cracks Plug Up?

On a rolling contact testing machine they developed, Monsanto researchers found early on that the fatigue life of cylindrical bearings (⅜" dia. pins) tested at 700,000 psi Hertz Stress, 300°F, 12,000 rpm gave the following far longer L_{10} fatigue life than three other standard lubricants:



In no instance has any meaningful comparative relationship shown up between traction coefficient and bearing L_{10} fatigue life. Why then the enormously longer use life, approximately double that obtained with any other lubricant? No better hypothesis has surfaced than one the Monsanto research team pencil-doodled out a year ago: it is possible that Santotrac *fills* the high stress, hairline cracks that are the precursors of bearing failure, *solidifies* under the high rolling pressure and inhibits crack propagations.

Some time, some how, some where – a way may be found to measure *that!*

To receive a copy of "Santotrac Lubricants", write or call:

F. H. Langenfeld
Monsanto Company Office B2SF-S
St. Louis, Mo. 63166
(314) 694-2523

Monsanto
the science company.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JUNE, 1926: "Surely it is not stretching a point too far to say that the automobile is the greatest economic factor in modern American life. The automobile bill of the American people is more than 14 billion dollars a year. Our annual investment in automobiles is greater than the annual value of farm crops and more than twice the annual investment in new buildings. Hence the automotive industry has risen to the position of the leading industry of the United States. This startling fact should not surprise us when we learn that the last annual registration of automobiles was about 20 millions. Not only is the industry stupendous in itself but also its effects on other industries are in proportion, in road-building, for example, where the programme now in the course of execution calls for the expenditure of about a billion dollars a year, and in the upbuilding of suburban communities where the new construction, in the case of many cities, is larger than that in the urban district itself."

"By July 1 the population of the continental United States will be 117,135,817, the Census Bureau estimates. This will be an increase of 11,426,197 since the last Federal enumeration in 1920. A gain of a million and three-quarters is recorded since last July."

"Cable dispatches from across the water have indicated that the Royal Society of London has been asked by one or more of its members to consider the advisability of requesting Sir Oliver Lodge to resign from the Society on the ground that his views on spiritualism are prejudicial to the interests of the Society. It is incredible that this great institution should assume the right of censorship on all expression of personal views by its members on general scientific subjects. On the other hand, since Sir Oliver, because of his wonderful facility in writing down to the understanding of the general public, has such a widely extended vogue, we think that he should be more careful to make it known that when he writes as a so-called spiritualist, he is giving merely his personal views, and that these views are not to be taken as an expression of the attitude and belief of the great Society to which he belongs."

"Man has never been able to penetrate much more than one mile beneath the earth's surface—a mere pinprick. Yet, while science does not claim positive assurance regarding the actual conditions in the earth's inner layers, there are certain things about it that we know are not so. One is that the earth's interior is a molten liquid. Probably nothing has tended so strongly to crystallize this widespread misconception as the existence of volcanoes, pouring out liquid lava. In the earth's center, however, the pressure is too great for the rocks to melt. Thus they are as rigid as steel. The seismologist can easily prove the latter statement, since he records earthquake shocks that travel through the earth's interior, and it is the characteristics of these shocks that prove the earth's interior to be rigid. Volcanoes are held by the majority to be of local, not deep-seated, origin."

"'You cannot eat your cake and have it too' is an old saying that embodies more truth than poetry. Yet that is the very thing the Government of the United States is now trying to do—cutting its income, cutting its expenses and still having more to show for its money. And how is it accomplishing this anomaly? By a scientific budget system. President Coolidge has explained the principle of it over the radio. 'Merely to reduce the expenses of the Government might not in itself be beneficial,' he said. 'No civilized community would close its schools, abolish its courts, disband its police force or discontinue its fire department. The underlying spirit of economy is to secure better education, wider administration of justice, more public order and greater security from conflagration, all through a superior organization. It is all reducible to a question of national efficiency.'"

SCIENTIFIC AMERICAN

JUNE, 1876: "As we go to press a special train is rushing across the continent. attempting to travel from New York to San Francisco in the unprecedented time of 88 hours, or in 60 hours less time than is now occupied by regular trains. The start was made from Jersey City at 1 A.M. on June 1, three cars being attached to a new engine, the tender of which carried 2,400 gallons of water and 136 bushels of coal. The 90 miles between Jersey City and Philadelphia were covered in 99 minutes. At 10:40 A.M. Pittsburgh was reached, and at 10 P.M. the train stopped at Chicago, having completed the distance of 907 miles in 20 hours and 57 minutes. The train reached Council Bluffs, Iowa, at 9:27 the next morning. The fastest time made since the train left New York is stated to have been on the Chicago and North-Western Railway, where 2½ miles were run in 2 minutes, or 1 mile in 48 seconds, being at the rate of 75 miles an hour. This is said to be the longest and fastest continuous run that has ever been made on any road in any country."

"It is generally admitted at the present time that the displacement of a luminous point receding from or approaching an observer produces an alteration in the length of the luminous wave proceeding from the point. If this principle is combined with the principles of spectrum analysis, it follows that the lines in the spectra of substances existing in a moving star will be displaced, and hence the movement of the body in one or the other direction may be recognized. Father Secchi has prepared a table in which he compares the results of determinations of the motions of stars obtained by various observers."

"Professor Haeckel, in his recent work *The History of Creation,* gives his opinion that the human race is a small branch of the group of *catarrhini,* and has developed out of long since extinct apes of this group in the Old World. He also points out a fact, necessary to be observed by unscientific people, namely that none of the manlike apes is to be regarded as the parent of the human race, and that the apelike progenitors of the human are long since extinct. In concluding his work, Professor Haeckel remarks on the desire of some who are not actually opponents of the doctrine of descent. 'They wait,' he says, 'the sudden discovery of a human race with tails, or of a talking species of apes.' Such manifestations, the author observes, would not furnish the proof desired, and unthinking persons would be provided with as satisfactory arguments as they nowadays employ in hurling their defiance against all who are evolutionists."

"Salicylic acid, since its introduction into the materia medica, has been very extensively employed as an antiseptic and a febrifuge. Its exact value in both of these offices is still *sub judice.* Over the past few months surprising results have been claimed for it in the treatment of acute rheumatism."

"A remarkably ingenious and novel application of the spectroscope in a case of alleged forgery was recently made by Dr. P. H. Van der Weyde, who appeared in court armed with a powerful microscope having a spectroscopic eye piece. Dr. Van der Weyde was given a variety of checks from the bank, together with the particular document on which the suit rested. The microscopic part of his instrument magnified the writing so highly that the dot on an i was sufficient for all purposes. The result was that the ink of the alleged forged signature and that of the bank's genuine writing gave exactly the same absorption spectrum, and hence were pronounced identical."

"The confession of the Boston belfry murderer, Piper, reveals an instance of a mental condition, the evil results of which seemingly defy any preventive measures that society can devise. The principal measures must lie in the hands of parents and instructors. It is for them to repress the first manifestations of an abnormally destructive nature, for during early youth that nature is most clearly exhibited, and it is for them likewise to see to it that such sources of moral contagion as society tolerates shall not influence the plastic and receptive minds of those entrusted to their care."

You can tell a lot about an individual by what he pours into his glass.

Bushmills.
The world's oldest whiskey.
Individuals have poured this
smooth mellow whiskey since 1608.

The "Mountain Climber" glass created for the Bushmills Collection by Henry Halem

A blend of 100% Irish Whiskies. 86 Proof. Bottled in Ireland. The Jos. Garneau Co., New York, N.Y. © 1974

THE AUTHORS

RICHARD G. NIEMI and WILLIAM H. RIKER ("The Choice of Voting Systems") are professors of political science at the University of Rochester. Niemi did his undergraduate work at Lawrence University and his graduate work at the University of Michigan. He joined the Rochester faculty in 1967 after obtaining his Ph.D. from Michigan. "A while ago," he writes, "I became bored on a drive between Rochester and Ann Arbor, and so I began dictating the article that appears here, stopping dictation whenever I passed big trucks." Riker's undergraduate degree is from DePauw University and his Ph.D. is from Harvard University. He taught for 14 years at Lawrence and 14 more at Rochester, where he is now Wilson Professor of Political Science. "One of the happiest developments from my work at Lawrence," he says, "is that one of my students there, Richard Niemi, is now my colleague at Rochester and my collaborator on this article."

GORDON A. THOMAS ("An Electron-Hole Liquid") is a member of the technical staff of Bell Laboratories. He received his B.S. from Brown University in 1965 and his Ph.D. from the University of Rochester in 1971. He is a coauthor, with T. Maurice Rice, John C. Hensel and Thomas G. Phillips, of *The Electron-Hole Liquid,* to be published this year by Academic Press. He has also done research on superconductivity, phase transitions in magnets and one-dimensional organic conductors.

WALTHER STOECKENIUS ("The Purple Membrane of Salt-loving Bacteria") is professor of cell biology at the Cardiovascular Research Institute of the University of California School of Medicine in San Francisco. A native of Germany, he obtained his M.D. from the University of Hamburg in 1950. Following a period of internship in the university's clinics, he turned to basic research, applying electron microscopy to the study of virus structure and composition at the Institute for Tropical Medicine at Hamburg. Two years later he transferred to the university's Institute of Pathology, where he worked on the electron microscopy of cell membranes. He came to the U.S. in 1959 for a one-year stint as guest investigator at the Rockefeller Institute for Medical Research and stayed on for eight years, becoming associate professor at Rockefeller University. He moved to San Francisco in 1967, "attracted not only by the city but also by an arrangement... that promised me money to pursue my research for the next seven years without having to write grant applications. This was fortunate, because I am convinced that I would not have been able to write a fundable grant application in the early stages of the work on the purple membrane."

WILLIAM G. MELBOURNE ("Navigation between the Planets") is with the Jet Propulsion Laboratory of the California Institute of Technology, where he heads the section on tracking and orbit determination. A graduate of the University of California at Los Angeles, he went on to acquire his Ph.D. in astronomy from Cal Tech in 1959. While a graduate student, he reports, "I took a summer job at the Jet Propulsion Laboratory, where interplanetary flight was rapidly evolving from fantasy to reality. The prospect so captured my imagination that I stayed at the Jet Propulsion Laboratory after graduation and have been there ever since.... Recently my professional interests have centered around the development of the radio telescope for intercontinental geodesy with the technique of very-long-baseline interferometry. Our hope is to measure relative positions on the surface of the earth to an accuracy of a few centimeters. At this accuracy the scientific information about the dynamics of the earth should be tremendous."

THOMAS W. JACOBSEN ("17,000 Years of Greek Prehistory") is professor of classical studies at Indiana University, where he serves as chairman of both the department of classical studies and the graduate program in classical archaeology. A Minnesotan by birth, he majored in classical languages and political science at St. Olaf College and also holds an M.A. in classics from the University of Minnesota. His Ph.D., in classical archaeology, is from the University of Pennsylvania. He taught classical languages and literature at Vanderbilt University before joining the Indiana faculty 10 years ago. Jacobsen spends most of his summers doing fieldwork at various prehistoric sites in southern Greece. Among his current responsibilities is the field directorship of the Argolid Exploration Project, sponsored jointly by Indiana University and the University of Pennsylvania, and the directorship of Indiana University's excavations at Franchthi Cave, which are described in the present article.

WILLIAM E. SPLINTER ("Center-Pivot Irrigation") is professor and chairman of the department of agricultural engineering at the University of Nebraska, where he earned his undergraduate degree in 1950. After obtaining his doctorate from Michigan State University in 1955 he did research for a time at North Carolina State University before returning to Nebraska to take up his present posts in 1968. His research has ranged over several areas of agriculture, including the mathematical modeling of plant growth, the growth dynamics and physical properties of various crops, the development of mechanical transplanters and harvesters and the electrostatic deposition of sprays and dusts. A long-time expert on irrigation systems (he was raised on an irrigated farm in North Platte, Neb.),

he serves as acting director of his university's Water Resources Research Institute.

F. RICHARD STEPHENSON and DAVID H. CLARK ("Historical Supernovas") work on different aspects of their subject at research institutions in Britain. Stephenson, who specializes in the historical aspect, is currently Earl Grey Memorial Fellow in geophysics and planetary physics at the University of Newcastle upon Tyne, his home city. He writes: "I was graduated as a physicist from the University of Durham in 1963, but my interest lay more in astronomy. As a postgraduate student I had an invaluable opportunity to further this interest in the study of the application of ancient astronomical observations in geophysics and astrophysics. I received my M.Sc. from Newcastle in 1964 and, after a time spent as a college lecturer, my Ph.D. from Newcastle in 1972. Since then I have continued to develop my research along similar lines." Clark, who is a radio astronomer, is at the Mullard Space Science Laboratory of University College London. A New Zealander, he received his degrees from the Victoria University of Wellington: a B.Sc. in 1966 and a Ph.D. in 1970. His research interests have shifted over the years from space geophysics to radio astronomical investigations of the X-ray emissions from galactic supernova remnants.

HENRY W. RYDER, HARRY JAY CARR and PAUL HERGET ("Future Performance in Footracing") are all residents of Cincinnati with a common interest in the physiology and psychology of running. Ryder is associate clinical professor of medicine at the University of Cincinnati College of Medicine. A graduate of Yale University, he received his M.D. from the University of Chicago in 1937. He describes himself as an internist in primary care with a lifelong professional involvement in clinical physiological research. In recent years he has climbed Mount Fuji in Japan and trekked to Kala Pattan in Nepal, looking, he says, "for a new challenge for marathoners who had recovered from myocardial infarction." When he is in Cincinnati, he adds, he jogs in the dark five to six miles every few days and commutes to his office and the local hospitals by bicycle. Carr, who acquired his M.A. in mathematics from the University of Cincinnati in 1970, works as a computer programmer for a local bank; he runs "between 15 and 20 miles a week." Herget is Distinguished Service Professor of Astronomy at the University of Cincinnati and director of the Cincinnati Observatory, where he has been a member of the staff since receiving his undergraduate degree in 1931. His Ph.D. thesis was on the computation of the orbits of newly discovered minor planets. Nearly every aspect of his career since then, he reports, "has been associated with the use of computers." During World War II he worked in the Nautical Almanac Office of the U.S. Naval Observatory. He has also served as a consultant on several U.S. space projects.

"What makes America work?...technology

by Jerome B. Wiesner,
President, Massachusetts Institute of Technology

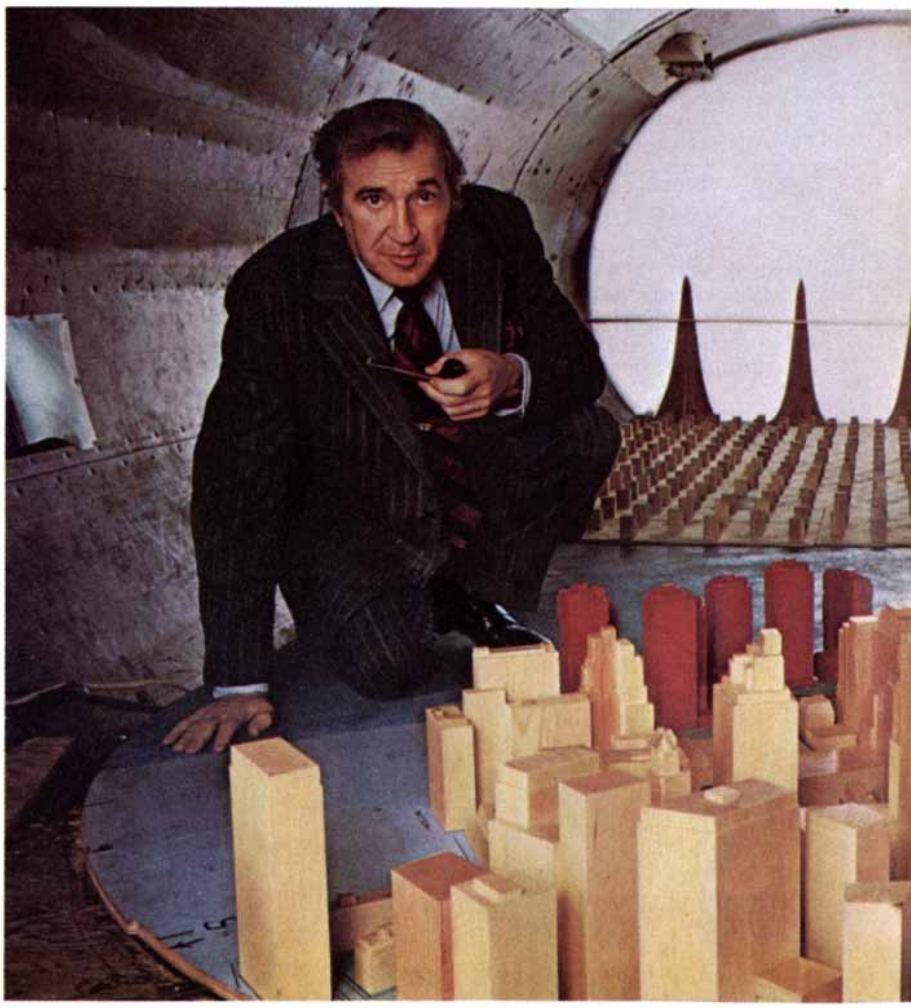
More than any nation in the world, the United States has the opportunity to lead mankind toward a life of greater fulfillment. This opportunity is based on benefits from our continuing advances in science and technology. It is significant that people everywhere look to the United States to provide the science and technology which they need as they, too, seek to improve their condition.

Yet the survival of our own abundant society is being doubted by many thoughtful people who share a powerful concern, a reasonable apprehension, about the impact of technology. In this enormously complex world, each large-scale technological advance has costs, side effects often unanticipated. Solutions to these in turn frequently create new and more perplexing problems, and because the scale of everything involved grows all the while at an exponential rate it becomes ever more difficult to modify the system.

It is obvious that many of our problems—pollution of the environment, the dangers of possible nuclear war or exhaustion of natural resources, especially energy—would not exist in a primitive society, but neither would our present levels of accomplishment and well-being. One of the greatest challenges facing Americans involves learning how to live with our own spectacular successes. Even if we wanted to, we could not reverse the forward thrust of the world. We have no choice but to continue to try to improve it—and we can—guided by our current appreciation of the potential dangers as well as benefits of new technologies.

Over the last half century, vast man-made systems have completely altered the human relationship to the natural world and even masked its dependence upon it. But there is no way for us to avoid being creatures of nature. So modern societies can only exist in a state of dynamic equilibrium that involves a continuing adaptation both to the changing man-made world and to nature. This requires new organizational forms, new knowledge, and new technologies. We need them now more than ever.

We must learn much more about how to manage a technological society: how to anticipate problems and to plan for long-range needs, how to set realistic priorities, how to judge the feasibility of projects and develop essential technolo-



–and we need it more than ever."

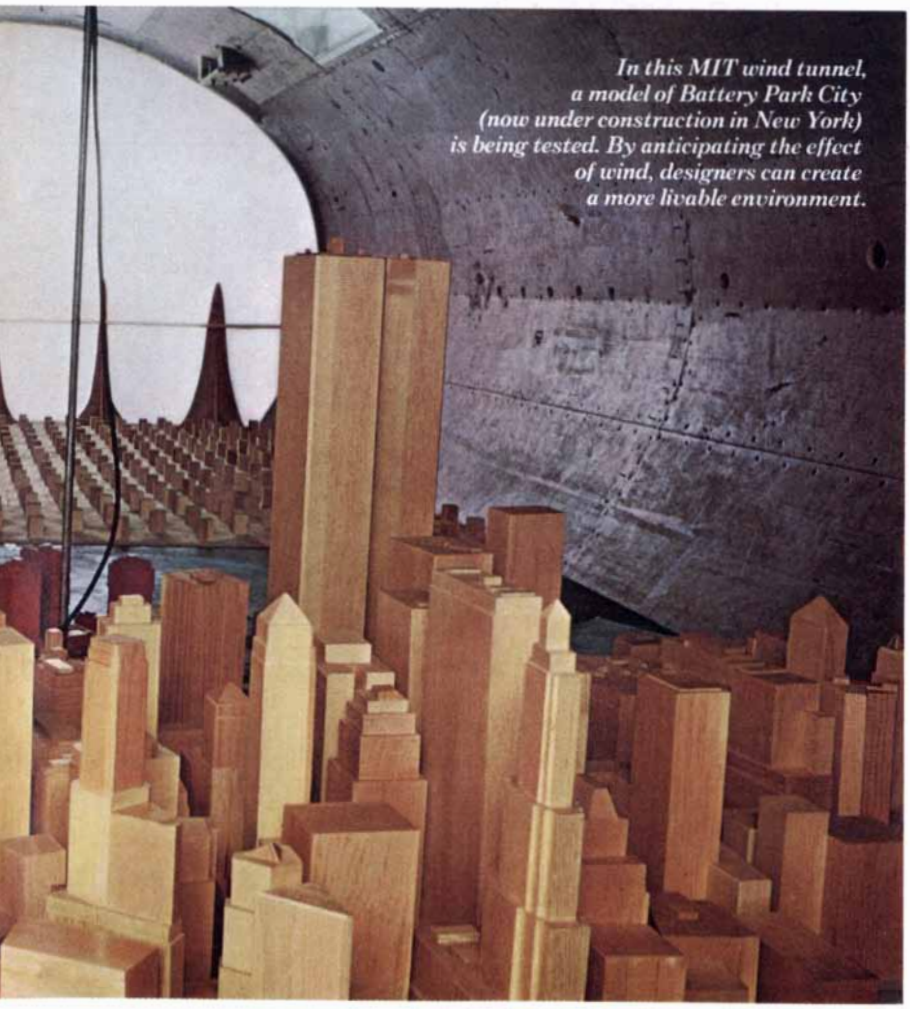
gies to be available when needed, and how to redirect ongoing work as needs and goals change.

Such a formidable job is by no means beyond our abilities. To do it properly, to evolve a self-correcting system which will produce the more equitable and humane society which we all seek, however, will require taking full advantage of the best of our industrial, educational and governmental institutions, interacting to exploit technology in wholly new modes.

★ ★ ★ ★ ★

Research at U.S. Steel

Learning to manage a technological society involves, among other things, the generation of realistic data that can be applied to the solution of existing and anticipated problems. One of the functions of the scientists and engineers at the U.S. Steel Research Laboratory is the development of such solutions within economic boundaries and a socially acceptable framework.



In this MIT wind tunnel, a model of Battery Park City (now under construction in New York) is being tested. By anticipating the effect of wind, designers can create a more livable environment.

For example, when it became apparent that supplies of high-grade iron ore, containing about 62 per cent iron in the dry state, were not inexhaustible, research programs were established to develop methods for concentrating the much larger reserves of low-grade ores, some of which contain as little as 15 per cent recoverable iron. These programs were highly successful.

Similarly, the effects of the energy shortage, both short-term and long-term, have been studied and six programs in twenty-five plant locations have been implemented. Alternative sources of "clean" energy... clean coke, coal gasification and nuclear energy, are also being given serious consideration for future operations.

Research programs of this nature will help to solve the problems that develop during the growth of our technological society.

United States Steel, 600 Grant Street, Pittsburgh, Pennsylvania 15230

USS TRADEMARK **We're involved.**

I always drink Cuervo Gold.

Now & then.

When the usual begins to feel a bit too usual, try some Cuervo Especial on the rocks.

Not to replace your scotch or vodka or whatever. But to substitute for it, now and then.

Of course you might cause a slight lowering of voices and raising of eyebrows.

Common folk tend to do that sort of thing.

Try it on the rocks.

CUERVO ESPECIAL® TEQUILA. 80 PROOF. IMPORTED AND BOTTLED BY © 1976 HEUBLEIN, INC., HARTFORD, CONN.

Established 1845 SCIENTIFIC AMERICAN June 1976 Volume 234 Number 6

The Choice of Voting Systems

Any voting system can lead to paradoxical results where losers are preferred to winners and winners become losers. In certain situations, however, some voting systems are better than others

by Richard G. Niemi and William H. Riker

Voting is one of the chief methods of making group decisions in all democratic forms of government. Cultural, social and political circumstances, however, have led to the development of a wide variety of voting systems. Perhaps the simplest means of arriving at the decision is the plurality method, where the candidate or proposal receiving the most votes wins. Another common method is majority rule, which requires that the winner get more than 50 percent of the votes cast (or sometimes be favored by 50 percent of the total eligible voters). And under proportional representation elective offices are allotted proportionately according to the popular vote for the competing political parties or candidates.

Is there a best system? Studies over the past two decades reveal that undesirable results are possible under any voting system. Improved methods of voting may be possible, but a perfect system does not exist.

It is well known that the system for electing the president of the United States can lead to a number of perverse results. Both the plurality method and the majority method are employed in presidential elections. The electoral college, which actually elects the president, is chosen by the plurality method within the states. All of a state's electoral votes, which are equal in number to its seats in Congress, are given to the presidential candidate who wins the most votes in that state. (Actually the votes are given to electors who are pledged to vote for that candidate, and on a few occasions an elector has voted for someone else.) Thus a candidate who obtains fewer popular votes than an opponent can win the presidency. This certainly happened in the presidential election of 1888 (Benjamin Harrison v. Cleveland), and it probably was the case in the elections of 1824 (John Quincy Adams v. Jackson, Crawford and Clay) and 1876 (Hayes v. Tilden). It may have happened as recently as 1960, when there was some uncertainty about the votes for unpledged and Kennedy-pledged electors in Alabama. Under the most extreme circumstances a candidate could currently win the presidency by gaining pluralities in the 11 largest states but not getting a single other vote.

If no candidate receives a majority vote from the electoral college, the House of Representatives elects the president. In this case a second-place candidate could be elected, as John Quincy Adams was in 1824. And a third-place or even a fourth-place candidate could exercise considerable influence over the choice of the president, as Henry Clay did in 1824 and George Wallace hoped to do in 1968.

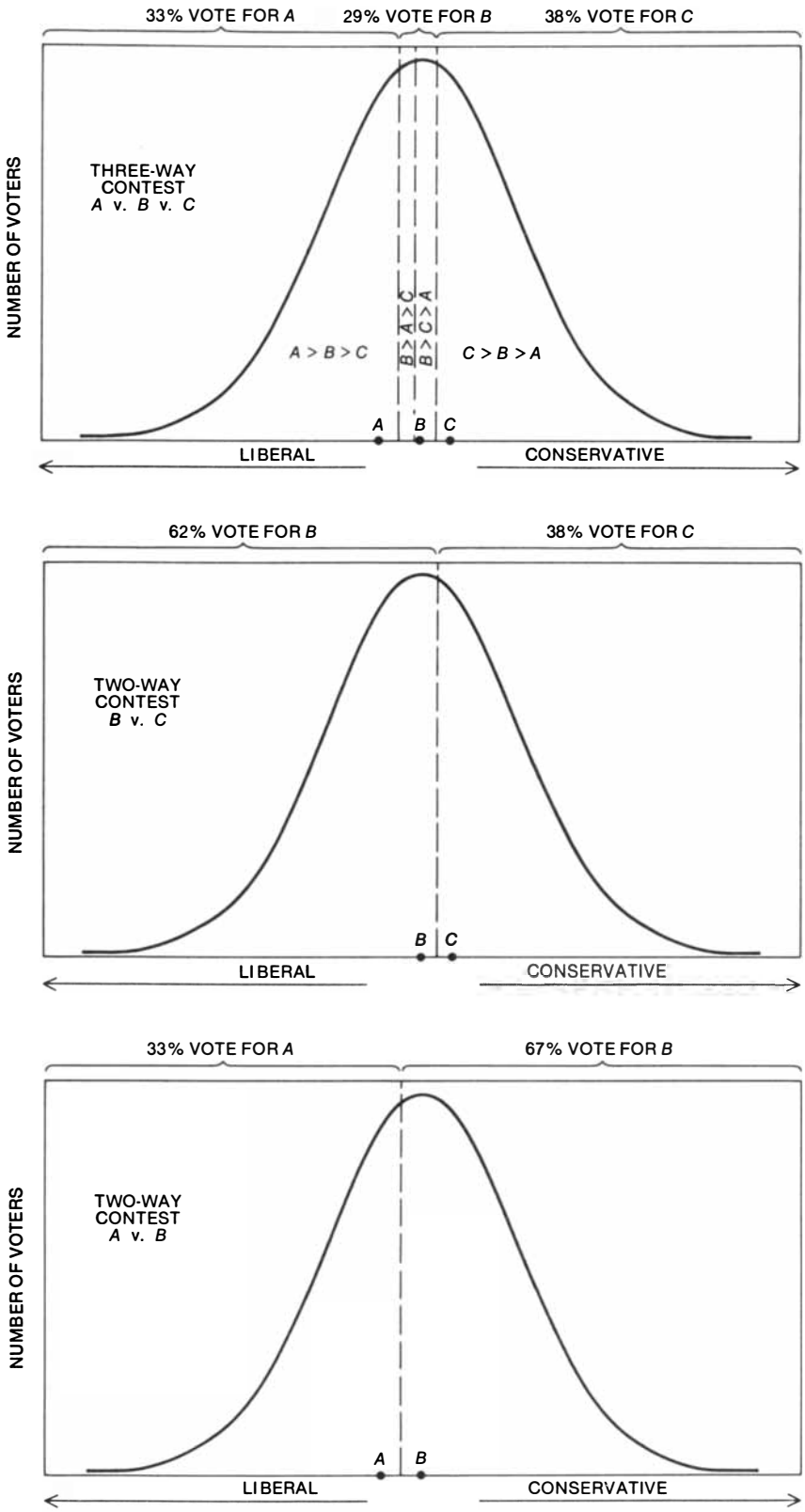
In spite of efforts to reform the system for electing the president, no proposal has emerged that is able to attract substantial support. Most of the proposals advocate the abolition of the electoral college and the direct election of the president by the voters. It is thought that only direct voting ensures that the candidate with the most popular support will win. But if by "most popular" we mean that more voters prefer the winning candidate to any other candidate in pairwise (two-candidate) choices, then direct voting does not ensure that the most popular candidate will win except in elections with only two candidates.

Suppose there are five presidential candidates in the election, four of whom are liberal and one of whom is conservative. The four liberal candidates might each get 18 percent of the vote, with the remaining 28 percent going to the conservative. Although it is true that the conservative would have outpolled all the other candidates, it is not unreasonable to suppose that a large majority of the electorate would have preferred any one of the liberal candidates to the conservative one. Or suppose there are 10 viable candidates in the election. A candidate could win with as little as 11 percent of the vote, and it could easily be the case that any one of the nine defeated candidates could beat the winning candidate in a pairwise contest.

Since such outcomes are possible in multicandidate elections, most proposals for the direct election of the president have a contingency feature to prevent the choice of a candidate who does not receive some minimum fraction of the vote (usually 40 percent). A common proposal is for a runoff election between the top two vote-getters. But does that solve the problem?

Suppose there is a presidential election with three candidates. One candidate is preferred by "liberals" (who are 43 percent of the voters), another by "conservatives" (40 percent) and a third by "moderates" (17 percent). Let us assume that the voters prefer the candidates in the order of their regard for them in terms of a hypothetical liberal-conservative dimension. That is, the order of preference of liberal voters is liberal-moderate-conservative and that of conservative voters is conservative-moderate-liberal. Moderates might prefer either the liberal candidate or the conservative candidate second, depending on the voters' and the candidates' exact location on the liberal-conservative dimension.

Suppose the moderates divide 13 percent with the order of preference moderate-liberal-conservative and 4 percent with the order moderate-conservative-liberal. In an initial three-way election the liberal candidate receives 43 percent of the votes cast, the conservative 40 percent and the moderate 17 percent. In the runoff election the liberal will win with 56 percent of the vote. Has the



MULTICANDIDATE ELECTION, where the candidate receiving the most votes wins, may eliminate a candidate who could defeat any of the other candidates in a two-way election. The voters are assumed to prefer candidates in order of their closeness to them in terms of a liberal-conservative dimension. For example, liberal voters prefer *A* to *B* to *C*, and conservative voters prefer *C* to *B* to *A*. Some moderate voters prefer *B* to *A* to *C*, and other moderate voters prefer *B* to *C* to *A*. In a three-candidate election (*top*) Candidate *C* would win, with Candidate *A* taking second place. *B*, however, could defeat *C* (*middle*) or *A* (*bottom*) in a two-way election.

candidate preferred by a majority of the voters won? The answer is no. If the moderate candidate were paired against the liberal in a two-way election, he would win because he would receive the conservative vote as well. The moderate candidate would also win in a two-way contest with the conservative candidate.

A more extreme example is a 10-candidate presidential election. The two candidates receiving the most votes might each receive only 11 percent of the vote, and there is no guarantee that either of them could defeat any of the other candidates in pairwise contests. All a runoff election ensures is that the winner is preferred to the loser by the majority of voters in the runoff election itself. A runoff election does not solve the underlying problem of finding the candidate who is preferred by a majority of the voters.

These results concerning presidential elections are important, but far more surprising things can happen as a consequence of voting. Let us begin by considering individual voters. It is reasonable to assume that the preferences of most voters are transitive. By this we mean that if a voter prefers *A* to *B* and *B* to *C*, then he will also prefer *A* to *C*. Indeed, we might well consider a voter to be irrational if he says he prefers Ford to Reagan, Reagan to Carter and Carter to Ford. Experimental studies suggest that observed intransitivities arise only because of a lack of careful thought by individuals. Thus individual intransitivity is likely to be found only when a voter is asked to make a number of successive choices between pairs of candidates. If the voter is asked to list all the candidates in the order of his preference, the temporary intransitivity disappears.

It therefore seems reasonable to assume that the preferences of individual voters are transitive. It is clear, however, that under majority-rule voting on pairs of candidates a group made up entirely of voters with transitive preferences can give rise to an intransitive situation such as one in which Ford can beat Reagan, Reagan can beat Carter and Carter can beat Ford. Such a situation is often referred to as the paradox of voting.

The paradox was noted by the 18th-century French philosopher and mathematician the Marquis de Condorcet. It also attracted the attention of a number of well-known 19th-century writers, including the Reverend Charles Lutwidge Dodgson, better known as Lewis Carroll. It is only recently, however, that a polyglot group of political scientists and economists have begun to discover the full implications of Condorcet's initial perception. These implications are often disconcerting to those who are proponents of voting as a method of making group decisions.

One implication is that when a legislature finds itself in a situation where the paradox of voting exists, the final choice made by

VOTER 1	VOTER 2	VOTER 3
PREFERS FORD TO REAGAN AND REAGAN TO CARTER, AND SO PRESUMABLY, PREFERS FORD TO CARTER	PREFERS REAGAN TO CARTER AND CARTER TO FORD AND SO PRESUMABLY PREFERS REAGAN TO FORD	PREFERS CARTER TO FORD AND FORD TO REAGAN AND SO PRESUMABLY PREFERS CARTER TO REAGAN

VOTING IN PAIRWISE ELECTIONS

	VOTER 1	VOTER 2	VOTER 3	GROUP RESULT
FORD v. REAGAN	FORD	REAGAN	FORD	FORD PREFERRED TO REAGAN BY TWO-TO-ONE VOTE
REAGAN v. CARTER	REAGAN	REAGAN	CARTER	REAGAN PREFERRED TO CARTER BY TWO-TO-ONE VOTE
FORD v. CARTER	FORD	CARTER	CARTER	CARTER PREFERRED TO FORD BY TWO-TO-ONE VOTE

PARADOX OF VOTING, also known as the Condorcet effect (after the Marquis de Condorcet), arises when the preferences of individual voters are transitive but the result is intransitive. In this context "transitive" means that if a voter prefers Ford to Reagan and Reagan to Carter, then he will prefer Ford to Carter. When the paradox of voting occurs, a group of voters with transitive preferences can still give rise to an intransitive outcome under majority-rule voting. For example, although Ford would win over Reagan, and Reagan in turn would beat Carter, Carter would win in a two-way contest with Ford. This phenomenon is sometimes referred to as a cyclical majority.

majority-rule voting may depend on the order in which the alternatives are voted on. In otherwise identical circumstances a legislature can pass an amended bill under one voting order, pass the unamended bill under another voting order and pass no bill under a third voting order.

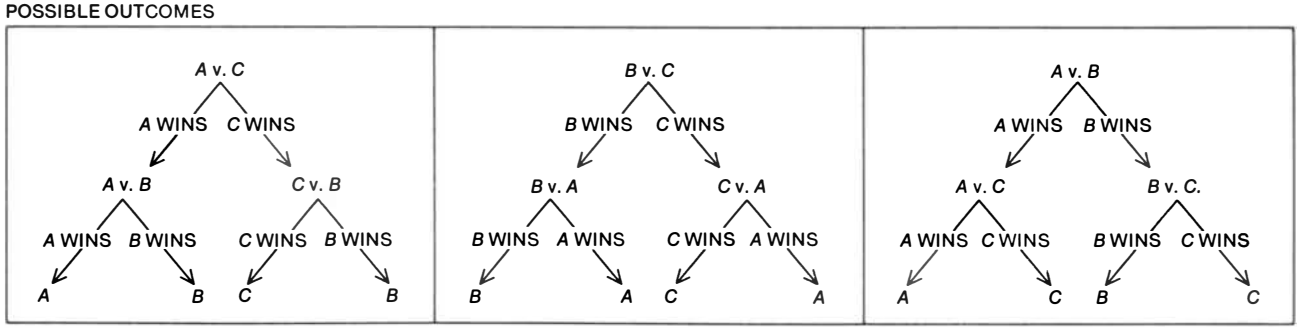
As if it were not enough that the choice may depend on the voting order, this fact can be used to twist the outcome of legislative processes. It may be possible to create a voting paradox such that no action is taken by the legislature even though a proposed bill would have passed prior to the creation of the paradox. A legislator could introduce an amendment to create such a paradox, and if the voting order were just right, the amended proposal would then be defeated.

That is not merely a theoretical possibility. One of us (Riker) has suggested that the passage of the 17th Amendment to the U.S.

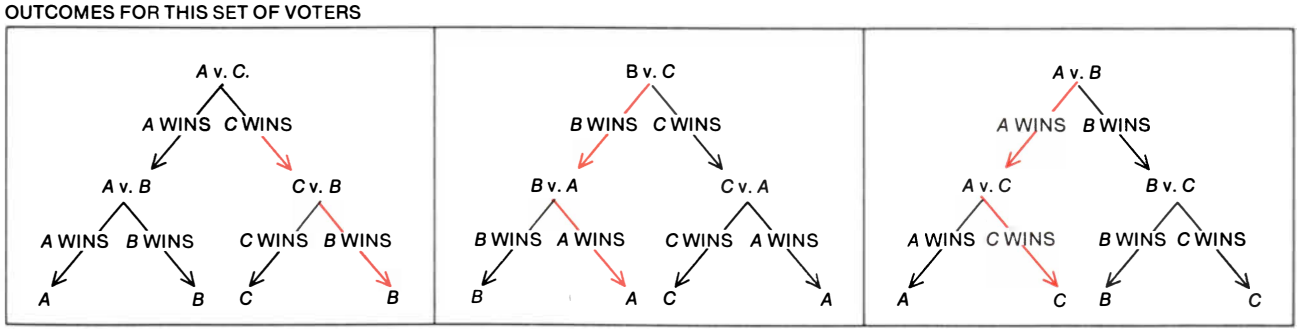
VOTER 1	VOTER 2	VOTER 3
PREFERS *A* TO *B* TO *C*	PREFERS *B* TO *C* TO *A*	PREFERS *C* TO *A* TO *B*

A = ORIGINAL BILL
B = STATUS QUO
C = AMENDED BILL

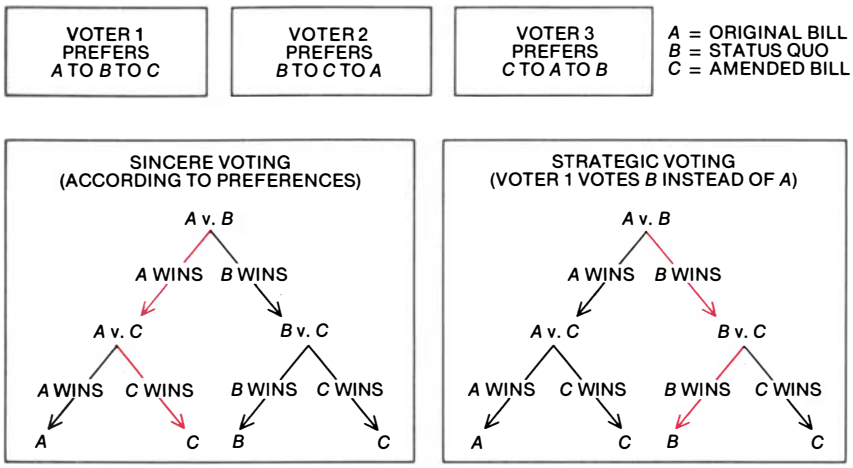
POSSIBLE OUTCOMES



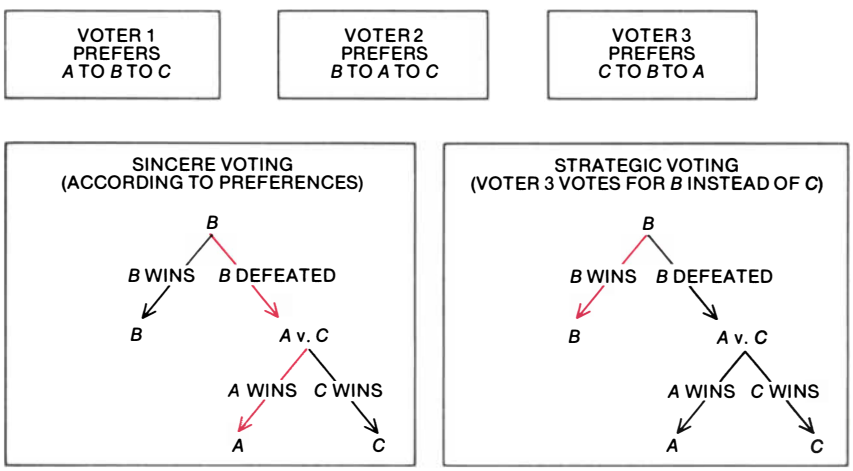
OUTCOMES FOR THIS SET OF VOTERS



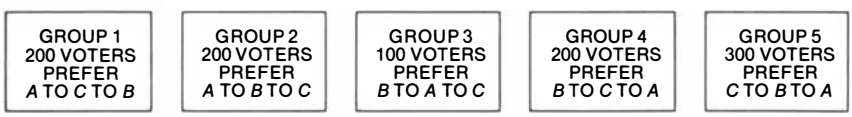
CHANGE IN THE ORDER OF VOTING on a bill and an amendment to the bill can alter the outcome. The voters are assumed to cast their vote for the alternative that is highest on their schedule of preferences. Depending on the order in which the alternatives are voted on, a legislature could pass the original bill, the amended bill or retain the status quo. Under these circumstances the later an alternative enters the voting, the greater its chances of being adopted are. Colored arrows depict course of voting for the possible voting orders.



STRATEGIC VOTING by a voter can alter the outcome in a more favorable direction. If Voter *1* casts his votes according to his preferences, his least preferred alternative will win (*lower left*). If, however, he votes in a "sophisticated" way, his middle preference wins (*lower right*).



SUCCESSIVE-PROCEDURE VOTING RULES followed in some Scandinavian legislatures call for one alternative to be voted on first. If the alternative wins a majority of the votes, the voting stops. If it does not win a majority, the remaining alternatives are paired against each other. In this case, where *B* is put up first, if the voters vote sincerely according to their preferences, *B* will lose two to one and then *A* will win in a two-way contest with *C*. If, however, Voter *3* votes strategically on first ballot, he can alter outcome so his middle preference wins.



	GROUP 1	GROUP 2	GROUP 3	GROUP 4	GROUP 5	OUTCOME
SINCERE VOTING (ACCORDING TO PREFERENCES)	200 VOTES FOR *A*	200 VOTES FOR *A*	100 VOTES FOR *B*	200 VOTES FOR *B*	300 VOTES FOR *C*	*A* WINS
STRATEGIC VOTING (SOME OF GROUP 5 VOTE FOR *B* INSTEAD OF *C*)	200 VOTES FOR *A*	200 VOTES FOR *A*	100 VOTES FOR *B*	200 VOTES FOR *B*	100 VOTES FOR *C* 200 VOTES FOR *B*	*B* WINS

STRATEGIC VOTING IN POPULAR ELECTIONS where the plurality method is used to determine the winner can lead to the defeat of the candidate who would win if all the voters cast their ballots according to their preferences. In this example, if enough voters in Group 5 are aware that their most preferred candidate is certain to lose, they can switch their votes to their middle preference. In this way they ensure the defeat of the candidate whom they least prefer.

Constitution, which provided for the direct election of senators, was delayed for 10 years by this device. An amendment to the proposed Constitutional Amendment was introduced and passed. Since the amendment to the Amendment was unfavorable to the South, however, Southern senators who had favored the original resolution voted against the Amendment as it was amended, and the Amendment was defeated. Only when the makeup of Congress changed did the Amendment pass.

A closely related line of inquiry concerns strategic voting. Under some circumstances a voter can vote contrary to his own preference and alter the outcome in a direction more favorable to himself. This phenomenon was first described by Duncan Black of University College of North Wales. Let us examine a simple case. Smith prefers *A* to *B* to *C*. Jones prefers *C* to *A* to *B* and Brown prefers *B* to *C* to *A*. If the voting follows common amendment-procedure rules, the winner of the first vote is set against the remaining alternative. Suppose the voting sequence is *A* v. *B*, with the winner set against *C*. If everyone votes sincerely, that is, in accordance with his preference at each step, *C* would be the final winner. If Smith votes against *A* on the first ballot, however, then *B*, his second choice, would be the ultimate winner. By voting in this "sophisticated" way, Smith ensures that his second preference rather than his third preference will win.

Sophisticated or strategic voting can also alter the outcome of a vote under other procedures. For example, it can be applied in the "successive procedure" rules followed in Scandinavian legislatures. Under this procedure when there are several competing proposals, one alternative is voted on first. If it wins a majority of the vote, the voting stops. If it does not win a majority, a second alternative is voted on, and if it wins, the voting stops. When there are only two alternatives remaining, they are paired against each other and voted on. Such a voting procedure obviously offers ample opportunity for strategic voting.

Strategic voting can also be applied in popular elections where the plurality method is chosen to determine the winner. And with the advent of public-opinion polls strategic maneuvering in elections has become much more feasible. The possibility of such maneuvering arose in New York during the 1970 election of a U.S. senator. There were three candidates, Richard L. Ottinger (Democrat), Charles E. Goodell (Republican) and James L. Buckley (Conservative). Both Ottinger and Goodell appealed to more liberal voters; Buckley was supported by conservative voters. As the election drew near Ottinger and Buckley led in the opinion polls. On the basis of these polls Ottinger supporters tried to persuade the Goodell supporters to vote for Ottinger on the assumption that they most preferred Goodell and least preferred Buckley. The appeal

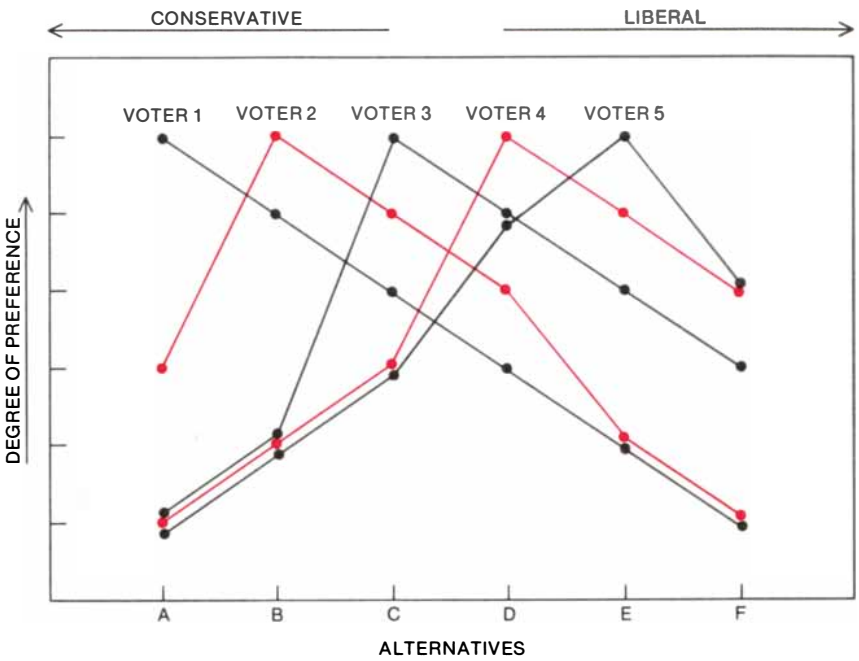
failed to persuade enough Goodell supporters, and Buckley won the election.

It should not be assumed that strategic voting is necessarily bad. Recently one of us (Niemi) and Richard McKelvey of Carnegie-Mellon University demonstrated that if one alternative can defeat all others in pairwise voting, that alternative will necessarily be selected by sophisticated voting. Sincere voting, however, may yield another choice. What this suggests is that strategic voting, although devious, does not merely serve individual ends. Under some circumstances it may actually force the selection of the best choice.

We have been describing the paradox of voting and strategic voting on a largely theoretical basis. One might wonder if these two aspects of voting are matters of genuine importance or are simply curiosities that do not turn up very often. Attempts to resolve the question have been inconclusive. Both empirical observations and mathematical simulations have been employed, but there are difficulties in each approach. Empirical observations are hindered by a lack of necessary data. Votes taken under most procedures do not make it possible to reconstruct the individual voters' complete preference orderings. Simulations must include assumptions about the distribution of voters' preferences, and there is no obvious way to verify those assumptions.

One fruitful approach is to ask whether or not there are conditions under which the paradox of voting and strategic voting are impossible. It turns out that there are such conditions, but some of them are so narrow that they are of no practical significance. There is, however, one well-known condition that prevents the paradox of voting from arising. This condition is "single-peakedness," and it was discovered by Black in the 1940's. It usually arises when the alternatives (candidates or issues) are evaluated in terms of a single criterion. For example, voters might evaluate candidates solely in terms of how they stand on the question of whether or not busing should be employed to achieve integration in the schools. Or the candidates might be judged on the degree of their liberalism.

Single-peakedness and its relation to the paradox of voting was an important discovery. It suggests why the voting paradox is less of a problem in the real world than it might be. If voters share a common frame of reference in their analysis of a problem, their preferences will often be single-peaked, although they may differ sharply in their most preferred alternative. An example is the liberal-conservative continuum often used to explain voting behavior. Although individual voters may differ in their first choice, the existence of a single dimension on which they can judge the alternatives usually implies single-peaked preferences. Under such conditions the voting paradox cannot occur.



SINGLE-PEAKED PREFERENCES can occur when voters evaluate the alternatives in an election in terms of a single criterion. If all the voters have single-peaked preferences, the paradox of voting cannot occur. The alternative preferred by the median voter (in this case Voter *3*) can beat each of the other alternatives in a two-way election. Because the preference orderings of voters are often single-peaked the voting paradox may be less of a problem than it could be.

In Black's original formulation every preference ordering had to be single-peaked if the paradox was to be avoided. One of us (Niemi) has extended the concept by a simulation approach. The simulation suggests that the paradox is made more unlikely the more closely single-peakedness is approximated. As a result if most but not all of the voters evaluate a problem by means of a common dimension, the paradox, although possible, is unlikely to arise.

The difficulty with single-peakedness and similar conditions is that they are not actual voting systems but restrictions on the preference orderings that can exist. Hence they do not solve the underlying problem that majority rule, operating by itself, fails to prevent the occurrence of the paradox of voting. In fact, Kenneth J. Arrow of Harvard University has shown that if, along with a few other seemingly innocuous conditions, individuals can hold any preference ordering of the feasible alternatives, there is no system that precludes the paradox.

Part of the reason for the susceptibility of voting systems to the voting paradox and to strategic voting is that voting mechanisms allowing one unweighted vote per person cannot take into account all the relations among three or more alternatives. To overcome this problem many students of it have proposed voting systems in which entire preference orders are summed. Although such systems vary in detail, they all involve some kind of weighting of positions in the preference order. For example, in one system, dealing with m alternatives, the most preferred alternative is given some maximum number of points, say $m - 1$, the next most preferred is given $m - 2$ and so on down to the last alternative, which is given zero points. For each alternative the points are summed over all individuals, and the alternative with the highest total is declared the winner. The advantage of such a system is that more than only first-place choices are considered in the voting. Analogous weighting systems are used to determine winners in some sporting contests, such as track-and-field meets. Five points may be given for a first place, three for a second, two for a third and one for a fourth. A team with one excellent performer in each event might win most of the events but lose the meet to a team with several strong contenders in each event.

Whatever the justification may be for such counting systems, there are difficulties with them. For one thing they are at least as vulnerable to strategic voting as unweighted systems. And they are perhaps even more prone than other systems to manipulation by the inclusion or exclusion of alternatives.

Where does all of this leave us in the evaluation of voting systems? Arrow's analysis has shown that the paradox of voting cannot be avoided without violating conditions that many, if not most, people would regard as being reasonable. Moreover, Allan F. Gibbard of the University of Pittsburgh, Mark A. Satterthwaite of Northwestern University and Satish Jain of the University of Rochester have recently shown that no reasonable system of voting

VOTER 1 PREFERS *A* TO *B* TO *C* TO *D*	VOTER 2 PREFERS *B* TO *A* TO *D* TO *C*	VOTER 3 PREFERS *D* TO *C* TO *B* TO *A*

COPELAND VOTING

	VOTER 1	VOTER 2	VOTER 3	RESULT
A v. *B*	*A*	*B*	*B*	*B* WINS
A v. *C*	*A*	*A*	*C*	*A* WINS
A v. *D*	*A*	*A*	*D*	*A* WINS
B v. *C*	*B*	*B*	*C*	*B* WINS
B v. *D*	*B*	*B*	*D*	*B* WINS
C v. *D*	*C*	*D*	*D*	*D* WINS

POINTS (WINS MINUS LOSSES)

A: 2 - 1 = 1
B: 3 - 0 = 3
C: 0 - 3 = -3
D: 1 - 2 = -1

FINAL RESULT: *B* WINS

COPELAND VOTING STANDARD compares alternatives with one another in pairwise contests. The score is calculated by subtracting number of losses for each alternative from its wins.

is immune to manipulation. Under every ordinal voting process there are some circumstances in which voters, by voting contrary to their preferences at some step, can obtain a more preferred outcome than the outcome that would be selected by sincere voting.

Ironically the consequence of the fact that all systems of voting are manipulable is that manipulability is partly diminished as an issue. If all systems can be manipulated, then systems cannot be distinguished by the property of manipulability. One might want to rule out the most easily manipulated systems, but among the less manipulable systems one should probably try to find some other standard for judging among them. That other standard must be the joint consideration of what we want voting to accomplish and how well the system under consideration is likely to accomplish it. The minimum we want voting to accomplish is a clear-cut decision. In order to make the choice of some alternatives over others palatable, however, there must be a consensus that the alternative chosen by the voting system is and ought to be the winner.

Condorcet was aware of the problem and proposed the following standard. The alternative that wins by a simple majority over all other alternatives in pairwise contests is declared the winner. Certainly the candidate who in pairwise contests beats everyone else has some claim to winning. Another standard, proposed in the 18th century by Jean Charles de Borda, involves positional weighting. Under the Borda standard, after the points for each alternative are summed for all the voters the alternative with the most points wins. The rationale is to incorporate into the decision the relative preferences of the voters among all the alternatives rather than merely first choices and the pairwise preferences.

These two standards do not always lead to the same outcome. In our judgment the Condorcet standard is more likely than the Borda one to give rise to a consensus, since the Borda standard can easily leave a majority of the voters dissatisfied with the outcome. The Condorcet standard, however, does not always yield an unambiguous winner. Whenever there is a voting paradox, the Condorcet standard yields no winner.

To remedy the defect in the Condorcet standard Black suggested that both standards be used: select a Condorcet winner if one exists, and if one does not, select the Borda winner. Black's hybrid appears to be reasonable until one examines the paradoxes involved in the Borda standard. For example, since the outcome of the Borda count is highly dependent on the position of alternatives in each voter's preference schedule, the results can be manipulated by excluding an alternative. Indeed, an example has been constructed by Charles R. Plott and John A. Ferejohn of the California Institute of Technology where the winning order of the candidates in an election is reversed by the removal of the last-place candidate. Surely there cannot be much of a consensus if the outcome is in some way a function of just what set of alternatives happens to be presented.

Black's hybrid therefore suffers from all the defects of the Borda method. Yet the problem remains that the Condorcet standard often does not yield a winner. Another voting standard was offered by the late A. H. Copeland of the University of Michigan. Under the Copeland standard all the alternatives to be voted on are paired with each other. The score for each alternative is calculated by subtracting the number of its losses from the number of its wins. The alternative with the highest score is the winner. The Copeland standard is thoroughly in the spirit of Condorcet. If there is a winner under the Condorcet rule, it will also be the Copeland winner. If there is no Condorcet winner in a particular situation, the Copeland method will select as the winner the alternative that can win the most pairwise contests.

There are situations, however, where the Copeland standard will select an alternative that loses in a pairwise contest with the

GROUP 1 200 VOTERS PREFER *A* TO *C* TO *B*	GROUP 2 200 VOTERS PREFER *A* TO *B* TO *C*	GROUP 3 100 VOTERS PREFER *B* TO *A* TO *C*	GROUP 4 200 VOTERS PREFER *B* TO *C* TO *A*	GROUP 5 300 VOTERS PREFER *C* TO *B* TO *A*

WEIGHTED VOTING FORMULA: $N_1 (m - 1) + N_2 (m - 2) + N_3 (m - 3)$

N_1 = NUMBER OF 1ST-PLACE VOTES
N_2 = NUMBER OF 2ND-PLACE VOTES
N_3 = NUMBER OF 3RD-PLACE VOTES
m = NUMBER OF CANDIDATES

SINCERE VOTING (ACCORDING TO PREFERENCES)	POINTS FOR *A*: 400 (3 - 1) + 100 (3 - 2) + 500 (3 - 3) = 900 POINTS FOR *B*: 300 (3 - 1) + 500 (3 - 2) + 200 (3 - 3) = 1,100 POINTS FOR *C*: 300 (3 - 1) + 400 (3 - 2) + 300 (3 - 3) = 1,000 *B* WINS
STRATEGIC VOTING (GROUP 1 VOTERS VOTE *C, A, B,* INSTEAD OF *A, C, B*)	POINTS FOR *A*: 200 (3 - 1) + 300 (3 - 2) + 500 (3 - 3) = 700 POINTS FOR *B*: 300 (3 - 1) + 500 (3 - 2) + 200 (3 - 3) = 1,100 POINTS FOR *C*: 500 (3 - 1) + 200 (3 - 2) + 300 (3 - 3) = 1,200 *C* WINS

WEIGHTED VOTING SYSTEMS take into account each voter's order of preference among three or more alternatives. In one such system, if there are m alternatives, the most preferred alternative is given $m - 1$ points, the second most preferred is given $m - 2$ points and so on. The winner is the alternative with the highest number of points. These systems, however, are particularly vulnerable to manipulation by strategic voting. For example, voters in Group 1 can defeat their least preferred alternative by listing *C, A, B* as preferred order instead of *A, C, B.*

Borda winner. This possibility somewhat dampens our enthusiasm for the Copeland method. Nevertheless, in view of the fact that no method is entirely satisfactory, our overall preference is for the Copeland standard because it seems to us that in most instances it minimizes the number of dissatisfied voters.

In the real world, however, the adoption of a voting system is determined largely by considerations of expense, convenience and familiarity. Most of the voting systems in service depart significantly from the voting standards we have described. For example, plurality voting, which is the least expensive system, is a truncated version of the Borda system where only first-place preferences are counted. Systems that call for a runoff contest between the two top vote-getting candidates in a plurality vote attempt to compensate for the serious defects of plurality voting by setting a Condorcet standard for the final winner. As we have pointed out, however, it is likely that the winner under this dual system could be defeated in a pairwise contest by one of the alternatives in the first election. Sequential voting in legislatures is a rough approximation of Condorcet voting, but since under most circumstances not every motion is set against every other motion, or even against the status quo, it is possible that a motion eliminated in the early voting could beat the final choice.

The deviations of actual voting methods from theoretical standards can be justified on the grounds of economy and familiarity. Nevertheless, it should be realized that there are better voting systems and worse ones. For example, the Chilean tradition of having the legislature choose as president the candidate with a plurality of the popular vote resulted in the election of Salvador Allende, who had only a little more than a third of the popular vote. If we assume that many of the voters who had not voted for Allende had him at the bottom of their preference schedule, it is perhaps not surprising that his election set the scene for prolonged political strife.

Given examples of this kind, there is good reason to choose a voting system that meets theoretical standards as closely as possible. What has prevented the full application of the Condorcet, Borda or Copeland standards is the technical difficulty of computing the results. With the aid of computers, however, that obstacle can now be overcome. For example, the installation of computerized voting in the U.S. Congress opens the way for changes in the voting system there. In addition it is now fully possible to use any one of the three standards we have outlined, or others, in general elections. The basic data required for all three are a complete schedule of preferences from each voter. Such a schedule could be obtained with voting machines or with punched cards that can be read by computer. The problem that remains is deciding which decision standard is the best.

GROUP 1	GROUP 2	GROUP 3	GROUP 4
300 VOTERS PREFER *B* TO *A* TO *C* TO *D*	300 VOTERS PREFER *B* TO *A* TO *D* TO *C*	300 VOTERS PREFER *A* TO *C* TO *D* TO *B*	200 VOTERS PREFER *A* TO *D* TO *C* TO *B*

CONDORCET VOTING

	GROUP 1	GROUP 2	GROUP 3	GROUP 4	RESULT
A v. *B*	300 VOTES FOR *B*	300 VOTES FOR *B*	300 VOTES FOR *A*	200 VOTES FOR *A*	*B* WINS
A v. *C*	300 VOTES FOR *A*	300 VOTES FOR *A*	300 VOTES FOR *A*	200 VOTES FOR *A*	*A* WINS
A v. *D*	300 VOTES FOR *A*	300 VOTES FOR *A*	300 VOTES FOR *A*	200 VOTES FOR *A*	*A* WINS
B v. *C*	300 VOTES FOR *B*	300 VOTES FOR *B*	300 VOTES FOR *C*	200 VOTES FOR *C*	*B* WINS
B v. *D*	300 VOTES FOR *B*	300 VOTES FOR *B*	300 VOTES FOR *D*	200 VOTES FOR *D*	*B* WINS
C v. *D*	300 VOTES FOR *C*	300 VOTES FOR *D*	300 VOTES FOR *C*	200 VOTES FOR *D*	*C* WINS

FINAL RESULT: *B* WINS

BORDA VOTING (WEIGHTED)

POINTS FOR *A*: $500(4-1) + 600(4-2) + 0(4-3) + 0(4-4) = 2{,}700$
POINTS FOR *B*: $600(4-1) + 0(4-2) + 0(4-3) + 500(4-4) = 1{,}800$
POINTS FOR *C*: $0(4-1) + 300(4-2) + 500(4-3) + 300(4-4) = 1{,}100$
POINTS FOR *D*: $0(4-1) + 200(4-2) + 600(4-3) + 300(4-4) = 1{,}000$

FINAL RESULT: *A* WINS

CONDORCET AND BORDA VOTING STANDARDS can yield different results. Under the Condorcet standard the alternative that beats all other alternatives in pairwise contests is the winner. In this example alternative *B* wins under the Condorcet standard because it beats alternatives *A, C* and *D*. Under the Borda standard of weighted voting, however, alternative *A* wins. The weakness of the Condorcet standard is that it will not produce a winner if paradox of voting is present. Winner under Borda count, however, may leave majority of voters dissatisfied.

GROUP 1	GROUP 2	GROUP 3
400 VOTERS PREFER *Y* TO *X* TO *A* TO *C* TO *B*	300 VOTERS PREFER *B* TO *C* TO *Y* TO *A* TO *X*	200 VOTERS PREFER *X* TO *A* TO *B* TO *C* TO *Y*

CONDORCET VOTING

X BEATS *A, B, C*
Y BEATS *X, A*
A BEATS *B, C*

B BEATS *Y, C*
C BEATS *Y*

FINAL RESULT:
Y, X, A, B, C, Y IN CYCLE
(NO WINNER)

COPELAND VOTING

POINTS FOR *X*: $3 - 1 = 2$
POINTS FOR *Y*: $2 - 2 = 0$
POINTS FOR *A*: $2 - 2 = 0$

POINTS FOR *B*: $2 - 2 = 0$
POINTS FOR *C*: $1 - 3 = -2$
FINAL RESULT: *X* WINS

BORDA VOTING

POINTS FOR *X*: $200(5-1) + 400(5-2) + 0(5-3) + 0(5-4) + 300(5-5) = 2{,}000$
POINTS FOR *Y*: $400(5-1) + 0(5-2) + 300(5-3) + 0(5-4) + 200(5-5) = 2{,}200$
POINTS FOR *A*: $0(5-1) + 200(5-2) + 400(5-3) + 300(5-4) + 0(5-5) = 1{,}700$
POINTS FOR *B*: $300(5-1) + 0(5-2) + 200(5-3) + 0(5-4) + 0(5-5) = 1{,}600$
POINTS FOR *C*: $0(5-1) + 300(5-2) + 0(5-3) + 600(5-4) + 0(5-5) = 1{,}500$

FINAL RESULT: *Y* WINS

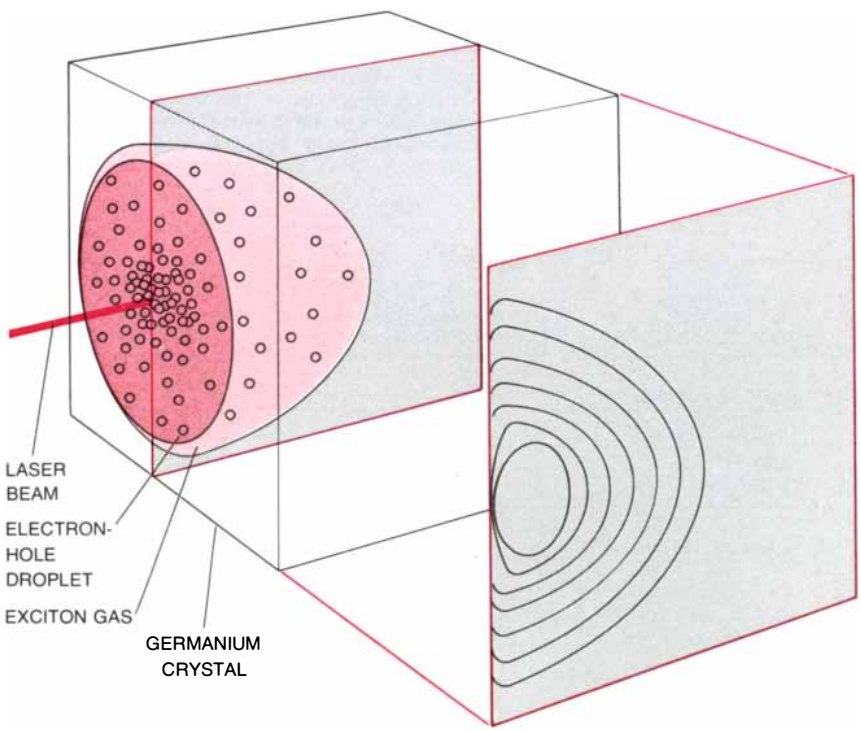
UNDESIRABLE RESULTS can occur under any voting system. In this example devised by Peter C. Fishburn of Pennsylvania State University there is no Condorcet winner because there is a voting paradox. Under the Copeland standard based on pairwise contests *X* is elected. Nevertheless, in a pairwise contest the winner under the Copeland standard loses to the winner under the Borda standard, *Y*. Thus in this particular case the Borda standard might be preferable.

An Electron-Hole Liquid

Inside a semiconductor crystal float glowing droplets of a liquid composed entirely of electrons and "holes." The properties of this quantum-mechanical fluid are revealed by the radiation it emits

by Gordon A. Thomas

Electricity was described by some early investigators as a kind of liquid that flowed through a conductor as a current. With the discovery of the electron this model of electrical phenomena was replaced by more precise theories, and although a few descriptive terms were retained, the notion of an electric fluid went out of fashion. Recently, however, it has been discovered that the carriers of electric charge inside a crystal can exist in a state that is a liquid in a modern sense. This new liquid is not the substance imagined by the early investigators, but it does have many of the properties associated with an ordinary fluid such as water.

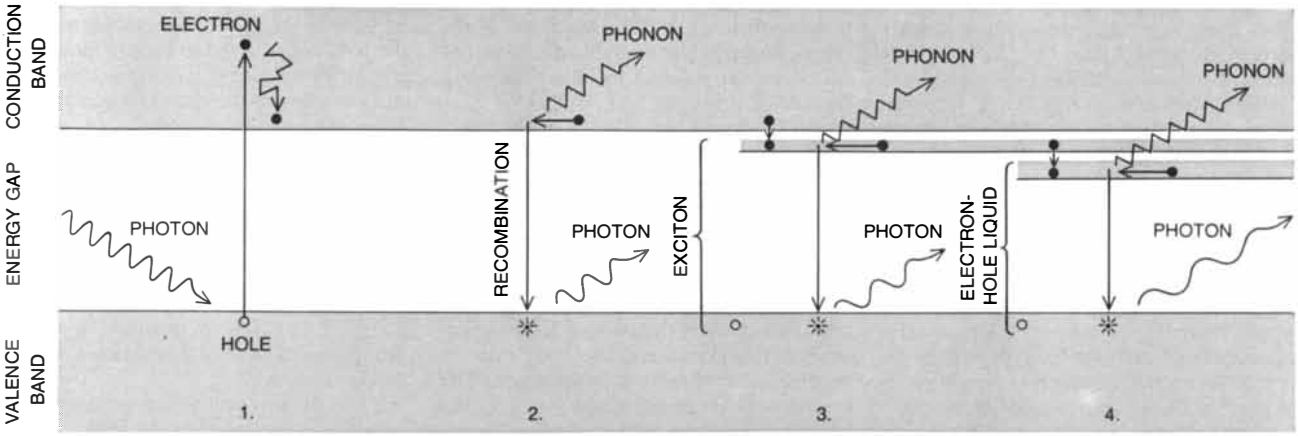


DROPLETS OF LIQUID in a crystal of the semiconductor germanium are composed of electrons and holes, the vacancies caused by the absence of an electron. The specimen of germanium is a very pure one, and it is maintained at low temperature (a few degrees Kelvin). Under these circumstances the electrons and holes are created in large numbers by infrared-laser radiation striking the surface of the crystal. An electron and a hole can bind together to form a two-particle system, analogous to a hydrogen atom, called an exciton. The excitons behave as a gas floating in the crystal, and when their "relative humidity" gets high enough, some of them condense to form the droplets of liquid. Evaporation and condensation continue at the surface of the droplets, maintaining a steady state. The concentration of droplets is greatest near the point where the laser radiation strikes the specimen; the section at right shows contours of constant droplet density measured by investigators at Bell Laboratories. In both the exciton gas and the liquid, electrons and holes recombine with each other continually, emitting infrared radiation observed as luminescence. The supply of electrons and holes is replenished by the laser radiation.

The charged particles that form the new fluid can exist as a vapor; when the relative humidity of the vapor becomes high enough, the particles condense. Like water, the new liquid evaporates and eventually disappears when it is heated or when the density of the particles in the surrounding gas is reduced. It will not form at all at temperatures higher than a critical value. It appears as a cloud of droplets, which scatters light as a fog of water droplets does.

Although the new liquid resembles water in many respects, it is in some ways most unusual. To begin with, it exists only inside a solid semiconductor, and it cannot be extracted from that environment; it is not a liquid that could be poured into a bucket or pumped through an ordinary pipe. Instead of the atoms or molecules of an ordinary liquid, it consists of electrons and "holes," the positively charged voids formed by the absence of an electron. The electrons and holes continually annihilate each other, and in the process they give off infrared radiation. As a result the liquid is an inherently unstable substance; it disappears in a small fraction of a second if the supply of electrons and holes is not continuously replenished. Finally, the new liquid is of particular interest because it is essentially a quantum-mechanical fluid. Effects that are negligible in conventional fluids have a major influence on the behavior of the electron-hole liquid. For these reasons the new liquid offers a unique testing ground for some of the fundamental principles of physics.

Semiconductors are materials whose electrical conductivity lies between that of good conductors, such as metals, and that of insulators. Among the best-known semiconductors are the elements silicon and germanium, and these provide a medium that is well suited to the creation and study of electron-hole droplets. Nearly perfect single crystals of these materials can be grown with dimensions of several centimeters. When the crystals are grown with great care, the crystal lattice is entirely free of dislocations, and the crystals are contaminated with an average of only one electrically active impurity atom in 10^{12} atoms



FREE ELECTRONS AND HOLES are created in a semiconductor when an electron acquires enough energy to cross a "gap" of forbidden energy levels. The energy can be supplied by a photon (a quantum of electromagnetic radiation) that promotes the electron from the valence band to the conduction band (*1*); although the electron may reach a level well above the minimum for the conduction band, it quickly falls back to a level just above the gap. The electron can recombine with a hole (*2*), giving up its energy in the form of a phonon (a quantum of vibrational energy) and a photon; in the process the electron annihilates the hole. Alternatively an electron and a hole can reduce their energy slightly by binding together to form an exciton (*3*). Under certain conditions excitons can reduce their energy further by condensing to form the electron-hole liquid (*4*). The particles that compose the liquid can also decay by the emission of a phonon and a photon. The photons are observed as infrared luminescence and their wavelengths are determined by the energy of the electrons and holes.

of silicon or germanium. In these respects their perfection and purity are unmatched by any other solid.

The fixed lattice of atoms in the semiconductor is an essentially passive environment for the liquid. In the theoretical case of a crystal at absolute zero (that is, at a temperature of zero degrees Kelvin) all the electrons in the crystal are confined in the bonds between the atoms; the electrons are said to be in the valence band, and that is their state of lowest energy. If the temperature is raised, or if energy is added in some other way, some of the electrons can be knocked loose from the valence band and excited to the conduction band. For the electron to make this transition it must absorb an amount of energy greater than a fixed minimum excitation energy because the valence band and the conduction band are separated by a "gap" of forbidden energies. An electron in the conduction band of a semiconductor is a mobile particle, free to wander throughout the crystal, but it differs from a free electron in empty space. For example, the effective mass of a conduction-band electron is much smaller than that of a free electron; in germanium it is about a tenth as great.

When an electron is excited to the conduction band, it leaves behind a vacancy in the valence band: a hole. The hole has a positive charge equal to the electron's negative charge, and like the excited electron it can migrate throughout the crystal. The hole moves by the reshuffling of electrons in the valence band: each time an electron from an adjacent atom "fills" the hole a new hole is created at that atom. In almost all its properties the hole behaves like a positively charged electron.

When electrons and holes are created in pairs, their numbers in a crystal are exactly equal. Both are essential to the creation of the new liquid. It is the collective electrostatic attraction between the electrons and the holes that binds the liquid together.

The electron-hole liquid was first observed in 1966 by the late J. Richard Haynes of Bell Laboratories, although he was not then able to interpret his discovery correctly. He observed luminescence at a wavelength that had not been seen before in a crystal of silicon at extremely low temperature. Two years later V. M. Asnin and A. A. Rogachev of the A. F. Ioffe Physical-Technical Institute in Leningrad, working with a crystal of germanium under similar conditions, saw an unexpected jump in electrical conductivity. These observations led L. V. Keldysh of the P. N. Lebedev Physical Institute in Moscow to the correct conjecture that the new substance is metallic and is composed of electrons and holes.

Haynes was able to observe the luminescence only when the semiconductor was cooled to within a few degrees of absolute zero and when it was illuminated. At such low temperatures a pure semiconductor contains almost no electrons in the conduction band (and hence almost no holes in the valence band). Some electrons and holes must be boosted to an excited state; this is accomplished when the semiconductor absorbs light or infrared radiation of an appropriate wavelength.

Light carries the energy necessary to excite an electron above the "forbidden" gap and into the conduction band; at the same time, of course, a hole is also excited in the valence band. The beam of light waves behaves as if it were a collection of particles—photons. The energy of each photon is determined entirely by the length of the corresponding wave: the shorter the wavelength, the greater the energy. Increasing the intensity of a light beam increases the number of photons in the beam but has no effect on the energy of the individual photons; more energetic photons can be obtained only by employing light with a shorter wavelength.

For a photon to excite an electron and a hole, the photon's energy must at least equal the energy of the forbidden gap. In germanium the width of the gap is about .7 electron volt, and that is the minimum energy a photon must have if it is to create an electron-hole pair. This energy is equivalent to a wavelength of roughly 18,000 angstroms and is in the infrared region of the electromagnetic spectrum. Photons with greater energy (and thus with shorter wavelengths) excite an electron-hole pair to an energy above the minimum level. Photons with less energy (longer wavelengths) cannot raise an electron above the energy gap to create a hole, and they are therefore not absorbed; germanium is transparent to wavelengths longer than about 18,000 angstroms.

Photons with enough energy to excite electrons in germanium or silicon are available in sunlight or from an ordinary light bulb, since all wavelengths of visible light carry greater energy than the energy gaps in these semiconductors. Large numbers of electrons and holes are required to form the new liquid, however, so that the source of the excitation energy must also provide high intensity (a large number of photons). The most suitable source is a laser, which can supply ample quantities of photons, all having the same energy and all focused in the same direction. A neodymium-glass laser, for example, generates an extremely intense beam of infrared radiation in which all the photons have an energy of about 1.17 electron volts.

When a powerful laser is focused on a crystal of germanium, unbound electrons and holes are produced copiously at the surface. Initially the energy of an electron may be far above the minimum required, but by a series of rapid transitions within

the conduction band the electron quickly returns to a level near the bottom of the band. The same process in the valence band returns the hole to an energy near the minimum value. In order to reduce its energy further the electron must then recross the gap in a single step, that is, it must find a hole and recombine with it. On recombination part of the energy of the system appears as a vibration of the crystal lattice; regular vibrations of this kind are described in quantum mechanics by saying that the excited electron and hole emit a phonon, a quantum of vibrational energy. Most of the particles' energy, however, is carried off by a photon. This photon, which has an energy only slightly less than that of the gap, can be observed as luminescence. After it has been emitted all the energy of the original excitation has been dissipated.

Prompt recombination is not the only possible fate of an electron and a hole. Under the appropriate circumstances it is not even the most likely one. In particular, if the semiconductor is pure enough and if the temperature is low enough, the electrons and holes can interact to form a gas or a liquid.

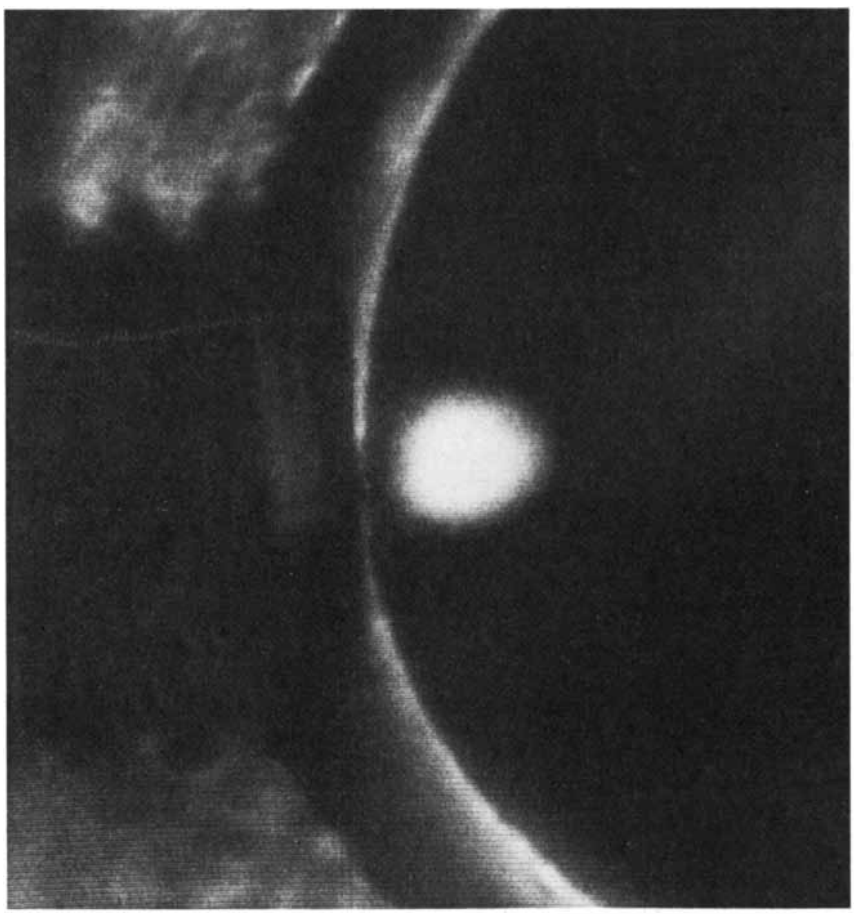
The gas is made up of "atoms" called excitons, each of which consists of a single electron and a single hole bound together by their electrostatic attraction. The exciton is analogous to a hydrogen atom in that it consists of a single positive charge and a single negative charge, but in the exciton the masses of the two particles are approximately equal and the total mass is much smaller than it is in hydrogen. The excitons are created continuously from the electrons and holes generated at the illuminated surface, and they rapidly diffuse into the interior of the crystal. In germanium they form a cloud whose shape is essentially that of a hemisphere, with a radius of about half a millimeter.

The exciton forms because it represents a state of slightly lower energy than the lowest possible energy of an unbound electron and hole. An exciton can survive for as long as 10 microseconds before the electron falls into the hole and the exciton is annihilated. As in the case of unbound particles, the recombination is accompanied by the emission of a phonon and a photon. The wavelength of the emitted radiation indicates precisely the energy of the carriers at the moment they recombine. Since the lowest energy of an exciton is slightly less than that of an unbound electron and hole, the wavelength of the emitted radiation is longer than that produced by the recombination of unbound particles.

If the intensity of the stimulating laser radiation is increased, more pairs of electrons and holes are excited and excitons are formed at a higher rate. Since the effective volume of the hemispherical cloud of gas is determined mainly by the fixed lifetime of the excitons, the additional excitons must be accommodated by an increase in density. When the density exceeds a threshold value, the exciton gas condenses. Droplets of a liquid appear abruptly, just as droplets of water form in the air when the relative humidity becomes high enough. This sudden condensation was first observed in 1969, in a specimen of germanium, by Yaroslav E. Pokrovskii and K. I. Svistunova of the Institute of Radio Engineering and Electronics in Moscow.

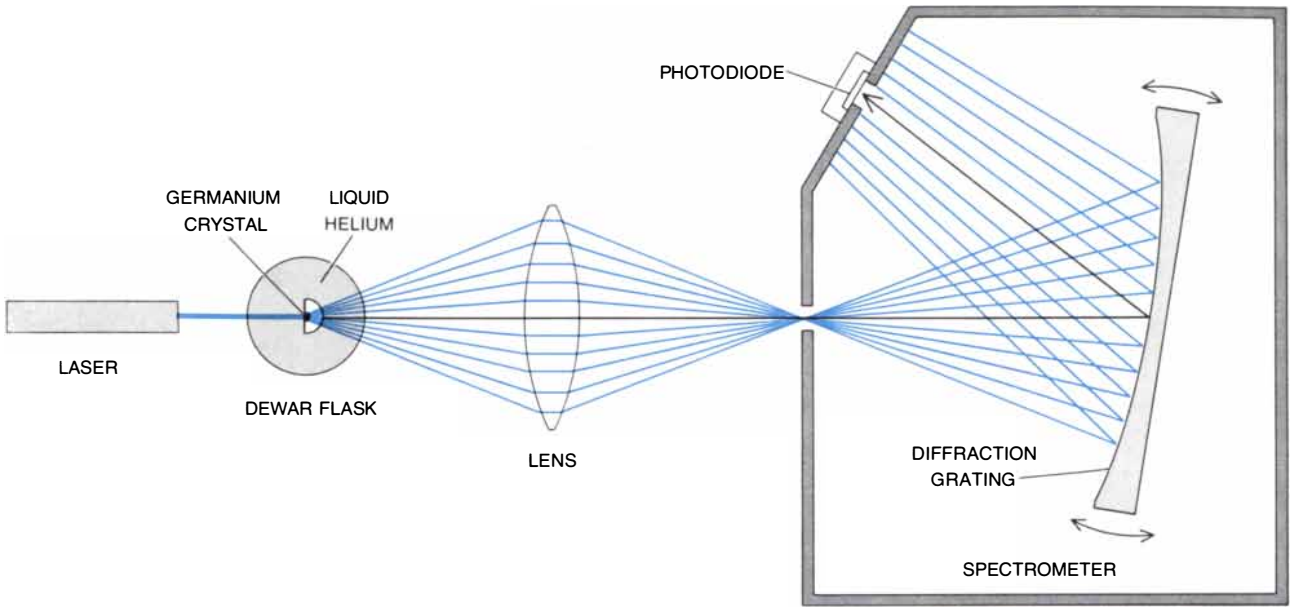
When an exciton enters the liquid state, the electron and hole give up their exclusive association and enter a "sea" of particles in which they are bound equally to all the other charge carriers in the droplet. Thus the liquid is made up not of excitons but of independent electrons and holes. For this reason the liquid is metallic and a good electrical conductor, whereas the exciton gas is an insulator.

Not all the excitons condense when the density of the gas reaches the threshold. The droplets remain surrounded by the gas and in dynamic equilibrium with it. Excitons are created by laser stimulation, and at the same time some are destroyed by recombination. As some excitons condense to form the droplets, others evaporate from the surface of the droplets to maintain the density of the exciton gas. Within the droplets unbound electrons and holes are annihilated by recombination and are replaced through further condensation. If the laser is turned off, this steady state is quickly destroyed. As we have seen, the excitons have a lifetime of only a few microseconds. At low temperature the droplets survive considerably longer, but even they disappear after about 40 microseconds.



GLOWING REGION in a disk of germanium reveals the presence of a fluid similar to the electron-hole liquid. It was photographed by a group of investigators from the University of California at Berkeley, who recorded the infrared luminescence from the fluid with a scanning device similar to a television camera. The specimen is maintained at about two degrees K. and stress is applied along one of the crystal axes by the screw at left. Electrons and holes are created by laser radiation striking the back of the disk and then move to the region of greatest strain, where they are confined. Under these conditions the electron-hole fluid is much less dense than it is in the absence of strain, and the charge carriers have a longer average lifetime before they recombine. The shape of the bright region is defined by the strain distribution.

Like the exciton gas, the electron-hole liquid forms because it represents a state of reduced energy. The combined energy of an electron and a hole in the liquid is slightly less than the energy of an exciton, and thus it is also below the minimum energy of free electrons and holes. One result of the liquid's lower energy is that the radia-



SPECTROSCOPIC METHOD extracts information about the electron-hole liquid from the luminescence emitted as the charge carriers recombine. The specimen of germanium is placed inside a Dewar flask filled with liquid helium and is illuminated with a neodymium-glass laser. The resulting luminescence is at the infrared wavelengths to which germanium is transparent, and it passes readily out of the crystal. A portion of the emitted radiation is focused into a spectrometer, where it is dispersed according to wavelength by a concave diffraction grating. The intensity at a given wavelength is measured by a photodiode; by rotating the grating all wavelengths can be examined.

tion emitted by recombining electrons and holes in the liquid has somewhat longer wavelengths than the radiation resulting from the decay of excitons. Indeed, it was this difference in wavelengths that made the first detection of the new state possible.

The relative energies of free electrons and holes, excitons and the electron-hole liquid can be understood qualitatively by considering the relation of energy to distance. The energy required to separate an electron from a hole is a function of the distance between them, just as the energy required to raise an object above the ground depends on the height to which the object is lifted. An electron excited to the conduction band is almost freed from the influence of its hole, and in the context of the semiconductor crystal it has been removed to an almost infinite distance. An exciton can form only when the unbound charge carriers come close enough together to "feel" each other's presence, and in the exciton itself they orbit at a distance of about 115 angstroms. This distance is clearly less than infinite, and the difference is sufficient to reduce the potential energy of the exciton to a level .004 electron volt below that of the gap. For this reason the energy of the photons emitted by recombining excitons is about .004 electron volt lower than that of the photons from free electrons and holes.

In the liquid state the density of the charge carriers is greater still, and so the distance between them must be smaller. They have an average spacing of roughly 100 angstroms, corresponding to a potential energy about .008 electron volt less than the exciton energy. It should be pointed out that even in the liquid the average distance between charge carriers is quite large when it is considered on the scale of the crystal lattice. In germanium the distance between atoms is about three angstroms.

The appearance of infrared luminescence at unusually long wavelengths is a sure signal that some new substance with an energy lower than that of the excitons is present in a semiconductor. How has that substance been identified as a liquid made of electrons and holes? This is an important question, which has been answered through three kinds of experiment. First, the scattering of infrared radiation by the new substance indicates that it is segregated from the gas in regions of a separate phase. This information alone, however, cannot reveal whether the substance consists of crystallites of a solid or droplets of a liquid. That it is actually a liquid is shown by observing the effects of heating or cooling the substance: its density varies like that of a liquid and it apparently does not freeze. Finally, it must be determined how the charge carriers are associated in the liquid, and that can be ascertained from the luminescence spectrum. The shape of the spectrum is what would be expected for a fluid made up of independent electrons and holes.

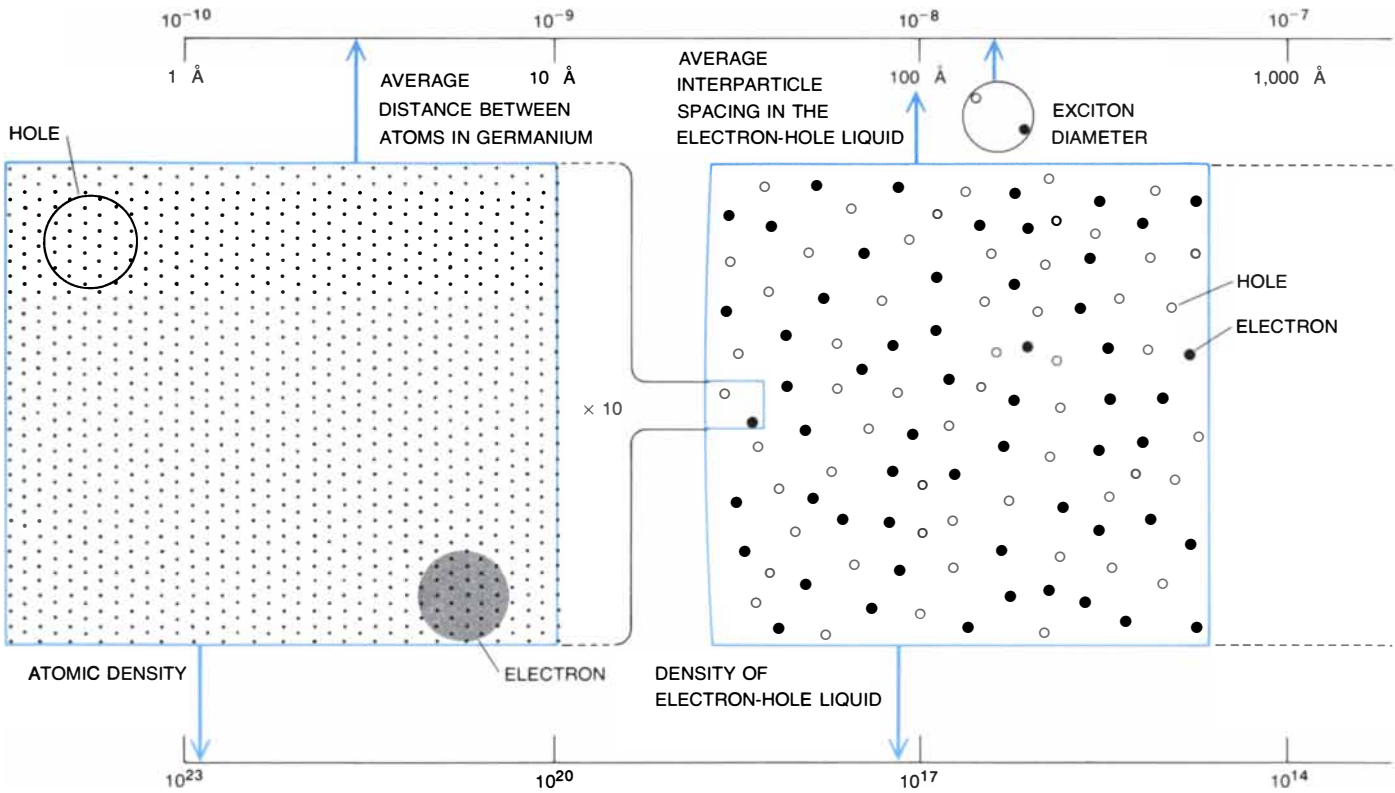
When Haynes first detected the luminescence, he thought he had discovered "molecules" made up of two excitons bound together. Such molecules had been predicted theoretically, whereas an electron-hole liquid had not even been considered. (Molecules probably do form in the exciton gas, but they are too few to be detected.)

The most convincing evidence that the new substance is in a separate phase was obtained in 1971, when Pokrovskii and Svistunova showed that the excited region of the semiconductor scatters a beam of infrared radiation. Such scattering is characteristic of small, distinct objects: it is the same effect that is observed when fog scatters the beams from the headlights of a car. Evidence that these objects have a spherical form would be a strong indication that they have surface tension and therefore that the substance is a liquid. The scattering measurements are not yet precise enough, however, to distinguish the shape of the objects.

As was first explained by Lord Rayleigh in 1871, the range of angles through which light or other electromagnetic radiation is scattered depends on the size of the particles; small particles scatter light through wider angles than large ones do. One can therefore calculate the size of the particles from the pattern of the scattered radiation. This technique reveals that the droplets are quite small: in germanium they have a typical diameter of .004 millimeter. In comparison the water droplets in fog are much larger; they vary widely in size but generally have a diameter of about .1 millimeter.

Scattering experiments have shown that a semiconductor can become densely populated with droplets. A typical hemispherical cloud may contain as many as 100,000. Since the diameter of such a cloud is only about a millimeter, it follows that the droplets must be packed tightly together, and in some cases they may be only .01 millimeter apart. The water droplets in fog, on the other hand, are rarely closer than a few millimeters.

Additional evidence that the new material is in a different phase from the gas was



**CHARACTERISTIC SCALE of electrons and holes suggests how droplets made of them can move freely through the densely packed atoms of a semiconductor crystal. The cloud of droplets formed by laser excitation (*far right*) has a diameter of about a millimeter and contains roughly 100,000 droplets. A single droplet (*second from right*) is .004 millimeter in diameter and consists of about 25 million electrons and an almost equal number of holes. Within a droplet the average distance between particles is about 100 angstroms (*second*

derived from the observation that the droplets absorb radiation in the far-infrared region of the spectrum, that is, at wavelengths considerably longer than those of the infrared absorption and luminescence. As the wavelength is increased the far-infrared absorption reaches a peak; then it declines again and approaches zero at the longest wavelengths studied. This pattern indicates that isolated particles are present.

Most of our information about both excitons and the electron-hole liquid has been acquired by studying the radiation they absorb and emit. By far the richest source of information is the luminescence spectrum. Although the infrared-scattering experiments and the far-infrared absorption suggested that a new phase had formed, it was the luminescence spectrum that completed the picture of the composition of the new substance. Some workers had speculated that the liquid might be made up of excitons or exciton molecules, but the shape of the luminescence spectrum indicated that the liquid is composed of dissociated electrons and holes.

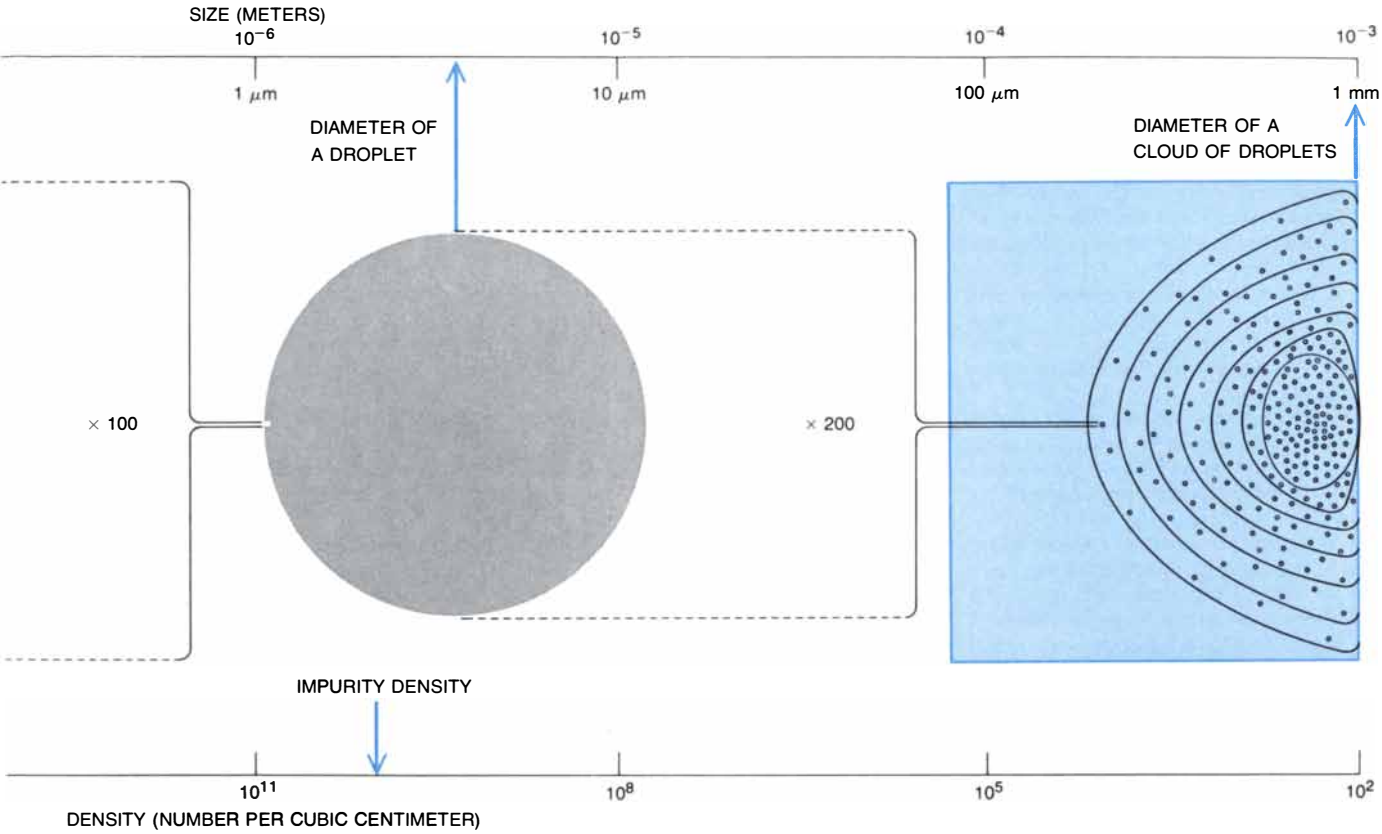
Since the luminescence of excitons and of the electron-hole liquid is in the infrared region of the spectrum, where germanium is transparent, the radiation can readily pass out of the crystal. The infrared radiation is collected by a lens and directed into a spectrometer, where a diffraction grating spreads the beam out according to its constituent wavelengths. A narrow slice of the spectrum, selected by a slit, is then focused on a detector sensitive to infrared radiation, and the intensity of the radiation at that wavelength is recorded. By rotating the diffraction grating the entire luminescence spectrum is swept past the slit.

The spectrum observed from a semiconductor containing electron-hole droplets is dominated by two peaks. One is narrow and is produced by the luminescence of excitons. The second peak is broader and begins at an energy slightly below that of the exciton line. The broader peak represents the infrared radiation emitted by the electron-hole liquid.

The interpretation of this spectrum requires the application of the principles of quantum mechanics and statistical mechanics. In particular, it is necessary to consider the distribution of energies occupied by the charge carriers.

In the vapor phase the excitons are relatively far apart and their collective behavior is approximately that of a "classical" gas. The properties of such a gas were described in the 1860's and 1870's by James Clerk Maxwell and Ludwig Boltzmann. They assumed that the particles of the gas move independently and that their velocities, or energies, are determined by random processes, such as collisions with other particles. The distribution of energies among the particles is then determined by a probability function, which states that the probability of finding a particle with a given energy falls off exponentially with increasing energy, so that most of the particles have comparatively low energies. Most real gases, including water vapor, are well described by this model; their molecules are said to obey Maxwell-Boltzmann statistics.

The excitons also obey Maxwell-Boltzmann statistics, and the probability that an exciton is moving with a given energy above the minimum energy decreases exponentially as the energy is increased. In a measurement of the luminescence spectrum, however, one observes a signal proportional to the number of excitons that recombine with energies within a finite energy interval: the interval fixed by the spectrometer slit. The probability of finding an exciton within such an interval depends not only on the statistical distribution but also on the number of ways an exciton with an energy in that interval can move (or in other words on the number of energy states within the interval). In the exciton gas, as in ordinary gases, the number of such states per unit energy increases with increasing energy. As a result of the combination of the statistical distribution and the density of energy states, few excitons move slowly (at the low-



***from left*); in an exciton the distance is slightly greater. Although these distances are small when they are considered on the scale of the cloud of droplets, they are large compared with the interatomic distances in the crystal lattice (*far left*). Both the electrons and the holes have a large effective diameter: although their size cannot be precisely defined, they are many times larger than a germanium atom. Because of the enormous difference in scale between the droplets and the germanium atoms the droplets float in the crystal lattice like fog.**

est energies). More excitons are found with slightly higher energies because there are more available states. At the highest energies the number of available states increases slowly, but the probability of their being occupied drops off rapidly because of the Maxwell-Boltzmann statistics. As a result the exciton luminescence line is distinguished by a sharp peak.

In the liquid phase the quantum-mechanical constraints on the behavior of the charge carriers are stringent. In each droplet of the liquid the electrons and holes are close enough together so that the entire droplet must be considered a single system. The electrons and holes do not obey the Maxwell-Boltzmann statistics; the probabilities of their occupying the available energy states are prescribed in a way first explained by Enrico Fermi and P. A. M. Dirac. They are said to obey Fermi-Dirac statistics.

The crucial restriction on the probability that a level is occupied in the electron-hole liquid is the exclusion principle introduced by Wolfgang Pauli in 1924 and further elucidated by Fermi and Dirac two years later. The exclusion principle states that no two particles that obey Fermi-Dirac statistics can occupy exactly the same quantum state. Its most familiar application is to the electron orbitals of an atom: if an orbital contains one electron, another can be added only if its spin is oriented in the opposite direction. Once the orbital contains two electrons (with opposite spins) no more can be added under any circumstances. If the atom were to acquire an additional electron, it would have to enter a different orbital.

The same rule applies to both the electrons and the holes in a droplet. The energy levels accessible to them do not form a continuum; rather, the particles must occupy infinitesimally small, discrete levels. The levels are too closely spaced for them to be resolved, but, like the orbitals of an atom, each level can hold only two electrons (or two holes). Once an energy level is filled all succeeding particles must go to higher energy levels. The exclusion principle as it applies to the electron-hole liquid can be restated by saying that the energy of an electron or a hole is determined entirely by its state of motion. The principle then states simply that no two electrons can have the same motion in the same place; in the context of quantum mechanics anywhere in the droplet is the "same place."

Because of the exclusion principle the energy states in the electron-hole liquid are filled in a simple serial manner, starting with the lowest state and continuing in sequence up to the highest state needed to accommodate all the particles in the droplet. In a crystal at absolute zero this uppermost filled state is called the Fermi level. Thus for particles that obey the Fermi-Dirac statistics the probability of finding a particle in a given energy level does not decline exponentially, as it does for particles that obey Maxwell-Boltzmann statistics. For the electron-hole liquid at absolute zero the probability remains constant at one for all states up to the Fermi level; then it falls abruptly to zero. In other words, all states below the Fermi level are certain to be occupied and all states above the Fermi level are certain to be vacant.

As in the case of the exciton gas, the energy distribution of the particles in the liquid can be employed to predict the shape of the luminescence spectrum. Once again there are few energy states available near the minimum energy, and their number increases at higher levels. In the liquid, however, there are two essentially identical energy distributions, one for electrons and one for holes. When an electron and a hole recombine, the difference in their energies is contributed to the emitted photon. The shape of the spectrum is deduced by assuming that any electron is equally likely to fall into any hole. Recombinations in which both particles have very high energy or very low energy will be observed, but they will be relatively rare. In the most common events the difference in the energies will fall near some median value. The result is a compar-

atively broad luminescence spectrum [*see illustration below*]. The predicted shape agrees quite well with the experimentally measured spectrum, offering the most convincing evidence available that the source of the luminescence is correctly interpreted as being a quantum liquid made up of independent electrons and holes. The luminescence spectrum also indicates the difference in energy between the liquid phase and the gaseous phase and the range of energy levels filled in the liquid. These quantities have been deduced theoretically by several groups of investigators, and they are in good agreement with the experimentally measured values.

Finally, from the luminescence spectrum one can calculate the average distance between an electron and a hole. As we have seen, the Fermi level is determined by the number of occupied energy levels in a droplet. It therefore measures the density of the liquid: it counts the number of charge carriers per unit volume. Since the Fermi level is reflected in the width of the luminescence peak, that width is a direct measure of density. In germanium the density both of the electrons and of the holes inside a droplet is 2.4×10^{17} per cubic centimeter. That density yields the average interparticle spacing of about 100 angstroms, and it implies that a typical droplet contains more than 50 million particles. When the electrons and holes are compared with the atomic lattice in which they lie, they are quite sparsely distributed. In germanium there are about 30,000 atoms for each electron or hole.

The electron-hole liquid can be created only in an extremely cold crystal. What happens when the temperature rises? As a rule an increase in temperature is equivalent to an increase in random motion. In the electron-hole liquid heating also induces random motion, but as in other processes the effect is strongly influenced by the quantum-mechanical nature of the fluid. As the temperature rises some particles are excited to energy states above the Fermi level, leaving vacancies in states below that level. The upper boundary of the energy distribution becomes blurred; this effect is apparent in the upper edge of the luminescence spectrum. Heat has the same effect on the electrons in an ordinary metal, but the change in energy is small compared with the Fermi energy of the electrons; in the electron-hole liquid it is drastic. Even a temperature of only four degrees Kelvin induces substantial random motion compared with the slow average motions of the electrons in the liquid. To produce a change of the same magnitude in copper the metal would have to be heated to almost 1,700 degrees K.

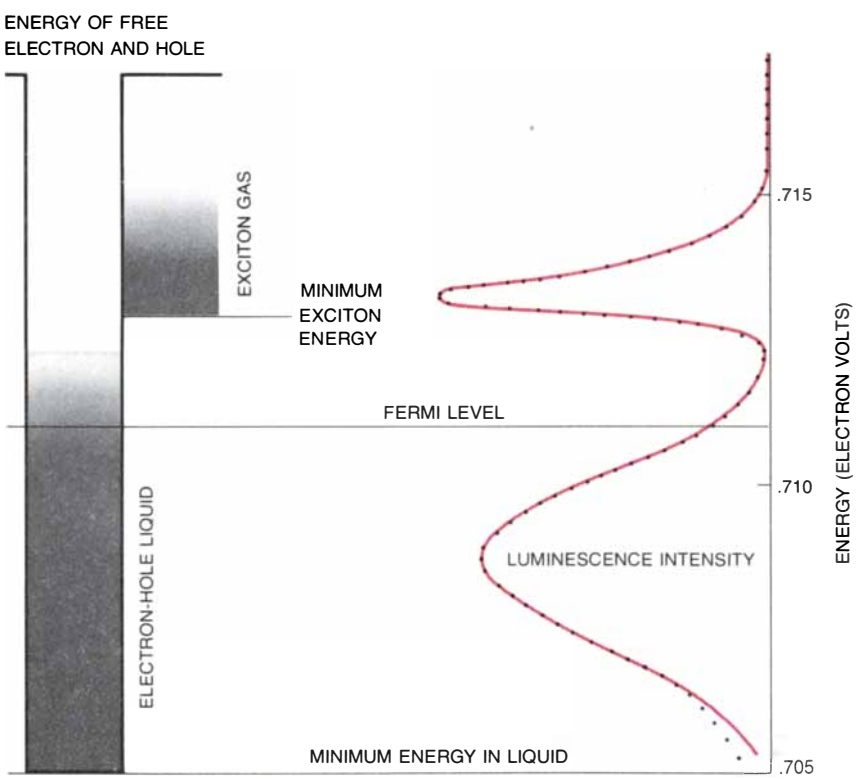
One effect of greater random motion is to increase the average distance between particles and thus to decrease the density: the liquid expands when it is heated, just as most materials do. Once again, however, a change that is of modest scale in an ordinary fluid is enormously magnified in the electron-hole liquid.

The thermal expansion of the liquid is observed as a change in the luminescence spectrum. Because the Fermi level measures the number of particles in a given volume, or the density, expansion must reduce the Fermi level. Since the width of the spectrum is determined by the Fermi level, the expansion produced by heating is observed as a decrease in the width of the peak even though the upper edge becomes broader. The liquid expands by 10 percent when its temperature is raised only four degrees; its coefficient of expansion is more than 100 times greater than that of liquid mercury near room temperature.

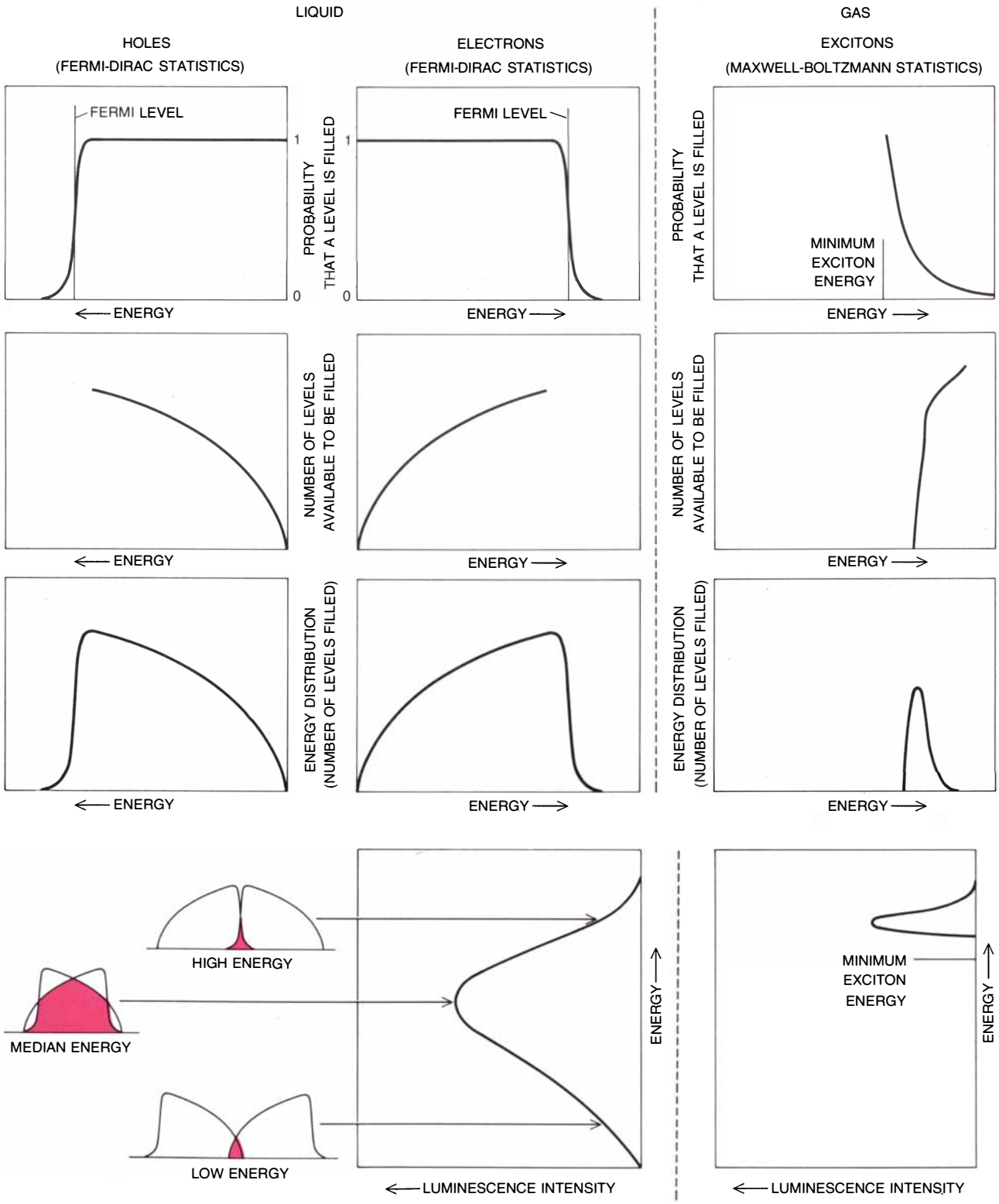
This enormous coefficient of expansion is a result of the nature of the cohesive forces that bind the electrons and holes together and resist random thermal motions. In a more conventional liquid thermal expansion results mainly from the random vibrations of the atoms or molecules and from the strong cohesive forces between them. In metals the free electrons do exert a thermal pressure like that of the electrons and holes, but it is effectively resisted by the large binding forces between the atoms. The electron-hole liquid, on the other hand, floats freely through the lattice of germanium atoms, and the atoms hardly restrain the motion of the charge carriers at all. Both the random motions and the restoring forces can come only from the electrons and holes; the effect is one of purely electronic thermal expansion.

The delicate cohesive forces between electrons and holes also make the liquid extraordinarily soft and compressible. The compressibility of liquid helium 3 (the isotope of helium that has two protons and one neutron) is generally regarded as being remarkably large: it is 10,000 times greater than that of water. The compressibility of the electron-hole liquid in germanium is 10,000 times greater still.

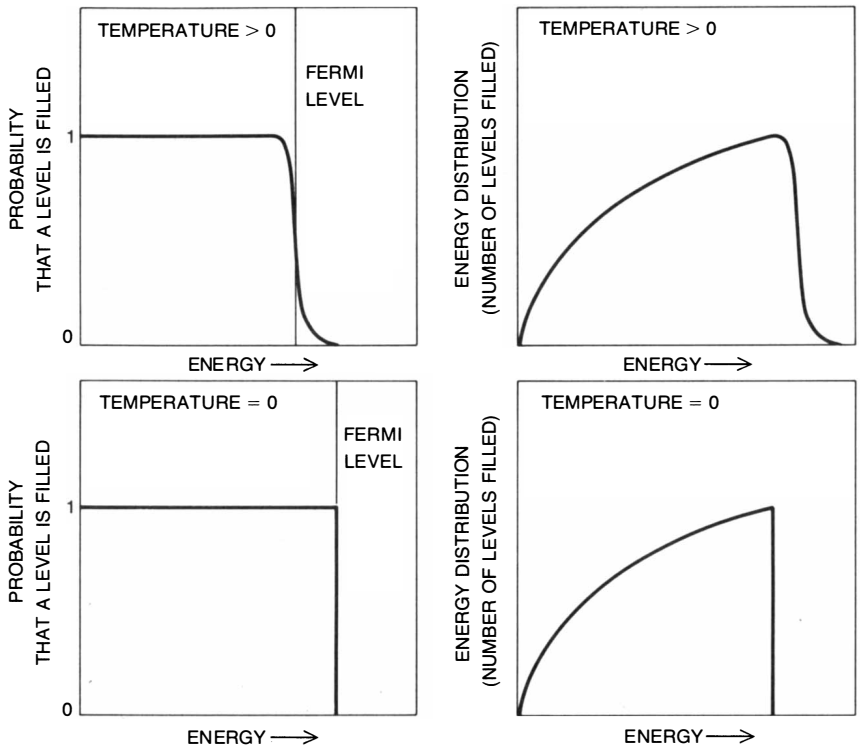
Higher temperature also increases random motion in the exciton gas. Moreover, as the temperature increases, so does the rate of evaporation from the surface of the droplets, so that more excitons are formed. The magnitude of this effect is suggested by the influence of temperature on the lifetime of the droplets. The low-temperature droplet lifetime of about 40 microseconds is determined almost entirely by the rate of recombination within the droplet. Following initial work by Pokrovskii and Svistunova, John C. Hensel, Thomas G. Phillips and T. Maurice Rice of Bell Laboratories demonstrated that at higher temperature evaporation drastically reduc-



LUMINESCENCE SPECTRUM is a rich source of information about the electron-hole liquid. The interpretation of the spectrum is based on the principle that the wavelength of a photon is determined by its energy. Excitons exist at higher energies than the electrons and holes in the liquid, and the luminescence from the exciton gas is therefore observed at shorter wavelengths than that from the liquid. Moreover, the range of energies the excitons occupy is smaller than the range occupied by the charge carriers in the liquid, so that the exciton luminescence peak is the narrower of the two. In the spectrum of the electron-hole liquid the luminescence intensity falls off rapidly above an energy called the Fermi level. For both phases of the fluid the theoretically predicted spectrum (*color*) is in good agreement with the observed one (*black dots*).



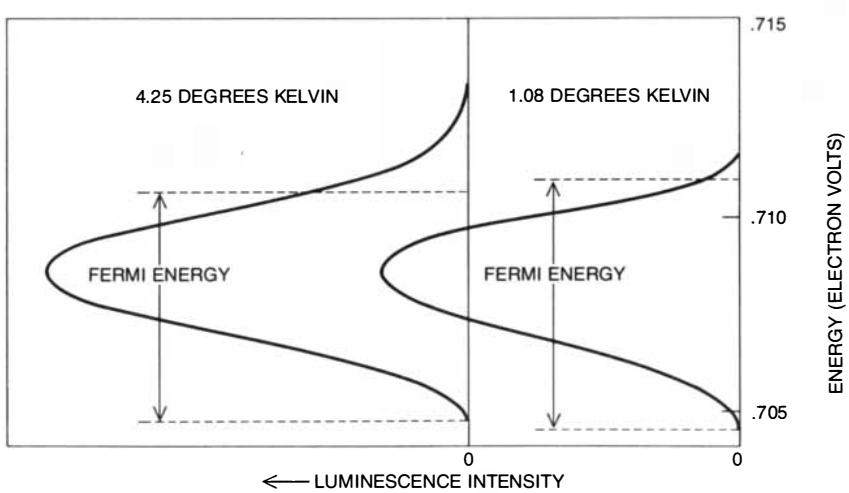
SHAPE OF THE SPECTRUM of radiation emitted by the exciton gas and by the electron-hole liquid can be explained by counting the energy levels occupied in each phase of the fluid. In the gas the probability that a given level is occupied declines exponentially with increasing energy, but the number of levels that are available increases in a unique way with energy; as a result the distribution of energies in the gas has a sharp peak and a distinctive shape, which are reflected in the luminescence spectrum. In the liquid the probability that a level is occupied is essentially constant (at unity) up to about the Fermi level, where it falls rapidly to zero. Again the number of levels available for both electrons and holes increases with energy, and their separate energy distributions are distinctively asymmetrical. Because a given electron can recombine at random with any hole, the shape of the liquid luminescence spectrum is determined by the probabilities of the various possible recombinations. These probabilities can be represented schematically by the areas of various intersections of the energy-distribution curves (*color*). Events in which both particles have very high energy are rare, so that the number of photons at short wavelengths is small; the same effect reduces the intensity of the long-wavelength luminescence from recombinations of two particles with very low energy. In the most common events the energies fall near a median value, so that the spectrum has a broad peak.



INCREASING TEMPERATURE has the effect of increasing random motion in the electron-hole liquid. In the theoretical case of a specimen maintained at absolute zero the probability that an energy level is occupied is one for all levels up to the Fermi level and zero for all higher levels. The resulting energy distribution has a sharp boundary: all available levels up to the Fermi level are occupied and all those above it are vacant. At higher temperature (a few degrees K.) a few particles can acquire enough energy from thermal motion to cross the Fermi level and occupy higher energy states. At the same time the Fermi level itself assumes a lower value.

es the lifetime. They found that at 4.2 degrees K. it is only a few microseconds, comparable to the lifetime of an exciton.

One way of summarizing our knowledge of a substance is to draw a phase diagram for it. Such a diagram, in one common form, indicates in what phase the material will be found at various combinations of temperature and density. The phase diagram for water, for example, shows the transitions between ice, liquid water and water vapor. A diagram that is in many respects similar can be drawn for the electron-hole fluid.



THERMAL EXPANSION of the liquid is observed through changes in the width of the luminescence spectrum. The upper edge of the spectrum grows wider when the temperature is raised because thermal motions impart additional energy to the charge carriers. The interval between the minimum energy and the Fermi level is reduced; this change takes place because the Fermi level is a measure of the density of the liquid and expansion reduces density. These effects are observed in ordinary metals, but in the electron-hole liquid they are greatly exaggerated.

The density of the exciton gas surrounding the droplets can be calculated from the luminescence spectrum. It is derived from the difference in energy between the liquid luminescence line and the gas line, since that is precisely the energy required for an exciton to escape from the droplet. In other words, the luminescence indirectly indicates the minimum density of excitons required for condensation of the droplets. The values of this equilibrium density at each temperature determine what is called the gas edge of the phase boundary.

As the temperature rises the density of the liquid decreases because of thermal expansion and the density of the gas increases because of evaporation. An obvious question is whether the graphs of the two densities meet. By analyzing a large number of luminescence spectra Hensel, Rice and I mapped the phase diagram and found that the graphs do meet: at a temperature of 6.5 degrees K. and at a density that represents an interparticle spacing of about 140 angstroms. At this point, called the critical point, the liquid and the gas become indistinguishable. At higher temperatures the separate liquid phase does not exist, and the density of the fluid can be increased smoothly, without phase transitions.

As Charles Cagniard de la Tour discovered in 1822, water has a critical point near a temperature of 650 degrees K., where the meniscus, or the boundary between the liquid and the gas, disappears. The fact that the critical temperature of the electron-hole liquid is 100 times lower merely reflects the fact that the energies and forces in the new liquid are much smaller than those in water. Because of the fundamental role played by the critical point in our understanding of a substance, at least six theoretical calculations of its value have been made. The reasonable agreement between these predictions and the measured value completes a theoretical description of the liquid that includes successful calculations of all the properties I have discussed here.

Nevertheless, there is one region of the phase diagram that is far from being understood. It is the region of density where the excitons may dissociate to form a gas made up of electrons and holes before condensing to form a liquid. Such a gas would differ from the liquid only in its density. It is possible that the transition can be described as one from an insulator to a metal, a process familiar in other contexts in the physics of semiconductors, but additional studies will be needed to clarify the behavior of the three possible phases of the fluid: the exciton gas, the electron-hole gas and the electron-hole liquid.

Another aspect of the phase diagram that is not understood is the way in which its shape and position change when stress is applied to the semiconductor crystal containing the liquid. One interesting experiment that provides insight into this question is the confinement of a collection of electrons and holes in a region of the crystal that is under strain. Among various tech-

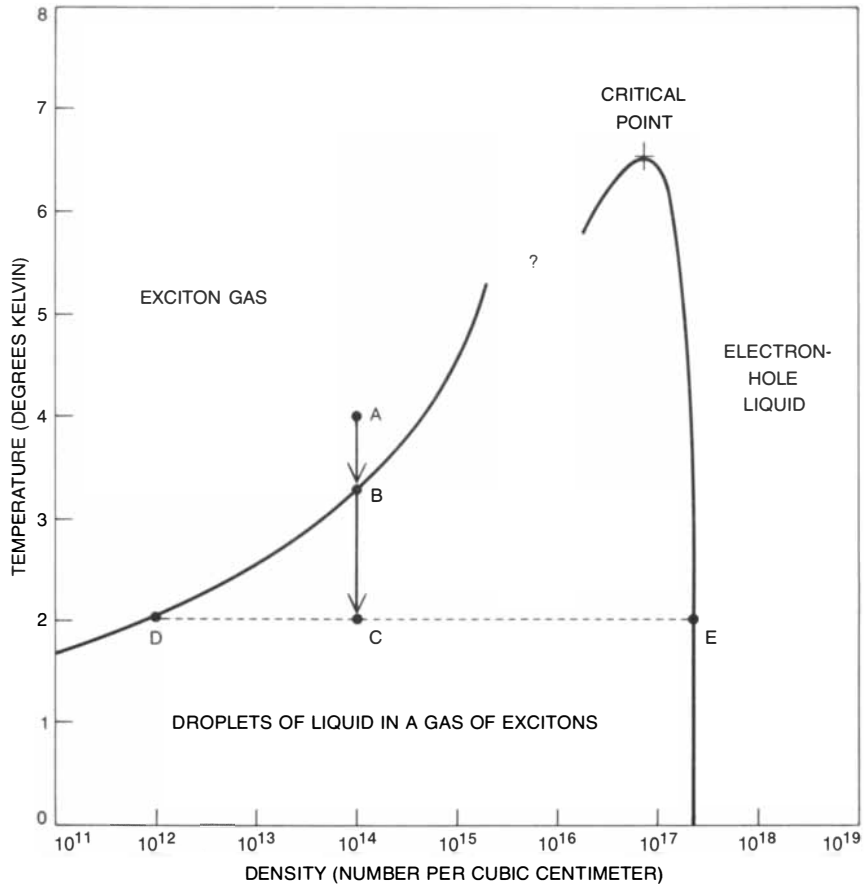
niques used to study the fluid under these conditions, one method is particularly vivid: the fluid has been photographed in its own light. The photograph was made by James P. Wolfe, Walter L. Hansen, Edward E. Haller, Robert S. Markiewicz, Charles Kittel and Carson D. Jeffries of the University of California at Berkeley. They employed an infrared vidicon—a device that is analogous in its operation to a television camera but that is sensitive to infrared radiation. Their semiconductor was a disk of germanium stressed along one of the crystal axes. When the disk was illuminated with laser radiation, the vidicon recorded a bright spot near the perimeter, indicating strong infrared luminescence.

The Berkeley investigators explain the photograph by suggesting that when the crystal is under stress, the cloud of droplets is replaced by a single large drop. V. S. Bagaev and his co-workers at the Lebedev Institute have shown that droplets tend to migrate toward the region of a specimen where the stress is greatest. In the strain-confined fluid, however, the density is much lower than it is in the electron-hole liquid, and the lifetime is up to 10 times longer. The reduced density of the fluid indicates that it differs from the ordinary liquid and that the phase diagram has shifted because of the strain. At present the basic form of the phase diagram under conditions of strain remains unknown.

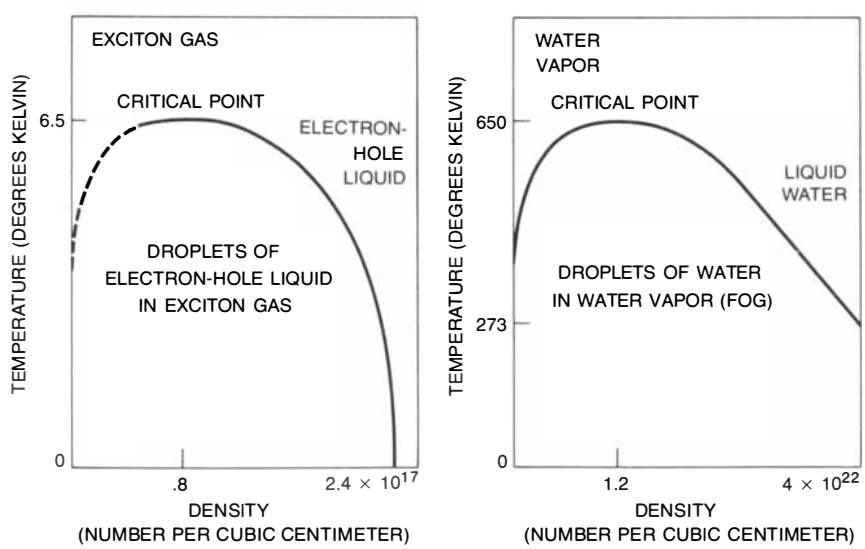
The comparison of the electron-hole phase diagram with that of water raises another obvious question. Water has three common phases, including a solid one. A corresponding solid phase of the electron-hole liquid is not expected. Because the electrons and holes are very light they will probably retain enough random motion to prevent the liquid from freezing no matter how much the temperature is reduced. If the liquid froze as it was cooled, a sharp break would be observed in the phase diagram at the freezing point. No evidence of such a break has been found.

In the decade since it was first observed the electron-hole liquid has been investigated intensively, and the work surveyed in this article by no means exhausts the list of noteworthy discoveries. For example, when a droplet crosses the junction of a diode, it bursts, giving rise to a pulse of charge. The behavior of the liquid is modified by impurity atoms in the semiconductor and by electric and magnetic fields.

Observations remain that are not at all well understood, of which the perplexing region in the phase diagram is only one example. For the most part, however, the evolution of our knowledge of the new liquid has been quite encouraging. It has been possible to predict in detail and with precision the properties of the liquid, and many of these predictions have been verified. This agreement between theory and experiment represents a substantial success in understanding the behavior of large groups of interacting particles.



PHASE DIAGRAM for the electron-hole fluid gives the boundary between the gaseous and liquid phases for various combinations of temperature and density. For example, if the temperature is four degrees K. and the density of electrons and holes is 10^{14} per cubic centimeter, then all the fluid must be in the form of an exciton gas (*point A*). As the fluid is cooled droplets begin to form at about three degrees K. (*B*). With further cooling more excitons condense; as a result the density of the gas is greatly reduced (*D*) but that of the liquid is little changed (*E*). The average density of electrons and holes remains constant at 10^{14} per cubic centimeter (*C*). The diagram has a critical point at 6.5 degrees K.; at higher temperature the liquid does not exist. In the region denoted by a question mark the behavior of the fluid is not well understood.



ELECTRON-HOLE LIQUID AND WATER have phase diagrams that are similar in shape but differ greatly in scale. The density of water (measured in numbers of molecules per cubic centimeter) is 100,000 times greater than that of the electron-hole liquid, and the critical point is at a temperature 100 times higher. Moreover, water freezes (at 273 degrees K.), whereas the electron-hole liquid probably cannot freeze no matter how much the temperature is reduced.

The Purple Membrane of Salt-loving Bacteria

The color is that of rhodopsin, the "visual purple" of the animal eye. In halobacteria rhodopsin serves as the pigment of a newly discovered photosynthetic mechanism that converts light into chemical energy

by Walther Stoeckenius

Life on the earth is based on photosynthesis, through which the energy radiated by the sun is converted to drive the metabolic processes of living organisms. Only plants, some bacteria and the blue-green algae are capable of photosynthesis because only they contain the critical chemical that captures the energy of light: chlorophyll. The absorption of light by chlorophyll initiates the transfer of electrons along a chain of other membrane-bound pigment molecules; the energy released in the course of electron transport is converted into the high-energy bonds of adenosine triphosphate (ATP), the primary energy carrier of living cells, and is thus made available for the synthesis of the earth's basic reserve of chemical energy in the form of starch, cellulose and free oxygen molecules.

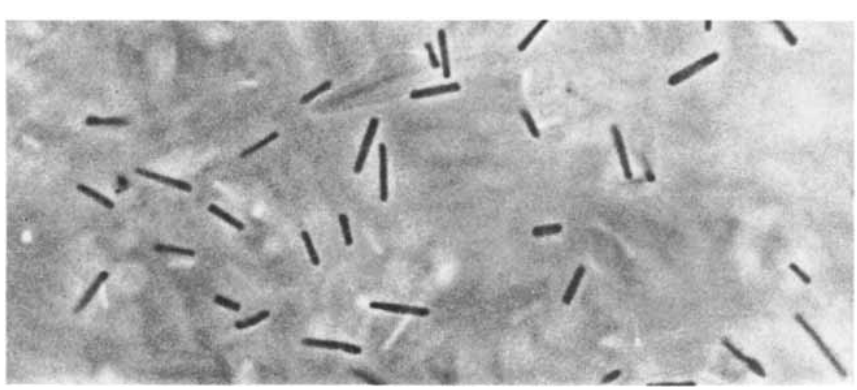
We have recently discovered another photosynthetic mechanism powered by a different pigment, one that is closely related to rhodopsin, which has hitherto been known only as a visual pigment in the eye of animals. This new system is found in the halobacteria: microorganisms that require high concentrations of sodium chloride for growth and that therefore proliferate in natural salt lakes and in salterns, where seawater is evaporated to produce salt. The pigment and the system it powers convert sunlight into the chemical energy the bacteria require for metabolism and for the maintenance of cellular integrity in an environment characterized by low oxygen concentrations, high light intensity and strong osmotic stress.

It was in 1964 that I came across a report that halobacteria lacked the tough outer wall that encloses the cell membrane of most bacteria, and that at low salt concentrations their cell membrane disintegrated into identical subunits of lipoprotein. I had been studying the structure and function of cell membranes for many years, and this sounded like a most unusual membrane; moreover, its composition seemed to contradict some of my basic ideas about how cell membranes are organized. And so, when Robert Rowen joined my laboratory at Rockefeller University later that year, we took a look at the halobacteria. It turned out that the report was in error: the bacteria do not lack a cell wall, and actually it was the glycoprotein cell wall, not the lipoprotein membrane, that was the source of the identical subunits. In checking on the original data, however, we made some new and intriguing observations, and we continued to study the halobacteria.

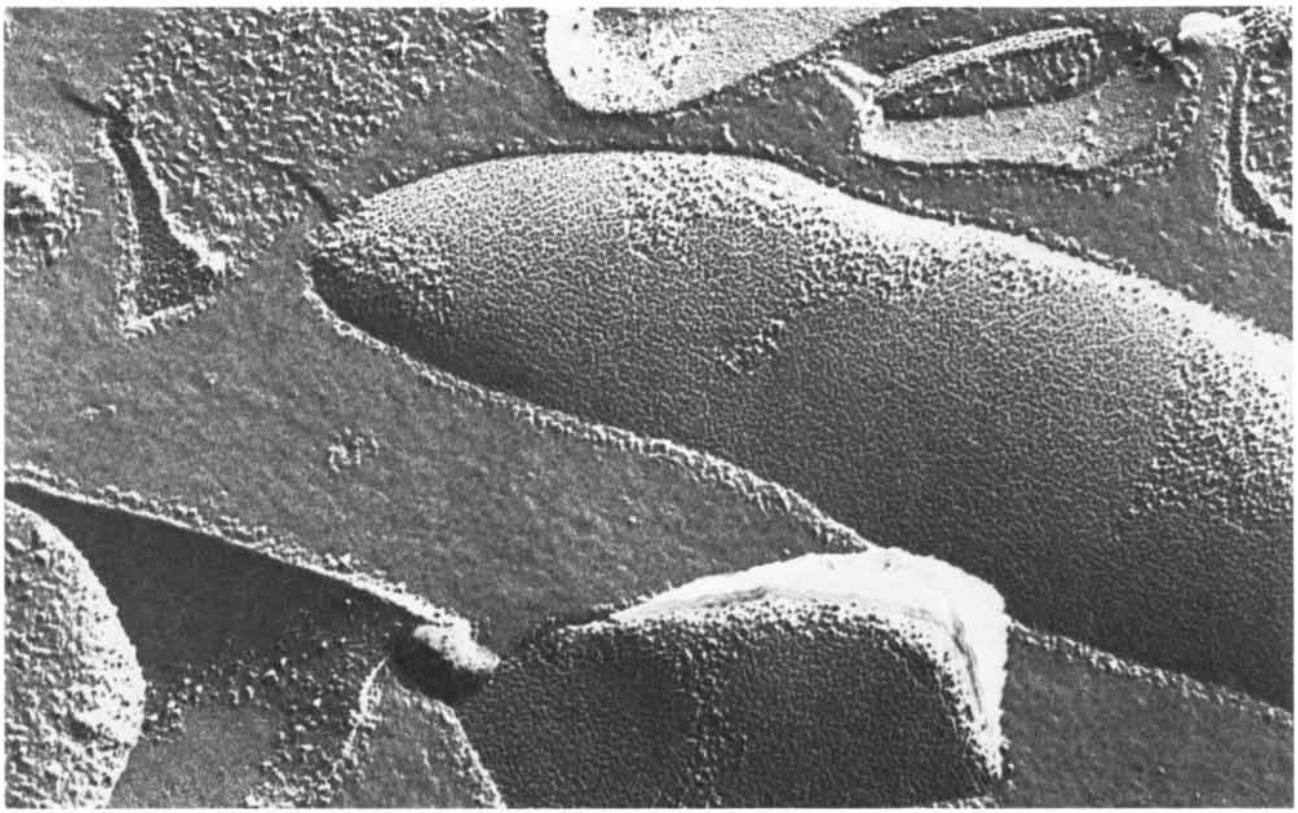
Halobacterium halobium, the species we mainly work with, is a rod-shaped cell about half a micrometer in diameter and perhaps five micrometers long. It grows best in a 4.3-molar solution of sodium chloride, that is, in 4.3 moles, or about 250 grams, of salt per liter of solution. (Ordinary seawater is about a .6-molar solution.) The bacterium is not viable at concentrations below 3.0 molar. In addition to nutrients such as amino acids or peptides (the constituents of proteins) the growth medium requires small concentrations of potassium, magnesium and calcium ions and traces of iron and manganese. The cells do not metabolize sugars anaerobically, and in the absence of such fermentation it was assumed that their only source of energy was oxidative phosphorylation: the process whereby animals and most bacteria convert the energy released during respiration by the oxidation of amino acids, fatty acids and carbohydrates into the high-energy bonds of ATP.

When we suspended *H. halobium* cells in a basal salt mixture (the growth medium without nutrients) and diluted the medium to expose them to decreasing salt concentrations, we found that they first lost their rod shape and then became quite permeable, so that most of the cell contents leaked out. At salt concentrations below 2.0 molar the cell wall began to disintegrate and dissolve. At 1.0 molar the empty sacs of membrane began to break up, but the fragments were not soluble; they ranged in size from about a hundredth of a micrometer to half a micrometer. We found that the way the membrane fragmented was determined primarily by the proteins in it rather than by the lipids, and that fact suggested some interesting possibilities for exploring the relation between structure and function in the membrane.

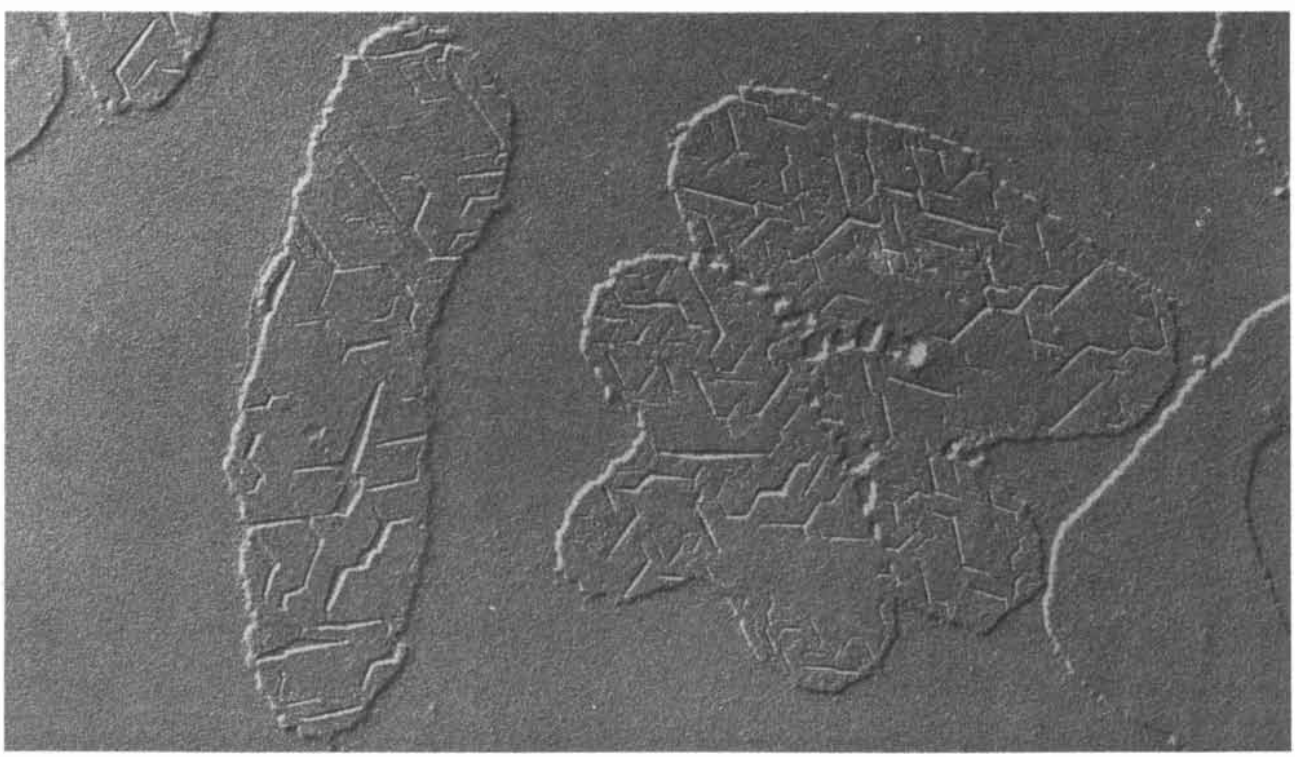
I had the notion, which was not very fashionable just then but is now widely accepted, that the structural backbone and main permeability barrier of cell membranes is a bilayer of lipid molecules. Protein molecules are inserted at various sites in the bilayer, and it is primarily the proteins that mediate the many functions of the membrane such as oxidative phosphorylation, the transport into and out of the cell of



HALOBACTERIA, here enlarged 1,850 diameters, are rod-shaped motile cells; the flagella are not visible. This is the mutant *Halobacterium halobium* R_1, with which the author works.



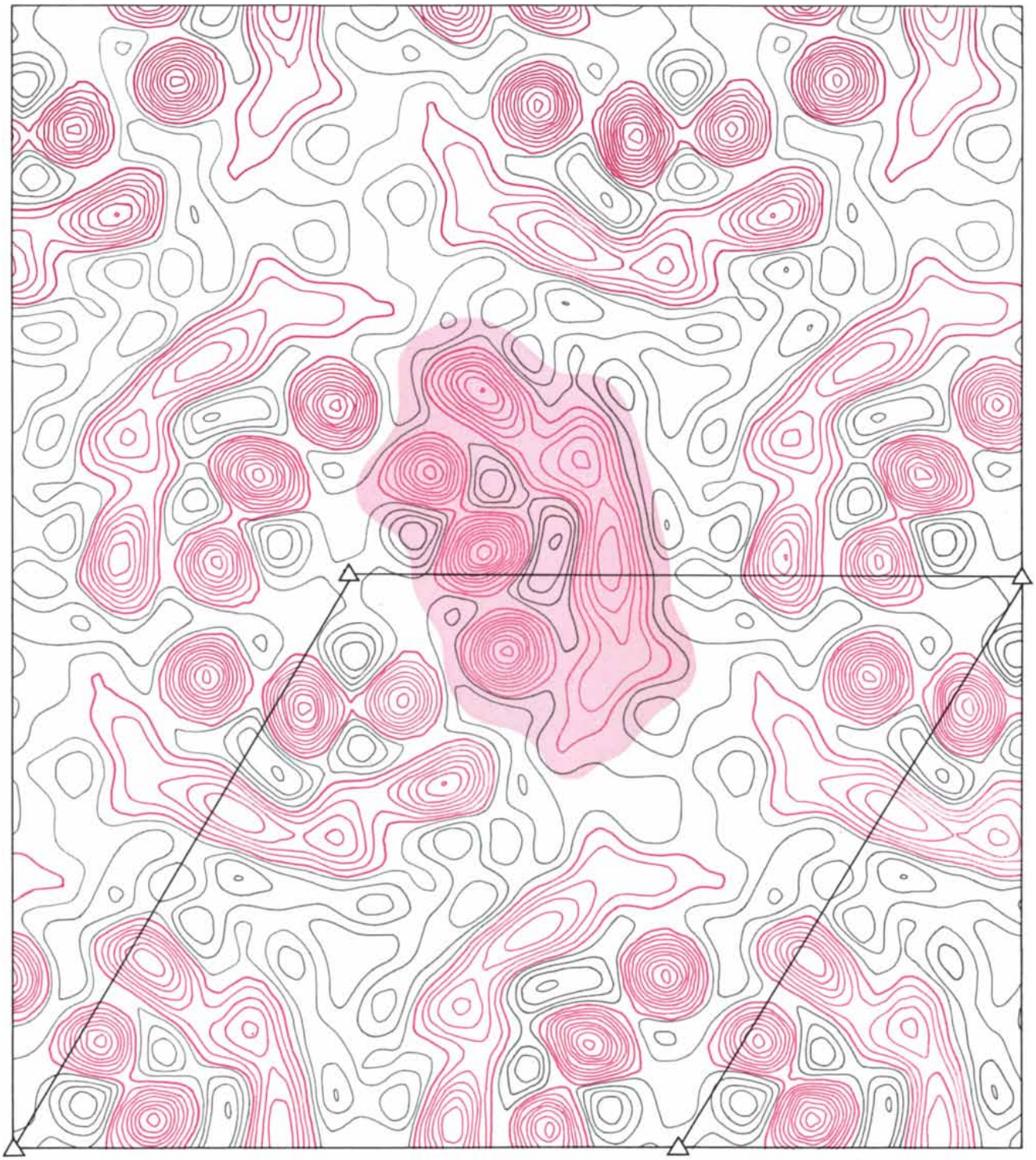
PARTICLES OF BACTERIORHODOPSIN, the protein in the purple membrane of *H. halobium,* are enlarged 87,000 diameters in a freeze-fracture electron micrograph of a cell. The membrane, having been frozen and then broken, has split along the central plane of its lipid bilayer. Since the cell is a cylinder with hemispheric ends, convex fracture faces belong to the cytoplasmic, or inner, half-membrane and concave faces to the exterior half-membrane. Proteins embedded in the membrane are revealed as particles on the fracture faces. Rough-textured regions are from the major portion of the cell membrane, the so-called red membrane. Fine-grained regions with a perceptibly hexagonal pattern are patches of purple membrane. They are on convex faces; corresponding regions on concave faces appear to be smooth, completely devoid of particles. Apparently the bacteriorhodopsin remains embedded in the cytoplasmic half after fracturing.



PURPLE-MEMBRANE SUSPENSION is enlarged 67,000 diameters in an electron micrograph. The suspension was sprayed on a smooth cleaved-mica surface. Cracks developed as the film dried; they follow hexagonal lattice planes intersecting at 60 or 120 degrees.

large molecules such as sugars and amino acids and the pumping activities that bring ions into the cell against electrical and concentration gradients; specific proteins and groups of proteins constitute functional sites in the membrane. The structural complexity arising from the presence of many different proteins and their particular functions makes it hard to analyze the cell membrane even when it has been obtained in a pure form. I speculated that because the proteins themselves control the membrane-disintegration process in halobacteria, they might introduce some specificity into the fragmentation, so that the various fragments would contain different functional sites. We decided to try to sort out the fragments.

Wolfgang Kunau, Gerrit Scherphof and I broke up the cells by various methods and then separated the fragments by centrifugation, which sorts them according to their size and buoyant density. There were three main fractions, which could be recognized



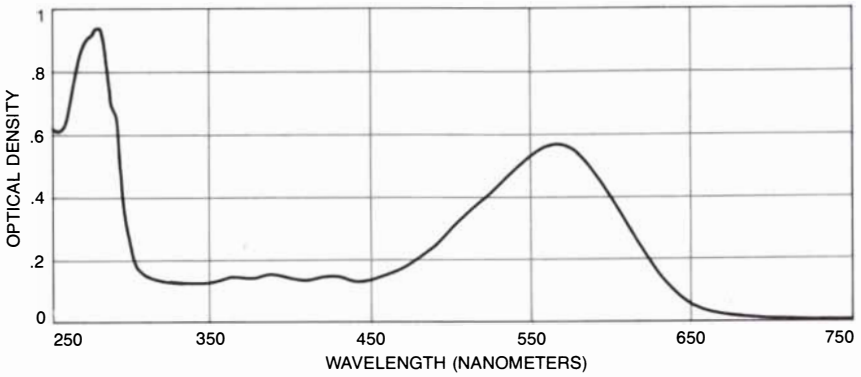
CONTOUR MAP of the purple-membrane lattice was derived by Nigel Unwin and Richard Henderson of the Medical Research Council Laboratory of Molecular Biology in Cambridge from electron micrographs. One unit cell is indicated by lines connecting four triangles that designate axes of symmetry. The colored contours represent alpha helixes of the pigment protein bacteriorhodopsin; approximate extent of one protein molecule is indicated by light colored area. The low-density regions between pigment molecules are presumably lipid.

by their color: a low-density orange-red fraction, a denser deep purple one and a still denser yellow one. Kunau and I found that the red fraction, which constituted the bulk of the membrane, contained fragments that differed widely in size and composition and that its color was due to a high content of bacterioruberin, a pigment that apparently protects the cells against the lethal effect of the high-intensity blue light to which they are exposed in their natural environment. The red fraction also contained cytochromes, flavoproteins and other components of the respiratory chain, which accomplishes oxidative phosphorylation.

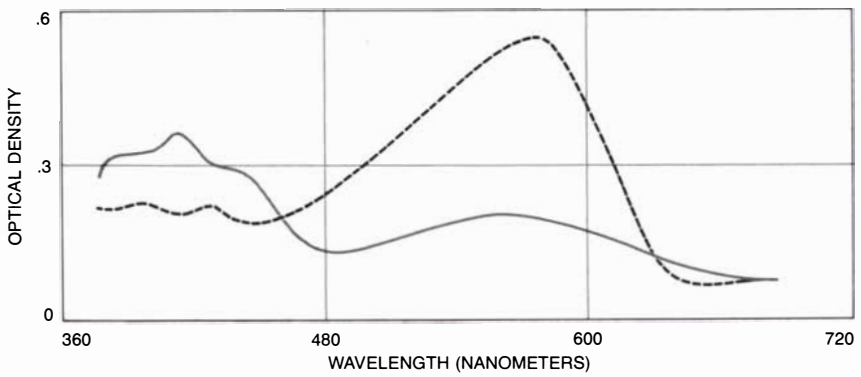
The other two fractions were more uniform. The yellow one consisted almost entirely of the walls of gas vacuoles, small gas-filled sacs in the cytoplasm that presumably help the cell to float at a certain depth. (It also included bits of the flagella that propel the bacterium, which are attached as two bundles at the ends of the cell.) The purple fraction was also rather uniform, but it was typically contaminated with bits of gas-vacuole membrane; we found we could avoid the contamination by working with a mutant strain, isolated by Rowen, that does not have gas vacuoles. The purple fraction turned out to be mostly protein; its low lipid content—only 25 percent—accounted for its high density. X-ray-diffraction studies yielded well-defined, high-resolution patterns, implying that the purple-membrane material had a structure approaching that of protein crystals in regularity. Electron micrographs also revealed a regular structure. By relating micrographs of the purple-fraction material to micrographs of whole cell membranes we finally established that the purple fraction was not diffused through the membrane; it was present in the intact cell as discrete patches set into and continuous with the cell membrane. We still had no clue to any possible function of this purple membrane.

In 1967 I moved from Rockefeller University to the University of California at San Francisco, and it was not until 1969 that Allen E. Blaurock, Dieter Oesterhelt and I resumed work on the structure of the purple membrane. Blaurock analyzed the structure of the fraction in detail by means of X-ray diffraction while I collected complementary electron-microscope data. We could discern a hexagonal lattice in the plane of the membrane; its unit cell, or smallest repeating unit, consisted of three molecules of protein and 40 molecules of lipid. That was consistent with Oesterhelt's finding that the purple membrane contained only one species of protein, with an apparent molecular weight of 26,000. The protein appeared to have a large proportion of hydrophobic, or water-avoiding, amino acids, which in turn was consistent with X-ray data indicating that a large part of each protein molecule must be buried in the hydrophobic interior of the membrane.

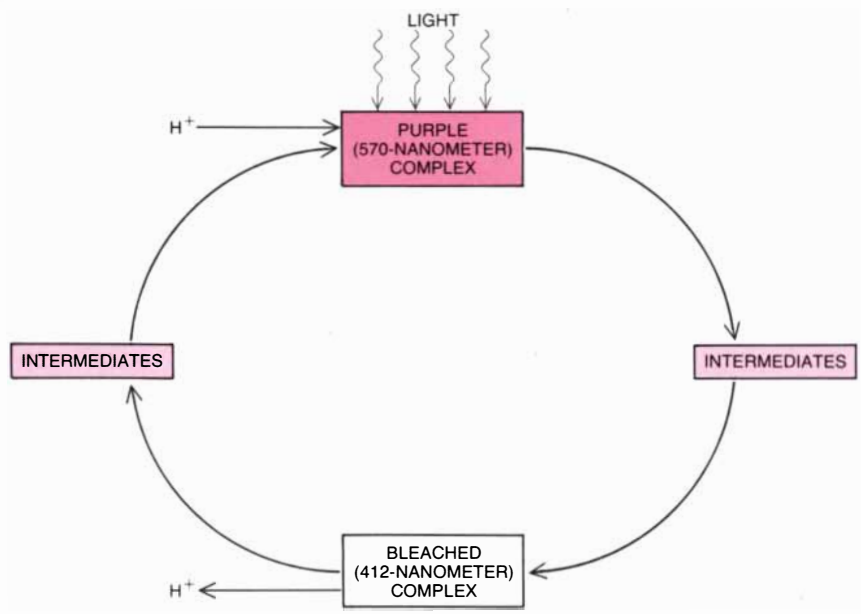
For a long time we were puzzled by the peculiar pigmentation of the purple membrane. One normally determines the wave-



ABSORPTION SPECTRUM of the purple-membrane fraction shows two major peaks: a broad band at about 570 nanometers and a narrow peak at about 280 nanometers. The broad band of absorption in the yellow-green part of the spectrum accounts for the material's purple color. Aromatic amino acid units in the protein are mainly responsible for the short-wavelength peak.



PHOTOREACTION, a reversible light-induced bleaching of the purple membrane, is made evident by low-temperature spectroscopy. Frozen in the dark, a membrane sample shows the usual bacteriorhodopsin spectrum except for a shift of the peak to 575 nanometers (*broken black curve*). A sample frozen while exposed to light is different, with a peak at 415 nanometers (*gray curve*). Light induces a cyclic bleaching and reversion to unbleached form. The return is the more sensitive to temperature, and so the bleached complex accumulates when the material is frozen; if it is warmed and then recooled in the dark, the 575-nanometer peak is restored.

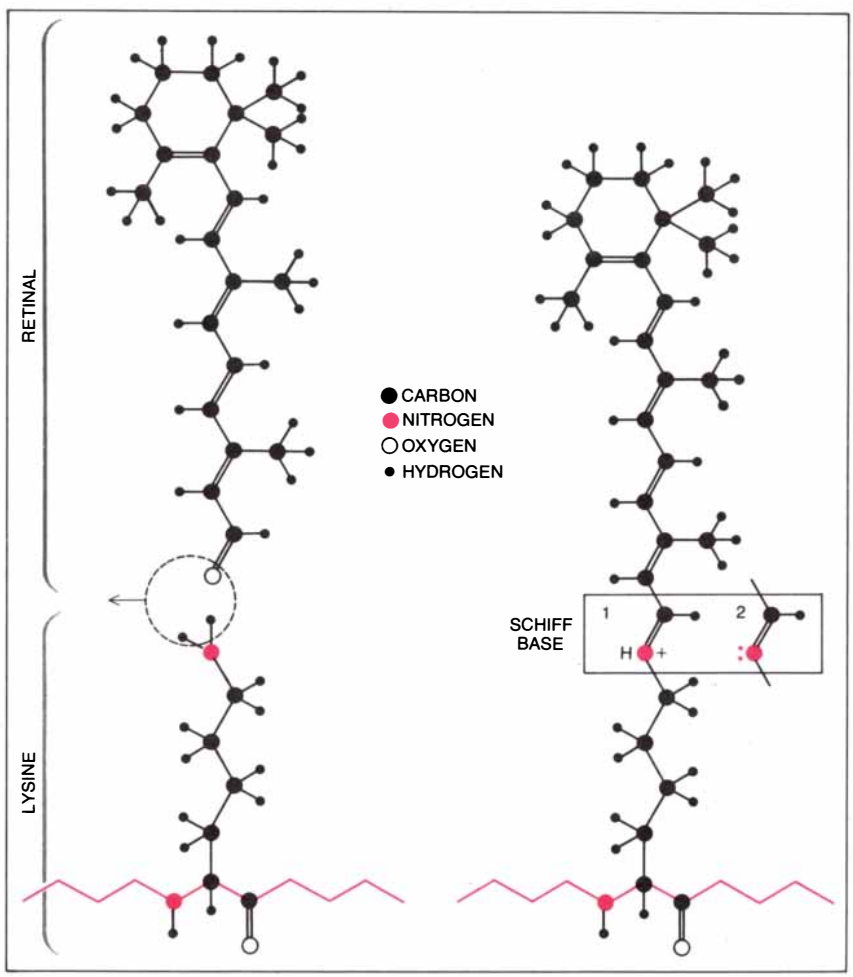


PHOTOREACTION CYCLE takes the bacteriorhodopsin from its unbleached form, with the 570-nanometer peak, to a bleached form with a peak at 412 nanometers (415 in the frozen preparation). There are intermediates along the way. The bleaching is a manifestation of a change in the protein's chemical structure: a proton, or hydrogen ion (H^+), is eliminated from the protein in the course of bleaching and is regained with reversion to the unbleached form.

lengths at which a colored biological substance absorbs light and then attempts to extract and purify the factor responsible for its absorption spectrum. A suspension of purple-membrane material showed a broad absorption band with a peak between 560 and 570 nanometers (in the yellow-green part of the visible spectrum) and a strong absorption peak, characteristic of proteins, at 280 nanometers in the ultraviolet region. Whenever we added a solvent or detergent, however, in an effort to extract the material's chromophore (the chemical group responsible for the light absorption and thus for the color), the color faded away.

It is hard in retrospect to trace the origin of the idea that finally solved the puzzle. It may have been the very fact of bleaching by solvents and detergents, or that rhodopsin was occasionally mentioned in the laboratory because Blaurock had once studied the structure of photoreceptor membranes in the eye, or that I had noticed a phototactic response in *H. halobium:* the cells reversed their direction of swimming when the intensity of illumination was decreased in the red part of the spectrum. Whatever the reason, Oesterhelt decided to see whether the color of the purple membrane might not be due to a rhodopsinlike chromoprotein. He followed the established procedure for analyzing rhodopsin and found that the purple-membrane protein does contain one retinal group per molecule. Retinal, complexed with various proteins, or opsins, is the chromophore of all visual pigments in animals; it is complexed with a particular opsin to form rhodopsin, the pigment of the rod cells of the human eye. We called the membrane protein bacteriorhodopsin.

Retinal is the aldehyde of vitamin A. In the visual pigments it is conjugated to a protein through what is called a Schiff base: a link between the aldehyde (CHO) group of retinal and an amino (NH_2) group on certain amino acid units of the protein. In rhodopsin, the best-known visual pigment, the retinal is linked specifically to the amino group of a lysine unit. The same is true of the purple-membrane protein [*see illustration on this page*]. Retinal itself does not absorb in the visible spectrum; it has a broad absorption band at about 380 nanometers, and in its protein-linked form it should absorb at about 370 nanometers. Yet the major absorption band for the purple-membrane material and for most visual pigments is at much longer wavelengths. This red shift is explained by the particular conditions under which the retinal is complexed with its protein: the amino acids to which it is joined appear to form a hydrophobic pocket that shields the Schiff base from the aqueous environment (whether of the eye or of the bacterial suspension), which would otherwise break the Schiff-base bonds by hydrolysis. We found that when this special conformation of the protein is destroyed by a solvent or a detergent, the absorption maximum does shift from about 570 nanometers to 370 nanometers, and then the retinal is released from the protein. That accounts for our failure to extract a 570-nanometer chromophore, and it also explains the bleaching.



RETINAL, the chromophore of bacteriorhodopsin, responsible for its light absorption, is linked to the polypeptide chain of the protein as is shown here schematically. Retinal and the amino acid (lysine) to which it becomes linked are shown at the left. When they combine, it is through the elimination of a molecule of water composed of the oxygen atom from the aldehyde terminal of retinal and two hydrogens from the amino group of lysine; what is known as a Schiff base forms between carbon and nitrogen atoms. The Schiff base loses and regains a proton in the photoreaction, alternating between two forms (*right*): protonated (*1*) and deprotonated (*2*).

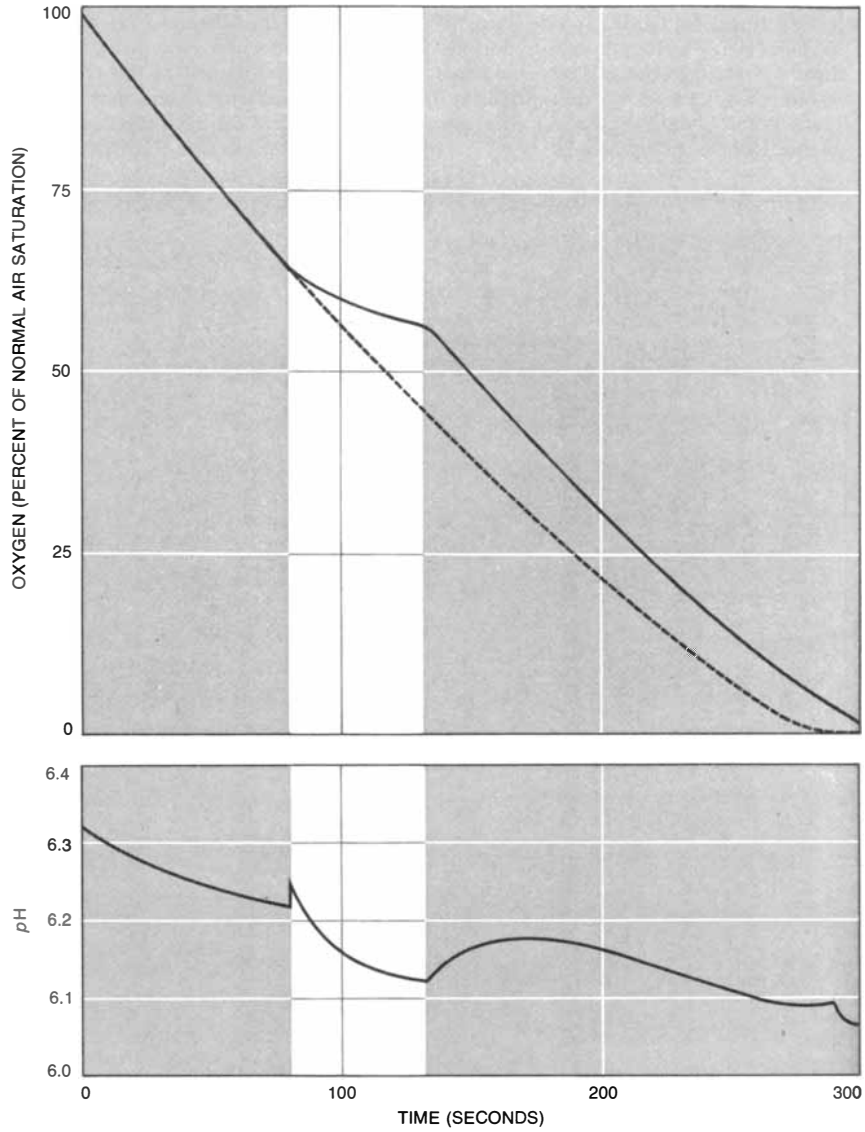
Bacteriorhodopsin's similarity to the visual pigments immediately suggested that its function in the cell was related to its light-absorbing properties. The function might be to trigger the phototactic response we had observed, but that seemed unlikely for two reasons. The first was the large amount of the purple material—enough to constitute up to 50 percent of the cell's surface membrane; phototactic responses are accomplished in other cells with much less pigment. The second reason was that bacteriorhodopsin did not seem to undergo any photoreaction that could effect a phototactic response. When rhodopsin and other visual pigments are exposed to light, they are bleached: they lose color and their absorption spectrum changes. The bleaching is evidence of the photoreaction that causes the visual excitation. Bacteriorhodopsin, on the other hand, seemed to show no significant change in absorption spectrum under illumination.

The large concentration of purple membrane in the cells suggested to me that if we did find a photoreaction, it might turn out to power some metabolic function rather than to trigger a sensory response such as phototaxis. A role in energy metabolism should be most readily detectable as an effect on the ATP content of the cells; light might also exert an observable "saving" effect on the cells' other energy-yielding process, that is, it might cut down on respiration. Working first with Eva S. H. Kirsten and then with Arlette Danon, I measured the ATP concentration of cells maintained under various light conditions, but at first the experiments yielded ambiguous results. On the other hand, I was soon able to convince myself that exposing the cells to light did reduce respiration.

It was an observation dating back to our early extraction attempts that ultimately served as the clue to the purple membrane's photoreaction and hence to its function. Oesterhelt had noted that treating a water

suspension of the membrane material with diethyl ether would bleach it and that the bleaching was reversed when the ether evaporated. Even though light did not appear to be involved, this bleaching was an intriguing reaction. And so one day I asked Martin Brown, who was then an undergraduate technician in my laboratory, to repeat the ether experiment—but to suspend the purple membrane in the high-concentration basal salt medium, which approximated its natural environment, rather than in plain water. Within an hour he reported that the membrane was now bleached by ether only when the suspension was illuminated by visible light and that the bleaching was reversed by near-ultraviolet radiation. (We later found that the ultraviolet was not needed; the light-bleached membrane would return to its original color and absorption state within a few seconds even in the dark.) We had the first indication of a photoreaction, albeit one involving a chemical reagent too.

I described the discovery in a letter to Oesterhelt, who had returned to the University of Munich. He visited the Max Planck Institute for Nutrition in Dortmund, and there, together with Benno Hess, whose laboratory was well equipped for such studies, he began to investigate the kinetics of the spectral change. The ether-bleached purple membrane's absorption peak, 412 nanometers, was found to be due to a modified bacteriorhodopsin that was formed and accumulated rapidly in the light and that decayed somewhat more slowly in the dark to re-form the 570-nanometer complex. Oesterhelt learned something more: the bleaching reaction and its reversal were accompanied by the release and uptake of hydrogen ions, or protons. The bleached complex was "deprotonated," and its reversion to the original 570-nanometer complex was accompanied by a "reprotonation." The movement of protons was reflected in a reversible change in the *p*H of the medium, which became more acid as protons were released from the membrane and then less acid as the membrane was reprotonated. Oesterhelt observed similar light-induced *p*H changes in suspensions of intact cells.



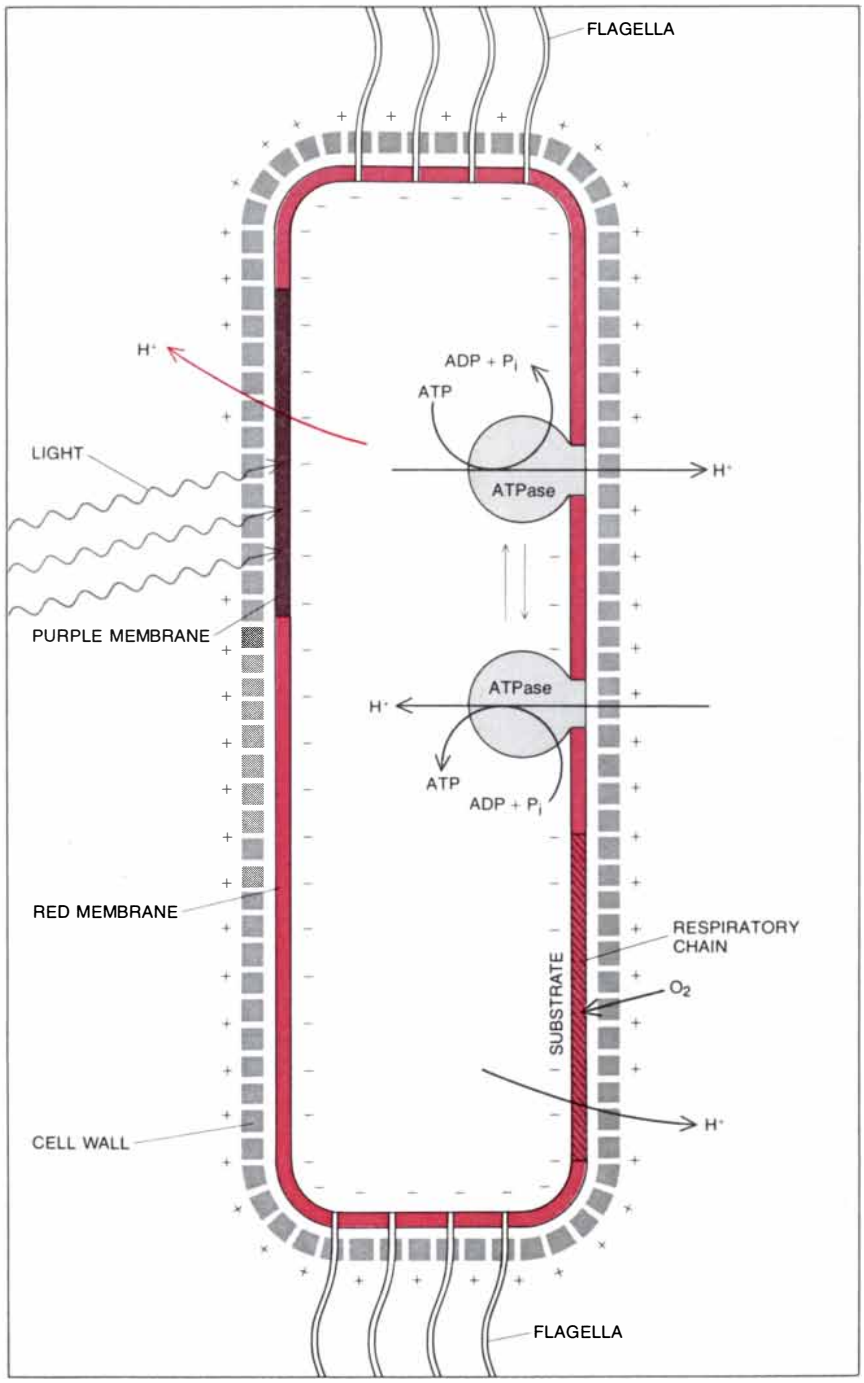
LIGHT AFFECTS the bacteria's oxygen consumption (*top*) and the acidity of the medium (*bottom*). The slope of the upper curve is a measure of the rate of oxygen consumption. In a closed chamber the cells consume oxygen rapidly in the dark (*shaded areas*). When a light is turned on (*white area*), there is a 76 percent decrease in the rate of oxygen consumption (*solid line*); the original rate is resumed when the light is switched off. The light also causes an increase in the acidity (a drop in the *p*H) of the medium as protons are expelled from the cells into the medium.

Meanwhile I was studying the purple membrane's change in absorbance with other instrumentation: by means of low-temperature spectroscopy with Richard Lozier in Warren L. Butler's laboratory at the University of California at San Diego, and by flash spectroscopy with Richard A. Cone at Johns Hopkins University. Both methods can detect transient changes in absorbance. They revealed that the light-induced changes we had hitherto observed only in the presence of ether also took place without ether in the purple-membrane material or intact cells suspended in basal salt medium and in the membrane material suspended even in plain water. We also discovered that we were dealing with a cyclic photoreaction involving several intermediates, of which the 412-nanometer complex was just one. The absorption of light in the 570-nanometer band drives the pigment through the cycle over and over again until the light is switched off; each cycle is completed in a few milliseconds and is accompanied by the release and uptake of protons [*see bottom illustration on page 41*]. Ether changes the kinetics of the cycle so that the bleached, deprotonated 412-nanometer complex accumulates and is readily observed. Without ether significant amounts of the bleached complex do not accumulate even under continuous illumination (except at very high light intensities), and so we had not detected it.

Oesterhelt and I published these results. Our key observations were a light-induced reversible acidification of the medium in cell suspensions, the release and uptake of protons by the purple membrane during the photoreaction and the fact that exposure of the intact cells to light slowed down their respiration. The release and uptake of protons by intact cells was clearly different from the change in protonation observed during the photoreaction of isolated purple membrane; in cell suspensions the change took place more slowly but attained a greater magnitude. We nonetheless thought that the two phenomena must be closely connected. We postulated that in the intact cell the purple membrane acts as a proton pump. The release and the uptake of protons, we suggested, take place on opposite sides of the membrane; there is a pathway through the membrane that enables protons to move from the uptake site to the release site but not in the opposite direction. Under steady illumination the pigment cycles continuously at a rate of several hundred times per second, taking up protons on the inner,

cytoplasmic side of the membrane and releasing them, on the outer side, into the medium. Hence a proton concentration gradient is established and an electric potential is generated between the cell interior and the medium. Light energy is converted into an electrochemical gradient.

This interpretation of our results was prompted by the "chemiosmotic" theory of energy transduction proposed some years ago by Peter Mitchell of the Glynn Research Laboratories in England. He held that energy derived from the flow of electrons along the electron-transport chain in either oxidative phosphorylation (respiration) or photophosphorylation (photosynthesis) is first stored in the form of an electrochemical proton gradient across the membrane containing the electron-transport chains. The energy stored in that gradient serves to synthesize ATP from adenosine diphosphate (ADP) and inorganic phosphate when the protons flow back across the same membrane through a particular enzyme, membrane ATPase. The mechanism of proton-pumping in oxidative phosphorylation and in photosynthesis is surely very different from the one we postulated for the purple membrane, but the result in all three cases is the same: the chemical free energy of some oxidizable substrate or of light is converted into the free energy of an electrochemical proton gradient. The primary function of the purple membrane, then, is to make light an alternative to respiration as an energy source for halobacteria. An observation consistent with this interpretation is that the cells synthesize a purple membrane only when they grow at low oxygen concentrations and in the light.



PROTON TRANSPORT is related to energy metabolism in a halobacterium cell as is shown here. The cell membrane contains patches of purple membrane and also respiratory-chain pigments. Either light or oxygen can promote the ejection of protons. The light-powered process is mediated by the purple membrane and the oxygen-fueled process by respiratory chains in the presence of oxidizable substrate. The resulting membrane potential (positive on outside) and proton concentration (larger on outside) drive a backflow of protons through the enzyme ATPase, synthesizing adenosine triphosphate (*ATP*) from adenosine diphosphate (*ADP*) and inorganic phosphate (P_i). The ATPase is reversible; it ejects protons when it hydrolyzes ATP.

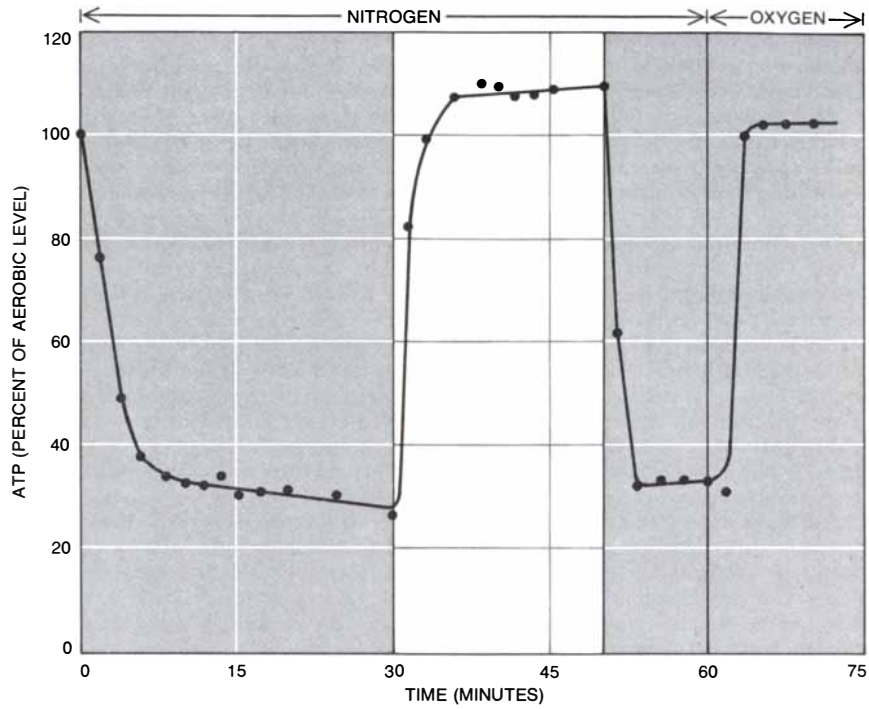
We still had to show that the cells can in fact synthesize ATP by means of the light-generated gradient. Our early experiments had been inconclusive, it developed, because we had failed to make the cells anaerobic enough. Arlette Danon and I continued to collaborate on the problem and she did the experiment successfully after returning to the Weizmann Institute of Science in Israel. She suspended the bacteria in a salt solution, without nutrients and in the dark, and bubbled nitrogen through the medium. Under these conditions the intracellular ATP concentration falls rapidly to about 30 percent of the original level and remains fairly constant; apparently the cells can shut down most of their energy-requiring processes when no external energy sources are available. When the cell suspension is illuminated, the ATP content of the cells rises rapidly up to or above the original level; in the dark it goes down again. Admitting air—which is to say oxygen—to the suspension has the same effect as light: the ATP content rises. When we measured ATP accumulation under illumination at various wavelengths, we found that only light of wavelengths absorbed by purple membrane is effective in ATP synthesis.

We went on to make observations that were at least consistent with more specific assumptions: that both respiration and light generate an electrochemical proton gradient across the cell membrane and that the gradient drives the synthesis of ATP by membrane ATPase. We did this by applying substances that interfere in specific ways with ATP synthesis. First we applied two chemicals that directly inhibit the activity of membrane ATPase, and both of them did prevent the accumulation of ATP driven by either light or respiration. Then we tried uncouplers: agents that allow electron transport to proceed but that in effect dis-

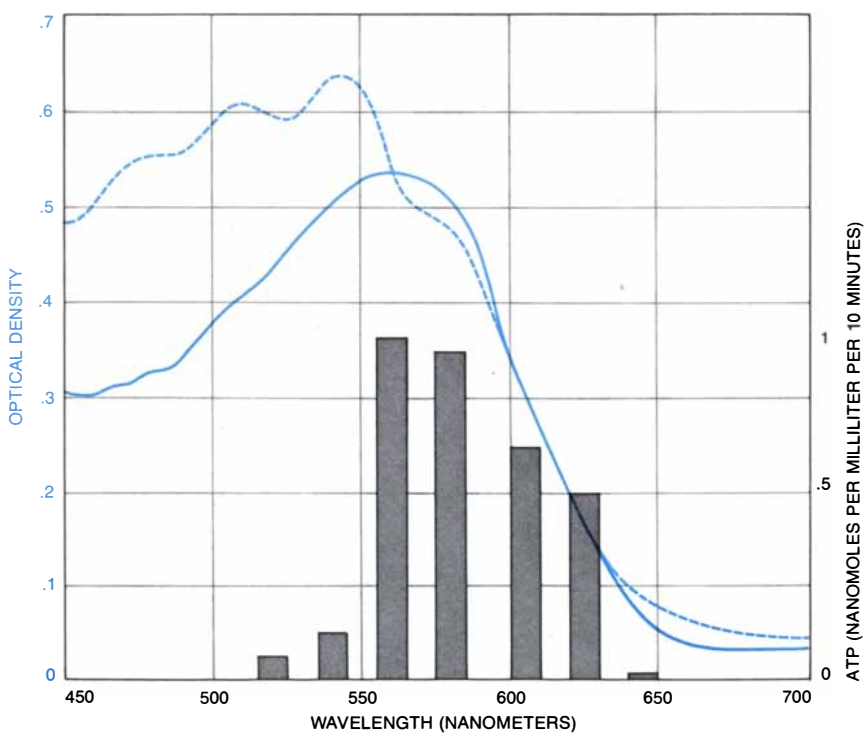
connect it from phosphorylation and thus from the ATP synthesis it usually accomplishes. Mitchell's theory holds that uncouplers make the membrane very permeable to protons, preventing the generation of an electrochemical gradient, and indeed model membranes treated with uncouplers do become permeable to protons. In halobacteria suspensions the uncouplers not only inhibited ATP accumulation driven by light or respiration but also prevented acidification of the medium, which reflects deprotonation. Finally we tried several substances that interfere specifically with the electron-transport chain, all of which blocked ATP accumulation driven by respiration but had no effect on the light-driven reaction. They also blocked respiration-driven, but not light-driven, acidification of the medium.

All these observations were in good agreement with our central postulate that the purple membrane is a light-driven proton pump. Even bacterial cells are enormously complex systems, however. Their energy metabolism is not completely worked out, and so it was possible to imagine plausible alternative explanations for our observations. One might assume, for example, that light absorbed by the purple membrane activated an enzyme that in turn catalyzed some energy-yielding process such as fermentation, which could therefore proceed only in the light. As for the proton-pumping, the release and uptake of protons that we observed did not necessarily imply that protons were being moved across the membrane. To be sure, constructing alternative explanations for all our data would call for some contrived assumptions, but we felt it would be difficult to rigorously exclude all such alternatives. In order to verify our conclusions it was better to try to achieve similar results in a system simplified enough to exclude alternative explanations.

H. *halobium* cells can be cracked open by alternate freezing and thawing, by applying shearing forces or by briefly reducing the salt concentration of the medium. Under such conditions the cell is emptied of most of its contents and the cell membranes reseal themselves to form empty vesicles, or sacs, that we call cell envelopes. Provided only that the purple membrane's orientation in the envelope was properly maintained, we found that these empty sacs responded to light by generating an electrochemical proton gradient. That excluded the possibility that enzyme systems in the cytoplasm were responsible for our various observations. There was still the possibility, however, that unidentified energy-yielding systems firmly bound to the red membrane of the cell might be playing some role. To deal with that possibility we had to remove the major (red) portion of the cell membrane and replace it with a permeability barrier that had similar properties but was metabolically inert. This kind of thing had been done before. Lipid bilayers form spontaneously when membrane lipids or syn-



ATP SYNTHESIS is shown to be driven by light or by oxygen. When *H. halobium* cells are deprived of oxygen (in a closed container through which nitrogen is bubbled) and light (*shaded region of chart*), the ATP concentration falls. When the light is turned on, the ATP level rises. It drops again in the dark, but it is restored when oxygen is bubbled through the suspension.



DEPENDENCE OF ATP SYNTHESIS on the purple membrane is demonstrated by comparing ATP production at various wavelengths with the spectra of the purple membrane and of a total cell suspension. The bars show the amounts of ATP synthesized when cells are illuminated through filters that pass narrow bands of wavelengths. The bars fit purple-membrane spectrum (*solid colored curve*) better than the spectrum for a total cell suspension (*broken colored curve*).

thetic copies of them are dispersed in water, and the bilayers tend to form tiny vesicles; various components can be combined with the lipid in an effort to reconstitute functional model membranes.

Efraim Racker of Cornell University, who had successfully reconstituted elements of the electron-transport chain of mitochondria by such methods, agreed to collaborate with me, and so I went to Ithaca with a preparation of isolated purple membrane. Our first experiment was successful. We combined purple-membrane fragments with man-made phospholipid vesicles by incubating the mixture with a little detergent or by agitating it with sound waves. When the resulting preparation was exposed to light, the medium underwent reversible changes in *p*H. The changes were much larger than could be accounted for just by a conversion of all the purple-membrane protein to the deprotonated state, and as a matter of fact the light intensities were such that only a little of the bacteriorhodopsin could have been deprotonated; the preparation had to be actively pumping protons. We found that the pumping action was sensitive to the effect of uncouplers, clearly indicating that the light was establishing a proton gradient across the lipid-vesicle membrane. Because bacteriorhodopsin was the only protein in these preparations and because it worked in vesicles made of various lipids, there was no longer any doubt that bacteriorhodopsin converts light energy to generate a proton gradient.

There was one striking difference between the effect of light on intact cells or cell envelopes and the effect on these lipid vesicles. In the case of the vesicles the light produced not a decrease in the *p*H of the medium (acidification) but an increase in *p*H (alkalinization); protons were being pumped from the medium into the vesicles. This suggested that the bacteriorhodopsin is oppositely oriented in cells and in vesicles—that in the vesicles the "cytoplasmic" side of the membrane faces outward toward the medium. We have been able to verify that with freeze-fracture electron micrographs. If a membrane is frozen and then broken up, it tends to split along the midplane of the lipid bilayer. We had seen that freeze-fractured membranes of intact cells presented two clearly distinguishable fracture faces. The convex faces, which must be on the inner, or cytoplasmic, half of the split membrane, were studded with closely packed membrane proteins; the concave faces, which are the fracture faces of the exterior half-membrane, were smooth [*see top illustration on page 39*]. Freeze-fracture preparations of the man-made vesicles show the opposite pattern, confirming that the protein is here oriented inside out [*see illustration below*].

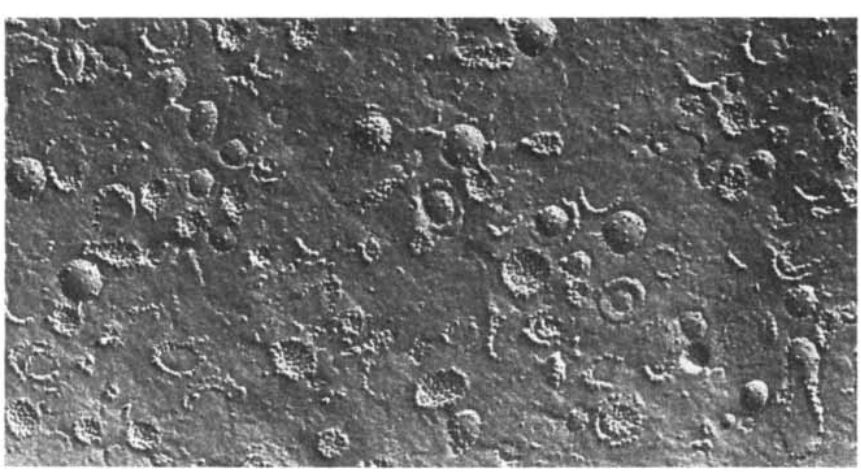
The observations on lipid vesicles have since been extended by San-Bao Hwang and others in my laboratory, by V. P. Skulachev and his colleagues at the Moscow State University and by workers in Racker's laboratory. One result deserves special mention because it constitutes a strong argument in favor of Mitchell's chemiosmotic theory. Racker had previously achieved ATP synthesis by incorporating parts of the electron-transport chain and mitochondrial ATPase in lipid vesicles and supplying oxidizable substrate. If, as the Mitchell theory holds, the electron-transport chain functions by generating an electrochemical proton gradient that in turn drives ATP synthesis catalyzed by ATPase, then it should be possible to substitute bacteriorhodopsin and light for the chain elements and the oxidizable substrate. Racker did that. He incorporated ATPase from beef-heart mitochondria in the bacteriorhodopsin-lipid vesicles; when he illuminated the preparation, it synthesized ATP from ADP and inorganic phosphate.

These investigations close the first chapter in the exploration of a new photosynthetic system. We have shown that an organism lacking chlorophyll can capture and convert light energy and utilize it to synthesize ATP or to drive other metabolic processes; it does so by means of a chromoprotein that closely resembles the visual pigment of animals. The pigment is incorporated in the cell membrane, where it converts light energy to transport protons across the membrane against an electric potential and a concentration gradient.

One problem to be tackled now is the nature of the intramolecular mechanism whereby absorption of light drives protons across the membrane; this is under intensive investigation in many laboratories, and progress continues to be rapid. At the Medical Research Council Laboratory of Molecular Biology in Cambridge, England, Richard Henderson and Nigel Unwin have exploited the crystalline arrangement of bacteriorhodopsin in the membrane to derive a three-dimensional map of the protein structure by a new technique of image reconstruction from a large number of electron micrographs made at extremely low beam exposures. This is the first membrane-protein structure to be determined at near-atomic resolution. Together with analyses of amino acid sequence, which are also being pursued in several laboratories, such mapping should tell us more about the pump mechanism. Our own investigation of the kinetics of the photoreaction cycle, together with spectrographic studies done with Aaron Lewis at Cornell, has revealed that it is specifically the Schiff base of the retinal that loses a proton and then gains one during the cycle and is therefore probably directly involved in the movement of protons. On the basis of these various data molecular models of the pumping mechanism are being developed that can be subjected to critical testing by new experiments. What is learned may improve our understanding of other ion pumps in many other membranes.

Another question is how the cell further converts the energy of the ATP and regulates the flow of energy through successive metabolic processes. Here the relatively simple envelope preparations are more amenable to analysis than intact cells. Investigators at the Ames Research Center of the National Aeronautics and Space Administration and at Cornell have found that light can power the uptake by the envelopes of sodium and potassium ions and of amino acids and that this uptake is independent of ATP synthesis.

The discovery of the purple membrane has begun to influence strongly a number of areas that are of high current interest in biology, notably the structure and function of membranes in general and of energy-transducing systems in particular. Examining the relation of bacteriorhodopsin and the visual pigments of the eye should provide a new perspective on both the mechanisms of vision and its evolution. Finally, we may eventually learn from these light-powered bacteria how we too can make better use of the sun as a source of energy.



FREEZE-FRACTURE electron micrograph enlarges man-made lipid vesicles incorporating purple membrane 77,000 diameters. Here the particles are predominantly on concave fracture faces, not on convex faces as in cells (*see top illustration on page 39*). That is, particles remain embedded in the outer half of the split membrane; the orientation of bacteriorhodopsin in vesicles is the opposite of what it is in bacteria. The protein pumps protons into vesicles, whereas it pumps protons out of cells; in light the vesicle medium becomes less, rather than more, acidic.

Quenching of Singlet Molecular Oxygen in Solution by Azomethine Dyes

Wendell F. Smith, Jr., William G. Herkstroeter, and Kenneth L. Eddy

Contribution from the Research Laboratories, Eastman Kodak Company, Rochester, New York, 14650. *Received August 16, 1974*

[Journal of the American Chemical Society **97**, 2764 (1975)]

Background: The two valence electrons of the O_2 molecule in its lowest energy state have parallel spins. This is unusual. Most molecules with an even number of electrons have singlet ground states. "Singlet" implies that every electron is paired with another of opposite spin. For the oxygen molecule, that state (written $^1O_2$) represents excitement. To attain it, radiation must be absorbed or energy must be acquired from some other excited molecule. Evidence has been accumulating in recent years that singlet molecular oxygen is involved in the formation of air pollutants and in such fundamental business as photosynthesis and the digestion of microbes by white blood cells. While the list of molecules which will react with $^1O_2$ is growing rapidly, the list of molecules which will quench $^1O_2$ back to O_2 is much smaller. This is because the excitation energy of $^1O_2$ is unusually low; a quencher molecule to relieve $^1O_2$ of this energy must have an even lower excited state.

Gist: A new class of $^1O_2$ quenchers has been discovered: azomethine dyes. Yellow azomethine dyes, absorbing light in the blue, are the least efficient $^1O_2$ quenchers. The most efficient are the cyan ones, absorbing in the red, the low-energy end of the visible spectrum. Magenta dyes, which absorb green light, are intermediate in their quenching efficiency. The quenching proceeds by the formation of an excited complex of the dye with $^1O_2$. The binding energy of the excited complex varies with the structure of the particular azomethine dye.

An inquisitive outsider to this field might ask these chaps:

Are azomethine dyes new?

Eddy: Not exactly. All three dyes in Kodak color prints and negatives and in Kodachrome and Ektachrome slides have been azomethines. An azomethine contains this structure:

Does efficiency in quenching singlet molecular oxygen influence dye stability as a practical matter?

Smith: Quenching of singlet oxygen by the dye is a reaction that competes with the reaction of dye destruction by singlet oxygen. Therefore the faster the quenching reaction relative to the chemical reaction, the better.

What is the significance of the binding energy of the excited complex?

Smith: Other things being equal, the greater the binding energy, the better the dye would be as a quencher.

What happens to the azomethine dyes after they have been quenched?

Herkstroeter: The quenching reaction with singlet oxygen raises them from their ground state to their lowest excited electronic state. Because of their ability in this state to twist about the carbon-nitrogen double bond, the return to the ground state is rapid. The overall result is to protect the dye.

SCIENCE AND THE CITIZEN

Uranium Prospecting

Nuclear power plants, which currently provide about 9 percent of the electricity generated in the U.S., last year consumed some 12,700 short tons of processed natural uranium ore (in the form of the oxide known as yellowcake, or U_3O_8). It is estimated that by 1985, when nuclear power is scheduled to account for perhaps a fourth of the total U.S. installed electric-generating capacity, the rate of consumption of fresh nuclear fuel will have reached at least 55,000 short tons of yellowcake per year (assuming that some of the spent fuel is by then being recycled in the form of reprocessed uranium and plutonium) or at most 62,000 short tons of yellowcake per year (assuming no recycling). And by the year 2000, when the nuclear share of the U.S. energy budget is expected to meet approximately half of the nation's electricity needs, the nuclear-fuel requirement is projected to reach anywhere from 127,000 to 164,000 short tons annually (again depending on the extent of recycling).

In cumulative terms (measured from a base date of January 1, 1975) the total amount of natural uranium fuel that would on this projection be consumed by 1985 would range from 339,000 to 366,000 short tons of yellowcake. The corresponding cumulative total by the turn of the century would be between 1,740,000 and 2,195,000 short tons.

In view of the fact that the proved domestic reserves of high-grade uranium ore contain on the order of 600,000 short tons of U_3O_8, and that the present exploration effort is yielding additions to reserves at the rate of approximately 12,000 short tons per year, the question naturally arises: Can the supply of natural uranium be expanded at a rate sufficient to keep pace with the projected demand?

According to a major new study of the problem of nuclear-fuel availability conducted by the Edison Electric Institute, the principal association of the electric-utility companies in the U.S., the answer is a tentative yes. The two-volume study, which took more than a year to prepare, concludes that "assurance of supply of nuclear fuels for the power plants now in operation or committed is generally satisfactory for the short term (1975–1980) and less satisfactory but manageable for the near term (1975–1985)." For the period up to 1985, the report says, "there are in existence or planned: adequate available reserves of natural uranium, adequate enrichment capacity and adequate conversion facilities." On the less satisfactory side of the near-term ledger "there is *in*adequate capacity or planning for: uranium mining and milling; spent fuel storage; radioactive waste management; large-scale demonstration of reprocessing and recycle, and transportation of spent fuel."

The supply of nuclear fuel for the longer term (1985 on), the report continues, "depends on the near-term deficiencies being corrected and timely decisions made on long lead-time items." Among the supply operations that should be accorded the "highest priority" soon, the Edison Electric Institute recommends that "ERDA [the Energy Research and Development Administration] should continue and accelerate its National Uranium Resources Evaluation Program. The utilities should assure aggressive exploration and supply initiatives in support of their individual requirements. Prompt consideration should be given to pooling of utilities' uranium supply interests, perhaps on a regional basis.... The U.S. Government should undertake to assure utilities timely access to foreign uranium supply."

Noting that uranium supplies to U.S. utilities during the period from 1985 to 1994 "could prove to be inadequate unless the present pace of domestic exploration effort is accelerated and major new discoveries are made and aggressively exploited, and/or supplies from foreign sources turn out to be more available than is now indicated," the Edison Electric Institute report goes on to state that "the uranium supply outlook in and beyond the mid-1990's is at present indeterminate because of gaps and inadequacies in the present knowledge of domestic and world uranium resources which make it impossible to assess the long-range prospects for continuing expansion of uranium production capability." In any event, the report advises, "the U.S. utility industry should take stock of the situation periodically."

A less optimistic view of the long-term uranium-supply picture is indicated by the results of a recent historical analysis of exploration statistics by M. A. Lieberman of the University of California at Berkeley. Writing in *Science,* Lieberman points out that if his estimates of undiscovered U.S. uranium resources are correct, and if the expansion of nuclear electric power proceeds as planned, "a serious shortfall in uranium supply will develop during the late 1980's." Referring to the widely held view that the advent of breeder reactors (nuclear reactors that produce more fuel than they consume) in the mid-1990's will be just in time to relieve the uranium-ore situation, Lieberman contends that "on the time scale discussed, the planned introduction of the breeder reactor can play no role.... Even a 'crash program' to develop the breeder will be unable to forestall the coming... uranium squeeze; it is far too late." He concludes: "If any significant growth in U.S. nuclear power is to be sustained over the next few decades, then either a marked increase in the rate of discovery of new, high-grade reserves must occur, contrary to past trends in exploration statistics, or the development of low-grade ores... must be undertaken immediately and vigorously."

The ultimate resolution of the long-term uranium-supply question is further complicated by the present state of uncertainty in the U.S. regarding the proper role of nuclear power in the nation's future energy budget. As H. A. Bethe pointed out recently in these pages ("Letters," SCIENTIFIC AMERICAN, April), "we are unfortunately in a vicious circle: public acceptance of nuclear power is uncertain, therefore there is not enough drilling for yellowcake, hence the proved reserve of uranium is insufficient, and this again diminishes public acceptance."

The Other Risks

The wide interest in the hazards associated with nuclear power plants has focused new attention on a broader area of concern, of which nuclear reactors constitute just one case: To how much risk is a population subjected by a wide range of man-made devices and practices, and to what extent is each risk offset by a benefit? Recently there have been attempts to apply techniques of risk-benefit analysis developed primarily to deal with nuclear power to some of the other hazards.

In testimony last fall before a committee of the California state legislature, David Okrent of the University of California at Los Angeles discussed reactor-safety standards and then asked: "How quantitatively does society know other risks to which it is exposed? From food additives and drugs? From earthquakes? From the failure of dams? From the shipment and storage of chemicals? From the burning of coal? From the discharge of mercury...?" Okrent reported on some U.C.L.A. studies. For example, his group calculated that there was one chance in 100,000 that in any one year a commercial aircraft would crash into the grandstand of the Hollywood Park racetrack while it was occupied by a large crowd, with a probable number of fatalities of between 3,000 and 8,000, and a maximum number of about 30,000. "On the average it is estimated that about five people on the ground within five miles of" Los Angeles International Airport and Hollywood-Burbank Airport "would be killed every 10 years" by commercial-aircraft crashes. Such results are statistical, Okrent pointed out: "We may have no such casualties at either airport in the next 20 years. Or we may have more."

The U.C.L.A. group studied the probability of a "sudden complete failure" of 10 California dams. The estimated probabilities of failure ranged from roughly one in 40 per year to one in 30,000 per year; the potential fatalities, from 14,000 to 260,000. "Are such hazards acceptable?" Okrent asked. He drew an implicit contrast with the standards proposed for nuclear reactors (in a referendum on which Californians will

"Unbelievable."

What would you say about a small sedan with the style of a fine European road car?

A small sedan with classic lines, a broad expanse of glass, a roomy interior that seats six comfortably and a long list of options including power seats, power windows, electric door locks, and automatic speed control. Very classy, indeed!

What would you say about a small sedan with a ride that rivals that of a full-sized car?

Aspen does it again! With a unique Isolated Transverse Suspension. It's rubber-isolated to reduce the noise and vibration transmitted to the passenger compartment. To give Aspen the comfortable ride you usually find in bigger cars.

What would you say about a small sedan with a small price that starts at only $3,371?

That's the manufacturer's suggested retail price for the base Aspen sedan (not shown), excluding state and local taxes, destination charge, and optional equipment, if any. The Aspen SE, pictured above, starts at just $4,400. That price includes an automatic transmission, power steering, a full vinyl roof, soft vinyl-upholstered 60/40 seats with recliners and center armrest, and an electric clock. The whitewalls, wire wheel covers, right side mirror, bumper guards, and rub strips shown are $167 extra.

What would you say about a small sedan that got an EPA estimated mileage of 27 MPG highway and 18 MPG city?

According to EPA estimated mileage results, that's what the Aspen sedan and coupe got with a 225 Six and manual transmission. (Your actual mileage may differ, depending on your driving habits, condition of your car, and optional equipment. In California, see your Dealer for EPA mileage results.)

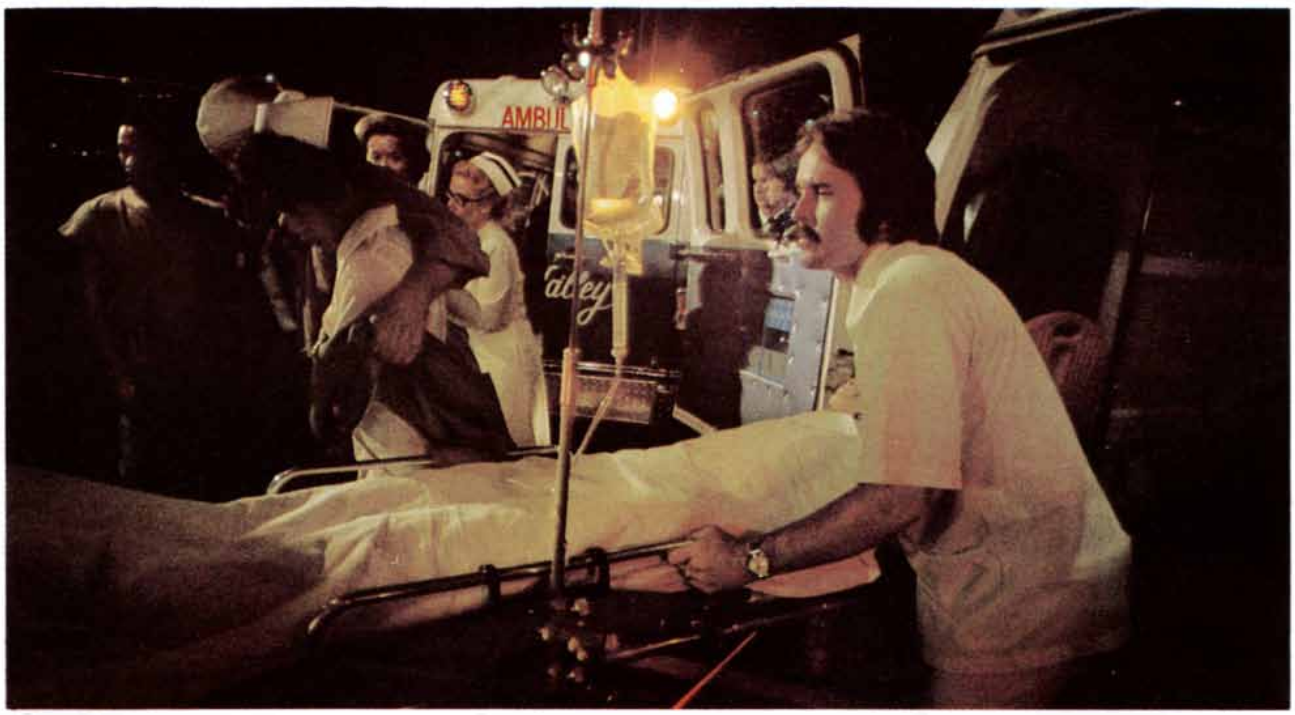
Dodge

ASPEN

The new Dodge Aspen.
For a small car at a small price, it's unbelievable.

Winner of the 1976 **Motor Trend** Magazine Car of the Year Award.

How to save hours when there isn't a second to waste.



When life depends on receiving blood quickly, filtering can become a fatal bottleneck. Now, those delays no longer have to happen.

Before a hospital patient receives a transfusion, the blood should be filtered. On a routine day, that's no problem.

But what happens when there's a rash of accidents over a holiday weekend?

Or a single massive accident?

The slow-but-sure filtering process can become a bottleneck that reduces the odds against survival in an emergency.

Why? Because each and every stainless steel filter casing has to be sterilized before it can be reused.

And sterilization can take several hours.

Precious time saved.

The answer to the problem is a pre-sterilized disposable blood filter casing made from a special plastic. An extremely tough, clear plastic called K-Resin®.

Today, these disposable, pre-sterilized blood filters are available to hospitals all over the world.

Disposable blood filter casings were made possible by strong, clear K-Resin plastic.

Interestingly enough, the same plastic that made these blood filter casings possible is opening new possibilities for safer toys.

Tonka Corporation, the toy truck manufacturer, is using K-Resin to make windshields for some of its products. (You've probably seen a Tonka toy truck on TV—it's the one that the elephant steps on.)

A stronger, safer plastic.

Tonka found that K-Resin plastic does not break as easily and is less likely to shatter on impact.

Not only that, it actually costs less than the plastic they had been using.

Wherever there is a need for a plastic that will retain its clarity and stand up to a lot of punishment, there's a need for K-Resin.

A stronger, safer plastic developed by the same people who make fine products for your car.

The people of Phillips Petroleum.

Surprised?

PHILLIPS 66

The Performance Company

be voting in June) by asking whether the state legislature would "determine that the safety of dams has been demonstrated by comprehensive testing in actual operation" of similar systems, or would "require complete insurance coverage or close down the dams."

Okrent's group sought to learn from U.S. chemical companies where certain potentially dangerous chemicals are stored, in what quantities and with what safety criteria. It could obtain no information. Apparently, however, "huge quantities of dangerous chemicals are stored within striking distance of population centers," Okrent said, and he asked whether the legislature had ever determined that "there exists no risk to public health and safety from the shipment, storage and use of chemicals employed in California's agricultural and food industries." As a final example of the non-nuclear accident potential Okrent cited the shipment and storage of liquefied natural gas. He concluded that the U.S. probably cannot afford "to raise the safety standards of existing non-nuclear technology" to the point where its risks do not exceed the risk from nuclear power.

Why, then, are public perception of technological dangers, and suspicion that they are not outweighed by benefits, focused almost exclusively on nuclear power? At a workshop on risk-benefit methodology at Asilomar in California last fall Edward L. Zebroski of the Electric Power Research Institute considered some factors that influence the perception of risk and the valuation of benefits. Vivid portrayal of the consequences of a risk increases its rank with respect to other risks, he said; a vast amount of good data is available on nuclear-fission effects, ranging from a small increment in radiation to the possibility of a major catastrophe, that make for dramatic extrapolation. The initial perception of a risk tends to be "anchored" in people's minds; in the case of nuclear energy the dominant image is the mushroom cloud. Public judgment of statistical validity tends to be unduly influenced by unusual occurrences, so that reported instances of equipment failure are considered "near misses" and accident-free experience is undervalued. The potential benefits of nuclear power, on the other hand, are often stated in abstract economic terms that have little personal impact. Moreover, any future energy shortage is generally assumed to be fairly small in magnitude, and "this is indeed a reasonable assumption of the probable consequences." Just as people must consider small-probability, large-consequence events in assessing risk, however, so they ought to examine the medium- and low-probability consequences of energy and economic shortfalls.

Messages from the Medium

As F. Richard Stephenson and David H. Clark comment in their article "Historical Supernovas," which begins on page 100 of this issue, the remnants of supernova

explosions are among the most unusual objects known in astrophysics today. They include expanding nebulas, powerful sources of X-ray and radio emission, neutron stars and perhaps even black holes. The study of such objects yields a wealth of information not only about the nature of supernovas but also about the medium between the stars of our galaxy. Now for the first time supernova remnants are being examined in a galaxy outside our own.

The remnants are in the Large Cloud of Magellan, one of a pair of galaxies that are satellites of our galaxy. J. Danziger and M. Dennefeld have examined the remnants with the one-meter Schmidt telescope of the European Southern Observatory at La Silla in Chile. One remnant, N 132 D, had actually been studied before, but Danziger and Dennefeld have specifically shown that it is the remnant of a supernova explosion resembling one in our galaxy that in 1700 gave rise to the powerful radio source Cassiopeia A. Furthermore, their investigation suggests that N 132 D is identical with a source of particularly intense X rays that has only recently been detected. If that is so, the fact that the Large Cloud of Magellan is so far away (some 200,000 light-years) would make N 132 D the most powerful X-ray emitter of all known supernova remnants.

One reason N 132 D might be such a strong emitter of X rays is that the interstellar medium within the Large Cloud of Magellan is denser than that within our galaxy. Both the Large and Small Clouds of Magellan appear to be in a stage of evolution quite different from that of our galaxy, and a comparatively large fraction of their mass is represented by interstellar gas. As the shock wave from a supernova explosion sweeps through that dense medium, it would excite the medium to emit more X rays than would be emitted by a more rarefied medium. By the same token the dense medium would retard the expansion of the remnant quicker than the more rarefied medium of our own galaxy would, thus affecting the remnant's entire evolutionary history.

Almost all the supernova remnants in the Large and Small Clouds of Magellan appear to be associated with H II regions (regions of ionized hydrogen). In our own galaxy few remnants are associated with such regions. It may be that intensive study of supernova remnants in the dense interstellar mediums of the Clouds of Magellan and other galaxies, together with theoretical work on supernovas and supernova remnants, will make it possible to utilize supernova remnants to probe the dense and mysterious medium in the nucleus of our own galaxy.

Gonorrhea Resurgent

In the decade since oral contraceptives became generally available in the mid-1960's the number of new cases of human infection with the bacterium *Neisseria gonorrhoeae* in the U.S. has risen at an annual rate of 12 percent. Today gonorrhea is an epidemic disease out of control. The infected are predominantly adolescents and young adults; at any given time there are some three million of them. The annual number of new cases reported is one million, or three times more than the total of all other reportable infectious diseases. Moreover, public-health officials estimate that only one case in eight is reported.

Since gonorrhea can be effectively treated with antibiotics, it may seem surprising that it should have reached epidemic proportions. The reasons lie in several characteristics of the disease. First, a gonorrhea infection confers little or no immunity against subsequent infections. Second, whereas the consequences of a prolonged infection can be grave, many who contract gonorrhea are not seriously inconvenienced by its symptoms and do not seek treatment. Third, about 50 percent of those who contract the disease have no noticeable symptoms and therefore are unwitting carriers. It is estimated that each infected individual is currently responsible for an average of three new infections per year.

Reversing the exponential increase in gonorrhea calls for a two-pronged attack: the identification of those who are indifferent to their infection or unaware of it (so that they can be treated) and the conferral of immunity (to break the cycle of reinfection). Recent work by Charles C. Brinton, Jr., and his colleagues at the University of Pittsburgh offers promise of achieving both objectives.

Studies of the *N. gonorrhoeae* bacterium show that it has two main phases: a "hairy" phase, when the surface of the bacterial cell is covered with pili, or fine threads, and a smooth phase, when the pili are absent. Only the bacteria with pili are infective. The pili, however, induce the manufacture of a specific antibody by the immune system of the infected individual. Brinton has found that a simple blood test based on purified suspensions of pili will show the presence or absence of an antigen-antibody reaction. Conventional methods of determining the presence or absence of a gonorrhea infection are in comparison time-consuming, imprecise and expensive.

Tests with chimpanzees and with human volunteers, Brinton has shown, indicate that the pili can also be utilized to manufacture a vaccine that confers immunity to reinfection by the specific strain of bacteria from which the pili are harvested. The production of a broad-spectrum vaccine against gonorrhea, however, awaits further research because different strains of the infective bacteria have different kinds of pili.

Mare Terrestrium

All the conspicuous features on the surface of the moon are the result of impacts. These features include not only the craters, which plainly advertise their origin, but also the great maria, or "seas," which are craters that filled with lava following the

If you could teach people who would listen, would you teach?

The Navy's Nuclear Power program is looking for a limited number of highly qualified instructors. People who have recently earned degrees in mathematics, physics, nuclear physics, engineering—or have practical knowledge of nuclear propulsion systems—and would like the opportunity to share that knowledge with others.

The young people you'd be teaching would be top students. They have to be. The Navy has over 20 years' experience in the nuclear field. The men we choose to command, operate and maintain our more than 140 nuclear reactors have to be the best.

But it's more than just a "teaching job", because you'd be a commissioned Officer in the U.S. Navy. The benefits are excellent—including good pay, housing and available medical care. Your four years would be spent at either Orlando, Florida, or Vallejo, California.

So if you've ever said to yourself, "I'd really like to try my hand at teaching", now's your opportunity.Contact the Officer Programs Officer at your local Naval Recruiting Office. Or call the following toll-free number—**800-841-8000.** Be someone special. Be an instructor in

The Nuclear Navy

impact of very massive objects. Most of the impacts took place during a relatively brief period about four billion years ago, when debris left over from the formation of the solar system was swept up by the planets and their satellites. The earth probably received as heavy a pelting as the moon did, and it therefore must have been densely cratered. It has recently been suggested that the earth may also have had maria analogous to those of the moon.

A crater on the earth would be unlikely to have survived for four billion years: it would have been obliterated by erosion or buried under sediments. Such surface features are also destroyed by the earth's tectonic activity: they are swallowed up at subduction zones, where segments of the earth's crust plunge into the mantle and are melted. Ancient maria too would have disappeared as recognizable surface features, but it is possible that some remnant of the maria material would have survived. This possibility is discussed in *Geology* by A. Y. Glikson of the Australian Bureau of Mineral Resources, Geology and Geophysics.

The maria would have formed in the basins created by impacts large enough to shatter or fracture even deep layers of the earth's primitive crust. Such impacts might stimulate activity in the mantle that would result in a massive upwelling of magma. Over a period of several hundred million years a substantial portion of the earth's surface might be flooded with lava. Glikson has pointed out that some specimens of a class of very old rocks have properties that suggest they could have been formed in such an event.

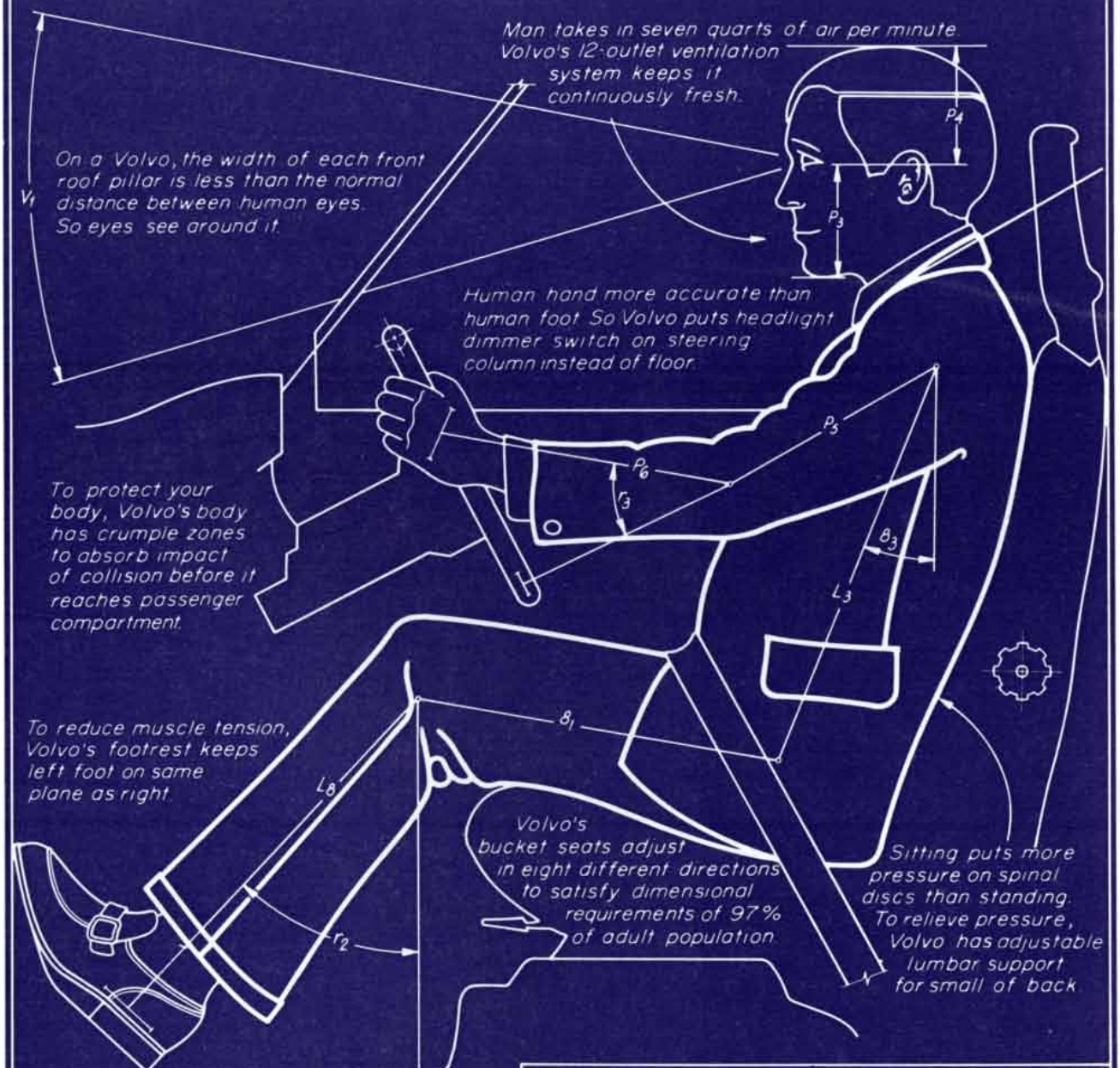
The rocks are called ultramafic-mafic volcanic rocks, which means that they are chemically basic and that they have a fine-grained or glassy structure. They are found as xenoliths—literally, foreign rocks—embedded in the ancient continental material known as greenstone belts. The greenstone belts contain the oldest rock known on the earth: its age, determined by the isotopic composition of the elements it contains, ranges from three billion to 3.8 billion years. Since the ultramafic-mafic xenoliths were incorporated intact in the greenstone at the time it formed, the xenoliths are necessarily somewhat older.

The ultramafic-mafic rocks are in many ways similar to the material of the modern crust underlying the oceans. The modern crust too has a magmatic origin: it is created by the upwelling of molten rock, primarily through a system of mid-ocean ridges. It is possible that an ancient ultramafic-mafic crust formed in the same way, but Glikson cites two items of evidence that argue against this conclusion.

First, if a system of mid-ocean ridges existed four billion years ago, then there must also have been sea-floor spreading, the development of large continents and the large-scale movement of crustal plates over the earth's surface. Actually there is evidence that these processes did not begin until somewhat later in the earth's history. Sec-

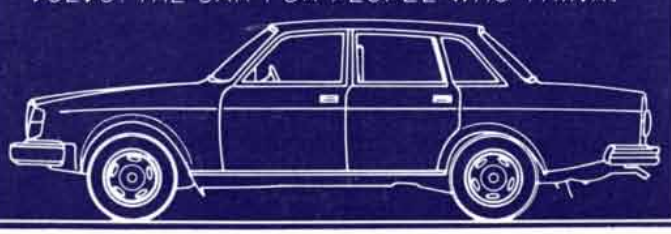
OF ALL THE THINKING THAT WENT INTO VOLVO, THIS IS WHAT WE THOUGHT OF MOST.

These days, you find a lot of car makers copying each other's designs. In building a Volvo, we're more influenced by yours.



Before the average driver can move his foot from the gas pedal to the brake in a panic situation, a car will travel 56 feet (at 55 m.p.h.). So Volvo puts power disc brakes on four wheels, not just two.

VOLVO. THE CAR FOR PEOPLE WHO THINK.



ond, from the structure and composition of the xenoliths it is possible to deduce the conditions under which they were formed. These conditions are unusual and quite unlike those thought to prevail at present-day mid-ocean ridges. The rock underwent a high degree of partial melting, a process in which it separated into fractions according to the melting points of its various constituents. Moreover, the melting was adiabatic, that is, it was not caused by the addition of heat. One way that rock might be melted adiabatically is through the release of pressure that would follow the deep fracturing of the crust.

Glikson emphasizes that this interpretation of the ultramafic-mafic xenoliths is only a hypothetical one; he notes that the data are "consistent with, although not necessarily indicative of, impact of extraterrestrial matter." Other ultramafic-mafic rocks were produced in volcanic events as recently as 2.6 billion years ago, and in some cases their geological context has been preserved so well that it can be concluded they definitely were not formed as a consequence of impact-related melting. Thus even if the xenoliths in greenstone belts are relics of ancient terrestrial maria, there must also have been some other process, endogenous to the earth, for producing similar rocks.

Low-Pressure Approach

When foods and other commodities of biological origin are stored or shipped in large quantities, they are notoriously subject to spoilage, partly as a result of the action of microorganisms. The spoilage is customarily retarded by such stratagems as drying, ensilage or refrigeration. A Florida plant physiologist, Stanley Burg, has introduced a new approach: lowering the air pressure.

In hypobaric storage, as it is called, products such as meats, vegetables and seedlings are kept in an insulated container at a pressure ranging from 1/76 atmosphere to .5 atmosphere, depending on the product. The low pressure is created and maintained by a vacuum pump of modest capacity. The preservative effect is due partly to the diminished amount of oxygen. The multiplication of aerobic bacteria and molds, which require free oxygen, is inhibited. The multiplication of anaerobic bacteria is retarded by continuously admitting a small amount of humid air, so that the atmosphere in the container is changed at a rate of about two volumes per hour. In effect the atmosphere is adjusted so that it is not favorable to either kind of bacteria.

Another effect is that gases evolved by the product dissipate quickly in the low-pressure environment. Stored fruit, for example, evolves ethylene, which is a ripening hormone. The combination of low pressure and changing air carries the gas off rapidly, which keeps the fruit fresh longer. Moreover, the fact that the environment is low in oxygen slows down the respiration of such perishables as fruit and cut flowers, thereby

Use this coupon to send us your Tricentennial idea.

Send your idea to:
Tricentennial
Atlantic Richfield Company
P.O. Box 2076
Los Angeles, CA 90053

From: _____

Address _____

City _____ **State** _____ **Zip** _____

My idea is:

We have always been a nation more interested in the promise of the future than in the events of the past.

Here at Atlantic Richfield we see the future as an exciting time. The best of times. And we know that all of us can achieve a splendid future by planning for it now.

We'd like your help. We need your vision. America will change a great deal by the year 2076. We want you to tell us what you think those changes should be.

What do you envision as the best way to solve our energy problems?

Should we have a universal language?

How do you think architecture of the future can improve on that of the present?

Or, if those topics don't appeal to you, pick one that does.

Please note that all ideas submitted shall become public property without compensation and free of any restriction on use and disclosure.

ARCO

Petroleum Products of
AtlanticRichfieldCompany

Celebrate America's Tricentennial 100 years early.

The September issue of SCIENTIFIC AMERICAN will be devoted to the single topic of FOOD

Neil MacLellan

SCIENTIFIC AMERICAN

415 MADISON AVENUE, NEW YORK, N.Y. 10017 (212) 688-3900

contributing to the preservative effect. The temperature in the container is held between 30 and 60 degrees Fahrenheit by refrigeration; the level depends on what is being stored.

These conditions, according to Burg, are unusually effective against spoilage. With some perishables, such as cut flowers, hypobaric storage offers a degree of preservation (lasting for a matter of months) that cannot be achieved in any other way. The container for hypobaric storage can be as small as a home appliance, although the current work involves much larger units. The system is now being tested commercially with 40-foot shipping containers. A warehouse employing hypobaric storage is in the planning stage.

Divine Wind

The role of the weather in determining the outcome of several military campaigns that were turning points of history is well known. The invasions of Russia by the armies of Napoleon and Hitler, for example, were brought to a halt by the weight of the Russian winter. It is less well known that two major invasions of Japan by the forces of Kublai Khan were both abruptly terminated by typhoons that smashed the invading fleet. The two 13th-century debacles are described in *Bulletin of the American Meteorological Society* by J. Neumann of Hebrew University in Jerusalem.

Kublai Khan, the grandson of Genghis Khan, became ruler of China in 1260. Between 1267 and 1274 he sent five missions to the emperor of Japan. The emissaries carried dispatches calling on the emperor to submit to the Khan or have his country invaded. None of the emissaries saw the emperor; all were summarily deported by the Japanese military regent. Kublai Khan then ordered the ruler of Korea, one of his vassals, to build a fleet of ships for an invasion and to raise a force to sail them.

The fleet consisted of some 300 large vessels and between 400 and 500 small ones. The ships were manned by 7,000 Korean and Chinese seamen and carried an army of 8,000 Koreans and 20,000 Mongols and Chinese. They set sail from Korea in November, 1274, and had begun to land in Hakata Bay, near Nagasaki, when a typhoon struck. More than 200 ships and 13,000 men were lost, and the rest of the invading force retreated to the mainland.

In 1275 Kublai Khan sent another embassy to Japan demanding submission. This time the emissaries were executed, and the Khan began preparations for a second invasion. In 1281 two new armadas set sail, one from Korea with some 40,000 Mongols, Koreans and Chinese and the other from southern China with 100,000 men. The two armies had again begun to land at Hakata Bay, and again a typhoon struck and destroyed nearly all the invading fleet. In the annals of Japan the storms that had repulsed the Mongol invasions came to be known as *kamikaze,* or "divine wind."

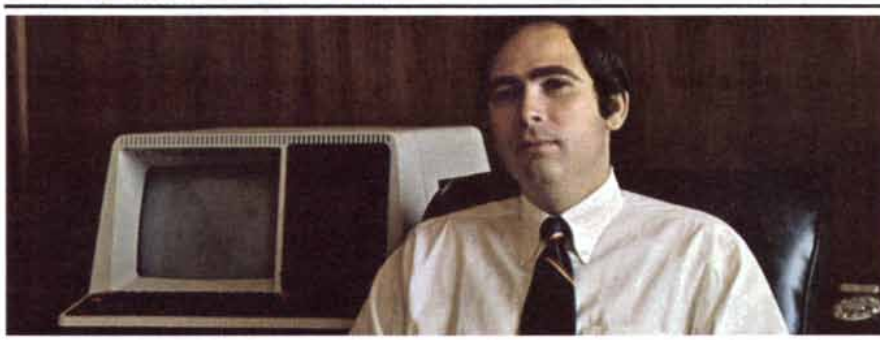
Digital credits these 3 men with saving $1.5 million in 18 months.

These 3 men credit Digital.

"Our DEC DATASYSTEMS will save us close to a million dollars in the next two years alone."

The Shawmut Bank of Boston recently tripled its timesharing computer power in order to expand their on-line computer services. Yet the bank is paying less for this service today than it did before turning to Digital.

Says Mr. Charles Hunt, Assistant Vice-President of Shawmut, "We were paying $15,000 a month for timesharing services. Then we installed two DEC DATASYSTEM 500's, which resulted in three times the computer capa-



"Six months after the DATASYSTEM went in, it was paying us back at the rate of $9,000 a month."

Gary Davis is president of the National Auto Parts Association's Pittsburgh group. Tracking down

"The purchase of our DECsystem-10's saved Ramada Inns $800,000 over Digital's nearest competitor."

Tom Castleberry, President of Micor, Inc., the Information Proc-

essing subsidiary of Ramada Inns of America, Inc., says, "In all, our company takes care of the reservations, accounting and property management for 2,100 hotels. Including 700 Ramada Inns. There's simply no doubt that our twin Digital DECsystem-10's give us the best way to handle this complex and voluminous data proc-

Saving managers millions.

Low cost computers that pay for themselves over and over again is what has made Digital one of the two computer companies in the world to install over 50,000 computers.

And what this means to you is that right at the time you need it most you have access to an enormous reservoir of experience and ideas for cutting costs while getting more done.

Digital is the pioneer in minicomputers, the leader in the development of small- to large-scale timesharing systems, and the company that has made many of the major breakthroughs in computerizing techniques.

Because of Digital, interactive computing, computer hierarchies and networks, data communications systems and new uses of computer systems are growing at an unprecedented rate.

And every year Digital places a remarkably comprehensive selection of computer tools within the financial reach of thousands and thousands of new users.

Today anyone who has a good applications idea for a computer can afford one. And just as important, they can get the help they

bility at a total cost of only $10,000 a month. We estimate the same capability would cost us $50,000 a month at a service bureau."

Thanks to Digital's flexibility, Shawmut has been able to apply this timesharing computer power to portfolio pricing, five-year bank forecasting, personnel management, bank location analysis and corporate financial planning. To do this, data communications terminals are located throughout the bank for instantaneous information.

"Our experience has been that our Digital DATASYSTEMS handle all this workload without bogging down on the job. We get the information we need in seconds."

Hunt concludes, "Money management is Shawmut's business. Computers are Digital's. When Shawmut can save approximately $40,000 a month using Digital computers, it shows we both do our jobs well." ■

inventory is something he knows a lot about.

"We have over 140,000 different auto parts spread over three large distribution centers. The headaches used to be enormous. Then we got our Digital DATASYSTEM. Not only has it made us more efficient, it has made us more competitive.

"Our service is faster and more accurate than ever. Our jobbers, many of whom have their own terminals, can order their parts through our computer 24 hours a day. We can tell a customer that the part he needs is in the warehouse, or when it will be—even when the warehouse is over 100 miles away. And that's service he can't get anywhere else.

"Yes, we looked at other systems. But we chose Digital for capability, cost savings, and besides, everyone we talked to said that Digital made the most reliable equipment." ■

essing load. The hardware is extremely cost-effective. And the ease-of-use gives us great economy. In fact, we only use two operators to do all phases of our computer work.

"Currently, we handle 9 million transactions a year. But we have the ability to go over 60 million. Today we are using less than 20% of our DECsystem-10 capacity for reservations. We have plenty of room for the future."

Right now, Digital's interactive teleprocessing system lets guests obtain hotel reservations in seconds. While it does on-line systems and applications programs for home office business accounting simultaneously.

The cost for all this efficiency? "Our dual DECsystem-10's cost us 50% less than the nearest competitor. In fact, with the money we saved, we covered the cost of 3 new buildings." ■

need from a company with over 4000 field representatives in 25 countries around the world.

To find out how Digital can help you save money, send for your free copy of "50,000 Computers Saving Managers Millions." Fill out the coupon below, or call or write Digital Equipment Corporation, Maynard, Mass. (617) 897-5111. European headquarters: 81 route de l'Aire, 1211 Geneva 26. Tel: 42 79 50. Digital Equipment of Canada, Ltd.

Please send me your free book "50,000 Computers Saving Managers Millions" to show me how I can get more work done at lower cost.
My computer application is _____

Name _____ Title _____

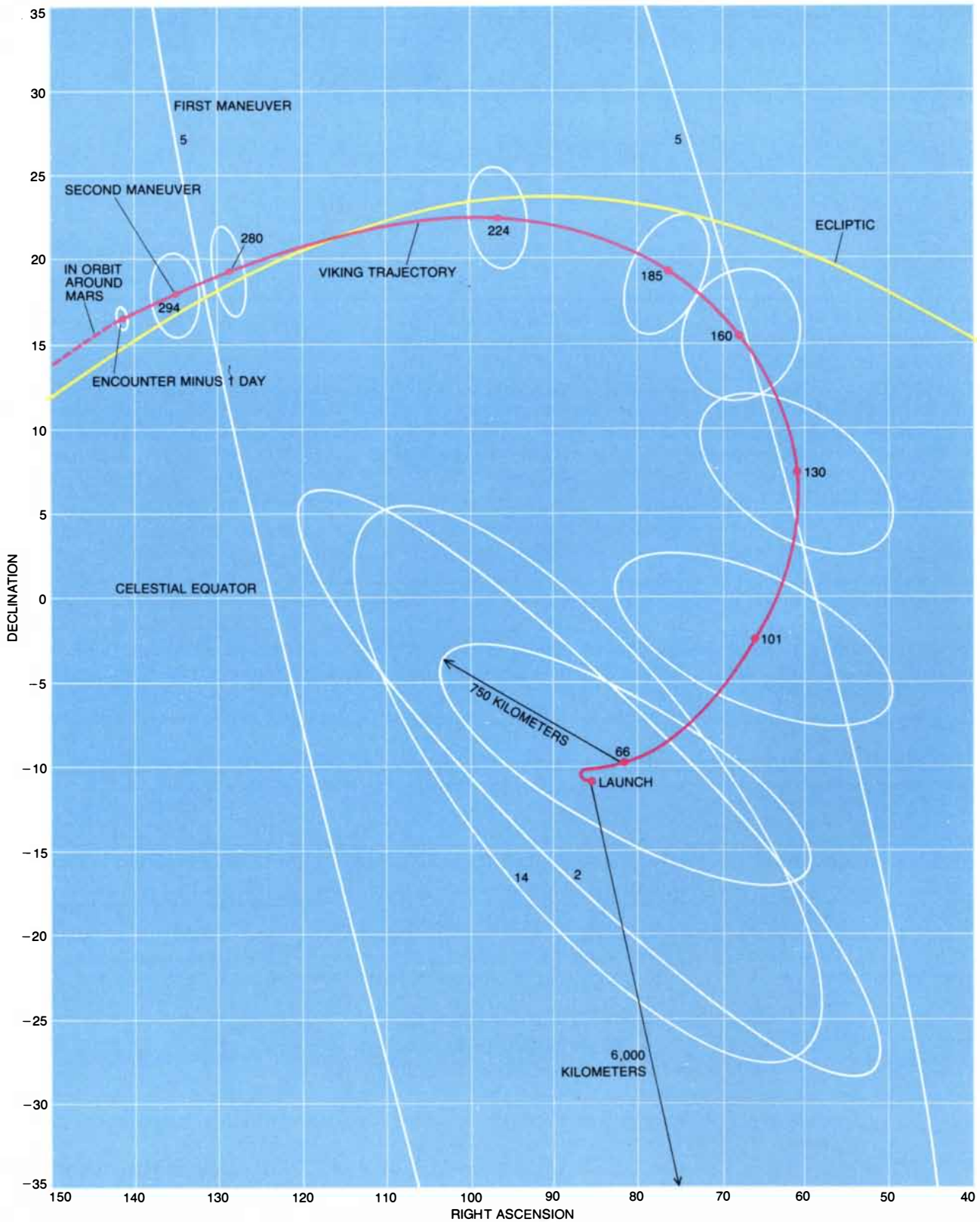
Company _____ Address _____

City _____ State _____ Zip _____

SA66

digital

50,000 computers saving managers millions.



NAVIGATION UNCERTAINTIES in the position and the velocity of *Viking 1* at a given point on its journey to Mars (traced here on the celestial sphere) can be translated into position uncertainties at the target planet, visualized as ellipses drawn in a plane perpendicular to the incoming trajectory. The numbers inside the uncertainty ellipses, all drawn to the same scale, indicate the number of days after launching. There is a 50 percent probability that the actual trajectory error at the target will fall within the computed ellipse. Velocity errors are the chief source of target error early in the flight but become less significant later on. The first mid-course correction, on the fifth day (designed to shift the aiming point at the target), introduces such uncertainties in velocity that the long axis of the uncertainty ellipse extends far beyond the edges of the illustration. The spacecraft appears stationary on the celestial sphere for several weeks after launching because it is traveling nearly radially away from the earth. Target-error uncertainties become smaller as more radio-tracking measurements are made. The target-error uncertainties continue to diminish until the effects of small random forces on the spacecraft begin to cancel information gained from the additional tracking. Total length of the spacecraft's journey from the earth to Mars is 700 million kilometers.

Navigation between the Planets

Advances in tracking and computing methods are such that when the first of the two Viking spacecraft reaches Mars late in June, its position should be known to within 50 kilometers

by William G. Melbourne

On June 19 *Viking 1*, the first one launched of two identical spacecraft now heading toward Mars with capsules designed to land on the surface and search for evidence of life, will reach the planet after a 10-month journey of 700 million kilometers. *Viking 1* is aimed at a target point 1,500 kilometers above the planet's surface. As the spacecraft encounters the target zone it will execute commands stored in its on-board computer that were sent earlier from the earth. Those commands will cause the spacecraft's retropropulsion system to fire for 43 minutes, subtracting enough velocity to place the spacecraft in an elliptical orbit around Mars that will vary in altitude from 33,000 kilometers at apoapsis down to 1,500 kilometers at periapsis. The orbital period will be 24.6 hours, so that *Viking 1* will circle the planet exactly once per Martian day.

The accuracy of the retropropulsion maneuver and ultimately the accuracy of the landing will depend on how well the position of the spacecraft is known at the time the commands are generated on the earth during the several days before the spacecraft arrives in the vicinity of Mars. The National Aeronautics and Space Administration flight team at the Jet Propulsion Laboratory that has been tracking *Viking 1* since its launching last August is confident that if the spacecraft continues to operate satisfactorily, the probability is greater than 50 percent that the error in determining its position before the encounter with Mars will be less than 50 kilometers. Since the first successful flight to Mars by *Mariner 4* in 1964 the accuracy of interplanetary navigation has improved by a factor of 100. It is realistic today to design planetary missions with a delivery accuracy of a few tens of kilometers for the terrestrial planets (Mercury, Venus and Mars) and a few hundred kilometers for Jupiter and Saturn.

The improvement in navigational accuracy, with the concomitant increase in scientific returns that was brilliantly demonstrated in the most recent photographic missions to Mars and Mercury, is the result of cumulative advances in key technologies. Radar measurements have greatly refined our knowledge of planetary distances and orbits. Over the past 15 years the stability of clocks has been improved by nearly four orders of magnitude. In the same period the area of the parabolic antennas for tracking space vehicles has increased sixfold, the sensitivity of radio receivers has been enhanced by a factor of 100 and antennas separated by nearly the diameter of the earth are now being used in pairs for sharpening tracking accuracy. There are also much improved techniques for calibrating the factors affecting the propagation of radio signals through the earth's atmosphere and over interplanetary distances. Finally, there have been steady refinements in the computer programs that process the tracking data and model the diverse and subtle forces acting on vehicles during their long free fall through space.

Navigation still means, as it did in the days of the Phoenicians, guiding a ship. It calls for a map, a travel plan, a means of periodically determining the ship's location and a method of selecting a new route when the ship has drifted off course. Interplanetary navigation satisfies this description by continually refining the map of the solar system, by devising a mission plan, by designing suitable spacecraft trajectories and by determining the trajectory en route and making corrections that enable the spacecraft to reach its interplanetary goal.

The Map

For interplanetary flight the map corresponds to a description of the time-varying positions, velocities and rotations of the earth and the other planets, together with such physical quantities as the planets' masses, sizes and shapes. A description of the orbital motion of the planets in the solar system is an ephemeris. Until about 1960 the accuracy of the planetary ephemerides was limited by the astronomical techniques for determining the positions of the planets (and in some cases the positions of asteroids) with respect to the background stars. From these measurements and gravitational theory the orbital positions of the planets were calculated and were commonly expressed in terms of the astronomical unit (essentially the mean distance between the sun and the earth). The resulting ephemerides were accurate to within a few parts per million, the equivalent of a few hundred kilometers for the terrestrial planets. Surprisingly, as late as 1960 the value assigned to the astronomical unit was in error by 66,000 kilometers, or nearly one part in 2,000. Since the astronomical unit is the scale factor needed to convert ephemeris positions into kilometers, if the value had been left uncorrected for the flight of *Mariner 2* to Venus in 1962, the spacecraft would have been aimed at a point on the wrong side of the planet.

The magnitude of the error was uncovered in 1961 when, with the development of large radio telescopes, high-power transmitters, ultrasensitive receivers and sophisticated signal-detection techniques, it became possible to detect radar echoes from Venus. The basic radar observation measures the two-dimensional distribution of the power of the returning echo signal with respect to its frequency and time delay. The rotation of the planet spreads the echo signal in frequency through the Doppler effect, and reflections from regions of the planet located at different distances from the earth cause the echo to be stretched out in time. From an analysis of this distribution of echo power the range to the spot on the planet closest to the earth (the subearth point) is obtained.

Radar systems have greatly improved over the past 15 years, so that today under favorable conditions the range measurement is accurate to about 50 meters for the terrestrial planets. Optical observations provide no direct means of determining range with any accuracy. Historically telescopes have been used to measure a planet's location along two axes perpendicular to the direction between the earth and the planet. For the terrestrial planets measurements of this kind have an accuracy of 500 kilometers. The ephemerides were formerly computed from a large number of such measurements over many decades.

The remarkable accuracy with which radar echoes yield range is somewhat misleading. As the planet rotates, the distance of the subearth point from the center of the planet changes because of the planetary topography. These variations affect the range measurement and must be mapped and accounted for to fully utilize the radar obser-

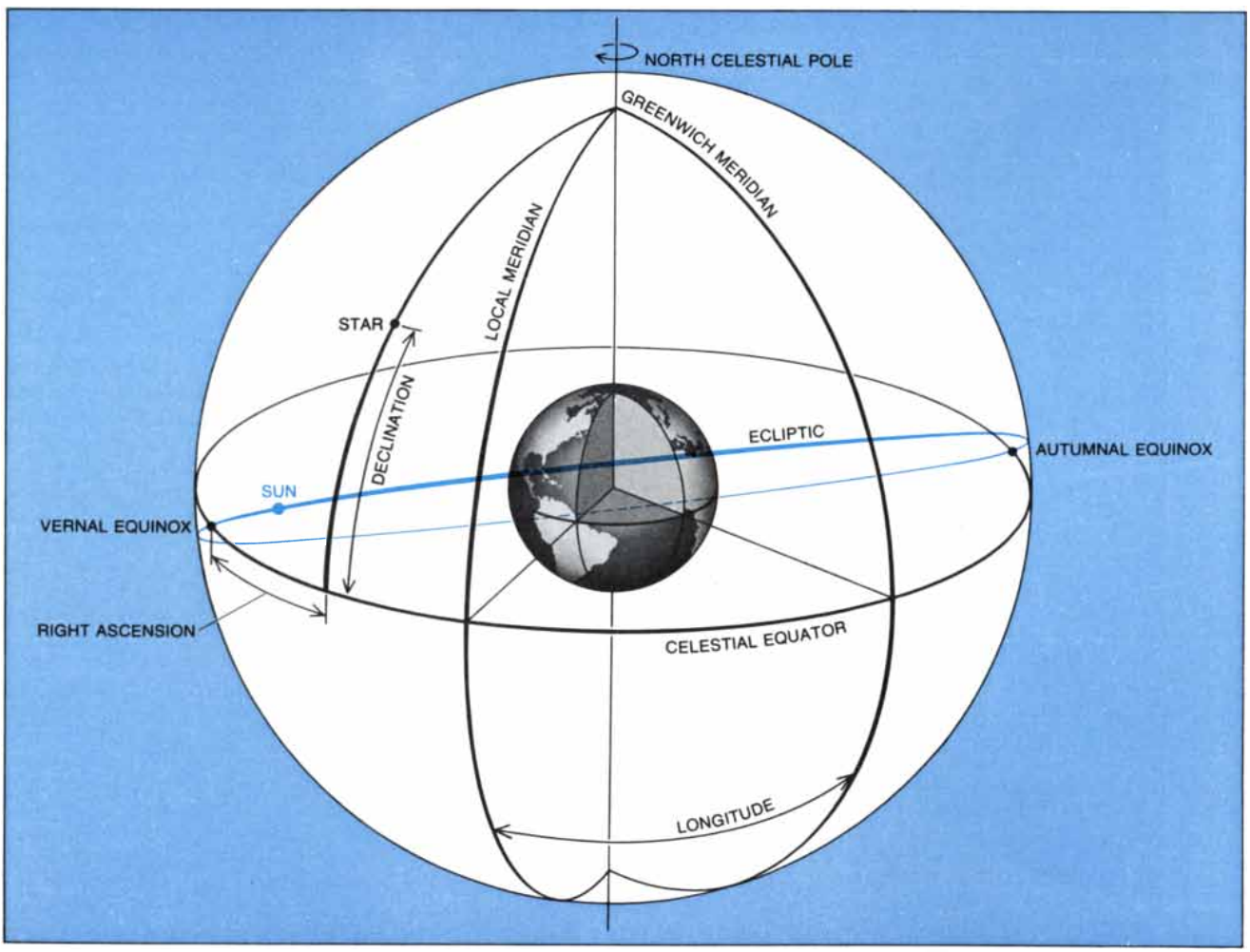
vation for improving the ephemeris. The ephemeris describes the motion of the center of the planet, whereas radar measures range to the surface. An equivalent range measurement to the center would be accurate to about one kilometer. The range measurement is blind to certain combinations of displacement in the positions of the planets. For example, a uniform shift in the orientation of the entire solar system with respect to the background stars would not be detected by the radar measurement. As a result radar observations must be supplemented with other types of observations that give the orientation of the solar system.

Also of concern is the fact that the terrestrial planets follow orbits around the sun that are tilted anywhere from 1.9 to seven degrees with respect to the plane of the ecliptic (the plane of the earth's orbit). Radar measurements are relatively insensitive to a planet's displacement perpendicular to the plane of the ecliptic because such displacement alters the range only slightly. Nevertheless, on the basis of several thousand radar observations of the terrestrial planets made over the past decade the accuracy of the planets' positions with respect to the earth has been improved to one or two kilometers in the components lying in the plane of the ecliptic and to 10 or 20 kilometers in the direction perpendicular to the plane of the ecliptic. The astronomical unit is now known to be 149,597,871 kilometers, with an uncertainty of about one kilometer.

Keeping Time

The problem of accurately determining the ephemeris of a planet is analogous to the problem faced by early explorers trying to determine the position of land masses on the earth's surface, which is apparent in the distortions in 16th-century maps of the New World. A close examination of such maps reveals two interesting characteristics. The latitudes assigned to places in the New World are usually accurate to better than one degree, or about 100 kilometers. The longitude values, on the other hand, are often badly in error, sometimes by as much as 30 degrees. Latitude, or distance from the Equator, is readily determined by observing the position of the celestial pole in the night sky: the point in the sky near the star Polaris (in the Northern Hemisphere) around which the stars appear to revolve because of the earth's rotation. The elevation of the celestial pole above the horizon is equivalent to the latitude of the observer.

The determination of longitude is more difficult because of the earth's rotation. Imagine that the earth is at the center of a hollow spherical shell of infinitely large radius and that the stars are fixed to the surface of the sphere [*see illustration below*]. The angular position of a star on the sphere is given by its right ascension and declination; right ascension is analogous to longi-



CELESTIAL SPHERE, the conceptual basis for describing the traditional astronomical coordinate system, is a westerly-rotating shell of infinite radius on which the stars are assumed to be fixed. The celestial equator is a projection of the plane of the earth's Equator. A plane through the earth's axis of rotation and Greenwich Observatory in England defines the Greenwich meridian. The angular position of a star (or of a spacecraft) is given by its right ascension and declination. Zero right ascension and declination, also referred to as the vernal equinox, is the point where the sun on its apparent northward journey across the sky crosses the celestial equator in the spring. The vernal and autumnal equinoxes therefore mark the points where the plane of the earth's orbit, the plane of the ecliptic, intersects the celestial equator. Right ascension is measured in an easterly direction along the celestial equator from the vernal equinox to the meridian passing through the star. Declination is measured along the meridian from the celestial equator to the star. The orientation of the celestial sphere at a given instant is provided by the right ascension of stars crossing the Greenwich meridian, which is equivalent to sidereal time.

tude and declination to latitude. The celestial sphere rotates around the celestial pole once a day, reflecting the earth's rotational motion. The orientation of the celestial sphere around its axis of rotation at any one instant is customarily given by Greenwich sidereal time, the equivalent of the right ascension of the stars just crossing the Greenwich meridian.

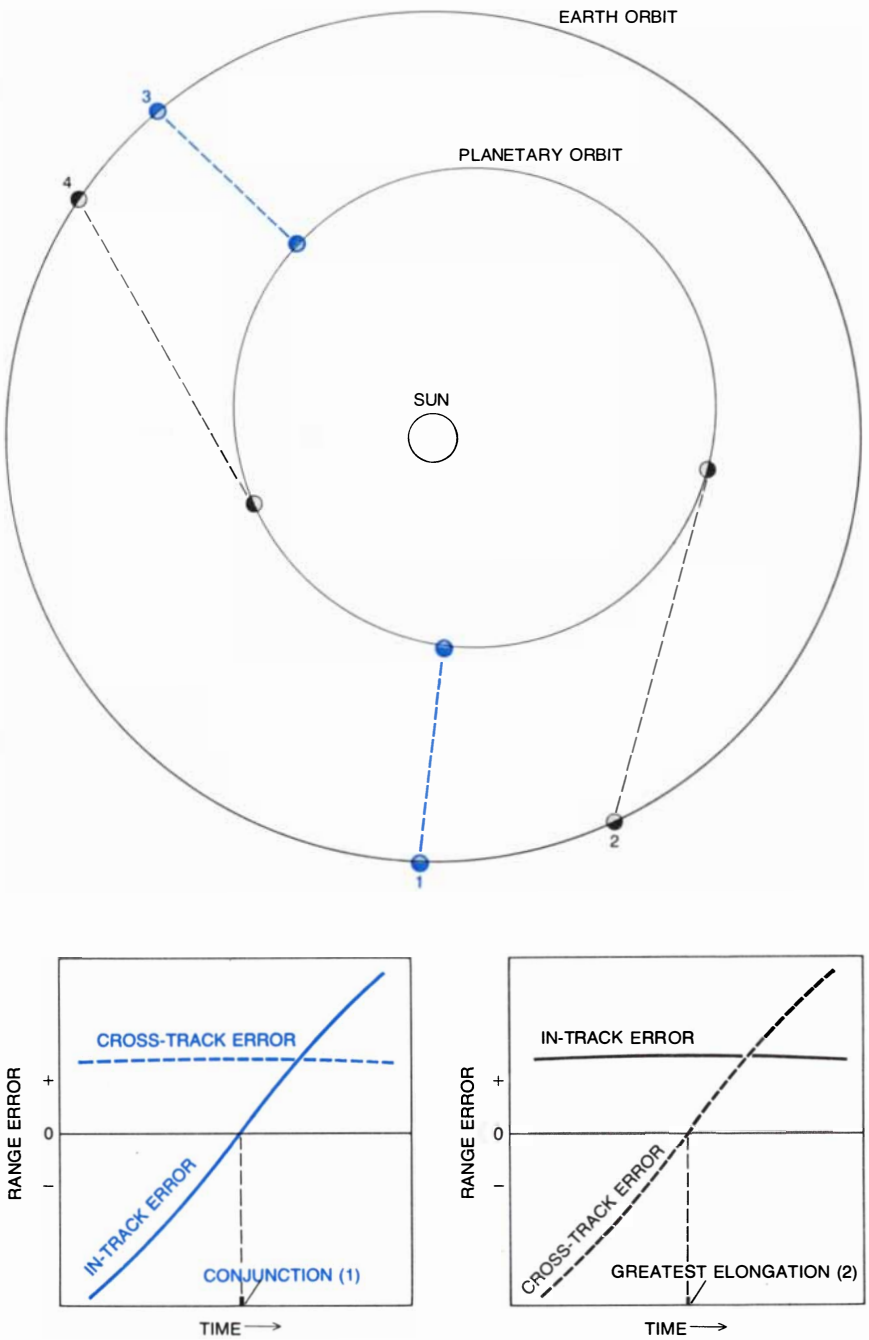
To obtain the longitude at night one observes the direction of one star or more and the time of the observation. With these observations and a catalogue giving the stellar positions, the right ascension of the stars crossing the local meridian at a given instant can be calculated. If the observer at that instant also knows the time in Greenwich, he knows the right ascension of the stars crossing the Greenwich meridian. The difference in right ascension between the stars on the Greenwich meridian and those on the local meridian is the longitude of the observer. Thus a critical element in navigating by the heavens is knowing what time it is in Greenwich when one is not there. Until the middle of the 18th century, when the first chronometers were developed and an alternate technique was introduced that made use of the moon's position on the celestial sphere, mariners had no way to keep time with any accuracy during voyages lasting several months; hence the large longitude discrepancies seen in the early maps.

The accuracy of timekeeping has improved by nine orders of magnitude over the past two centuries, with more than half of the improvement coming in the past 25 years. A good 18th-century chronometer could run for a week or two without gaining or losing more than a second. By 1945 quartz-crystal clocks had been developed that would maintain one-second accuracy for several years. Today's hydrogen masers are stable to better than two parts in 10^{15} over 24 hours. If the error in the rate of such a maser remained constant, a gain or loss of one second would take more than 10 million years.

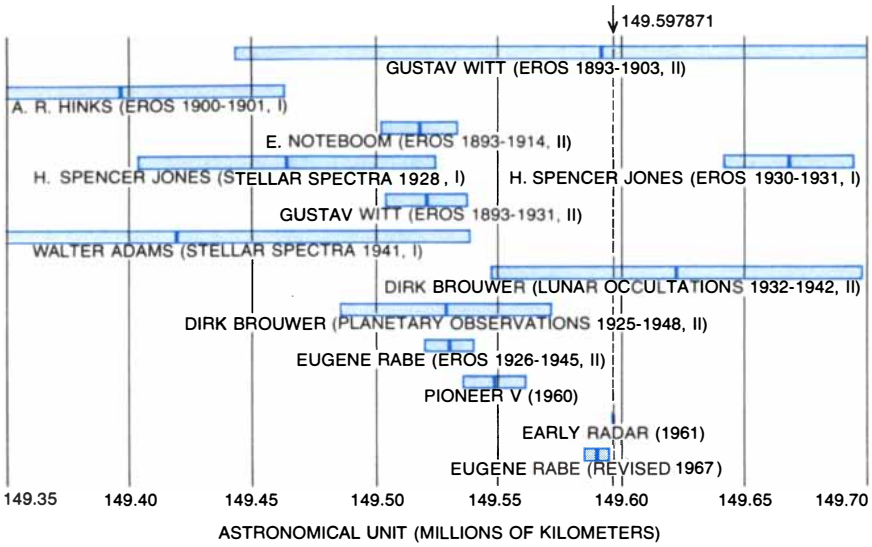
Interplanetary navigation makes use of time in two different ways. Time gives the epoch associated with an event, such as a specific observation. For that purpose it is sufficient to know the correct time to the nearest tenth of a millisecond, since planets and spacecraft move only a few meters during such an interval and the earth rotates through only a few centimeters. The more critical role of time is in the making of radio observations of the range and radial velocity of a spacecraft. That calls for keeping track of the motion of a radio signal traveling at the speed of light with an accuracy of a few decimeters (10^{-1} meter). For this purpose one must be able to measure time intervals to an accuracy of nanoseconds (10^{-9} second) for periods of several hours. The task requires clocks that are stable to about one part in 10^{13}.

Space Communications

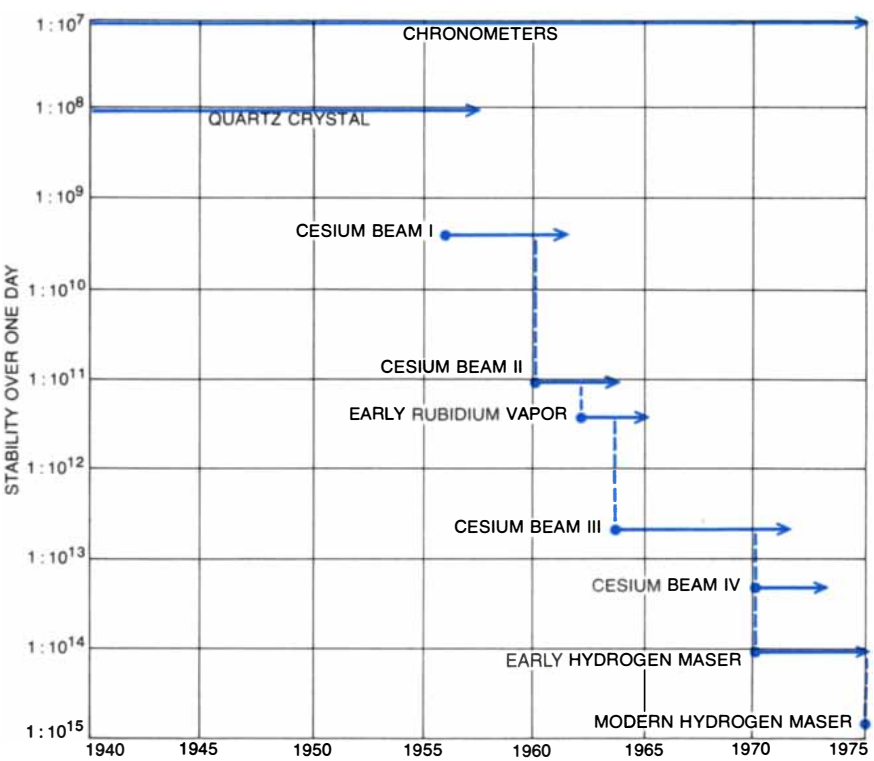
The measurements used to navigate U.S. spacecraft are derived from the radio link



PLANETARY ORBITS have been determined with great accuracy over the past 15 years by measuring the round-trip travel time of radar echoes from the planets at many points in their orbit. The range to the point on a terrestrial planet's surface that directly faces the earth is now measured to an accuracy of 50 meters. In practice a weighted least-squares procedure is used to find the values of well over 100 parameters that describe the orbital motion and physical properties of the planets. The computation leads to the best fit between the theoretically predicted observations and the actual ones. An error in a given parameter gives rise to a corresponding error in the predictions that develops a characteristic "signature" over a period of time. The magnitude of the signature and its degree of distinctiveness compared with the signatures of other parameters determine the accuracy with which the parameter can be evaluated from the observations. In this simplified case the orbital parameters and the radius of an inner planet in the same plane are to be determined. When the planet and the earth are at conjunction (*1*), an in-track error (an error along the planet's path) is perpendicular to the earth-planet line of sight and would not give rise to a range error. As the planets move away from conjunction the in-track error begins to show up as a range error that develops a unique *S*-shaped signature over a few weeks of tracking (*bottom left*). The slope of this signature is very sensitive to an in-track error and provides an accurate determination. A cross-track error (one perpendicular to the path) gives rise to a nearly flat signature that resembles a radius error. At the position of "greatest elongation" (*2*), when the angle between the planet and the sun is at a maximum, the signatures of the in-track and cross-track errors are reversed, allowing the latter to be isolated. Later, when the planets have lined up again at *3* and *4*, signatures are slightly modified by eccentricity and orientation of orbit and provide additional information about these parameters.



ASTRONOMICAL UNIT (A.U.), which is essentially the mean distance between the sun and the earth, was difficult to measure accurately before the advent of planetary radar. Preradar measurements were based either on direct (*I*) or dynamical (*II*) methods. One direct method involved triangulating the distance to the asteroid Eros on occasions when it passed within about 20 million kilometers of the earth. Another direct method exploited the Doppler shift in stellar spectra to determine the velocity of the earth. Dynamical methods use the observed perturbations caused by the earth in the orbits of other bodies to obtain the ratio of the sun's mass to the earth's mass. This ratio plus the known value of the universal gravitational constant times the earth's mass yields a value for the astronomical unit through a defining relation. The most successful application of the perturbation method involved observations of the orbit of the asteroid Eros. Successive determinations culminated in the value obtained in 1950 by Eugene K. Rabe. Subsequently Rabe discovered a computational error, which led to the revised 1967 value. The value based on the flight of *Pioneer 5*, the first spacecraft determination of the astronomical unit, was probably corrupted by effects of solar radiation pressure. Planetary radar, which determines to an accuracy of one part in 10^8 the time required for radio waves to travel one astronomical unit, yields the value of 149,597,871 kilometers, with an uncertainty of one kilometer.



STABILITY OF CLOCKS has increased by about six orders of magnitude, or a factor of one million, over the past 30 years. Stability is expressed as the relative change in frequency averaged over 24 hours. Recent advances in timekeepers have exploited the extremely sharp radiation frequencies of certain molecules to regulate vibrational frequency of crystal oscillators.

between the spacecraft and the earth. The current Viking missions to Mars provide a typical example. In flight the Viking spacecraft are stabilized in orientation so that the panels that convert sunlight into electric power face the sun. The directional antenna of the spacecraft's radio system points toward the earth. The antenna transmits a phase-modulated radio signal that carries scientific and engineering data from the spacecraft to the radio tracking stations on the earth. The antenna also receives commands and information carried by radio signals sent from the tracking stations.

The earth-based part of the radio-communication system is a network of large radio telescopes, or tracking stations, located geographically so that the spacecraft is continuously in view. There are three principal tracking-station sites: one in California in the Goldstone Lake region of the Mohave Desert, one in Spain near Madrid and one in Australia near Canberra. Each site is equipped with antennas that are 26 meters and 64 meters in diameter. The tracking stations maintain a radio-communication link with the Space Flight Operations Facility at the Jet Propulsion Laboratory in Pasadena, Calif. The facility originates commands to the spacecraft and receives data from the spacecraft by way of the tracking stations. Thus the planetary mission is actually remotely controlled from Pasadena.

The need for the large antennas is twofold: to radiate a signal strong enough to be picked up across interplanetary distances by the spacecraft's small antenna and to collect the weak signal returned by the craft's necessarily low-power transmitter. For the purposes of transmission the larger the antenna, the narrower the beam of electromagnetic energy radiated toward the spacecraft. For example, when the Goldstone 64-meter antenna is operated at a frequency of 2.2 gigahertz (2.2 billion cycles per second) in the *S* band of the radio spectrum, half of the power in the transmitted beam falls within a circle with a diameter of about 1,000 kilometers at the distance of the moon. At 300 million kilometers, which will be the approximate distance between the earth and Mars when the Viking spacecraft have reached their goal, the diameter of the beam is about 700,000 kilometers. Although the total power contained in the beam is still half of the amount originally transmitted, which can reach 400 kilowatts, the spacecraft's 1.5-meter antenna can intercept only a minute fraction of it, about 10^{-12} watt.

The radio signal emitted by a Viking spacecraft is only 16 watts, and since the dish on the spacecraft is small, the return beam has spread out to a diameter of many millions of kilometers by the time it has arrived at the earth. Nevertheless, a Viking spacecraft could be successfully tracked even if it were 50 astronomical units away, or beyond the orbit of Neptune. A 64-meter antenna with a cryogenically cooled ruby maser in the first amplifier of its receiver is capable of tracking the phase of carrier signals as weak as 2×10^{-21} watt, which is

roughly equivalent to the radiant power the antenna would collect from a lighted match on Mars.

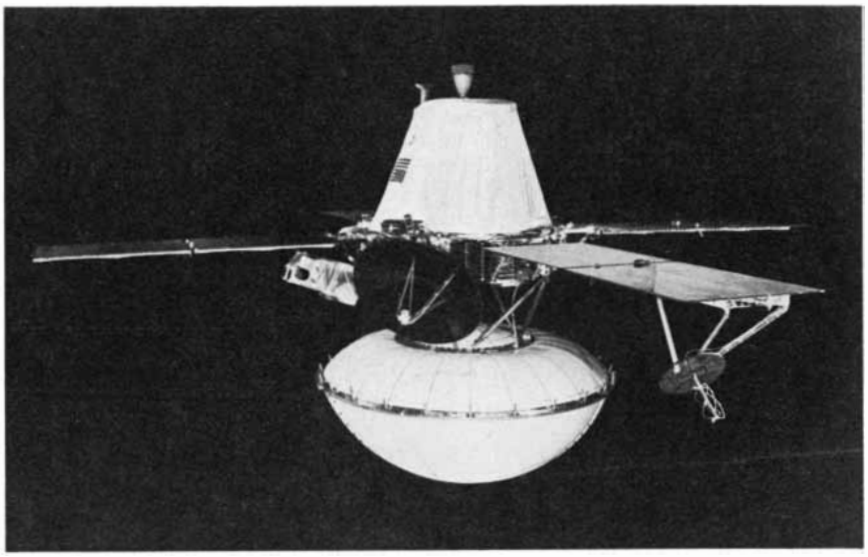
The tracking stations are designed to transmit and receive simultaneously. The downlink frequency from the spacecraft to the earth is about 10 percent higher than the uplink frequency of 2.2 gigahertz in order to eliminate interference between the outgoing and the incoming signals.

Navigation Measurements

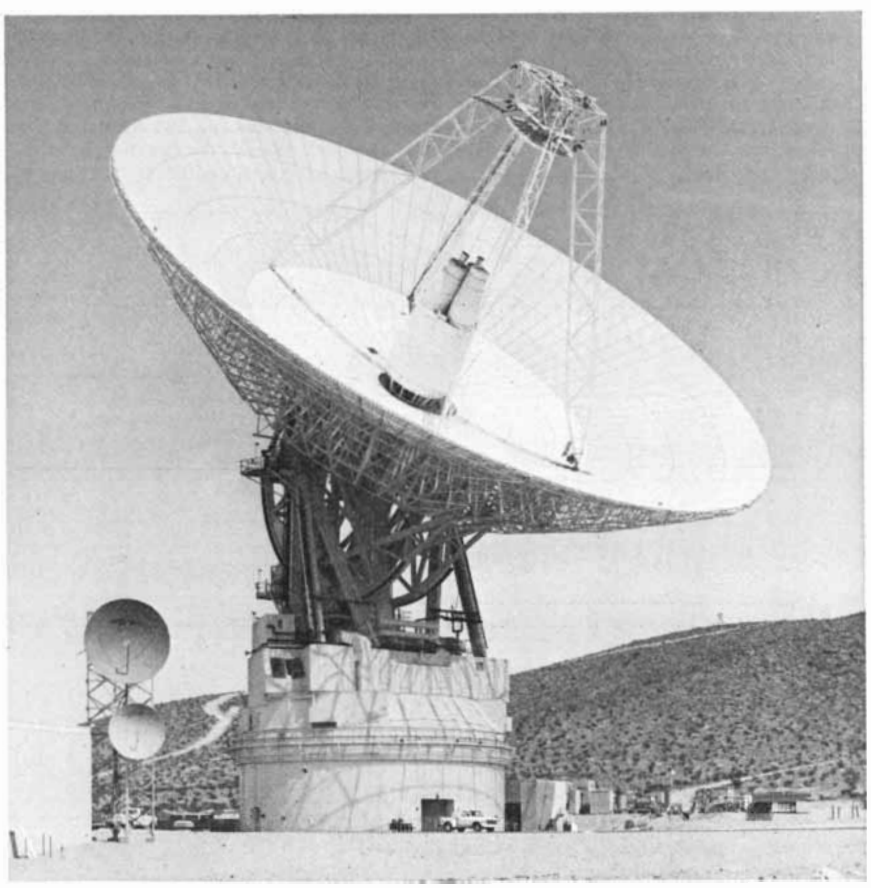
The navigation measurements that are obtained by the radio system are range and the rate of change in range, or radial velocity. The radial velocity is obtained from the Doppler shift in the frequency of the radio signal. The simplest Doppler measurement is the downlink, or one-way, mode, in which the frequency of the signal received by the tracking station is compared with the best estimate of the frequency of the signal sent by the spacecraft. The oscillator a spacecraft can carry is not as stable as it might be and its frequency is not accurately known; hence it cannot be relied on for measuring the craft's radial velocity. In the two-way coherent mode, which is the one that is normally used, the tracking station continuously transmits and receives. The signal transmitted from the earth is received by the spacecraft's radio system, which faithfully replicates the received phase, increases its frequency by 10 percent and transmits it back to the earth. In this mode the downlink signal is said to be coherent with the received uplink signal.

Because the spacecraft is moving with respect to the earth the craft's radio system "sees," or receives, an electromagnetic wave whose frequency differs from the one transmitted from the earth. Thus if the distance between the spacecraft and the tracking station is increasing, the craft's radio system retransmits a signal whose frequency is shifted to a slightly lower value by the Doppler effect. On the downlink trip from the spacecraft to the earth the frequency of the signal received by the tracking station is again shifted downward because of the increasing downlink distance. As a result a double Doppler shift is registered in the two-way mode. The difference between the frequency received by the tracking station (compensated for by the 10 percent downlink offset in frequency) and the frequency transmitted one round-trip time earlier is the actual Doppler frequency. In practice the measured Doppler frequency is obtained by comparing the frequency of the received signal with the frequency of the signal currently being transmitted instead of the signal transmitted one round-trip time earlier. The error this introduces is negligible because the stability of the transmitting frequency, being better than one part in 10^{13}, prevents any significant buildup of frequency error during one round-trip time interval.

The Doppler tone that results from comparing the received and the transmitting frequencies is typically on the order of 100



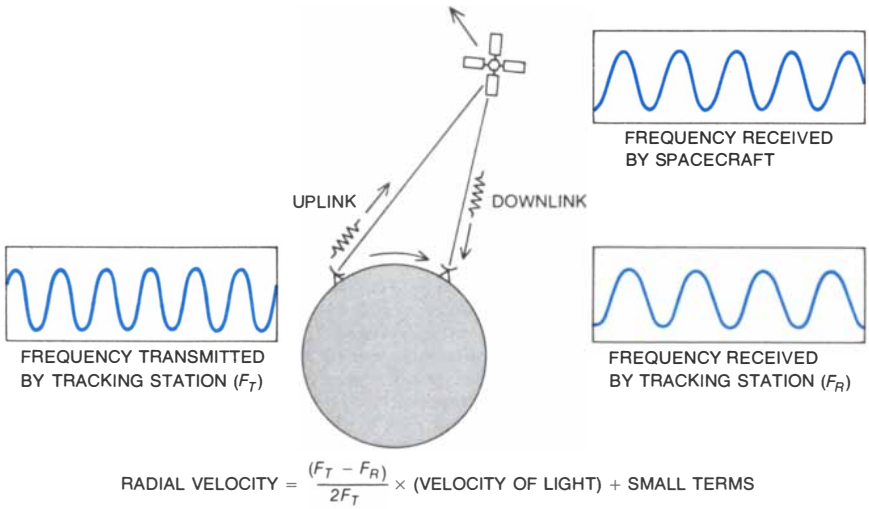
TWO VIKING SPACECRAFT now en route to Mars were launched on August 20 and September 9, 1975, with Titan III-E Centaur vehicles. *Viking 1* is scheduled to reach Mars and go into orbit on June 19; *Viking 2* will follow in orbit on August 7. Each craft weighs 3,500 kilograms, measures 9.7 meters between the tips of its solar panels and carries a 1,060-kilogram lander capsule. The two vehicles will conduct an intensive reconnaissance of the planet from orbit before releasing their landers, which will examine the Martian soil for signs of life and conduct other experiments. The Viking project is managed by the Langley Research Center of the National Aeronautics and Space Administration. Orbiters were built by the Jet Propulsion Laboratory of the California Institute of Technology, landers by the Martin Marietta Corporation.



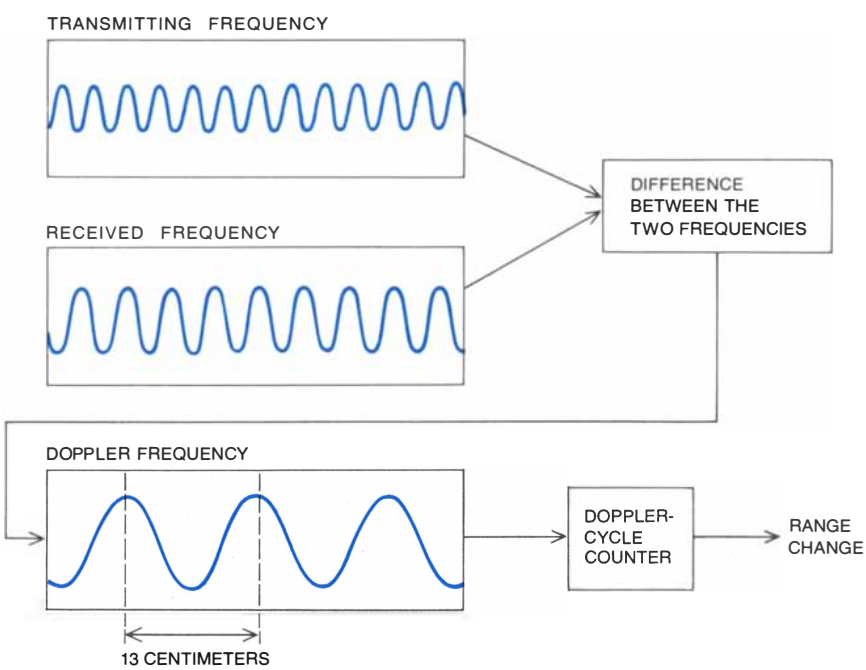
SIXTY-FOUR-METER PARABOLIC ANTENNA at Goldstone, Calif., is operated by the Jet Propulsion Laboratory for radio communication with interplanetary spacecraft. Similar tracking antennas located near Madrid and Canberra enable the Jet Propulsion Laboratory's Space Flight Operations Facility to maintain an unbroken communications link with space vehicles. Antennas are designed to transmit at *S*-band frequencies (2.2 gigahertz) while simultaneously receiving signals from spacecraft at *S*-band and *X*-band frequencies (8.5 gigahertz).

kilohertz. A Doppler-cycle counter keeps track of the number of successive wave crests of the Doppler tone that have accumulated with time, and it resolves the phase with a precision of a hundredth of a cycle. During the time interval between two successive Doppler wave crests the phase of the received radio wave has slipped one full cycle with respect to the phase of the transmitting radio wave or, what is equivalent, the combined uplink and downlink distance has changed by exactly one wavelength of the transmitting signal. Since the wavelength in the *S* band is about 13 centimeters, the spacecraft has moved radially about 6.5 centimeters during the full cycle. By reading the Doppler-cycle counter at successive times one obtains the change in two-way range over that time interval. Doppler accuracies are often given in terms of radial velocity, although the fundamental measurement in this Doppler system is range change. The accuracy of the average radial-velocity measurement over a one-minute counting interval is about .1 millimeter per second under good signal conditions.

The other radio-navigation measurement, range, is obtained by determining the round-trip time of the radio signal. The current ranging accuracy is about 20 nanoseconds in round-trip time, which is equivalent to about three meters in one-way range. Although this accuracy is remarkable, it does not supplant the Doppler measurement. The Doppler system continuously tracks the relative change of phase of the received radio wave over a period of time, whereas the ranging system at discrete times measures the round-trip time that is determined by the velocity of the radio signal. For measuring "high frequency" range-change behavior over a short time interval, say several hours, the continuous phase-tracking provided by the Doppler system tends to be more accurate, to at most a few decimeters of error in range change over 12 hours. On the other hand, successive range observations made once a day over a period of many days or weeks provide a more accurate measure of "low frequency" range-change behavior.



DOPPLER DISPLACEMENT of the uplink and downlink signals is used to determine the spacecraft's radial velocity, that is, its velocity along a line drawn between it and the earth. If the distance between an outbound spacecraft and the tracking station is increasing, the uplink signal is Doppler-shifted to a slightly lower value when it is received by the spacecraft. The radio system on board the craft amplifies the received signal and coherently retransmits it back to the earth, where it is received with a second Doppler displacement. The spacecraft's radial velocity is given by the equation. The phrase "small terms" refers to the fact that the radial velocity also depends on small but important effects that have been omitted in this illustration.



DOPPLER-CYCLE COUNTER keeps track of the number of cycles, and precise fraction of cycles, by which the phase of the received signal falls behind the phase of the transmitting signal in the course of time. The wavelength of an *S*-band signal is about 13 centimeters. Accordingly if the received wave falls exactly one cycle behind the transmitting wave during some measured time interval, the two-way change in range between the tracking station and spacecraft is one wavelength, or 13 centimeters. The one-way change is half that, or 6.5 centimeters.

Orbit Determination

The spacecraft and the planets follow orbits through space that almost exactly obey the gravitational laws of motion. If a spacecraft's position and velocity in space (specified by a set of six numbers) are known for a particular instant of time, a unique trajectory can be computed. For every combination of position and velocity values the range and the Doppler shift vary with time in a unique way that is characteristic of the resulting trajectory. From the range and Doppler observations of a particular spacecraft one obtains curves of range and radial velocity plotted against time from which one can compute the actual trajectory of the spacecraft. The process of determining that trajectory, together with the values of any relevant parameters affecting either the trajectory or the observations, is termed orbit determination. The accuracy of orbit determination depends on the accuracy of the measurements, the geometric characteristics of the trajectory, the time span over which the observations are made and the accuracy of the earlier information about the values of the parameters appearing in the mathematical structure of the dynamical and observational models used to analyze the observations.

The dynamical model must include, in addition to the flight path of the spacecraft, the orbital positions and physical characteristics of the planets that affect the trajectory, the acceleration caused by the pressure of solar radiation on the spacecraft, the velocity changes produced by the propulsion system of the spacecraft and so on. The observational model must include the positions and velocities of the tracking stations

PLYMOUTH VOYAGER

NO OTHER WAGON WORKS HARDER ON A GALLON OF GAS.

As a wagon buyer, you're a special breed with special needs. Try Voyager on for size. It comes with two different wheelbases, plus an extended body version that seats 15 people. Which is three more than any other wagon. Obviously, the more people room, the more cargo room.

26 HIGHWAY MPG, 18 CITY MPG.
That's according to the latest estimated EPA test results on Voyager's Slant 6 with manual transmission. If you don't think those are incredible figures, just compare them with those of any other wagon built anywhere in the world. Your actual mileage may differ depending on how and where you drive, the condition of your wagon and its optional equipment.

A TURNING DIAMETER 3-FEET SHORTER?
That's right. 3 feet shorter than Ford or Chevrolet. Which means tight U-turns and easier parking. Especially with optional power steering. Plus standard power front disc brakes and independent front suspension.

HERE'S "THE CLINCHER"
For the first 12 months of use, any Chrysler Corporation dealer will fix, without charge for parts or labor, any part of our 1976 passenger cars we supply (except tires) which proves defective in normal use, regardless of mileage. The owner is responsible only for normal maintenance like changing filters and wiper blades.

CHRYSLER
Plymouth

In California, see your dealer for mileage data for California equipped vehicles.

The lens of the SX-70 Land camera is a window on unnoticed worlds.

Through its frame you can discover the imprint of Nature on a single leaf. Or the inexplicable beauty of a great machine. Or nobility in a care-worn face.

The SX-70 is automated to an exquisite degree so it never diverts your concentration from your subject. And its instant picture tells you, on the spot, whether you have captured the elusive quality that illuminates the commonplace.

Polaroid Corporation

and the spacecraft, together with small factors such as relativistic effects and the time delays incurred by the radio signal as it travels through the radio system of the spacecraft and through the ionosphere and troposphere of the earth. The parameters of the dynamical and observational models are adjusted to generate theoretical range curves and Doppler curves that match the set of observations as closely as possible. The difference between an actual observation and a theoretical value is termed a residual. If the residuals are small and have a fairly random pattern, the agreement is considered satisfactory and the models are used to predict the values of the observations that are to be made in the future.

This is the "moment of truth." If the agreement with later observations is poor, one must adjust the model and try the fitting process again. Usually only the trajectory parameters of the spacecraft need to be adjusted, but occasionally the residuals can be satisfactorily reduced only after other parameters of the model are reexamined. For example, the trajectory of a spacecraft in orbit around a planet is gravitationally affected by the nonspherical shape of the planet and by the slightly irregular distribution of its mass. Conversely, the residuals that result from those effects can be analyzed to gather information about such planetary characteristics.

Trajectory-Correction Maneuvers

The next step in the navigation process is optimizing the flight path. If the residuals continue to behave satisfactorily, one acquires confidence that the model is generating a trajectory that closely coincides with the actual trajectory. One must then decide whether or not to change the course. If the spacecraft is close to the desired flight path, nothing needs to be done. If, however, there is substantial uncertainty about the actual path, additional tracking may be required before further action is warranted. Even if the spacecraft is close to its preplanned course, the mission team may decide that a modified flight path may yield a higher return of scientific information or may better satisfy other operational considerations. Such analyses continue throughout the flight, and the activity becomes intense as the spacecraft approaches its target.

Once a new trajectory is selected the correction in spacecraft velocity needed to achieve the new trajectory is calculated. For example, a few days after the launching of *Viking 1* the first trajectory-correction maneuver had as its objective the removal not only of the normal velocity errors resulting from launching but also of errors intentionally introduced in order to eliminate the risk that the Centaur stage of the launching vehicle, which trails the spacecraft, might collide with Mars and biologically contaminate the planet. If the velocity errors introduced at launching were left uncorrected, they would cause the spacecraft to miss Mars by several hundred thousand kilometers. The velocity adjustment needed to cor-

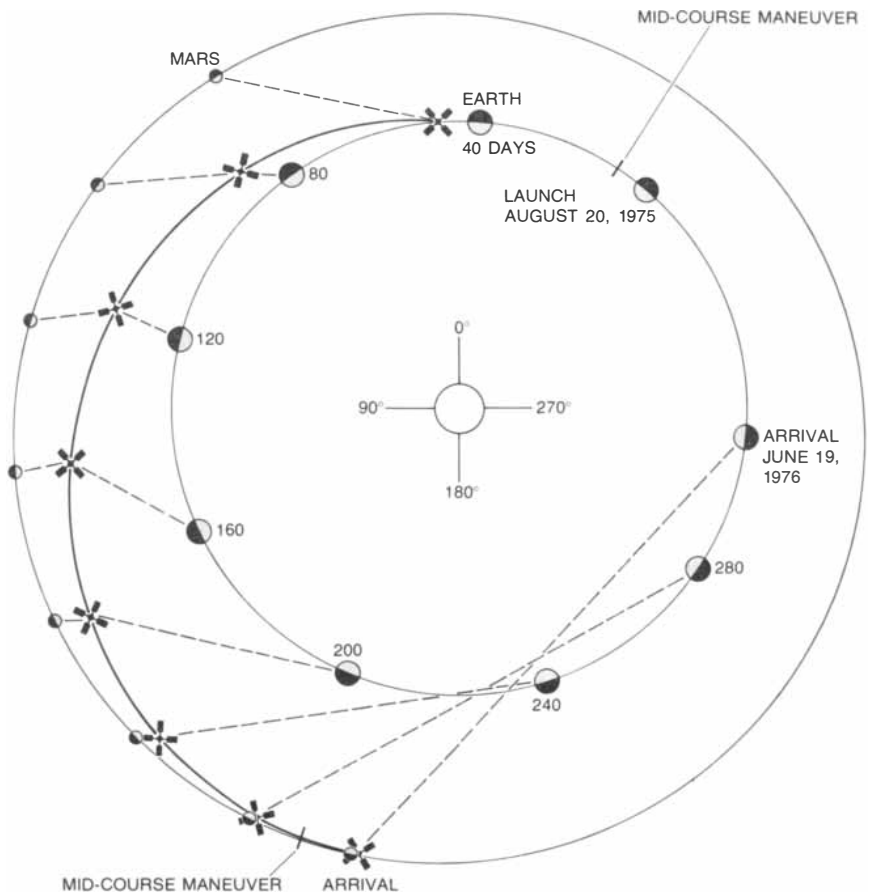
rect a miss of this magnitude is about 10 meters per second. The maneuver, which takes less than a minute, will normally achieve better than 99 percent of the required correction.

The residual velocity error, together with remaining uncertainties involved in orbit determination, is likely to cause the spacecraft to miss its target by several hundred kilometers. Although prior knowledge of the spacecraft's position is not affected at the time of the maneuver, knowledge of the spacecraft's velocity is corrupted because the maneuver is imperfectly executed, and it cannot be exactly reconstructed from engineering data sent from the spacecraft. Immediately before the first trajectory correction the spacecraft's velocity is usually known to an accuracy of about one kilometer per day. Immediately after that maneuver the velocity accuracy may be degraded by a factor of 10 or more [*see illustration on page 58*].

The results of a maneuver need to be evaluated as soon as possible both to verify that it was executed within normal limits and to obtain a measure of the resulting trajectory errors that must be corrected by maneuvers later in the flight. The new velocity components must be remeasured along three axes. It is convenient to choose mutually perpendicular axes along the line between the earth and the spacecraft (the radial direction), along the direction in which the spacecraft appears to be moving across the celestial sphere (the cross-velocity direction) and along the direction perpendicular to the other two.

The abrupt change in radial velocity during the maneuver can be directly measured with Doppler data to an accuracy of about .01 kilometer per day. For maneuvers executed early in a mission this uncertainty translates into an error at the target planet on the order of one kilometer. The component of the maneuver in the cross-velocity direction can be observed because it changes the centripetal acceleration of the spacecraft in relation to the tracking station through the change in cross velocity. The change in centripetal acceleration causes the radial velocity to drift linearly with time away from its behavior before the maneuver. After several hours of Doppler tracking, the linear drift in the residuals is usually sufficient to establish the slope of the drift, and hence the cross-velocity component of the maneuver, to an accuracy of about one kilometer per day. After several days of continued tracking the accuracy improves to about .1 kilometer per day. The change in the third component of the maneuver, which is nearly perpendicular to the plane of the ecliptic, is the most difficult to establish. Since this velocity component is perpendicular to both the radial direction and the cross-velocity direction, it does not initially change the radial velocity or the radial acceleration. Many days of tracking are needed before its effect becomes observable.

The determinations of the three components of spacecraft position vary widely in accuracy. The spacecraft's radial position, that is, its location on a direct line drawn from the earth, can be determined from one range measurement to an accuracy of about three meters, or three orders of magnitude more accurately than our knowledge of the positions of the planets. The determination of the two position components perpendicular to the radial position calls for considerably more effort because the immense distances involved make the range measurement insensitive to perpendicular displacements. For example, when a spacecraft is at a distance of 100 million kilometers, a displacement of 1,000 kilometers perpendicular to the earth-spacecraft direction results in an increase of only five meters in range. Fortunately the gravitational field of the sun and the diurnal motion of the tracking station furnish important information about these two perpendicular components.

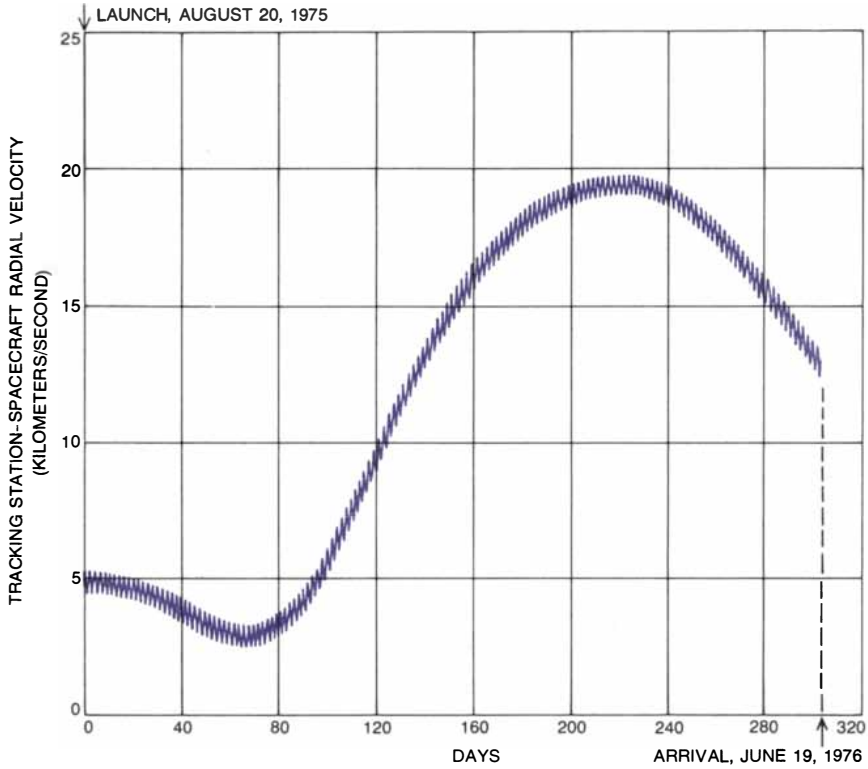


INTERPLANETARY TRAJECTORY OF VIKING 1 shows the positions of the earth, Mars and the spacecraft at 40-day intervals following the launching last August 20. *Viking 1* will reach Mars on June 19, after traveling 700 million kilometers in 305 days. The distance between the earth and Mars on the arrival date will be 313 million kilometers. One mid-course maneuver was made a few days after launching. A small velocity correction will be made June 9.

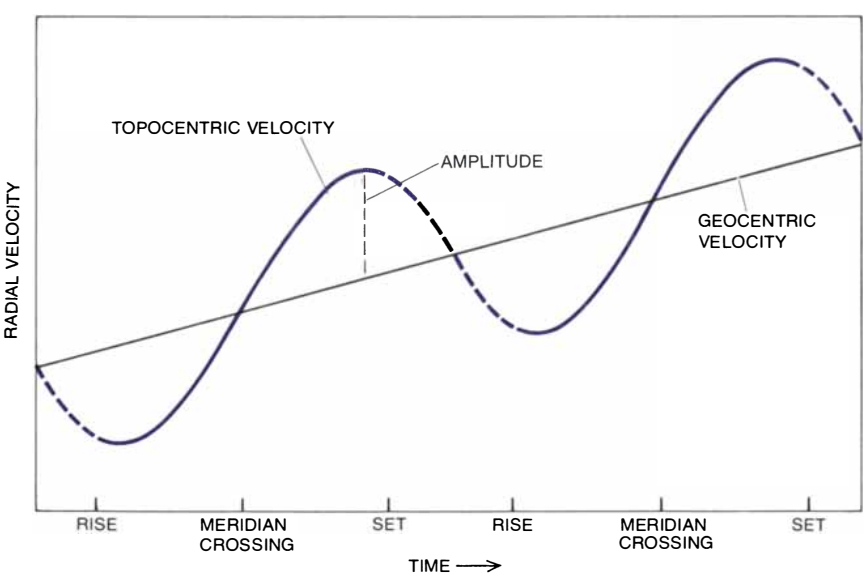
Determination of Position

The gravitational attraction of the sun is responsible for an acceleration of the spacecraft's radial position that varies slowly over the several months of the flight. This acceleration depends on the spacecraft's cross velocity and on the distance of the craft from the sun and from the earth. The three bodies lie at the vertexes of a triangle, two sides of which (the line from the earth to the sun and the line from the earth to the spacecraft) are known almost perfectly from the ephemeris of the earth and from one spacecraft range measurement. It follows that only the angle between the two known sides is required to calculate the length of the third side: the distance of the spacecraft from the sun. Thus the acceleration and its rate of change with time obtained from successive range measurements over many days furnish information about the sun-earth-spacecraft angle and about the cross velocity.

On the other hand, the behavior of the range with time is insensitive to a change of the spacecraft's position in the direction perpendicular to the plane of the triangle. That direction is usually almost perpendicular to the plane of the ecliptic because trajectories between the earth and the other planets cannot deviate very much from that plane. As a result where distant spacecraft are concerned the components of position and velocity perpendicular to the plane of the ecliptic cannot be accurately determined from solar gravitational effects. If it were not for the information provided by the rotation of the earth, as I shall now explain, these trajectory parameters could



VIKING 1'S RADIAL VELOCITY has varied over a considerable range during its 10-month voyage. The rapid increase between Day 80 and Day 200 primarily resulted because the earth in its orbit is traveling faster than *Viking 1* in its larger orbit, and in effect is leaving the spacecraft behind. Around Day 220 the radial velocity of the spacecraft began to decrease because an increasing fraction of the earth's higher velocity appeared as a component perpendicular to the earth-spacecraft direction, as is shown in the preceding illustration. A small sinusoidal variation is imposed on radial-velocity curve by the motion of tracking station around the earth's axis. On May 26 (Day 280) *Viking 1*'s heliocentric velocity was 19.7 kilometers per second.



DAILY SINUSOIDAL VARIATION in the earth-spacecraft radial-velocity curve results because the tracking station is being carried toward the spacecraft as it rises in the east and away from the spacecraft after it crosses the meridian and sets in the west. The sinusoidal curve is the topocentric radial velocity, from which the geocentric radial velocity is computed. For determining the spacecraft's orbit the observed phase and amplitude of the topocentric curve provide information about the spacecraft's right ascension and declination. The slope of the geocentric curve is related to the spacecraft's cross velocity (the component of velocity perpendicular to radial velocity) and to position of spacecraft in relation to the sun and the earth.

not be satisfactorily determined without many months of observation.

Since the tracking stations are on the earth's surface, the direct measurements they provide of the spacecraft's radial parameters (range, velocity and acceleration) are topocentric rather than geocentric, that is, the values are related to a point on the surface of the earth rather than to a point at the center of the earth. To obtain the geocentric equivalent of these values, which are normally more fundamental for trajectory calculations, one must allow for the tracking station's location on the surface of the rotating earth. The topocentric values can nonetheless serve for securing important information that would be difficult to get any other way. The topocentric radial velocity of the spacecraft includes a small daily sinusoidal variation because the tracking station is being carried toward the spacecraft by the earth's rotation when the spacecraft is ascending in the eastern sky (like a planet or a star) and away from it when it is descending in the western sky.

From Doppler-tracking measurements of the spacecraft from its rising to its setting it is possible to extract accurate estimates of the spacecraft's right ascension and declination. To do this one in effect plots against time the observed sinusoidal topocentric radial-velocity curve on top of the essentially smooth curve for the geocentric radial velocity [*see bottom illustration at left*]. The point where the two curves intersect gives the time the spacecraft crossed the tracking station's meridian. Knowing the time of crossing and the longitude of the tracking station is equivalent to knowing the right ascension of the stars also crossing the tracking station's meridian at that instant and therefore is equivalent to knowing the right ascension of the spacecraft itself. The amplitude of the topocentric curve (its maximum deviation from the geocentric curve) depends on the spacecraft's declination. Actually the amplitude is the product of three quantities: the rotational rate of the earth, the distance of the tracking station from the axis of rotation of the earth and the cosine of the spacecraft's declination.

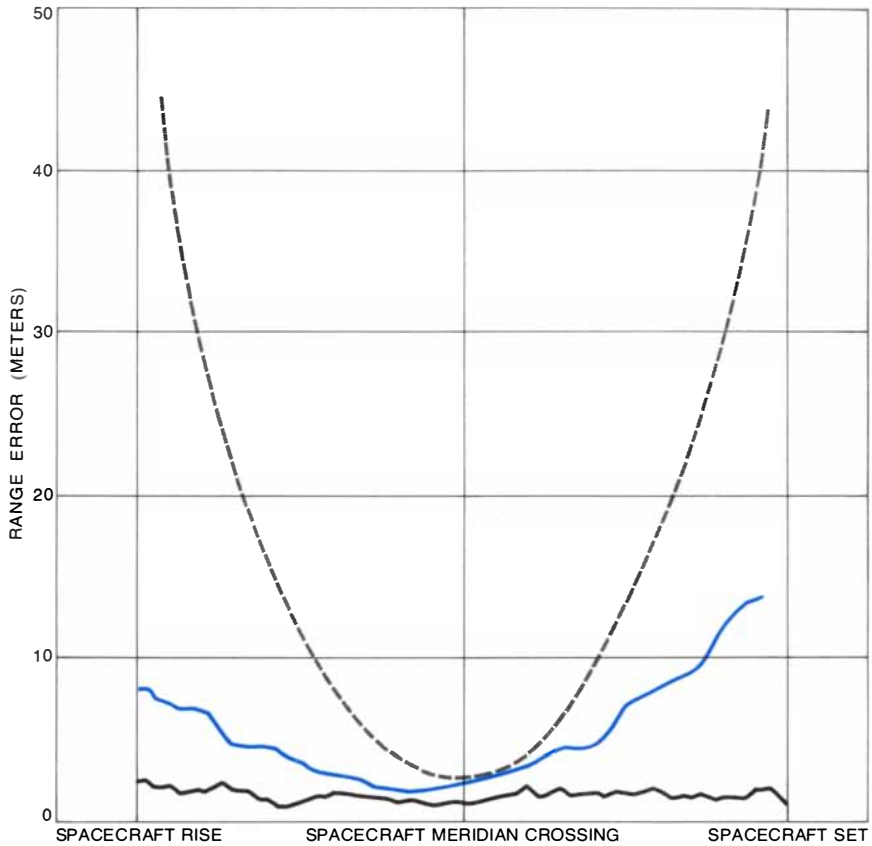
Doppler tracking therefore complements range tracking by providing information about the spacecraft's position in the direction perpendicular to the plane of the ecliptic. It also provides information about position in the plane of the ecliptic, and in particular about the sun-earth-spacecraft angle, much earlier than such information could be obtained from range tracking alone. Just two days of radio tracking usually yields estimates of a spacecraft's angular position to an accuracy of a few tenths of a microradian, or a positional accuracy of better than 100 kilometers for flights to the terrestrial planets.

This positional information also makes it possible to derive the cross velocity from the observed acceleration of the spacecraft. Of the six trajectory parameters (three of position and three of velocity) usually only

the component of the spacecraft's velocity perpendicular to the plane of the ecliptic remains insufficiently determined after two days. It is derived primarily from successive determinations of declination and usually requires a few weeks of tracking. It is ironic that the principal function of Doppler tracking is to furnish angular information and that range measurements over successive days provide the best information about radial velocity.

Tracking-Station Locations

The accuracy of values for a spacecraft's right ascension and declination obtained by Doppler tracking is highly dependent on accurate knowledge of the tracking station's longitude and its distance from the axis of rotation of the earth. Any error in the assumed value for this distance introduces a corresponding error in the declination value. The locations of the tracking stations have been determined by tracking spacecraft that are "anchored" to a planet, either by passing close to it or by being permanently in orbit around it. Under such conditions one can make a good determination of the spacecraft's position in relation to the planet. Since the ephemerides already provide an accurate planetary direction, the spacecraft's right ascension and declination can usually be computed to a few tenths of a microradian. The amplitude of the topocentric Doppler curve and the intersection time are then implicitly utilized in reverse to obtain the locations of the tracking stations with an accuracy that is typically of one meter.



MEASUREMENTS OF SPACECRAFT'S RANGE must be corrected for variations in the speed of radio signals as they pass through the earth's troposphere (*broken curve*), the earth's ionosphere (*color*) and the plasma of interplanetary space (*black*). The variation in the resulting range errors with time degrades the Doppler measurements from which the spacecraft's right ascension and declination are computed. The curves show typical effects on *S*-band signals.

At that level of accuracy it is necessary to take into account variations in the rate of rotation of the earth and in the geographic location of the poles that result from changes in the earth's moments of inertia. For example, if variations in the location of the poles are neglected, the error in the assumed location of a tracking station can be as much as 10 meters. Classical astrometric techniques make it possible to determine the orientation of the earth to within a meter. New techniques now being introduced should improve on this level of accuracy by an order of magnitude over the next few years. For several years large radio antennas throughout the world, including those of the tracking network, have been used in pairs as very-long-baseline interferometers to study the structure of extragalactic radio sources. A by-product of this work is the determination of the baseline vectors between pairs of antennas, which is equivalent to a determination of the orientation of the earth. By the same token the relative locations of the antennas have also been accurately determined. Improvements in key elements of the interferometer system are now being made and should yield new determinations with an accuracy measured in centimeters.

Another promising technique is measuring the round-trip travel time of ultrashort laser pulses beamed to earth satellites and particularly to the special mirrors, known as corner reflectors, placed on the moon by American astronauts and by unmanned Russian spacecraft. Laser ranging to these reflectors from the McDonald Observatory of the University of Texas has been done for the past six years and appears now to have an accuracy of a few centimeters. A corresponding accuracy for the determination of the earth's orientation should be realized as soon as one or two additional widely separated laser sites become operational.

When the spacecraft is very close to the celestial equator, the declination is difficult to determine. Because the amplitude of the topocentric Doppler curve depends on the cosine of the declination it is insensitive to variations of the spacecraft's declination when the declination has a value near zero. Moreover, in such cases small errors in tracking-station locations that would change the amplitude of the topocentric radial velocity require a large change in the estimated declination to compensate for them. In other words, when the spacecraft is near the celestial equator, the declination determination is extremely sensitive to errors that have not been taken into account. Fortunately the direction of the spacecraft usually changes rapidly on flights to the terrestrial planets, so that it rarely remains near zero declination for more than a few days. For missions to the outer planets, however, this sensitivity is a more serious problem because the spacecraft slows down considerably as it recedes from the sun and can stay near zero declination for months. Indeed, that is what will happen in 1980, when the Mariner flight to Saturn will be in the critical two months just before the spacecraft encounters the planet.

For the zero-declination case a promising approach is simultaneous or nearly simultaneous ranging from the Northern and Southern hemispheres, making use of the Goldstone and Canberra tracking stations. The difference between the round-trip times of radio signals received at these two stations gives a direct measurement of the angular position perpendicular to the line of sight in the plane containing the spacecraft and the two tracking stations. The large north-south separation of Goldstone and Canberra makes the differential time-delay measurement particularly strong in the declination direction. If the difference in range errors at the two stations can be kept to within three meters, it should be possible to measure the declination of a spacecraft in the vicinity of Saturn with an accuracy of about 400 kilometers, or about .3 microra-

dian in angular accuracy. Angle measurements from a single antenna are limited by diffraction effects to an accuracy of about one milliradian.

Propagation Effects

Another source of error in radio tracking is the variation in the speed of the radio signal as it propagates through space and through the earth's ionosphere and troposphere. The variations in speed cause small distortions in the Doppler curve with corresponding errors of up to a few parts per million in the determination of the right ascension and the cosine of the declination. The tropospheric effect is the largest, but it can usually be corrected with considerable accuracy by taking temperature, pressure and humidity measurements at the tracking station. The variations in the propagation velocity caused by free electrons in the ionosphere and in outer space are more difficult to correct for because of the partly random variation in the number of such electrons along the path of the radio signal.

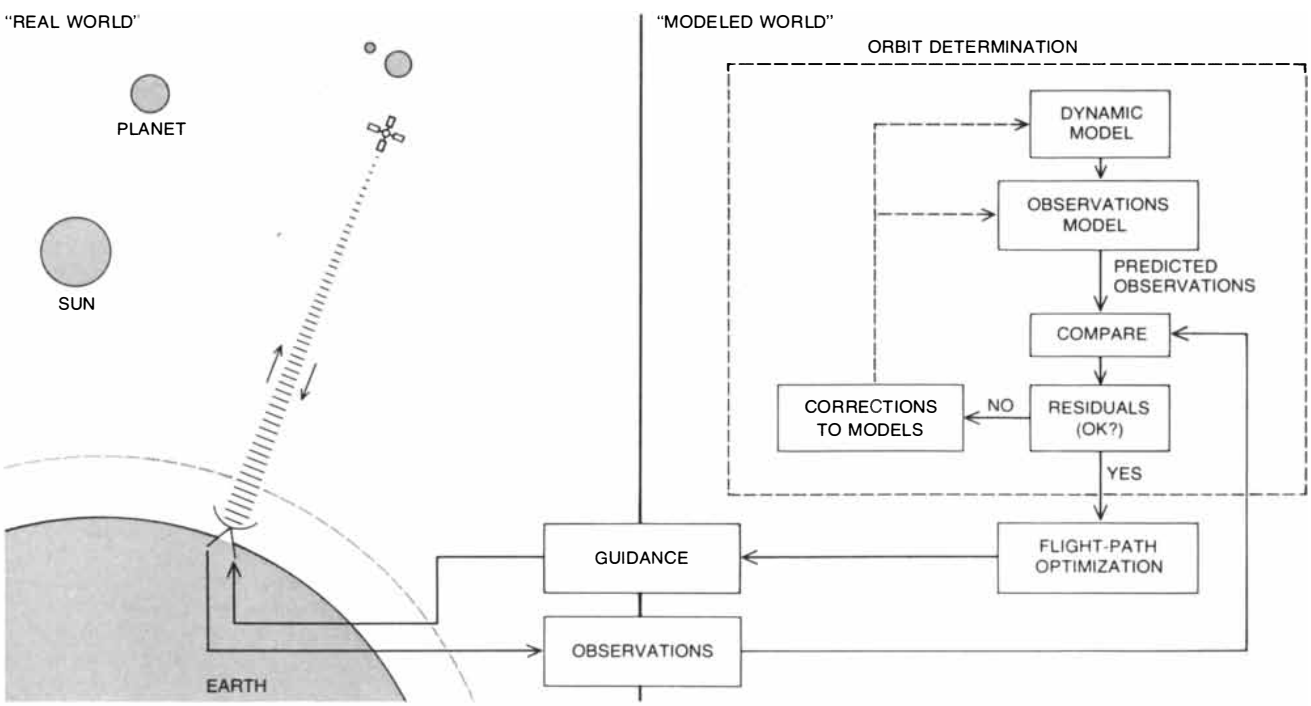
The error in range caused by the electrons is inversely proportional to the square of the carrier frequency of the radio signal. Accordingly most spacecraft now transmit to the earth a second radio signal at an *X*-band frequency (8.5 gigahertz) that is coherent in phase with the primary *S*-band signal of 2.2 gigahertz. Phase coherence means that the phase of the transmitted *X*-band wave is at all times a precise and constant multiple of the phase of the transmitted *S*-band wave. Any departure from phase coherence shows up as a difference in the time delay and the Doppler shift between the two signals received at the tracking station. Such discrepancies provide a measure of the number of electrons and its rate of change along the downlink path from the spacecraft to the earth. With this information one can make suitable corrections in the round-trip time and in the Doppler measurements.

Nongravitational Forces

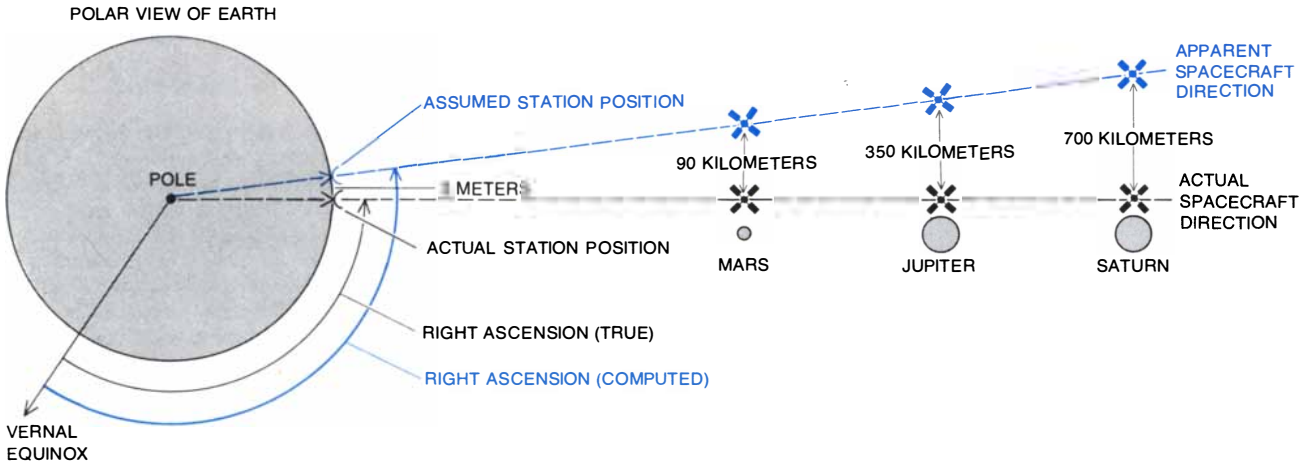
In general, knowledge of a spacecraft's position and velocity slowly improves over many weeks of continuous tracking. This improvement tends to be offset by a variety of small nongravitational forces acting on the spacecraft. Such forces are difficult to include in the orbit-determination models. It is not difficult, however, to include the pressure of solar radiation on the spacecraft; indeed, the effect of radiation pressure was allowed for in the earliest planetary missions. On a flight to Mars, for example, the pressure of solar radiation will have added about 10,000 kilometers to the spacecraft's heliocentric distance by the time the target is reached. What causes trouble is that the precise value of the solar pressure is altered by the steady bombardment of protons and other particles in the solar "wind." During the flight these particles change the reflectivity of the various materials covering the spacecraft by as much as 1 percent, and the changes are difficult to predict.

In order to maintain a desired orientation in space the spacecraft itself expends gases that give rise to small accelerations. Such force fluctuations and variations in solar-pressure acceleration are partly random and cannot be completely modeled in a deterministic way. The fluctuations are usually small, causing a displacement from the gravitational path of only a few meters per day, or only a few kilometers for the entire flight. These random forces effectively lead to a loss of information by breaking down the deterministic link between the old and the current observations that is provided by the gravitational laws of motion. The concern here is not how much these effects drive the spacecraft off course but how they affect the determination of the spacecraft's orbit.

The estimate of a poorly or indirectly observed component of position or velocity is particularly vulnerable to unmodeled effects. Typical random accelerations can cause trajectory errors at the target planet ranging from 100 to 1,000 kilometers if they are ignored. Serious errors in the estimation of trajectory parameters can arise in a direction that is not easily accessible to observation because the dynamical model "thinks" it knows the path of the spacecraft in the well-observed direction on the basis of past observations and the deterministic properties of the gravitational laws. As a result the model strongly resists being tinkered with in the well-observed direction. In compensating for the small changes in the observed range and the Doppler shift induced by the unmodeled accelerations the orbit-determination process is likely to greatly magnify the error in the estimate of the spacecraft's position or velocity in a poorly observed direction, because a large adjustment in that direction is needed to significantly reduce



NAVIGATION PROCESS includes the flow of observations into computer data files, orbit determination, flight-path optimization and trajectory correction. Orbit determination is an iterative procedure for testing hypotheses aimed at obtaining the most likely values of all physical parameters affecting the observations. The procedure includes models of the physical processes that affect the motion of the spacecraft, the motion of the observer and the path of the radio signal. Flight-path optimization is a subset of the mission optimization effort, a large and complex iterative activity that is designed to maximize scientific information that can be obtained by spacecraft.



POLAR VIEW OF EARTH shows how a small error in the longitude assigned to the tracking station leads to a much magnified error in computing the spacecraft's right ascension. If the error in the location of the tracking station is only three meters, when the spacecraft is at the distance of Mars, its actual position will be 90 kilometers away from its inferred position, based on two days of tracking. Similarly, the error would be 350 kilometers at Jupiter and 700 kilometers at Saturn. Tracking over many weeks reduces such errors by means of the information furnished by the solar gravitational effects. Tracking-station locations are known to an accuracy of one meter.

the error in the residuals and because the model does not resist change in that direction.

Some relief is obtained by relaxing the model so that it is not as rigid about predicting the future on the basis of the past. The relaxation is accomplished by restructuring the model so that the possibility that the spacecraft is being subjected to stochastic, or random, forces is taken into account. One tries to include a priori information on the statistical characteristics of these stochastic forces, such as their expected magnitude and their correlation with time. The result is that instead of generating only one estimated trajectory the model develops a probability corridor in position-and-velocity space. The model assigns a probability measure that the spacecraft is at any one place in the corridor at any one time. On the basis of the observations it also provides estimates of the statistical parameters of the actual stochastic forces acting during the observational period.

This approach, known as sequential estimation, is not a cure for the vulnerability of poorly observed parameters to unmodeled effects, but it helps greatly. The technique demonstrated its value in the flight of *Mariner 10* past Venus and then on to Mercury. That flight was the first where the gravitational field of one planet was harnessed to deflect a spacecraft toward another. The angle of deflection depended critically on the spacecraft's position as it reached Venus; any error would be magnified several thousandfold by the time the craft arrived at Mercury. In order to conserve trajectory-correction propellant so that there would be enough left to accomplish the three successive encounters with Mercury, the potentially very large trajectory correction just after the Venus encounter had to be minimized by navigating the spacecraft as accurately as possible before it arrived at Venus. Postflight studies have shown that *Mariner 10* was navigated by Venus within 17 kilometers of the preplanned path. Without sequential estimation the error could have been between 100 and 200 kilometers.

Planetary Encounter

So far I have been discussing the strictly interplanetary phase of the flight, where the gravitational effect of the target planet is small. As the spacecraft approaches the target, however, one can detect in the tracking data that the planet's gravitational field is beginning to accelerate the spacecraft along its trajectory. For the terrestrial planets the acceleration begins between one day and three days before encounter; for Jupiter it begins at about a month. Because at this point the spacecraft is traveling in a nearly straight line toward the planet, the change in acceleration due to the planet's gravity provides very accurate and early information on the distance remaining between the spacecraft and its target. As the spacecraft gets closer the planetary field begins to bend the trajectory, and the bending rate can be well determined.

With this information and knowledge of the spacecraft's velocity with respect to the planet (based on the planetary ephemerides and on orbit determination during the interplanetary phase) one can predict the distance of the spacecraft's closest approach. The third position coordinate, perpendicular to the plane of motion of the spacecraft in relation to the planet, is not well determined by the planet's gravity until the spacecraft is in the immediate vicinity of the planet. Information acquired during the approach phase is valuable in making final plans for operating the spacecraft's scientific instruments. The refined position information acquired by radio tracking during the planetary-approach phase usually comes too late to be useful in calculating the final velocity maneuver needed to place the spacecraft in the desired position at encounter with the planet. In the Viking missions, for reliability reasons, the retromaneuver at encounter that will put the spacecraft in orbit around Mars will be based on commands sent from the earth at least a day earlier; the information obtained during the last day before encounter will not be used for this critical maneuver. These are the reasons why accurate planetary ephemerides and the precise determination of a spacecraft's interplanetary orbit are important.

The orbit of the Viking spacecraft around Mars is determined on the basis of information in the tracking data that is different from the information used to determine the spacecraft's interplanetary orbit. The dominant factors are the known gravitational attraction of Mars and the relative motion of the earth and Mars along their respective orbits. The gravitational field of Mars is used to determine the size and shape of the spacecraft's orbit (to an accuracy of about one part per million) and the position of the spacecraft along the orbit. It also makes it possible to determine the orientation angle of the plane of the orbit around any axis that is perpendicular to the line of sight between the earth and Mars to within one microradian, or 10 meters at periapsis. The relative motions of the two planets cause the aspect of the spacecraft orbit as it is viewed from the earth to change slowly with time, making it possible to determine the orientation angle of the plane of the spacecraft's orbit with respect to the line of sight between the earth and Mars to within a milliradian, which corresponds to an accuracy of 10 kilometers in the position of the spacecraft at periapsis.

The irregular distribution of a planet's mass gives rise to perturbations in the spacecraft's orbit, particularly near periapsis, that degrade our ability to predict the position of the spacecraft along the orbit. The accurate prediction of position several orbits ahead is obviously important for

landing the spacecraft at a specified location on the surface of the planet. One of the navigation activities during the two weeks when the Viking spacecraft will be in orbit around Mars before landing will be to determine the effects of mass perturbations in the regions over which the spacecraft will fly and thereby to improve our ability to make orbital predictions. Judging by our experience with the Mars orbiter *Mariner 9* in 1971, we expect to predict the Viking orbital period several orbits in advance to within one second per revolution. Insofar as errors in the determination of orbit are concerned that should introduce an error of no more than 10 kilometers in the landing point of the spacecraft.

Flights to the Outer Planets

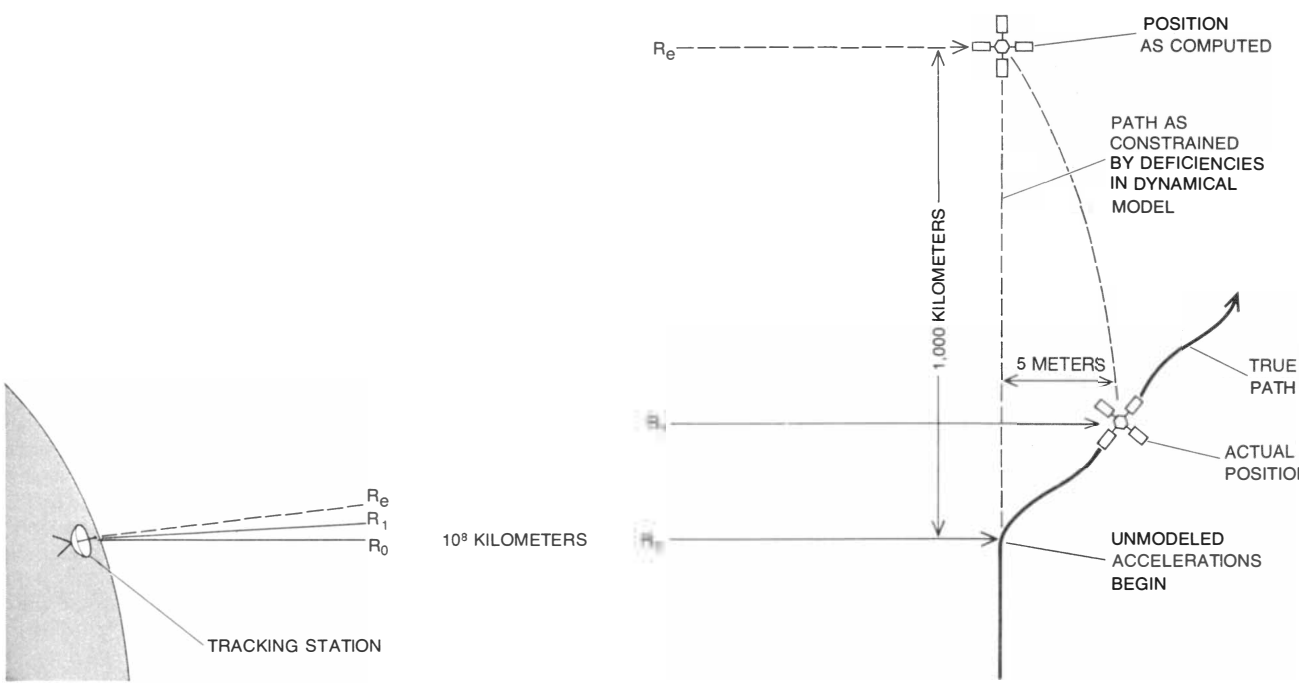
The first flights to the outer planets began with the spacecraft *Pioneer 10* and *Pioneer 11*, which flew by Jupiter in 1973 and 1974 respectively. Next year two Mariner spacecraft will be launched toward Jupiter, and with gravitational assistance from the planet at encounter the spacecraft will be deflected on toward Saturn. The second spacecraft will be targeted to fly past Saturn and on toward Uranus. Some of the larger satellites of Jupiter and Saturn will be investigated on these missions. At present the ephemerides of the two satellite systems are not good enough for accurate targeting. The satellites interact with one another gravitationally in such a way that they give rise to resonance phenomena that make it hard to calculate their motions years in advance. Because the satellites are so far away from the earth (between five and 10 astronomical units) earth-based optical telescopes are unable to supply direction measurements accurate enough for the navigation of spacecraft. Although radar range measurements of some of the largest satellites of Jupiter and Saturn can now be made with the recently upgraded 305-meter radio telescope at Arecibo in Puerto Rico, the ephemerides of the satellites are likely to remain a problem for several years to come.

To compensate for these difficulties it is now planned to employ the television system on the Mariner spacecraft to augment earth-based radio tracking. The television system, whose primary function is to return detailed pictures of the planets and the satellites, will record the images of the satellites and the background stars when the spacecraft is still a long way from the target planet. Since the stellar positions are known, the observations will provide the right ascension and declination of the center of the recorded satellite image. By combining a series of such observations with the orbit-determination information obtained during the interplanetary phase of the spacecraft's flight, it should be possible to make corrective maneuvers early enough to navigate the spacecraft to within a few hundred kilometers of the desired position in relation to the satellites of the two giant planets. To lay the groundwork for making use of this navigational technique it was tested with *Mariner 9* and the two small Martian satellites, Deimos and Phobos. It will be used with the Viking spacecraft.

Prospects

Interplanetary navigation presents something of a paradox. On the one hand nature can be regarded as being somewhat perverse in requiring that many dynamical and observational processes be modeled with great precision if high navigational accuracy is to be attained. On the other hand a windfall of scientific information about the solar system has been extracted from the small effects that many of the same processes have on the spacecraft radio-tracking data. Knowledge of unprecedented accuracy has been obtained about the mass of the planets, the distribution of mass within the planets and the overall configuration of the planets. Valuable information has also been gained on the physical properties of the planets' atmospheres and ionospheres from their effect on radio signals.

Refinements in interplanetary navigation are planned over the coming five years that will improve its accuracy by an order of magnitude. Thus it should be possible to extend to Jupiter and Saturn the navigational capability already demonstrated with the terrestrial planets. Improved navigational accuracy should also make it possible to conduct definitive tests of the validity of some of the alternative theories of gravitation and in particular to refine by an order of magnitude the determination of effects predicted by the general theory of relativity.



MAJOR ERROR IN COMPUTED POSITION can result if a spacecraft is unexpectedly subjected to small random forces that are not modeled in the orbit-determination program. The program is particularly sensitive to error in a poorly observed direction if the unmodeled forces shift the spacecraft along a well-observed direction. The hypothetical spacecraft depicted here has been traveling along a straight path whose perpendicular distance from the earth is 10^8 kilometers (R_0). Before the disturbing forces begin, the orbit-determination model from preceding observations correctly computes R_0 as 10^8 kilometers. After a while the small unmodeled forces cause the spacecraft to deviate from the previously straight path by five meters, and a new range observation is taken (R_1). The only way the deficient model can simultaneously satisfy the observation and the constraint of straight-line motion resulting from the preceding observations is to greatly magnify the error in new estimate (R_e) of spacecraft's position, misplacing it by 1,000 kilometers in perpendicular component.

Vivitar Series 1 35-85mm f2.8 variable focusing lens.

For many photographers a lens of this focal length range is an "ideal" lens covering wide angle, through normal, to medium telephoto. But for professional and scientific photographers the requirements of an "ideal" lens include more than a versatile range of focal lengths.

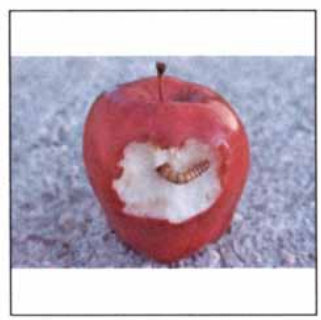
The assignment given by Vivitar to the Series 1 Research and Development Team included many complex design specifications, some conflicting, such as the need for a relatively high-speed lens and the insistence on compactness.

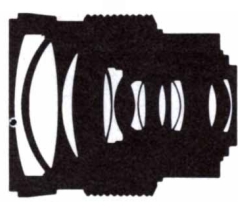
The specifications included the following: (1) A focal length range from 35 to 85mm. (2) Close focusing capability. (3) An f2.8 maximum aperture, so the lens could be used in place of a normal lens in low-light areas. (4) Length under 4 inches. (5) Weight under 30 ounces. (6) Contrast and resolution equal to or superior to comparable lenses. (7) Smooth but rugged mechanical operation, (8) Multicoating

To accomplish all these requirements in one lens the team spent two years working with some of the world's largest computer banks. And instead of using the customary zoom lens configuration, which entailed some sacrifices in resolution at wide angle close focusing, the designers chose a variable focus solution. Changing the focal length is accomplished by movement of three independent floating groups in the lens, thus allowing close focusing to 4.3" from the front element (in the 35mm position), without sacrificing resolution. All done with a single-touch control.

The mechanical motion of these three groups is controlled by cams milled into sleeves. In order to achieve the extremely close tolerances specified by the Japanese mechanical designers, the cams had to be machined on specially-made, numerically-controlled lathes designed and built in Germany and Switzerland.

The extreme compactness of the lens itself necessitated the use of a nested series of cams and operating sleeves. Again, extremely close tolerances had to be maintained on the five concentric sleeves to eliminate centration errors. Even with the use of the most advanced optical equipment available today, to comply with the specifications each lens is individually adjusted by an engineer to ensure optimum optical alignment. The result is a valuable example of international cooperation, representing the combined efforts of Japanese, German and American specialists.





Optical Specifications

Construction:	12 elements 9 groups
Angle of view:	28° to 63°
Minimum focus distance	
From film plane:	10.2 in. (25.9cm)
From front element:	4.3 in. (10.9 cm)
Maximum reproduction ratio:	1:3.5
Focal length ratio:	2.4:1
Mechanical Specifications	
Filter size:	72mm
Weight:	27 oz. (770 gr.)
Length:	3.6 in. (9.1cm)
Max. diameter	3.19 in. (81mm)
F/number range:	f2.8 to 16

Slip-on lens hood included. Available in mounts to fit Nikon, Canon, Minolta, Konica, Olympus OM, and Universal Thread Mount cameras.

Vivitar Series 1 Program

In recent years, remarkable progress has been made in solving some of the classic problems of optical design. Intensive work in the field was spurred on by the demands of space exploration and military applications and vastly aided by the growing sophistication of computer technology.

Vivitar optical designers, working with programs devised for highly specialized optical tasks, have used computer-generated designs to develop for Vivitar a new series of lenses capable of performance unreachable until now.

Each Vivitar Series 1 lens so far introduced represents a breakthrough in optical design. The 200mm f3 and the 135mm f2.3 are among the fastest automatic telephoto lenses in their respective focal lengths. They are unusually light and compact. Each has a uniquely positioned rear compensating element that automatically corrects aberrations at all points from the closest focusing point to infinity.

The 70-210 f3.5 automatic zoom lens and the 35-85mm f2.8 auto variable focusing lens are likewise definite advances in lens design, both offering extremely close focusing capability and remarkable compact configuration.

Most recently introduced are a 90mm f2.5 macro, a 28mm f1.9, and a 600mm f8 solid catadioptric lens. To come, 800mm f11 solid catadioptric and 1200mm f11 solid catadioptric telephoto lenses. There eventually will be a complete optical system of more than twenty Vivitar Series 1 lenses, each representative of advanced technology and demonstrably superior to other lenses currently available.

Vivitar®

Marketed in the U.S.A. by
Ponder&Best, Inc.
Corporate Offices: 1630 Stewart Street, Santa Monica, CA 90406.
In Canada: Precision Cameras of Canada, Ltd., Quebec.

17,000 Years of Greek Prehistory

Excavations at a site in the Peloponnesus show evidence of human habitation from the Ice Age through Neolithic times. They reveal the basic economic foundation of the Classical Greek civilization

by Thomas W. Jacobsen

The contribution of Classical Greece to modern civilization is so well known that Athens is virtually a synonym for high culture. The earlier Bronze Age cultures of the Mycenaean Greeks and the Minoans of Crete are also familiar to many. Until recently, however, little was known about still earlier stages in the human history of Greece: the Old Stone Age and the New. Now archaeological fieldwork at a site in southern Greece has uncovered a remarkable record of those earlier stages. Stratified deposits formed in late Paleolithic times, more than 20,000 years ago, are overlain by a virtually uninterrupted sequence of remains that extends to the end of Neolithic times, only 5,000 years ago. The site, a cave and its immediate surroundings on a rocky headland named Franchthi on the coast of the Argive Peninsula, is unique in Greece for the span of cultural development it preserves. Its archaeological record documents some of man's earliest efforts to come to terms with an environment that was often hostile and testifies to his increasingly complex interactions with the plant and animal life and the land and sea of Greece. It was these interactions that provided the foundations of Greek civilization.

The rich stratigraphic record of Franchthi Cave enables us to ask seminal questions about conditions in Greece during and immediately after the last major glaciation in Europe. For example, how was the area affected by the retreat of the European ice sheet some 10,000 years ago? What can we learn, beyond a mere descriptive inventory of material remains, about the daily life of the people who lived in the area? How did the environment that confronted a Neolithic Greek farmer differ from that encountered by a Paleolithic hunter-gatherer in the same area thousands of years earlier? For that matter, how did man's activities over the millenniums affect the environment?

Environmental studies suggest that the earliest human activity known at Franchthi approximately coincided with the coldest phase of the Würm glaciation, the last of the European ice ages. In other parts of southeastern Europe, in certain areas of the Middle East and apparently in much of Greece the climate was cold and dry. Vegetation was sparse; the landscape was open and resembled a steppe. Shorelines were some kilometers seaward of today's coast because much of the earth's water was stored in the great ice sheets.

This cold and dry period, when the annual range of temperatures in southern Greece may have been much like that in mountainous areas of northern Greece today, was followed by a time of gradually increasing warmth and moisture. As the Würm ice sheet shrank, the sea level slowly rose and the coastline moved inland. Trees invaded the sparsely vegetated landscape until the steppe became a region of woodland and open glades. As we shall see, this climatic change seems to have had a marked effect on the lives of the hunter-gatherers in the Argive.

In the Franchthi area the surface rocks are heavily eroded limestones that rest on a substrate of igneous rocks. Here and there, where the erosion has removed the limestone cover and exposed the harder substrate, the soils that have developed are particularly fertile. On the fringes of these exposures, where the limestones and the impermeable igneous rock meet, springs are fairly common. Elsewhere the terrain is marked by features such as caves and sinkholes that are characteristic of what geologists call a karst landscape. Franchthi Cave, which shelters much of the archaeological deposit at the site, is one such feature. In this semiarid environment, notable for its long, dry summers, the perennial springs in the vicinity of the cave must have been a major attraction to man.

The cave lies at the western tip of the Franchthi headland. Today its mouth is no more than 15 meters above sea level and only about 75 meters from the shore. The cave is now about 150 meters deep, but originally it was much larger. Rockfalls in the interior have blocked off most of the sheltered area, and one of them left a window in the roof of the cave. This catastrophic event, possibly caused by an earthquake, seems to have taken place about 3000 B.C. It may have been the main reason the cave was abandoned as a living site late in the Neolithic period.

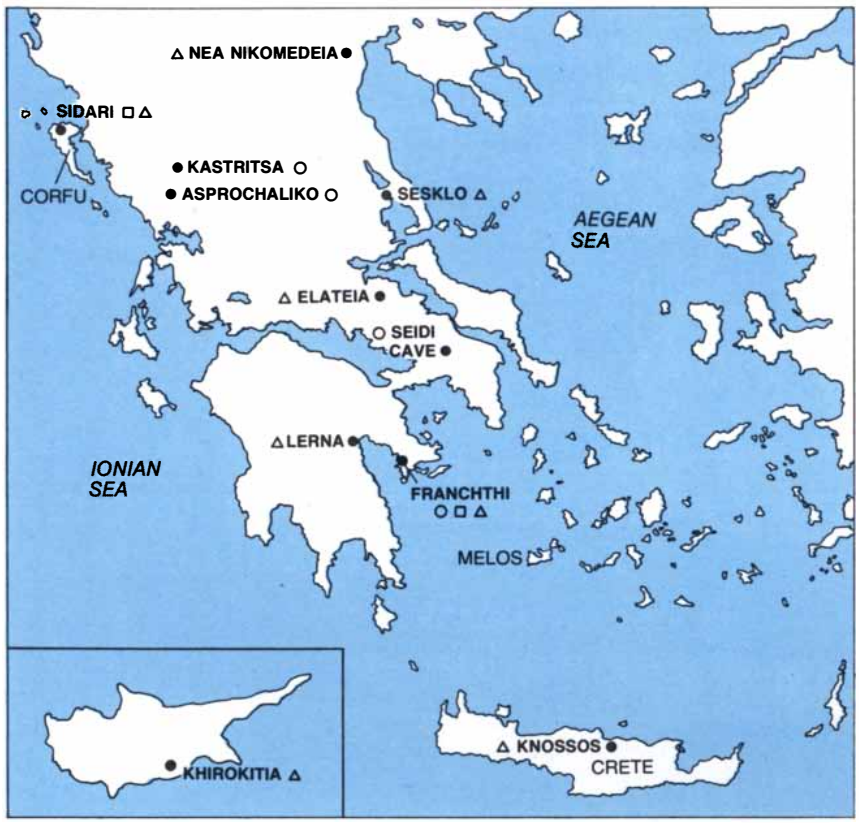
The archaeological excavations at Franchthi were begun in 1967. Since then there have been five additional seasons of fieldwork, the most recent in 1974. Our group from Indiana University has conducted the investigation in collaboration with workers from the University of Pennsylvania and other American and European institutions; the work has been done under the auspices of the Greek Archaeological Service and the American School of Classical Studies at Athens. In the opening seasons our interests were focused on the cave, where we made several trial soundings. Two of them, designated *H/H*-1 and *F/A* on our site plan, were ultimately excavated to a depth of nine and 11 meters respectively. They provide the stratigraphic framework for the site as a whole.

The two soundings revealed well-stratified and mutually complementary sequences of archaeological remains that document successive human occupations of the cave. The earlier of the cultures represented are reminiscent of the Upper Paleolithic and the Mesolithic elsewhere in Europe. (The Mesolithic is the transitional period between the Paleolithic and the Neolithic in the Old World.) Thanks largely to the efforts of the radiocarbon laboratory at the University of Pennsylvania we now have more than 50 carbon-14 age determinations from different human occupation levels in Franchthi Cave. The two earliest determinations, based on samples found near the bottom of the Paleolithic deposit, yield dates around 20,000 B.C.

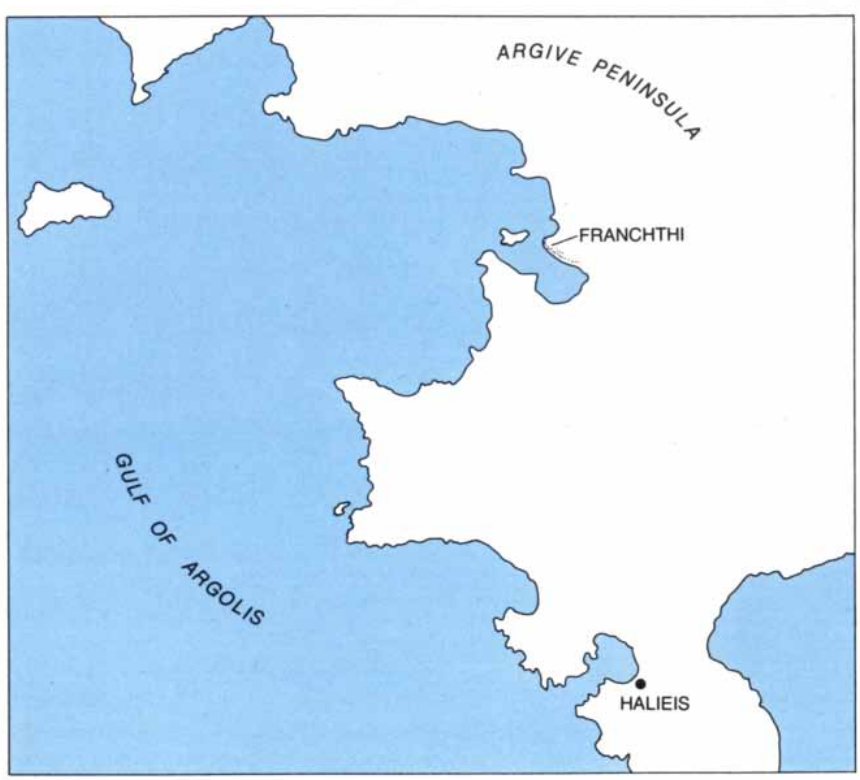
It is these age determinations that place the earliest indications of human activity at Franchthi in the time of the final Würm glaciation. The first occupants of the cave, probably a small band of seasonal visitors, seem to have been engaged principally in hunting. The animal remains at this level are dominated by the bones of a species of horse, probably a wild ass. The seasonal occupants of the site seem also to have gathered wild plants, but studies of plant remains from these levels are incomplete. The occupants' tools, made from flint or chert,

FRANCHTHI CAVE appears as a low, shadowed arch halfway up the steep hillside that rises from the water's edge. When the first bands of Paleolithic hunters used the cave as a campsite, the sea level was lower, fresh water was nearby and the coast was kilometers away.

MOUTH OF CAVE (*right*) and one of two deep excavations, *H*/*H*-1, are visible in this photograph looking across the interior of the cave. Most of the cave floor is now buried under substantial falls of ceiling rock (*left*); the falls may have led to the abandonment of the cave.



PRE-BRONZE-AGE SITES in and near European Greece include six where only Neolithic or later remains were found: four on the mainland and one each on Crete and Cyprus (*triangles*). Three mainland sites contain Paleolithic remains (*circles*). A site on Corfu contains late Mesolithic (*rectangle*) and early Neolithic levels. Franchthi contains remains of all three periods, including Mesolithic and Neolithic tools made of obsidian from Melos, an Aegean island.



FRANCHTHI, a rocky headland on the eastern shore of the Gulf of Argolis, is in a region of limestone where erosion has produced many caves and sinkholes. The archaeological remains of the Paleolithic, Mesolithic and Neolithic periods accumulated in one such cave at Franchthi.

include "backed" (that is, blunted on one edge) bladelets of a type well known from Upper Paleolithic contexts elsewhere in the Old World.

As the Würm ice sheet retreated and warm and moist conditions came to prevail, the successive inhabitants of the site appear to have adapted successfully to the changing environment. Among the remains of the larger animals the bones of the red deer and the bison begin to outnumber those of the horse. The bones of a fourth animal, probably a wild goat, also appear. Plants too were being collected in a wider variety; we find the remains of wild pulses such as vetch and lentil. Other food resources were exploited, as is indicated by the shells of land snails and marine mollusks. The first fishbones appear, revealing that small-scale fishing has begun.

Tools of flint and chert have become abundant. Their shape and their small size are typical of many stone-tool industries in the eastern Mediterranean during the Final Paleolithic, some 12,000 to 10,000 years ago. Backed bladelets are still present; we also find small disk-shaped scrapers and microliths, some with "geometric" shapes such as triangles and trapezoids. Denticulated (notched) pieces, whose function is uncertain, are particularly common in the latest Paleolithic strata.

Our analysis of the material from the lowermost strata at the cave is not yet sufficiently advanced to enable us to reconstruct how these seasonal hunter-gatherers exploited the Franchthi area in terms either of how much ground they covered or of when during the year they were present. It may be significant, however, that we have found two open-air Paleolithic sites within six kilometers of the cave. One is near the mouth of a magnificent gorge to the east of Franchthi where water is available the year round. This would have been an ideal place for the hunters to ambush game. Their prey may have been animals that funneled into the narrow passage during annual migrations or resident animals that came to drink.

As for how long the visitors stayed at Franchthi, preliminary paleotemperature analyses of marine-mollusk shells suggest that during the Final Paleolithic the site was occupied at least throughout the summer. There is no clear-cut indication of winter activity at the cave. This negative finding may be due at least in part to the inhospitable dampness of the cave during the winter rainy season.

The Paleolithic strata in Franchthi Cave are overlain by a long record of the activities of two successive Mesolithic occupations. The earlier Mesolithic strata in the cave, deposited during the late ninth and early eighth millenniums B.C., yield some evidence of continuity with the Final Paleolithic strata just below them. The Mesolithic occupation can nonetheless be distinguished from the preceding occupation in several ways. Perhaps the most striking is that the inventory of animal bones no long-

SCIENCE/SCOPE

Pictures from NASA's Landsat-1 satellite are helping Alaska's natives -- some 100,000 Indians, Eskimos, and Aleuts -- select the best lands from the 99 million acres set aside for them from the federal public domain. Doyon, Ltd., one of 12 regional native corporations, asked the University of Alaska to recommend the best land in a vast, roadless tract in central Alaska. Using the Landsat pictures and the limited ground and aerial data available, university scientists mapped seven million acres. The maps show potential farm land, marketable timber, and hard-rock mineral deposits. The Landsat pictures were taken by the Hughes-built multispectral scanner.

More than 20 countries have selected the U.S. Army's TOW (tube-launched, optically-tracked, wire-guided) missile as their infantry heavy assault anti-tank weapon. Hughes/Tucson has produced more than 120,000 TOWs to date. Test firings from 150 lots -- governed by the Army's rigid "fly before buy" program -- have resulted in 100 percent acceptance. Missile reliability in all firings, foreign and domestic, is 97.6 percent. TOWs fired from U.S. Army AH1Q Cobra helicopters have been equally reliable. TOW can also be installed in several other helicopters, foreign and domestic.

The Smithsonian Institution's new National Air and Space Museum in Washington, D.C., will have an automatic central control system manufactured and installed by Hughes' microelectronic products division. The high-speed, wide-bandwidth system will control exhibits, fire safety, security, environment, remote inquiry and display terminals, and two-way closed-circuit television. It also will transmit 464 channels of high-fidelity audio to locations throughout the museum for individual exhibit sound tracks and visitor information announcements.

Engineering opportunities at Theta-Com, a Hughes subsidiary which manufactures microwave and VHF-distribution equipment for the CATV industry, include: CCTV Applications Engineer experienced in security surveillance applications, to prepare proposals, design installations, supervise construction. Some travel....Microwave Relay Equipment Design Engineer, intermediate level.... Experienced Digital Communications Engineer. All positions require BS or higher degree. Send resume and salary history to: Director of Engineering, Theta-Com, P.O. Box 9728, Phoenix, AZ 85068. An equal opportunity employer.

A new type of high-voltage DC circuit breaker -- the first to be placed in commercial operation -- will be designed and built by Hughes Research Laboratories under contract to the Electric Power Research Institute. The breaker will be installed in the Pacific Intertie system. It will permit automatic transfer of current from an earth-return mode (which is used when station equipment of one pole of the normally bipolar system is rendered inoperative) to a metallic-return mode. The disconnect switch the Hughes breaker will replace requires about an hour to make this transition, during which the Intertie system is completely shut down.

Dannon Yogurt. If you don't always eat right, it's the right thing to eat.

Every day, millions of people give up eating. For snacking.

Well, if you find yourself doing more eating on the run than at a table, make sure you're eating Dannon Yogurt.

Our label shows you that Dannon is high in protein, calcium and other things nutritionists say are good for you.

It also shows that, unlike so many snack foods, Dannon is low in fat, contains no starch, no gelatin or other thickeners. And none of those hard-to-pronounce additives. Because Dannon Yogurt is 100% natural. Not just "natural flavor," but natural everything. No artificial anything.

Dannon is reasonable in calories, too. Especially when you consider how satisfying and nutritious it is.

What's more, Dannon gives you the benefits of yogurt cultures. They make yogurt one of the easiest foods to digest, and have been credited with other healthful properties too.

Oddly enough, not all yogurts have any yogurt cultures to speak of. In some brands—mainly pre-mixed or Swiss Style—the cultures are often deactivated by the processing.

We created a whole culture of yogurt lovers.

Dannon outsells all other brands. For a number of good reasons.

For example, we go out of our way to get the best natural ingredients: to Eastern Europe for strawberries, to the West Coast for boysenberries, and we go to Canada for blueberries. (Maybe the reason that other yogurts don't come close to the taste of Dannon is that other yogurt makers don't go quite as far.)

And it's the yogurt delivered direct to your store "from Dannon to dairycase." So if it tastes fresher, that's because it *is* fresher.

Dieters aren't the only people who are big on Dannon®.

Today, almost everybody's eating Dannon. It makes a quick, delicious breakfast, a light but filling lunch, and of course you can't beat it as a high nutrition dessert or snack. Spoon it out of the cup as is, or mix with cottage cheese, fresh fruit, peanut butter, honey, or what-have-you.

A suggestion for beginners: since plain yogurt may be a bit tart, start with Dannon fruit yogurts—strawberry, blueberry, red raspberry, and others.

For more facts, including some unexpectedly delicious ways to eat Dannon, write for our booklet, "Yogurt and You." It's free.

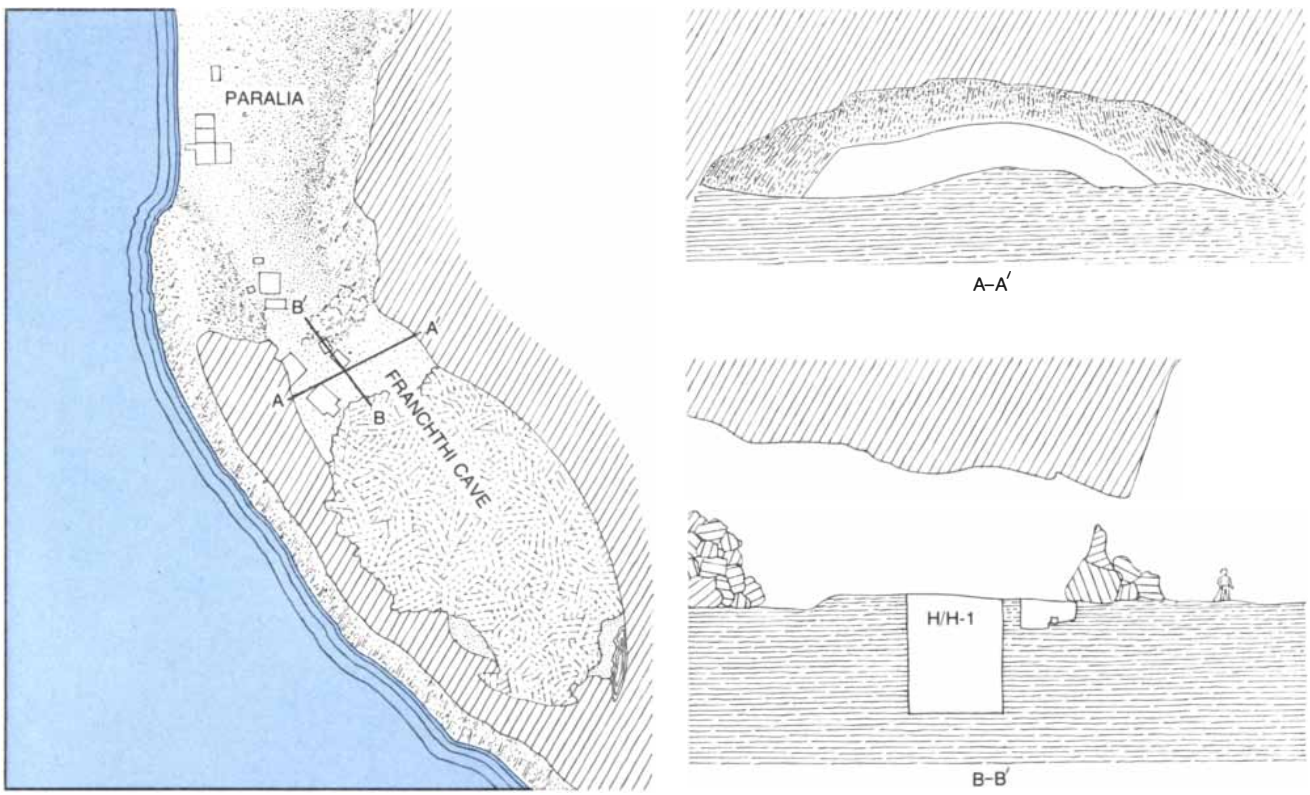
If you are a dieter, by all means write for our 48-page guide, "Dieting, Yogurt and Common Sense." Just send 25¢ for postage and handling to Dannon, 22-11 38th Avenue, Long Island City, New York 11101.

It will give you more reasons why Dannon is the right thing to eat—whether you are counting calories or not.

78B

er includes those of the wild horse and the wild goat. Now it is the bones of the red deer, which had become the favored prey in Final Paleolithic times, that make up the majority of the remains of larger animals.

The prominence of the red deer in the early Mesolithic levels at Franchthi could of course represent a change in the hunting preference or practices of the inhabitants. It is also possible, however, that a gradual change in the environment was accompanied by a greater abundance of deer. It is at this time that the landscape seems to have been transformed into open forest. Part of the evidence for the transition is the remains of plant foods at the site. Small quantities of wild pulses (now including the pea) continue to be found in the early Mesolithic strata. At the same time the shells of pistachios and almonds suddenly become abundant. However forested or unforested



HILLSIDE CAVE AND SHORELINE at Franchthi (*left*) are where the author and his colleagues unearthed a consecutive series of occupation strata deposited between 20,000 and 3000 B.C. The shoreline deposit, exclusively Neolithic in age, was some two meters in depth; the two main excavations within the cave, yielding artifacts from Paleolithic to Neolithic in age, were nine and 11 meters deep. Solid lines inside the cave (*left*) indicate the orientation of the two profiles of the cave (*right*). Most of the cave is filled with fallen rock.

	WILD ANIMALS									WILD PLANTS		DOMESTIC ANIMALS				DOMESTIC PLANTS	
	ASS	GOAT	RED DEER	CATTLE/BISON	PIG	LAND SNAILS	MARINE MOLLUSKS	SMALL FISHES	LARGE FISHES	PULSES	PISTACHIO	ALMOND	SHEEP/GOAT	PIG	CATTLE	PULSES	CEREALS
FINAL NEOLITHIC			H	?	?	G	G	F	?		G	G	h	h	h	f	f
LATE NEOLITHIC			?	?	?	G	G	F	F		G	G	h	h	?	f	f
MIDDLE NEOLITHIC			?	?	?	G	G	F	F		G	G	h	h	h	f	f
EARLY NEOLITHIC			H	?	?	G	G	F	F		G	G	h	h	?	f	f
UPPER MESOLITHIC			H	H	H	G	G	F	F	G	G	G					
LOWER MESOLITHIC			H	H	H	G	G	F		G	G	G					
UPPER PALEOLITHIC	H	H	H	H	H	G	G	F		G							

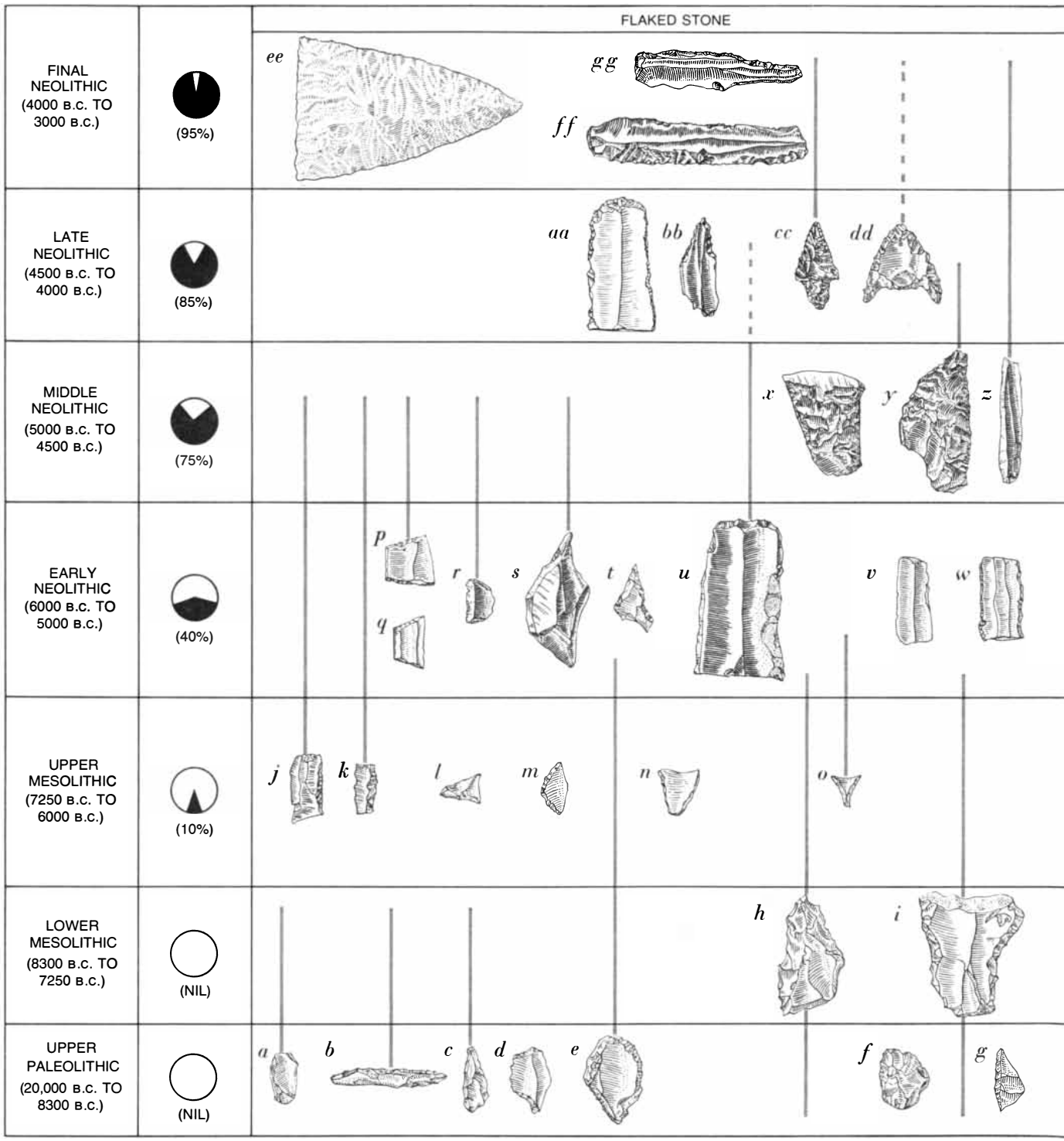
PROCUREMENT METHOD

HUNTING H
GATHERING G
FISHING F
HERDING h
FARMING f

FOOD RESOURCES in the vicinity of Franchthi were exploited in differing degrees by seasonal visitors to the site. Both Paleolithic and Mesolithic visitors hunted game, gathered wild plant foods and fished, but their hunting and gathering choices were often limited by environmental conditions. For example, the Paleolithic climate evidently did not favor pistachio and almond trees but suited wild ass, wild goat and bison. Mesolithic conditions seem to have favored animals more at home in a woodland setting, such as red deer and wild boar. Visitors during both periods caught fish and gathered shellfish and land snails. Before upper Mesolithic times, however, only small fishes were caught. Hunting and fishing continued during Neolithic but generally fell off as domesticated plants and animals appeared.

the landscape may have been, the trees that bore these fruits were certainly present.

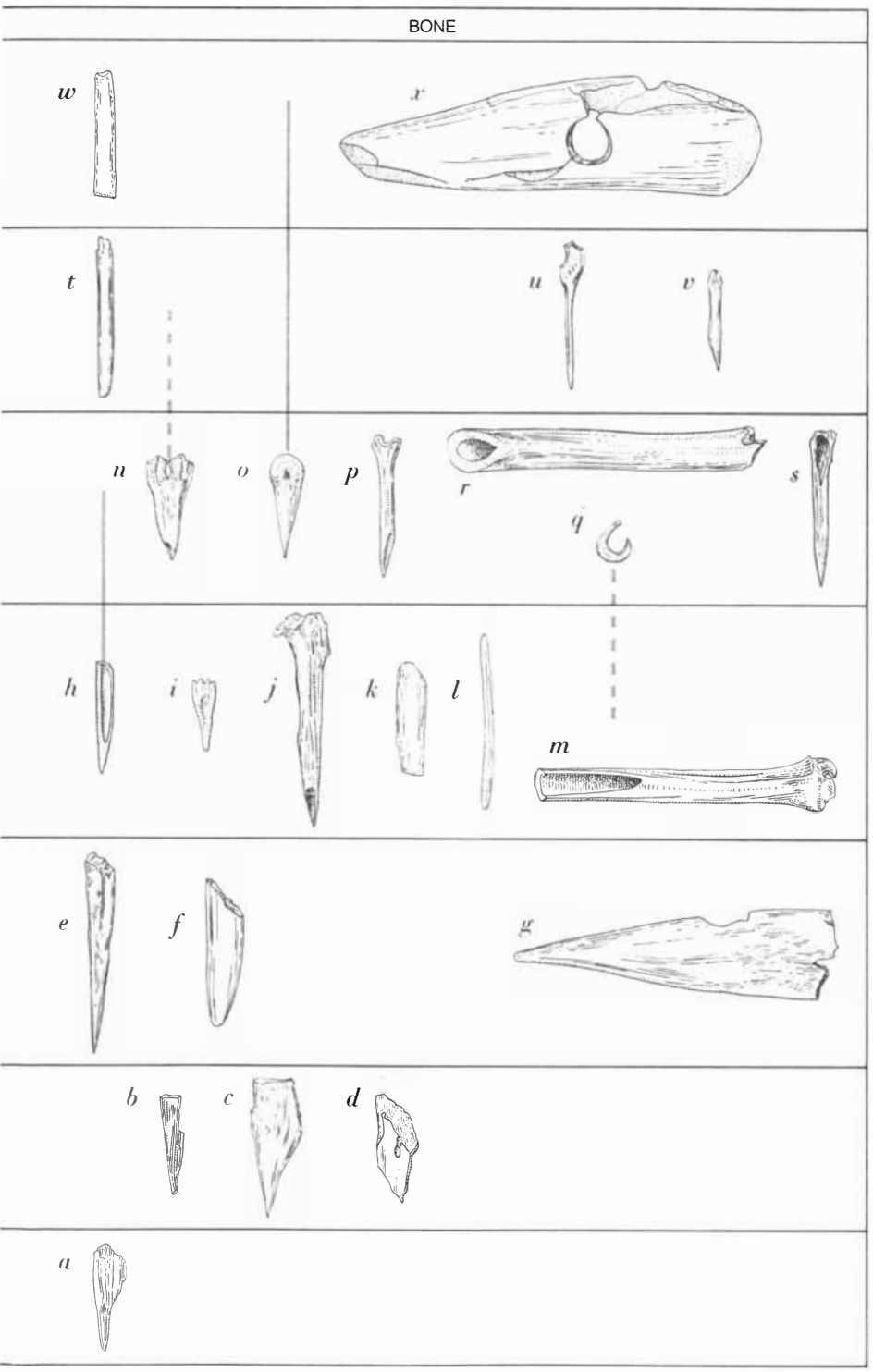
The occupants of the site continued to do some fishing and to supplement their diet with land snails and marine mollusks. For that matter, the early Mesolithic hunting and gathering activities were in the main still comparable to the economic strategies that had been pursued by the occupants' Paleolithic predecessors; the differences in the archaeological evidence are more reflective of an environmental adjustment than of a major change in the manner of subsistence. We expect, however, that with further study it will be possible to estimate more accurately the relative importance of each component of the overall subsistence pattern at any given time in the history of the site. Once that is accomplished we should be better able to understand the changing relations between man and his environment



SELECTED ARTIFACTS from successive levels at Franchthi are made from flint and obsidian (*above and right*) and from bone (*far right*). The gray lines indicate when similar objects are found in more than one of seven temporal intervals (*far left*); pie charts show the increasing preference for obsidian as a raw material. Upper Paleolithic and Mesolithic tools include "backed" bladelets (*a–c, j, k*), microburins (*d*), small scrapers (*e, f*), microliths (*g, l–o*) and denticulated pieces (*h, i*). Similar forms (*p–r*) were still produced in the earlier phases of the Neolithic but are often slightly larger. Perhaps the most typical Neolithic tool is the simple blade (*z*), sometimes worked further to produce a variety of specialized forms (*v, w, aa, bb, ff, gg*). Projectile points—transverse (*x*), shouldered (*y*), tanged (*cc*) or

during the long period of human occupation at Franchthi.

The later Mesolithic levels at the site, the earliest of which may be dated at about 7250 B.C., are marked by two notable innovations. One is the appearance among the animal bones of fish vertebrae that are much larger than any found in the earlier levels. The vertebrae are comparable in size to those of the modern tunny, a fish that typically weighs several hundred pounds. The bones clearly point toward an increased exploitation of marine resources in the vicinity of the site. The second innovation provides strong confirming evidence for the first. It is the sudden appearance of a new tool material: obsidian, a handsome and distinctive volcanic glass. The oldest obsidian at Franchthi antedates by at least 1,000 years the earliest-known occurrence of the material anywhere else in the Aegean basin.



tanged and barbed (*dd*)—may have been used in both fishing and hunting; large "lance heads" (*ee*) appear in final Neolithic strata. Bone tools, particularly common in the Neolithic, include many points that vary in form and workmanship (*a, b, e, h, n and u are typical*). Bone points could have been applied to many kinds of task. Other forms typical of the Neolithic are scoops or gouges (*m, r*) and fishhooks (*q*). Perforated hoes or mattocks made from antler (*x*) are uncommon in the deposits, and they are confined to the later phases of the Neolithic period.

Obsidian is superior to flint in some respects; for example, although it is not as strong, it lends itself to a sharper cutting edge. As a result the material was intensively sought in later prehistory, not only in Neolithic times but also during the Bronze Age. Colin Renfrew of the University of Southampton and his colleagues have located many of the main sources of raw obsidian in the eastern Mediterranean. Thanks to their cooperation we now know that the Mesolithic obsidian found at Franchthi probably came from a deposit of the volcanic glass at Adhamas, on the northern side of the island of Melos, which is separated from Franchthi by some 150 kilometers of open sea.

The sudden and simultaneous appearance of large fishbones and obsidian in the later Mesolithic strata at Franchthi, taken together with the close quantitative correlation between these two kinds of material in subsequent Mesolithic strata, strongly suggests that the acquisition of obsidian from Melos was intimately connected with the new fishing practices on the Argive Peninsula. It remains for us to identify those practices. For example, we can only guess at the nature of the craft used by these pioneer voyagers. Until quite recently, however, the lobstermen of Corfu put out to sea in simple boats made of reeds. Vessels of this kind should have been within the technological competence of Mesolithic fishermen; they might also have been adequate for the Melos voyage.

Apart from a comparative abundance of microlithic implements—now made out of both flint and obsidian—and a smaller number of bone tools, little has survived to give a picture of the material culture of the Mesolithic residents of Franchthi. The microliths, many of them geometric, would have been equally useful to hunters as projectile points, to fishermen as harpoon barbs and possibly even to the collectors of plants as a cutting edge for primitive sickles. Studies of the microliths' working edges for evidence of different kinds of wear may eventually enable us to choose among such possibilities, but the required analyses remain to be carried out. The bone tools are largely in the form of points and could have served a variety of purposes, but again detailed study lies in the future. Simple items of personal adornment such as pebble pendants and pierced shells, which are quite rare in the Paleolithic strata, are somewhat commoner in the Mesolithic ones.

We have found one complete burial belonging to the Mesolithic period. It contained the oldest entire skeleton yet uncovered in Greece. The burial provides at least indirect information about the ideology of the people then living at the site. A man

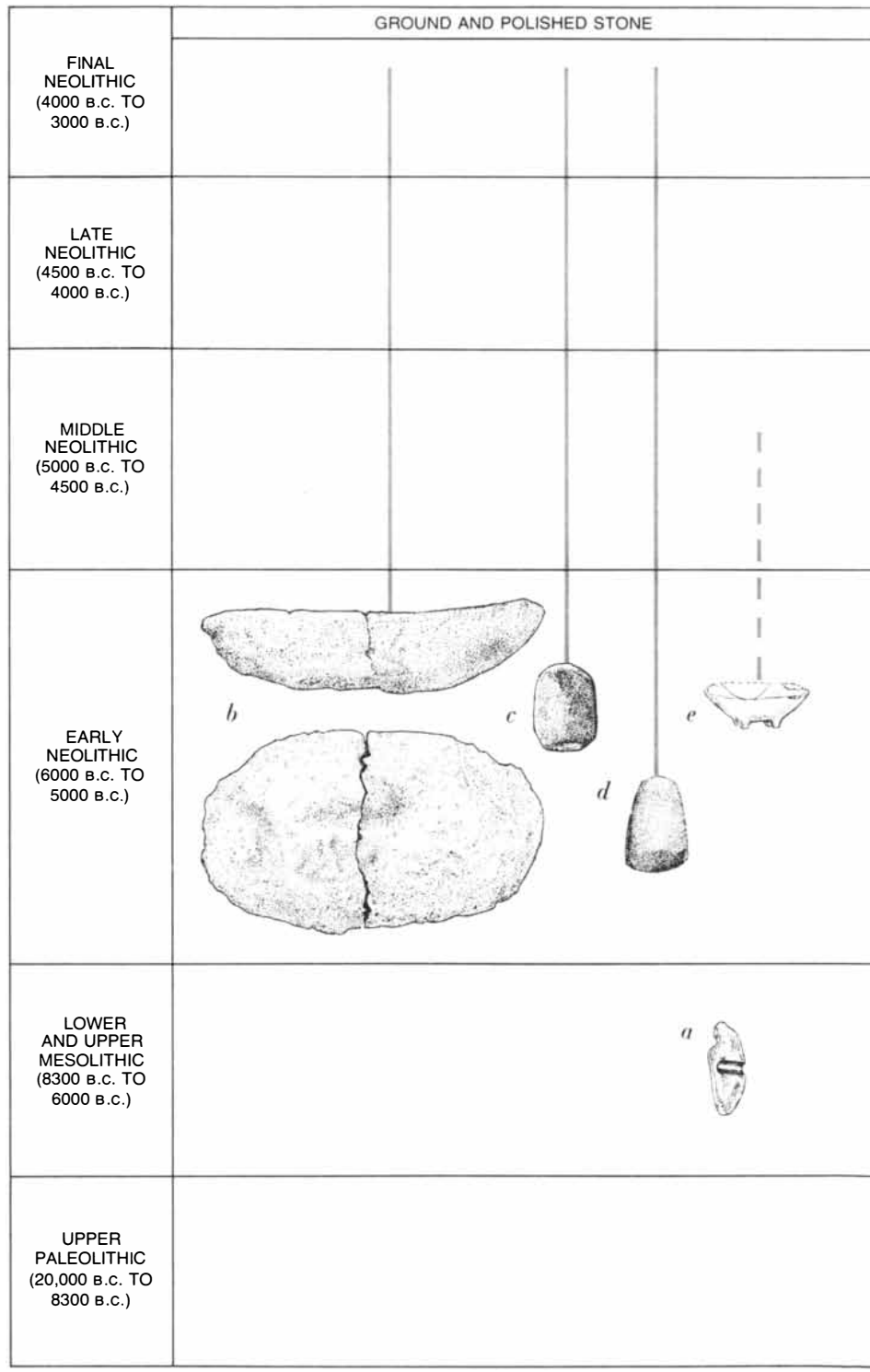
aged about 25 had been interred in a shallow depression just inside the present mouth of the cave. The body had been put into the grave with its knees drawn up and had been covered with a pile of fist-sized stones. The young man seems to have died from severe blows on the forehead, but it is possible that he had already been near death from malaria. That is the diagnosis of J. Lawrence Angel of the Smithsonian Institution, based on his analysis of certain bone abnormalities. We take the lack of any recognizable grave offerings as reflecting the absence among the Mesolithic occupants of Franchthi of any significant concern about life after death, or at any rate the absence of any clearly developed sense of personal property. As for the possibility that the young man had been suffering from malaria, it may be relevant that the postglacial rise in sea level could well have created marshy coastal areas suitable as breeding grounds for malarial mosquitoes.

About 6000 B.C. or shortly thereafter something new began to happen at Franchthi. There is no marked discontinuity in the occupational sequence, but the materials excavated from these levels differ in many respects from anything found in earlier ones. For example, the animal bones tentatively identified in Paleolithic strata as being those of some kind of wild goat were entirely absent from Franchthi during the 2,000-year Mesolithic occupation of the site. About 6000 B.C., however, the bones of goats (and sheep) reappear in substantial numbers. Moreover, they seem to be the bones not of wild forms but of domesticated ones. Sheep and goats are the domesticated animals most commonly associated with Neolithic cultures in this part of the Old World, particularly in southwestern Asia and southeastern Europe.

At about the same time two new plants—wheat and barley, both probably domesticated—make their first appearance at Franchthi. This combination of zoological and botanical evidence, strongly suggesting that the inhabitants of Franchthi had begun to engage in animal husbandry and agriculture early in the sixth millennium B.C., is supported by the appearance of new kinds of stone tools in the same levels of the cave.

The new tools include celts: ax heads made of hard stone that are given their final shape by grinding and polishing instead of flaking. Mounted on a handle of wood or antler, an implement of this kind can be used to clear the land of small trees and brush or to loosen the soil for planting. Coarse millstones also appear for the first time, along with flint blades that show evidence of having served as the cutting edge of sickles.

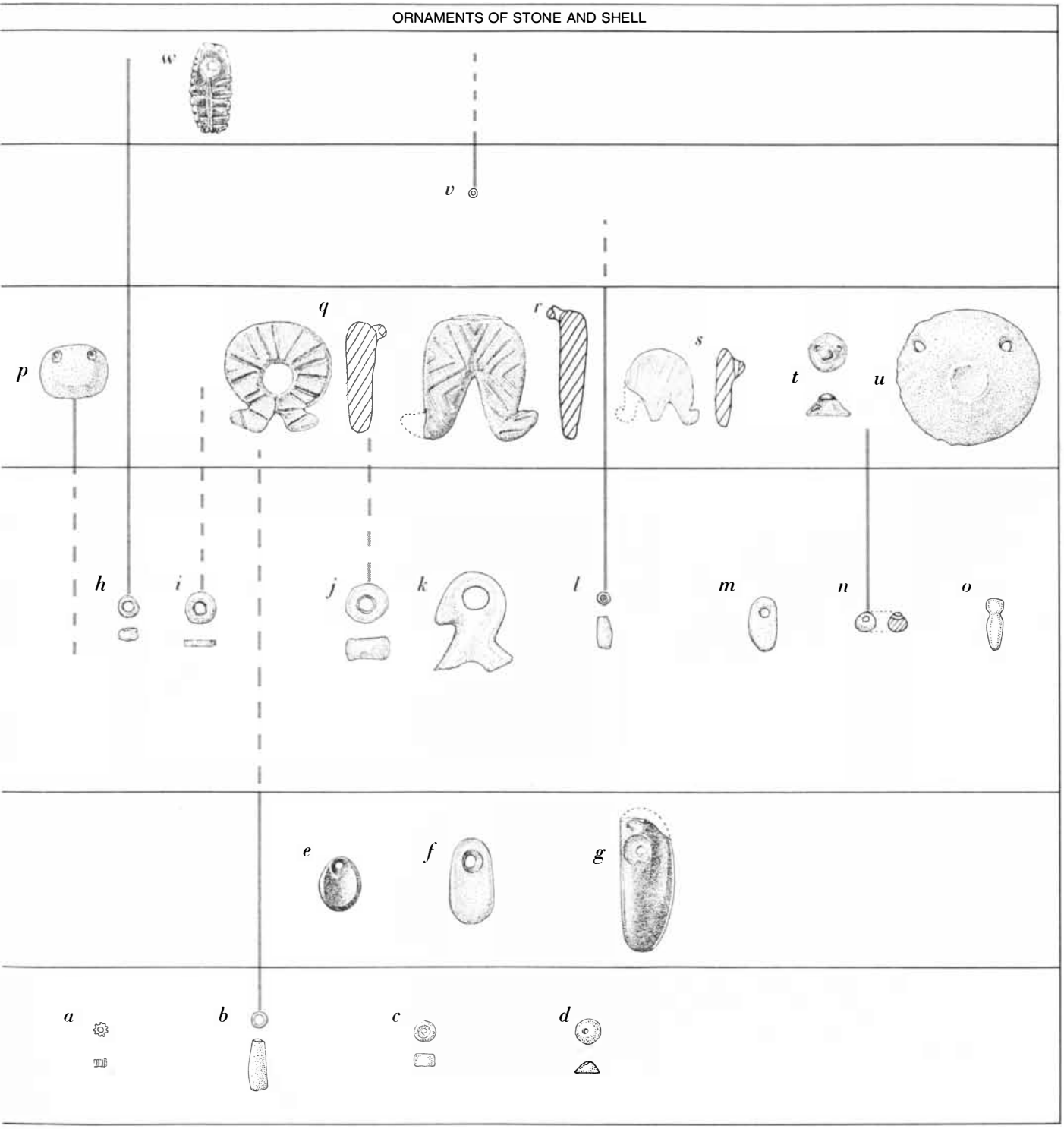
Our present correlation of these strata within the existing framework of carbon-14 dates at Franchthi is not precise enough to settle the question of whether or not the first evidence for pottery making at the site is exactly contemporaneous with the evidence for animal husbandry and agriculture. It is reasonably certain, however, that if there was any delay between these first indications of the transition from a Mesolithic to a Neolithic way of life and the start of pottery making, it was not a long one. We should perhaps have expected this; among early agricultural societies elsewhere the need for durable nonporous containers for the storage of food seems to have led to experimentation in the shaping and firing of clay vessels. The earliest pots at Franchthi, which are among the earliest found anywhere in Greece, were made without benefit of a potter's wheel; their forms are simple and



OTHER FRANCHTHI ARTIFACTS include stone objects (*above*) finished by means other than flaking, such as grinding and polishing. Most are Neolithic but one (*a*), a deeply grooved stone, is from a Mesolithic level. It may have been used for making bone points. The Neolithic pieces include a quern (*b*) for milling grain, a rubbing stone (*c*) to be used with the quern, a polished stone celt (*d*) and a tiny marble bowl (*e*) found in the grave illustrated on page 84. It is the

they are generally undecorated. In these respects they resemble the pottery found in a number of comparably early sites elsewhere in the eastern Mediterranean.

This innovative period at Franchthi coincides with the first extension of the area of occupation beyond the shelter of the cave. The new area was to the northwest of the cave, running along the present shoreline. Here in 1973 and 1974 we uncovered a stratified Neolithic deposit that turned out to be some two meters deep; its bottom level rests on virgin soil. The Paralia deposit, as we call it, contains some remnants of structural stonework. Sea erosion and the breakdown of the craggy slopes above the deposit, together with the limited extent of our excavations in the area, leave the exact nature of the structures obscure. We have considered the possibility that they are the remains of agricultural terraces, but the most



most complete of a number of stone vessels, usually found in fragments, from the earlier Neolithic levels. Ornaments (*above*) range from simple Paleolithic and Mesolithic articles to more complex Neolithic ones. Beads (*a, b*) were made from naturally fluted and smooth seashells; others (*c, d*) were made from stone, as were the Mesolithic pendants (*e–g*). Among the Neolithic ornaments stone studs (*o*) and bossed pendants (*t, u*) are noteworthy, as are zoomorphic (*k*) and anthropomorphic (*q–s*) amulets. The last, representing half-trunks and legs, may be fertility charms; they have not been found elsewhere in Greece. One pendant (*w*) had ocher in its grooves.

probable explanation is that they were retaining walls on the uphill side of a small settlement. There may be vestiges of dwellings here, but much of the seaside Neolithic settlement could now be submerged in Franchthi bay.

The lowest Neolithic strata in the Paralia deposit contain the earliest evidence for pottery making at Franchthi. Until the age of these strata can be more precisely determined we are placing it a little later than 6000 B.C. With the possible exception of a hiatus around 4000 B.C. the deposit is thereafter made up of a succession of strata that continue to the end of the Neolithic period in Greece, about 3000 B.C. This coincides generally with the Neolithic sequence in the cave itself, and so we can add some three millenniums of Neolithic settlement to the long record of the Paleolithic and Mesolithic occupation of Franchthi.

These must have been millenniums of considerable growth of population throughout Greece, as the establishment of the Paralia settlement itself suggests. Nevertheless, a slight puzzle remains. In spite of the implication of permanence normally associated with an agricultural way of life and in spite of the presence of stone structures, we still lack evidence of definite overwinter occupation at Franchthi. Such evidence may become available as our study of animal remains continues, but we must accept the possibility that a seasonal regime governed the social and economic life of these farming people much as we think it governed the life of the people of earlier times. One point to be considered in this connection has been suggested by our preliminary study of the botanical remains: the cultivation of olive trees, fig trees and grapevines, a major winter activity of the Greek farmer today, was unknown to the Neolithic farmers of Franchthi. Indeed, evidence from elsewhere in Greece indicates that the cultivation of these important Mediterranean staples was not fully established in the Aegean basin until the Bronze Age.

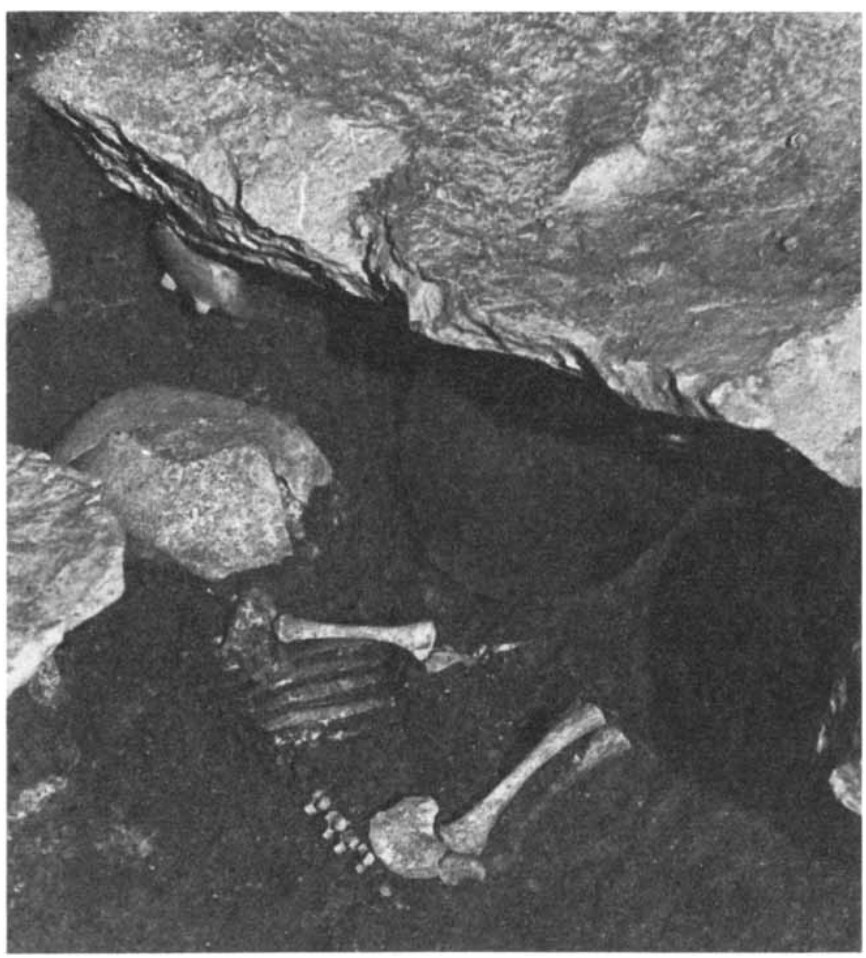
Many of the differences between the Mesolithic and Neolithic ways of life at Franchthi are apparent in the contrasting elements of material culture. For example, to the extent that Mesolithic ideology was reflected in personal ornament its expression was limited to a few simple pendants and beads made from pebbles and shells. The Neolithic deposits are far richer in such artifacts. Among other objects, they have yielded a number of clay figurines of both animals and human beings. Most of the figurines are represented only by fragments, but the anthropomorphic figures are mostly female, and many of them emphasize sexual characteristics.

The Paralia excavations have yielded a good deal of evidence concerning an important component of the social and economic life of these early farmers: their craft activities. We may not be justified in speaking of craft specialization at that early time, but the presence of an industry for the production of shell beads is certainly noteworthy. We first recognized the possibility that such an industry existed when during the 1973 season at Paralia we unearthed a large number of small drill-like implements, from two to three centimeters long, flaked out of a reddish flint of poor quality.

The most popular raw material for the production of flaked stone tools at Franchthi during the Neolithic was obsidian; indeed, by the end of the Neolithic obsidian was virtually the only material so used. Nevertheless, the special flaking qualities of the reddish flint seems to have made it more suitable than obsidian for drills, even though it is good for little else. We were at first surprised when we found the poor-quality flint was a far commoner raw material in certain Paralia deposits than obsidian. The reason became apparent when we found finished beads by the score together with bead "blanks" in various stages of completion. Evidently the flint drills are associated with the production of shell beads.

Our excavations at Paralia indicate that the bead industry was active at the site for a short time in the latter part of the sixth millennium B.C. This marks one of the rare occasions in Aegean prehistory when an early craft has been identified by the discovery not only of the finished product but also of the tools that were used to make it. Whether or not these objects were made by an individual or a group whose special occupation it was, the absence of either beads or drills in other Neolithic contexts at Franchthi suggests that the bead industry was localized in the settlement along the shore.

Although stone beads and pendants had been made at Franchthi as early as Mesolithic times, their number and quality increase markedly in the Neolithic levels. Several of the pendants are of particular interest because they represent only the lower half of a human body. They could have been intended as fertility charms; at least this would have been in keeping with the ideology of a society that placed a social and economic emphasis on the fecundity of man and nature alike.



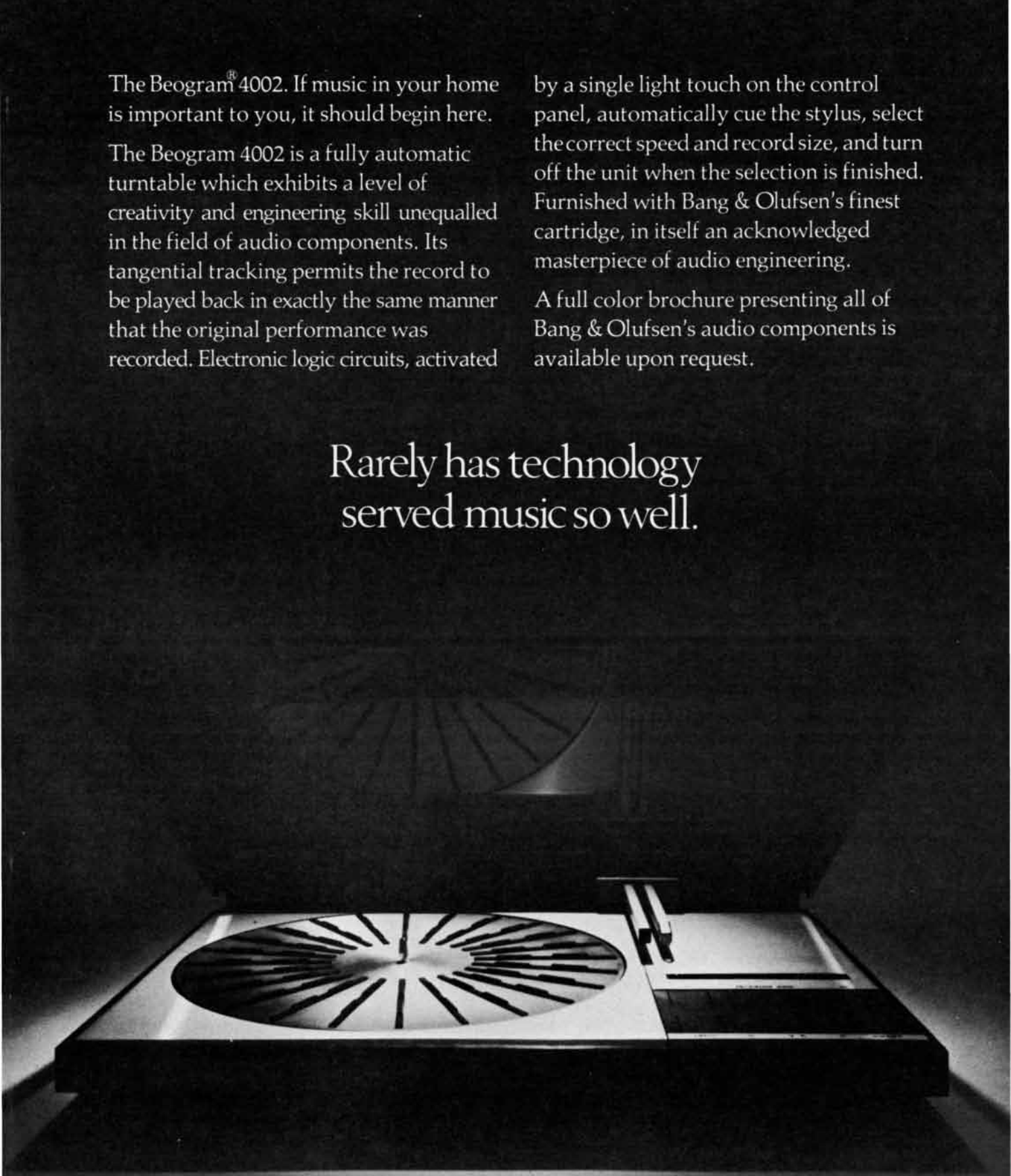
INFANT BURIAL, the only one of eight such child burials of the Neolithic period at Franchthi to contain grave offerings, was inside the cave. To the right of the forearm, partly concealed by one of the stones that covered the burial, is half of a broken clay pot; it is suggestive of the widely followed practice of "killing" pottery that was used as a grave offering. Also under the rock, above the infant's skull, is a small marble vessel that was nearly intact when it was found; the vessel is among the artifacts that are depicted in the illustration on preceding two pages.

The Beogram® 4002. If music in your home is important to you, it should begin here.

The Beogram 4002 is a fully automatic turntable which exhibits a level of creativity and engineering skill unequalled in the field of audio components. Its tangential tracking permits the record to be played back in exactly the same manner that the original performance was recorded. Electronic logic circuits, activated by a single light touch on the control panel, automatically cue the stylus, select the correct speed and record size, and turn off the unit when the selection is finished. Furnished with Bang & Olufsen's finest cartridge, in itself an acknowledged masterpiece of audio engineering.

A full color brochure presenting all of Bang & Olufsen's audio components is available upon request.

Rarely has technology served music so well.



Bang & Olufsen

Bang & Olufsen of America, Inc., 515 Busse Road, Elk Grove Village, Illinois 60007

1. LANCIA BETA COUPE
a sporty, front-wheel drive car with a 5-speed transmission, rack-and-pinion steering, plus real leather bucket seats.

2. LANCIA BETA SEDAN
technically identical to the Coupe yet also has an adjustable steering column, reclining bucket seats, and deep, plush carpeting.

3. LANCIA BETA HPE ESTATE WAGON*
a genuine sports wagon. Flip down the rear seats and toss in almost anything — from a pair of skiis to a six piece set of luggage.

4. LANCIA BETA SCORPION*
the most exciting Lancia of all. It's a 5-speed, mid-engine, 2-seater sports car. And it's the result of 68 years of European racing.

LANCIA

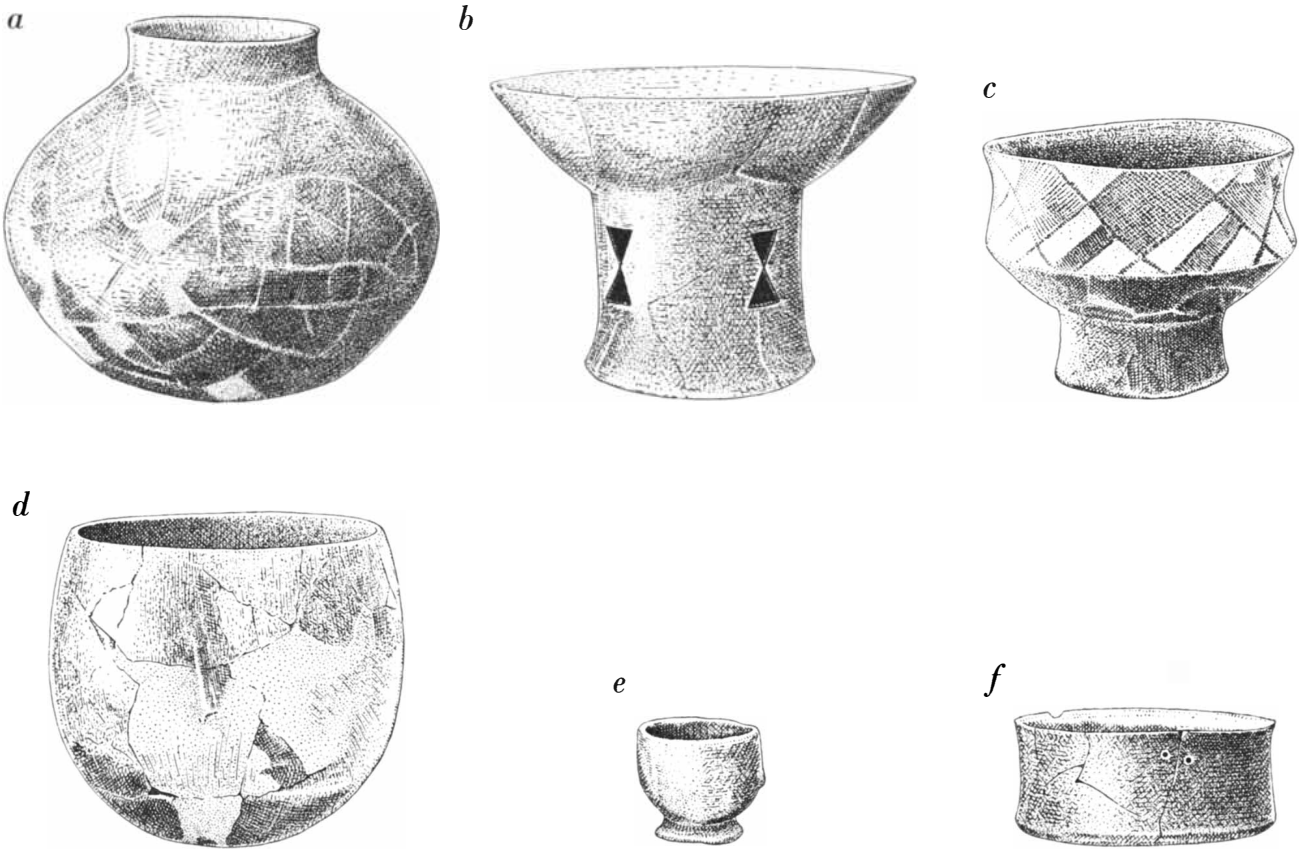
The intelligent alternatives.

The one thing that's come between most people and a European grand touring car has been an outrageous price tag. Fortunately, there's now an intelligent alternative. Four to be exact. Because Lancia, one of Europe's oldest and most respected car manufacturers, has recently introduced to America a complete new line of cars. Each one is a classic example of sophisticated engineering as well as unhurried European craftsmanship. And each is realistically priced. So now that the financial barrier between you and a European grand touring car has been removed, we suggest you visit your nearest Lancia dealer soon.

* Subject to availability.

For the name of your nearest Lancia dealer call toll free (800) 447-4700 or in Illinois (800) 322-4400.

Overseas delivery, leasing and daily rental arranged through your participating dealer. Lancia of America, Division of Fiat Distributors, Inc., 155 Chestnut Ridge Road, Montvale, N.J. 07645.



NEOLITHIC POTTERY from the earlier levels at Franchthi (*d, e*) is undecorated but skillfully made. The deep hemispherical pot has a burnished surface, as does the footed piece, a cup or small bowl of sophisticated design. The potter's skill reached its height in the middle Neolithic. The collared jar (*a*) is simple in form, but its lustrous finish results from the chemical composition of the paint that was applied before firing. Other middle Neolithic decorative devices include cutouts (*b*) and painted geometric designs (*c*). The holes in the rim of the shallow basin (*f*) were made after the piece had been cracked during use; the break was probably repaired with a tightly tied thong.

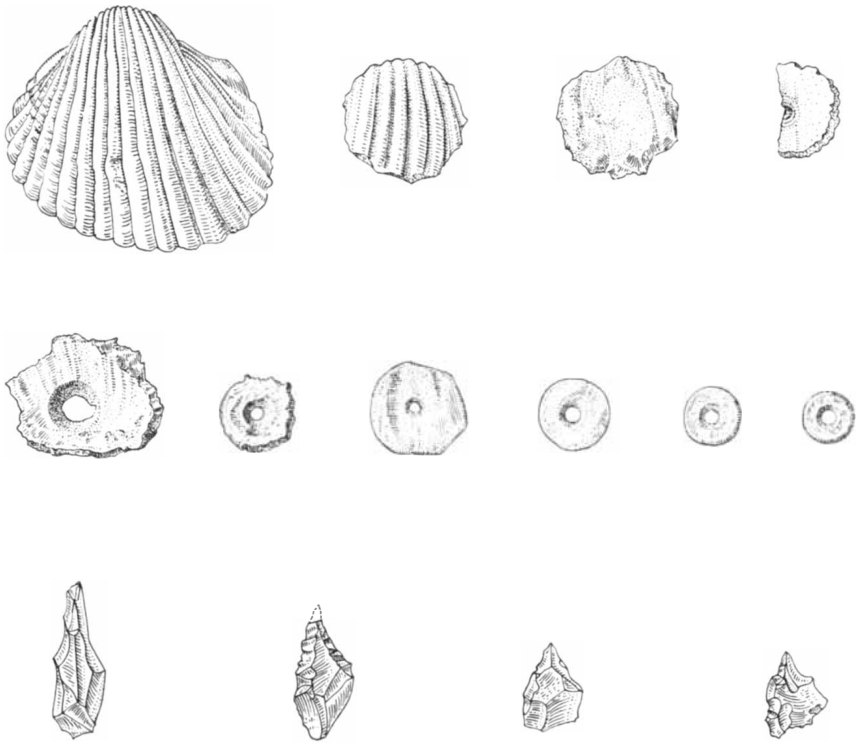
It should not be supposed, however, that such an emphasis implies any neglect of the collateral economic activities of the society: hunting, gathering and fishing. It is not unlikely that women continued to collect wild plants in season. Moreover, entirely new kinds of artifacts—bone fishhooks and tanged projectile points made of both flint and obsidian—show that fishing and hunting were still important activities.

Burial customs are almost always a reflection of ideological attitudes, and the Neolithic burials we uncovered at Franchthi are significant in this respect. Generally the Neolithic practices follow what little we know of the Mesolithic tradition; the burials are distributed seemingly at random within the confines of the settlement (inside the cave or along the shore), the graves are shallow pits and the bodies were buried with their knees drawn up. At the same time certain innovations are evident. For example, the custom of secondary burial seems to have arisen in the later phases of the Neolithic. This practice called for letting the soft parts of the body decay by either leaving the body exposed or burying it temporarily. Thereafter the bones were bundled together and deposited in their final resting place.

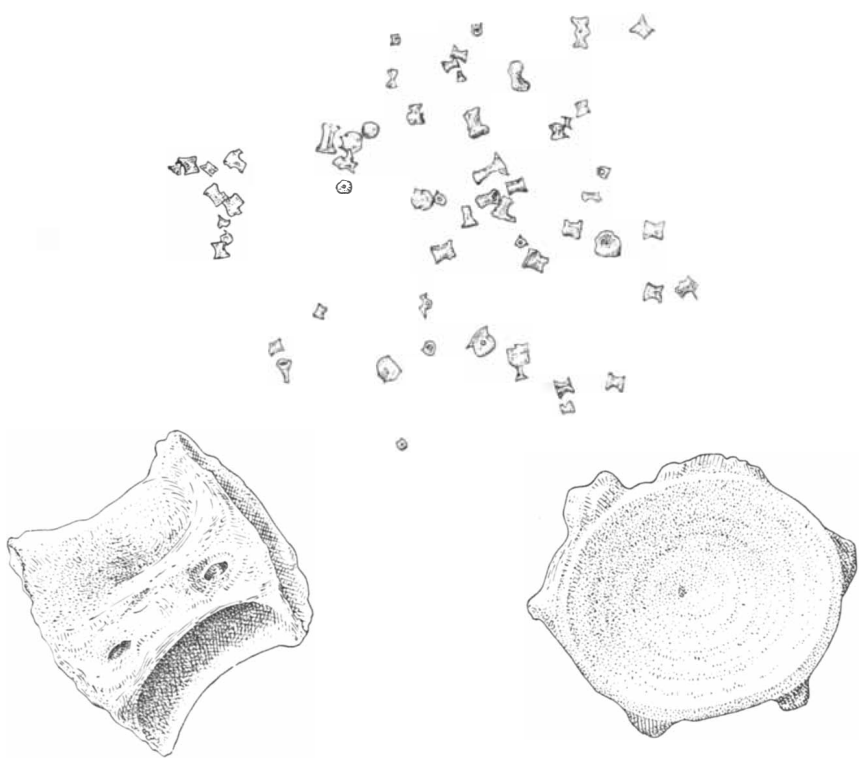
Grave offerings are rare in the Neolithic burials at Franchthi. The two exceptions are worth mentioning. The first is an infant burial of the early Neolithic that we found inside the cave during the 1973 season. It was one of a total of eight Neolithic burials of infants or children unearthed at Franchthi, four inside the cave and four outside. It was also the only one of the eight to include grave goods. Although the infant was only a few weeks old when he died, an unusually fine vessel made of marble and almost exactly half of a broken clay pot were buried with him. One wonders why, in a time when grave goods were so rare, such a young person should have been honored with any. One also wonders whether the pot had been broken symbolically ("killed," as some ethnologists say) during a funeral ceremony, as was still customary in the vicinity of Franchthi until just a few years ago. These questions must remain unanswered, but it seems safe to guess that the infant or his parents had some kind of special status in the community.

The other noteworthy Neolithic burial is that of a 40-year-old woman whose skeleton was wedged so tightly into a small pit in the Paralia settlement that it almost certainly represents a secondary burial. A complete but well-worn pot that shows signs of having been mended was buried with her; its style suggests that the woman died sometime before 4500 B.C. Among the other grave goods is a group of bone tools that includes several well-made points and a number of obsidian blades. Such a set of tools would be appropriate for a person engaged in any one of several crafts. Whether the objects were gifts indicative of the woman's status in the community or were merely personal possessions is hard to ascertain. It seems likely, however, that among the later Neolithic inhabitants of Greece a sense of personal property was growing.

The Neolithic settlement at Franchthi is in most respects typical of early village-farming communities throughout the eastern Mediterranean. Once established it flourished for some 3,000 years. Over that span the inhabitants of the cave and the shoreline settlement seem to have had increasing contact with other parts of the Aegean basin and even with areas well beyond the Aegean. That, however, is another story, and perhaps I should conclude this account by discussing a key question: To what extent does the Neolithic settlement at Franchthi represent a natural or inevitable



BEAD INDUSTRY, utilizing cockleshells as the raw material, was practiced only at Paralia during a few score years of the Neolithic occupation of Franchthi. Drill-like tools, made from a poor grade of flint (*bottom*), heavily outnumber obsidian tools in strata at Paralia representing a short span of time in the late sixth millennium. Their function as bead-making artifacts became apparent when they were found together with unfinished beads, blanks and complete beads.



IMPROVED FISHING is dramatically demonstrated by the sudden appearance in upper Mesolithic strata of fish vertebrae (*bottom*) much larger than the vertebrae (*top*) found in Paleolithic and lower Mesolithic strata. The appearance of the larger fishbones coincides with the appearance of obsidian at Franchthi. Obsidian came from Melos, a 150-kilometer voyage away.

development of the local Paleolithic and Mesolithic occupations? A clear answer to the question would contribute greatly to our understanding of early Greek prehistory and would help to clarify the relations between this area and the prehistoric cultures of neighboring regions to the north and the east.

To review the record briefly, the most dramatic innovations over some 17,000 years at Franchthi came at the end of the Mesolithic, about 8,000 years ago. The physical limits of the settlement were enlarged, the size of the local population may have increased accordingly and a host of new elements can be perceived in the archaeological record. The most striking of these elements is the sudden appearance of species of plants and animals previously unknown in the area. The new plants and animals appear to be domesticated, and so it is unlikely that they represent the culmination of a long period of local experimentation in plant and animal husbandry. The evidence thus leads us to conclude that a basically agricultural economy, with all the term implies from the technological, social and ideological points of view, was introduced from elsewhere and imposed on a native Greek Mesolithic substratum. Whether or not this cultural revolution was accompanied by an influx of new people cannot now be established. All things considered, however, at least a modest infusion of new blood seems likely.

It might be noted that the situation at Franchthi bears a striking resemblance to that which is thought to have prevailed in much of Europe when agriculture first arose. It would be incautious, however, to suggest that the circumstances at Franchthi were necessarily typical even of Greece. To be sure, there is no evidence from elsewhere in Greece to contradict such a suggestion, but for the present it would be prudent to insist that something approaching a true picture will emerge only after considerably more excavation and research have been conducted throughout the region.

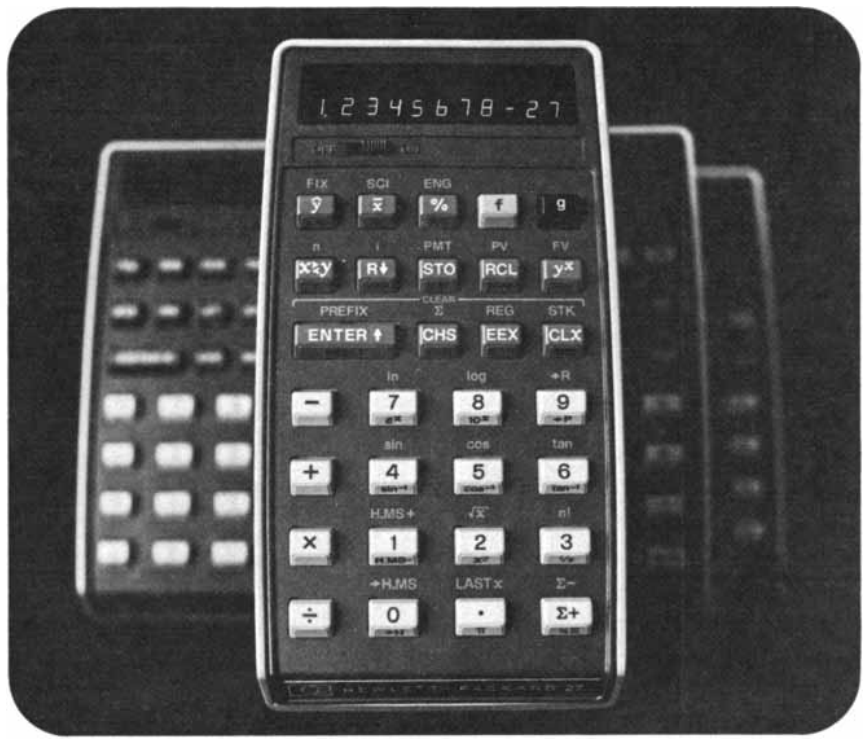
One conclusion that does seem to be justified at this point is that the establishment of the village-farming stage at Franchthi (and apparently at most other sites in Neolithic Europe) fits the generally accepted reconstruction that perceives agriculture as beginning in the "nuclear" Middle East. A growing body of evidence indicates that experimentation with potentially domesticable plants and animals was in progress at several locations in the Middle East during the millenniums represented at Franchthi by the late Paleolithic and Mesolithic occupations. This early experimentation was greatly facilitated by the availability in that part of the Old World of those same plant and animal species (indeed, many of the same species that eventually appeared at Franchthi). It is this fact more than any other that distinguishes conditions in the Middle East from those that seem to have prevailed in Greece and elsewhere in con-

FEMALE FIGURINE, made late in the fifth millennium B.C., is remarkable both for its seated posture and for the painted suggestions of clothing. Figurines representing human beings and, less frequently, animals first appear in the Neolithic levels. Most are in fragments, which may indicate deliberate destruction. All the figures of identifiable sex are female.

tinental Europe before the advent of agriculture.

Given this chronological priority, it has usually been assumed as a corollary that the village-farming way of life spread from the Middle East to Greece by some process of diffusion, probably at the same time bringing the first farmers (and even possibly the first human inhabitants) to Cyprus and Crete before ultimately reaching the southern shore of Europe. Such may well have been the case; at least we finally have some evidence that generally supports the hypothesis. Before dismissing our question with a simplistic "Noah's ark" answer, however, we must recognize that not all the problems have been solved and that many related questions are still unanswered.

For example, much more needs to be learned about the complex of factors that must bear on the origins of agriculture in the Middle East itself. To mention one unanswered question, archaeologists have only now begun to deal with the problem of why it all happened there when it did. As for Greece, the picture is even cloudier. Our preliminary findings at Franchthi constitute an important beginning, but they are only a beginning. Did the Neolithic of southern Greece really come into being as abruptly as it now appears it did? What factors led the Mesolithic hunter-gatherers to modify their long established way of life so markedly? What was happening elsewhere in the Greek peninsula at that time? These and other questions will occupy us as we press on with our study of this significant site.



The new HP-27 Scientific/Plus is the HP Professional for most reasons, including math, statistics, and finance.

For the scientist or engineer whose responsibilities include targeting, budgets, statistical analysis, or other financial or forecasting concerns, the HP-27 is uncommonly effective as a total, professional calculating system. This is because of its range of preprogrammed functions and some nifty features such as—

Ten addressable memories, in which you can perform register arithmetic and store constants and intermediate answers, let you make complex calculations with ease and confidence.

Selective memory clearing operations allow you to preserve useful data while opening up other registers for new calculations.

HP's special logic system with four-register stack almost completely eliminates the need to re-enter data, and lets you see all intermediate data and evaluate problems without worrying about parentheses or heirarchies.

You have a choice of display modes: fixed decimal, scientific notation, or engineering notation (values with exponents that are multiples of three).

For mathematical calculations, the HP-27 solves the most-used exponential, log, and trig functions—including sines, cosines, tangents and their inverses in three angular modes; natural and common logs and antilogs; coordinate and angle conversions; pi, exponents, and reciprocals.

For statistical calculations, the HP-27 lets you analyze data, allocate resources, forecast costs, and solve the problems most often encountered in science and business—including normal distribution, correlation coefficient and variance; two-variable mean and standard deviation; linear regression, linear estimate, and factorials.

For financial calculations, the numbers that you'd otherwise have to search for in bulky books and tables—and their manipulations—are built into the HP-27's circuitry. These time-value-of-money calculations include internal rate of return for different cash flows; discounted cash flow; compound interest; and annuities.

Price is $200* for the calculator with battery pack, recharger/AC adapter, case, and 216-page owner's handbook with 99 pages of applications in math, statistics, navigation, surveying, and finance.

Along with the other HP Professional hand-held and portable printing calculators, the HP-27 is available at selected department stores and college bookstores. Call 800-538-7922 in the U.S.A. (in California, 800-662-9862) for detailed information and the name of the HP retailer nearest you.

New HP Current Tracer: a missing link in digital troubleshooting.

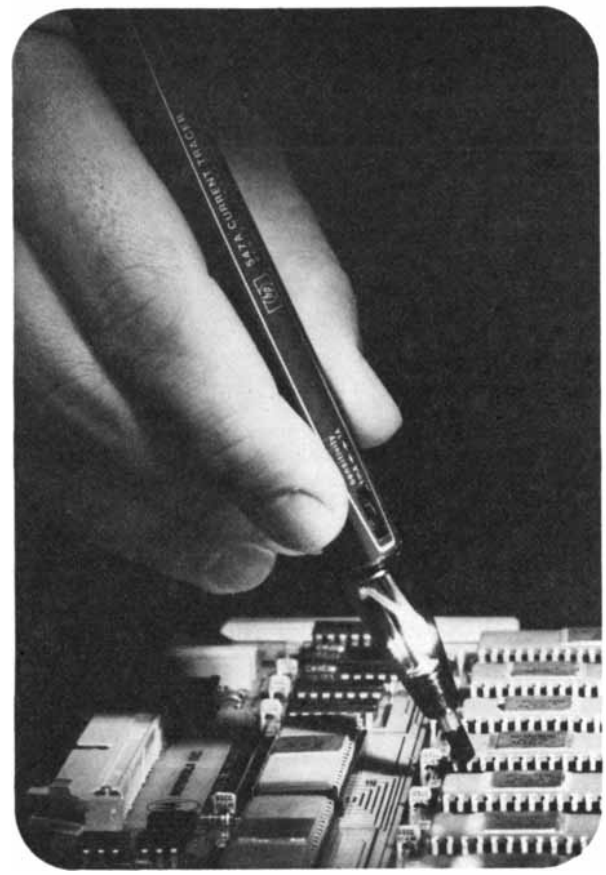
This little probe precisely locates low-impedance faults in digital circuits by tracing current pulses to sources or sinks—thus pinpointing hairline solder bridges or backplane shorts, and solving tough troubleshooting problems in *wired-AND/OR* and *three-state* busses.

A few years ago, HP's IC Troubleshooters—the Logic Probe, Pulser, Clip, and Comparator—revolutionized digital troubleshooting. These voltage-based instruments have greatly simplified the task of detecting logic faults in digital circuits and isolating them to a particular node. In the majority of cases, they lead you directly to the failure: you need only replace the bad element and the circuit is back in business.

But when a node shared by several components is stuck in one logic state, it is often impossible to identify the faulty component using only voltage-based test instruments. In such cases, the usual remedy is to replace all the IC's that share the faulty node—a costly procedure that often results in a damaged PC board.

And, it turns out, one should *not* give up testing at this point—at least not if one owns a Current Tracer. Even though a node is stuck in one logic (voltage) state, the driver may still be delivering current pulses to the node as it tries to change state. What the troubleshooter needs to know is where the current is flowing.

And that's precisely what the HP 547A Current Tracer does. Able to operate on all logic families, it provides the "missing link" of *current* information which, when combined with the *voltage* information from HP's other IC Troubleshooters, allows one to trace the fault directly to the failed gate or circuit element without cutting traces or unsoldering components.



Deceptively simple in appearance, the Current Tracer is actually a sophisticated test instrument. Besides a precision inductive pickup that is sensitive only to AC current with fast transitions (200 ns), it also incorporates a wide-band 200 MHz amplifier with adjustable sensitivity of 1 mA to 1 A, and a simple one-lamp readout that unambiguously displays relative current levels along the circuit. It costs $350*.

For more information on these products, write to us, Hewlett-Packard, 1505 Page Mill Road, Palo Alto, California 94304.

*Domestic USA prices only.

HEWLETT hp PACKARD

Sales and service from 172 offices in 65 countries.

For assistance call: Washington (301) 948-6370, Chicago (312) 677-0400, Atlanta (404) 434-4000, Los Angeles (213) 877-1282, Toronto (416) 678-9430

Mail to: Hewlett-Packard, 1505 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) HP-27 Scientific/Plus
( ) HP 547A Current Tracer

Name _____
Company _____
Address _____
City _____ State _____ Zip _____

00645

Center-Pivot Irrigation

From the air over the U.S. high plains one can now see thousands of large circular fields. They are irrigated by a system in which water is applied from a central well by a novel rotating machine

by William E. Splinter

Traveling in orbit around the earth at an altitude of some 270 miles, the Skylab astronauts rotated their spacecraft to establish a bearing on one of their principal check points: a cluster of several hundred round green spots, each half a mile in diameter, arrayed in an orderly pattern on the earth's surface below them. What they were viewing was a dense concentration of circular irrigated fields in north-central Nebraska, a pattern easily identified from space. Passengers on commercial jet airliners increasingly notice the same sight over many other areas of the continental U.S., including eastern Colorado, central Minnesota, the Texas Panhandle, the Pacific Northwest and northern Florida. Now the proliferating green circles can be seen even in the middle of the Sahara. What is being observed is perhaps the most significant mechanical innovation in agriculture since the replacement of draft animals by the tractor.

These circular green fields, most often found in arid or semiarid country, are being watered by the world's first successful irrigation machine. Invented by Frank Zybach of Columbus, Neb., some 25 years ago, the machine typically consists of a series of water sprinklers of the impact type mounted on a six-inch pipe that is in turn supported by a row of seven or more mobile towers. Water is pumped into the pipe from a source at the center of the field, and the towers carry the system around the center pivot. The rate at which the towers and the pipe advance is set by the speed of the outermost tower; an alignment device detects any laggards and moves each tower to line up with the one beyond it. Thus an advance by the outermost tower sets off a chain reaction of advances beginning with the second tower from the end and progressing toward the center of the circle. Zybach developed his machine while farming in Colorado near the town of Strasburg, east of Denver. After many trials and adjustments the system was made to work, and a U.S. patent was granted in 1952.

The first commercial center-pivot irrigation systems, introduced a year later, relied on a mechanical device called a Trojan bar acting on the tower wheels to ratchet the tower ahead. The mechanism was actuated by a linkage to a cylinder and a piston driven by the pressure of the water in the supply line; the water was bled from the supply line for the motive power of each tower. Many units on the market today still operate on this system.

A number of variations on the original design have also been tried. One system, mounted on slides rather than wheels, is advanced by a movable foot actuated by cables running the length of the system and connected to a hydraulic cylinder. Most of the systems, however, are mounted on large steel or rubber wheels and are driven by electric or hydraulic motors at each tower. Many of the systems are reversible so that the farmer can back them out of mudholes or reposition them when necessary. The delivery pipe is usually supported about eight feet above the ground. The sprinklers are spaced so that the water is applied at an increasing rate with distance outward along the pipe. Since the greatest amount of water must be delivered at the circumference of the circle, the losses due to hydraulic friction and the increasing rate of application along the pipe make the design of the center-pivot irrigation system an interesting engineering challenge, best handled with the aid of the computer.

How has the development of this novel irrigation system affected agriculture? In the first place, the center-pivot system enables the farmer to irrigate large tracts of land automatically. Once the system is set for a given application, it advances continuously in a circle, applying irrigation water without need for further attention other than monitoring and repairing occasional breakdowns.

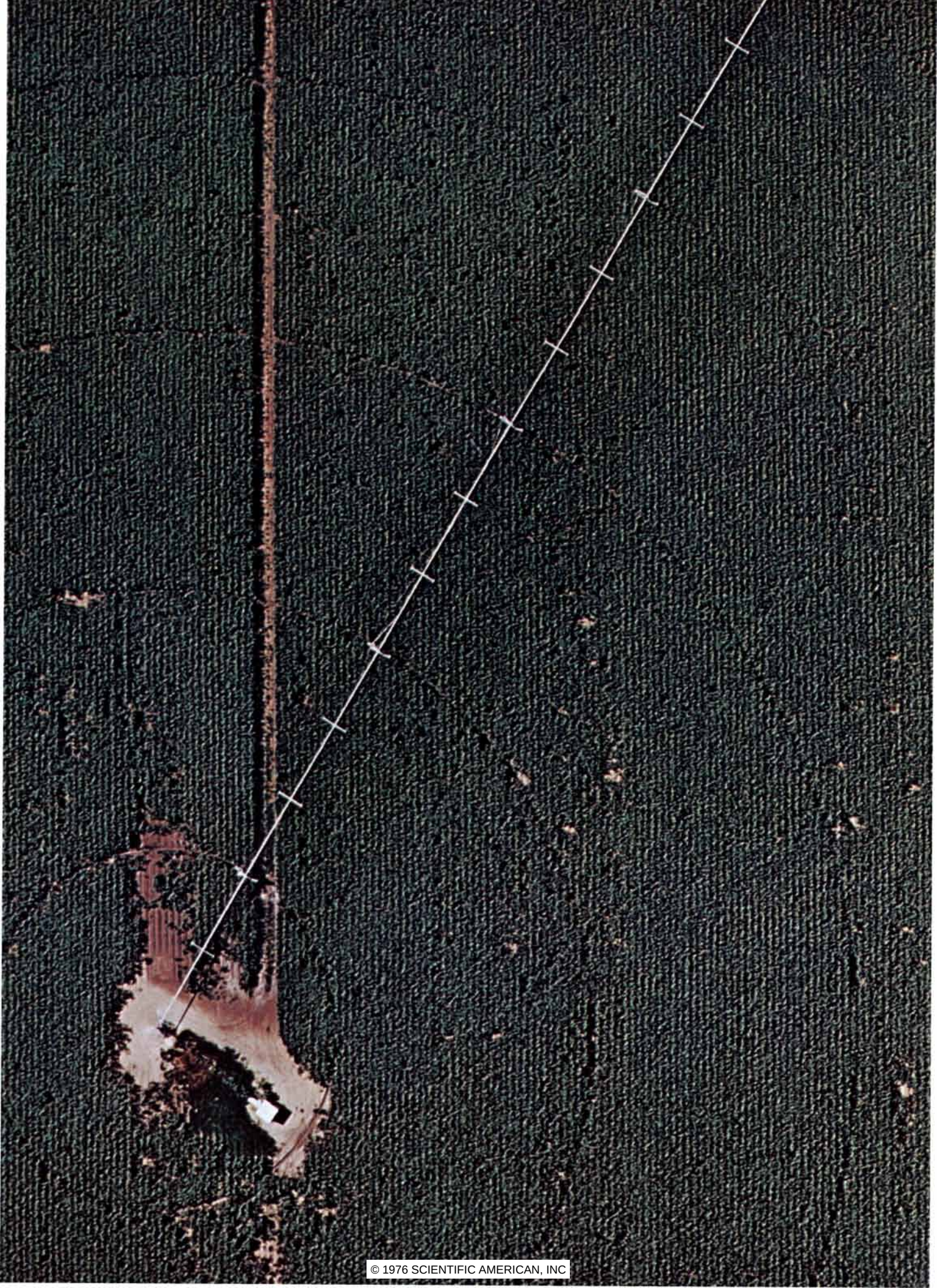
In the past irrigation has been synonymous with intensive human labor. Since prehistoric times men have had to divert streams to irrigate their crops. Such diversion involved the building of dams, canals and lateral ditches. Even in modern times conventional irrigation takes monumental amounts of time and energy. Once the system was installed the farmer was faced with the continual drudgery of channeling the water to the fields, shoveling rows open, closing off rows that no longer needed water and continually maintaining the system against erosion or rodent damage.

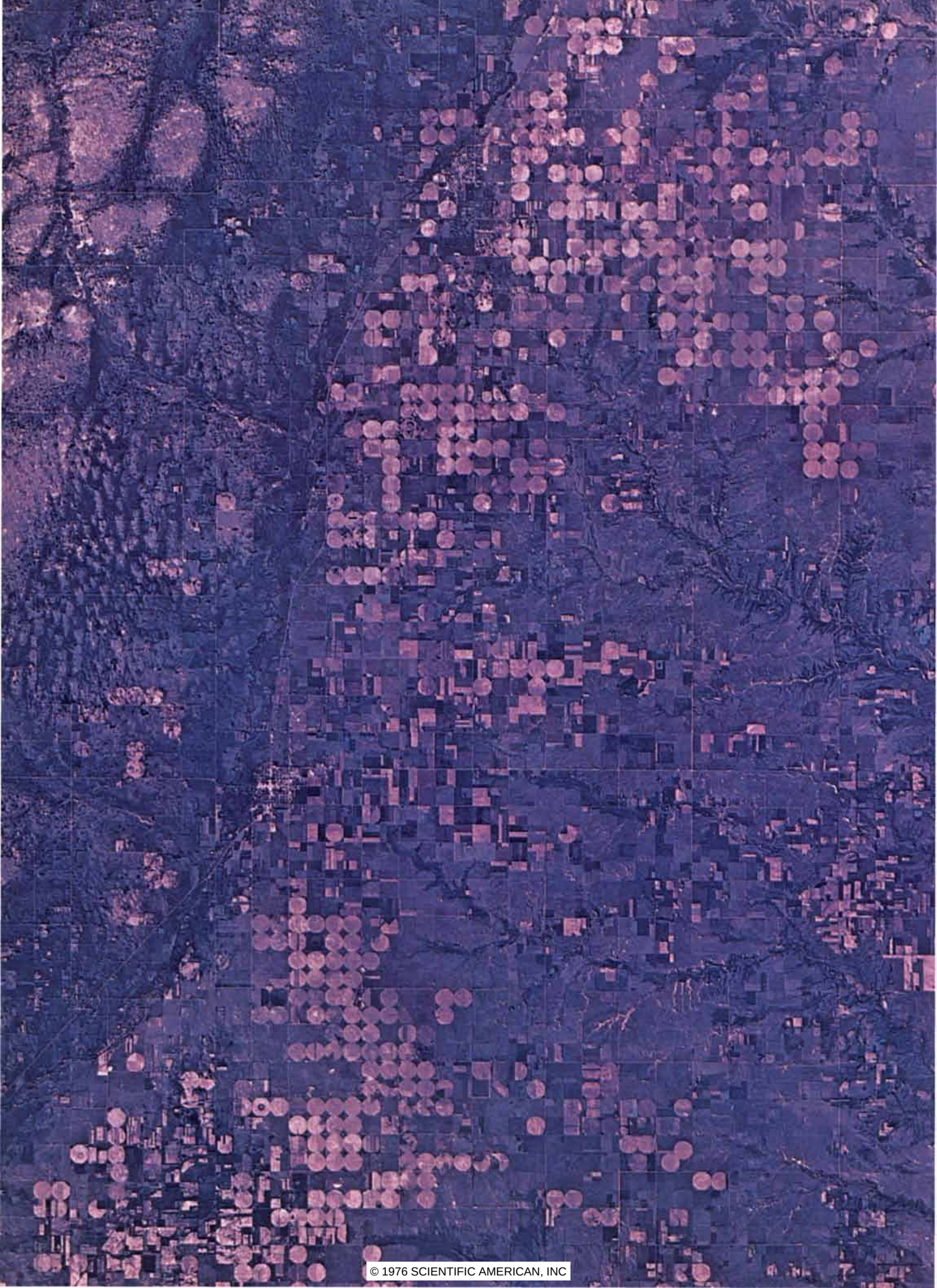
The development of siphon tubes in the 1940's reduced some of the shoveling, but setting the tubes was still a backbreaking job. More recently the development of "gated" pipe (that is, pipe with multiple, controllable outlets along its length) has further reduced the amount of labor involved in applying water to individual rows, but the bulky pipes must still be moved and the individual gates adjusted for each irrigation.

The technology of sprinkler irrigation has evolved primarily since World War II, with the advent of rotating sprinkler nozzles and the availability of comparatively cheap aluminum pipe. In the first systems lengths of aluminum pipe were coupled together and the sprinkler nozzles were held above the crop by risers. Water was supplied at a pressure ranging from 40 to 80 pounds per square inch. After a given application of water the entire system was dismantled and moved by hand to the next location, where it was reassembled.

Some savings in labor were achieved by mounting the pipe on large-diameter wheels, which were advanced every 12 to 24 hours across a field. Water was supplied by a main line, and the pipe had to be disconnected and reconnected whenever it was moved. The low clearance of the pipe limited the use of these "side roll" devices to alfalfa and other short crops. Stronger couplings were then developed for the standard irrigation-pipe systems, so that quarter-

AERIAL CLOSE-UP on the opposite page shows a center-pivot irrigation system in northeastern Colorado photographed from an altitude of about 1,000 feet. The concentric circles are made by the wheels of the mobile towers that carry the sprinkler pipe around the center pivot. The small building at lower left is the pumping station. The crop is corn. The field shown here is in the vicinity of the one seen from a higher altitude on the cover.



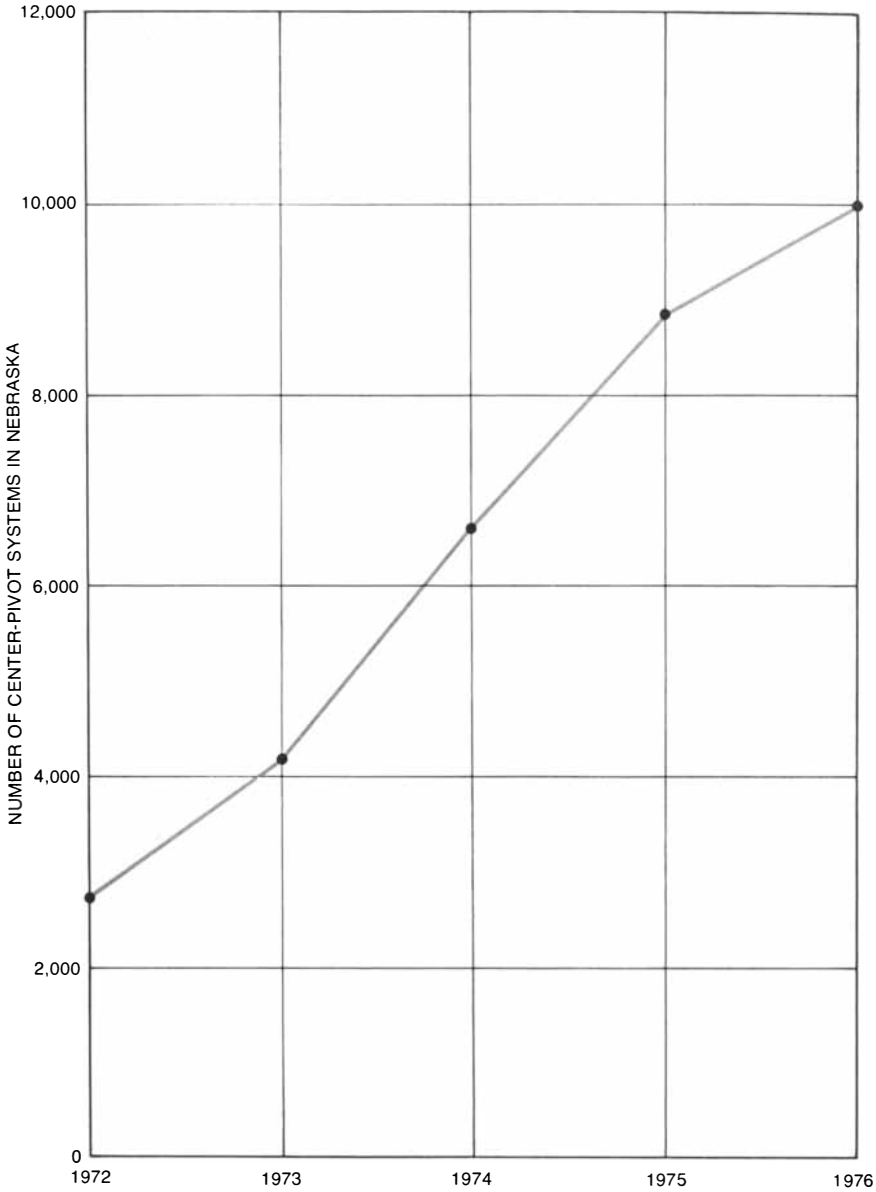


mile lengths could be towed back and forth with a tractor. This system, called skid-tow, reduced labor, but there were still problems in pulling the pipe back and forth through a field of corn or sorghum.

Now everything is automatic. The center-pivot system is designed so that the farmer can apply small amounts of water every few days. A minimum time for a center-pivot system to make a circular traverse is about 12 hours. The farmer can apply a larger or smaller amount of water by operating the outer tower at a lower or higher rate of advance. Most center-pivot regimes call for one traverse every three or four days, with the application of about an inch of water for each revolution.

Most of the systems are designed to fit the conventional unit of agricultural land in the U.S.: the quarter section, or 160 acres. The circular pattern leaves out the corners of the field, so that only 133 of the 160 acres are irrigated. (Close-packing of the circles in a hexagonal array would of course increase the ratio of the irrigated land to the unirrigated.) The land in the corners is commonly utilized for pasture, farmsteads, feedlots, grain storage, trees or dry-land crops. Larger units have been designed to irrigate 220 acres, and several capable of irrigating 530 acres have been installed. These larger units carry half a mile of eight-inch pipe on 20 towers and irrigate a circular field with a diameter of a mile. One manufacturer has developed a special hinged outer tower that follows a buried wire electronically and swings out to irrigate the corners left dry by the circular design.

In the late 1960's fully automatic sprinkler systems with buried water-supply lines and programmed, electrically operated valves were installed in eastern Colorado. Although such systems are versatile, they are limited to high-value crops because of the high cost of their installation. Recent research has resulted in the development of valves that make feasible the automation of surface irrigation with gated pipe. This new system will be less energy-demanding than sprinkler systems, and some commercial development is starting. The center-pivot system, however, is by far the predominant method of automatically irrigating field crops available today.



RAPID PROLIFERATION in the number of center-pivot irrigation systems installed in Nebraska in the past few years, plotted in this graph, is monitored by the author and his colleagues at University of Nebraska with the aid of a device that electronically scans satellite imagery.

The ability of center-pivot systems to apply water lightly and frequently opened up the second revolutionary feature of this new technology. Many agricultural areas are limited in productivity because the soils are coarse-textured or sandy. Such soils hold little water: less than an inch per foot of soil depth compared with two or more inches per foot of soil depth for fine-textured or loamy soils. As a result coarse or sandy soils are characteristically dry and can usually serve only for rangeland or at most for marginal farming.

The light, frequent application of water from a center-pivot sprinkler system replenishes the moisture in the root zone sufficiently to allow intensive cropping on these soils. In research studies at North Platte, Neb., my colleagues in the departments of agricultural engineering, agronomy and animal science at the University of Nebraska found that pasture irrigated with a center-pivot system produced between 700 and 900 pounds of live beef per acre per year on cool-season grasses. This compares to a normal productivity of 27 pounds per acre per year for open range in the Nebraska sandhills.

Farmers and agricultural researchers have found that corn, sorghum, sugar beets, wheat, potatoes and other crops can be raised on this sandy soil under center-pivot irrigation. The yields are comparable to those on silt or clay. The main crop raised under center-pivot irrigation in Nebraska is corn.

Coarse or sandy soils not only can hold little moisture but also have scant capacity for soil nutrients. The third revolutionary feature of the center-pivot system is that by injecting fertilizers into the water-supply line one can administer nutrients only as

DENSE CLUSTER of center-pivot irrigation systems appears in the false-color terrain photograph on the opposite page, made by the Skylab astronauts on June 9, 1973, from an altitude of about 270 miles over north-central Nebraska. Most of the circular irrigated fields in the photograph are planted with corn. This semiarid region, which lies between Niobrara and Elkhorn rivers near O'Neill, Neb., was formerly used for pasture, hay meadows and marginal farming.

they are needed by the growing crop. Our studies have shown that fertilizers applied to sandy soils leach readily through the soil and pass below the root zone. Once they are below the root zone they are carried down to the water in the aquifer underlying the soil. Applying the fertilizer only as the crop utilizes it greatly reduces the hazard of contamination of underground waters by nitrates and other mobile nutrients. Obviously such an approach will enable the farmer to apply only what is needed, thereby saving him many thousands of dollars in production costs.

Our research extended this feature to include the application of herbicides. Weed-control chemicals can be applied through a center-pivot system operated at a high rate of advance, applying about a third of an inch of water to control broadleaf weeds. In effect, the system operates successfully as a giant field sprayer.

The first center-pivot irrigation systems were designed to adjust to the terrain by allowing the pipe to flex between the support towers. The development of flexible couplings at each support tower now makes it possible for the center-pivot system to adjust to quite rolling terrain. This greatly reduces the cost of the preparation of land for irrigation, since the cost of grading alone can equal the cost of all other inputs to the development of a surface-irrigation system. Grades of up to 30 percent can be accommodated by the towers, although it is recommended that the grades not exceed 10 percent because of surface erosion and the chance of gullying.

It is because of these features—automatic operation, control of application rate and frequency, accommodation to rolling terrain and to coarse or sandy soils and precise application of fertilizers and herbicides—that center-pivot systems are being rapidly adopted throughout the U.S. and around the world. A major manufacturing industry has developed to supply the expanding demand for these systems. From satellite observations we have determined that more than 9,000 center-pivot systems are in operation in the state of Nebraska alone. Many more have been installed elsewhere in the U.S. In addition units are being installed in Libya, Australia, Hungary, France and the Middle East.

Like almost all other new technologies, center-pivot irrigation has its problems. First, there must be a source of water of a quality suitable for application to a growing crop. Many of the major sandy areas do have water below the surface that has accumulated over thousands of years. The sandhills of Nebraska, for example, have an immense pool of water lying a short distance below the surface. These waters could not be used for irrigation with conventional systems because of the rapid infiltration of water into the soil and the soil's poor water-holding capacity. The automatically applied light irrigation of the center-pivot system suits the development of these areas precisely.

In some areas water can be pumped from rivers. Several large projects in Oregon and Washington pump water from the Columbia River for irrigation through center-pivot systems. By far the majority of center-pivot irrigation systems, however, are supplied by deep wells into the underground aquifer.

In many areas, for example in the Texas Panhandle, there is a risk that the aquifer will be mined to the point of exhaustion. States such as Colorado have severely restricted the drilling of new wells. A research effort in Texas and Nebraska is aimed at developing a technology to recharge the aquifer with surface water that is available in surplus intermittently from rainfall or melting snow. In Nebraska off-season stream flows are being evaluated for their recharge potential. If an economic technology could be developed to replenish the aquifer with water of acceptable quality, the hazard of depletion could be greatly reduced.

In addition to the problem of the depletion of underground water reservoirs there is the fact that center-pivot systems require a good deal of energy. In our studies we have found that 43 percent of the energy devoted to agriculture in Nebraska goes to pumping water for irrigation. In Nebraska a typical center-pivot system consumes about 50 gallons of diesel fuel per acre per year in applying roughly 22 inches of water. That is 10 times the fuel needed to till, plant, cultivate and harvest a crop such as corn.

The average center-pivot system in Nebraska draws water from a depth of 180 feet at a rate of 900 gallons per hour. The pressure required to move the water along the pipe and supply the impact nozzles ranges from 50 to 90 pounds per square inch. While one center-pivot system is operating it consumes water enough for a city of 10,000 people. Most systems in the middle latitudes, however, operate for only two months of the year; hence the total amount of water pulled from the aquifer each year would be only enough to supply a town of 1,000 people. These are nonetheless large amounts of water and energy.

At present most of the center-pivot systems in Nebraska are powered by diesel engines; some are driven by natural-gas-powered engines and others by electric motors. Conversion from diesel fuel or natural gas to electric power is not immediately feasible because center-pivot systems place a heavy load on electric-generating capacity for a fairly short period during the summer, a load that is not matched over the rest of the year. Thus generating capacity, transmission-line capacity and substation costs limit the development of electrically powered pumping plants.

In Nebraska we have worked to decrease the electrical peak load by scheduling the operation of center-pivot systems so that the electrically driven systems are turned off during periods of peak use by other consumers. In midafternoon, when air conditioners are operating at their maximum capacity, and early in the evening, when housewives begin preparing dinner, the nonirrigation peak is particularly high. By monitoring the soil moisture in an area closely one can advise farmers when to apply water and how much should be applied. In one area of 3,300 acres in central Nebraska the scheduling of electrically driven center-pivot systems to avoid heavy-usage times reduced the electrical peak by 667 kilowatts.

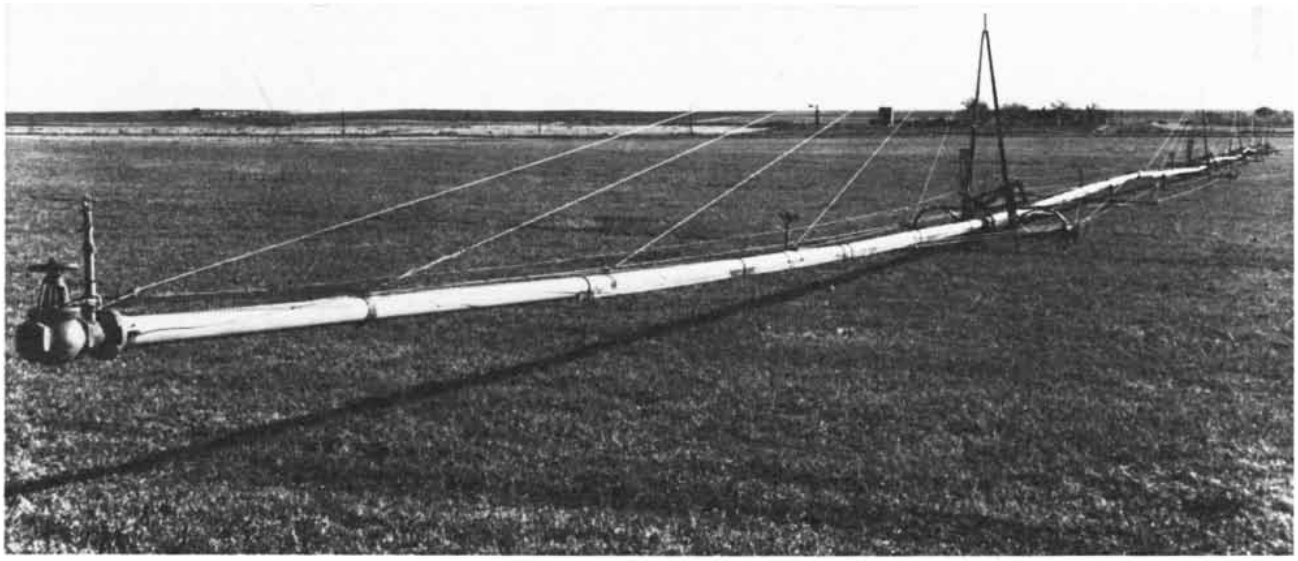
There are other problems in switching from diesel fuel or natural gas to electric power. Hydroelectric power has been developed about as much as it can be in the U.S. As supplies of oil decrease, coal and nuclear fuels will have to take up the slack, at least between now and the end of the century. The expansion of coal-fired and nuclear-powered generating plants, however, is encountering both social objections and technical problems.

There is some potential for utilizing part of the product of center-pivot agriculture to power center-pivot systems, either through direct combustion or conversion to methane. An average field of corn produces 3.5 tons of stover (stalks, leaves and cobs) per acre. One or two tons should be left on the field to protect the soil from erosion by wind and water, depending on the soil type and the slope. Since stover has a fuel value of about 6,000 B.t.u.'s per pound, the remaining material could be used for fuel. Theoretically there is enough energy in one ton of stover to drive a center-pivot pumping plant, provided that an economic method of energy conversion can be developed.

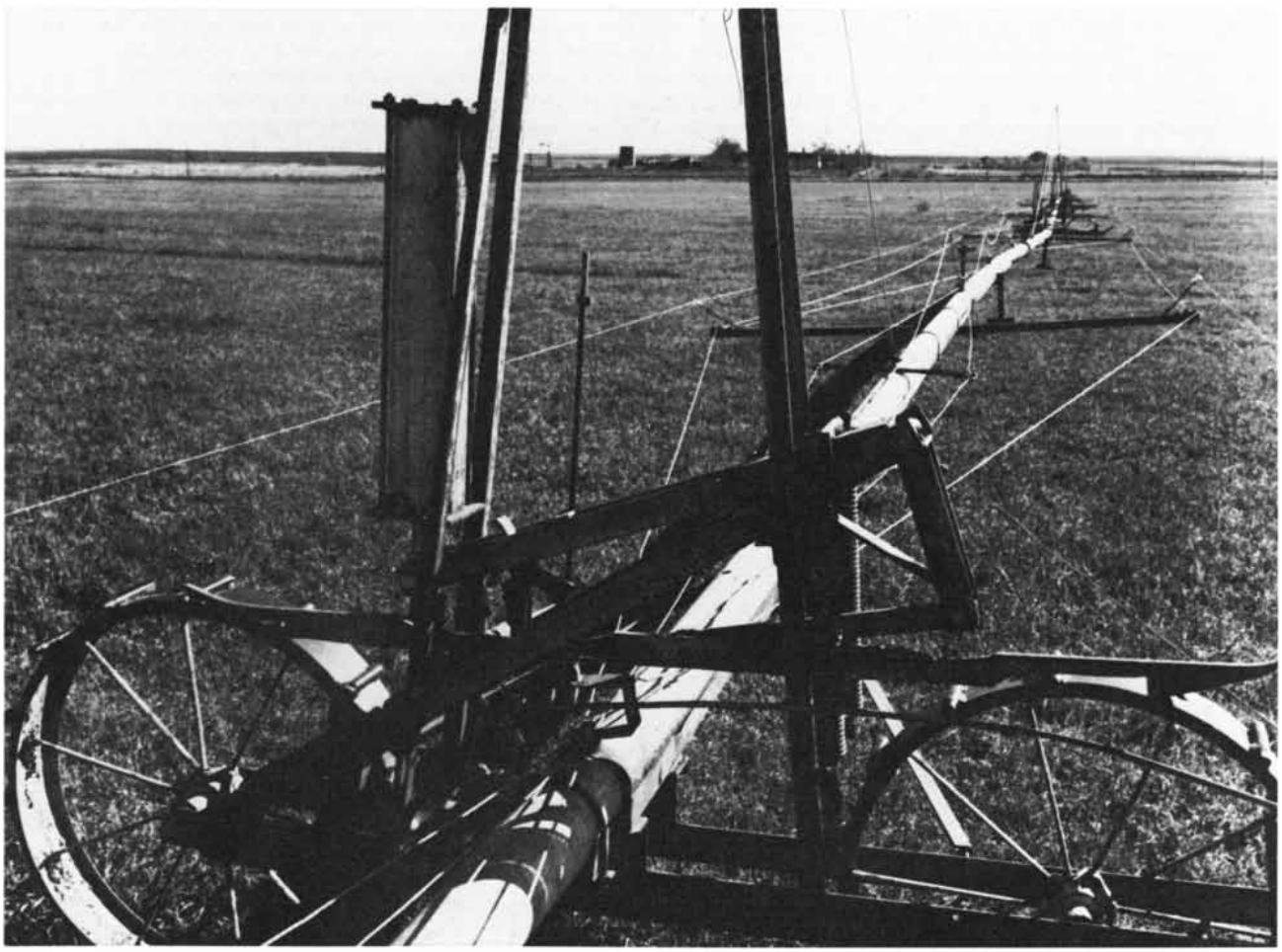
Everyone who has lived on or traveled across the Western plains knows that they are rich in wind. Wind energy has been harnessed to pump water for centuries in Europe and the Middle East. Wind has served to pump water for livestock throughout the Western plains since late in the 19th century. About 1900 a considerable number of designs for wind-powered machines were developed and tested throughout Nebraska. The introduction of the windcharger in the 1930's brought to remote farms and ranches enough electricity to power radios and a few light bulbs and appliances. Therefore it is often asked: Why not drive irrigation systems with wind power?

In 1947 members of our department mounted a four-blade windcharger with a diameter of 12 feet on a 101-foot tower in order to determine what power could be extracted from wind energy over an extended period. The machine was monitored for 27 months. The average output over that interval was four kilowatt-hours per day. Assuming that the output could be stored, that amount of electricity would operate a one-horsepower electric motor at full load for about four hours per day.

If one wanted to power a 40-horsepower



ORIGINAL CENTER-PIVOT MACHINE, invented by Frank Zybach some 25 years ago, introduced two key innovations that led to the successful development of a completely automatic irrigation system. In Zybach's scheme the pipe supplying water to conventional sprinkler nozzles is supported by a series of A-frame towers. The system is propelled by means of water bled from the supply line (*see illustration below*). The second key idea in the scheme was the development of a mechanical means to keep the system in alignment by detecting any deflection of the pipe at an intermediate tower and causing the next-inner tower to advance. As the outermost tower proceeds in a circular path, each of the inner towers moves intermittently in response to pipe deflection caused by movement of next-outer tower.



CLOSE-UP of one of the mobile towers of the original center-pivot machine shows the parts of the system with greater clarity. Water taken under pressure from the supply line powers a piston, which ratchets the tower ahead by means of a mechanical device called a Trojan bar that engages lugs on both support wheels. The rate of advance is set by the flow of water into the piston at outermost tower.

electric motor for irrigation (which would be a fairly small pumping plant), one could scale up from the windcharger experiment. One would need 240 windchargers of the type we studied or one unit with a blade 180 feet in diameter. The level of technology in the design of windcharger blades in the 1940's allowed the extraction of 42 percent of the wind's energy. The theoretical maximum for extracting energy from the wind is 59 percent. Hence there is some room for improvement, and one such improved blade design is being tested at Oklahoma State University. So far, however, the scale of any device planned for powering irrigation systems through wind power is so large that it is a serious deterrent to development.

What is the economic potential of center-pivot irrigation systems? Today development costs are about $350 per acre for the well, the power plant, the pump and the center-pivot hardware. The systems have generally been installed on unimproved, marginal lands, so that land investments have been comparatively low. Production costs would be equal to those of other irrigation systems plus the cost of the energy needed to pump the water, which may run from $10 to $20 per acre per year. Yields of 140 bushels per acre of corn can be reasonably expected, so that at today's prices a gross return of $460 per acre could be realized. Studies by economists indicate that for every acre placed under irrigation the gross income to the state of Nebraska is $500 per acre per year, taking into account the activities of all economic sectors. The rapid proliferation of center-pivot systems clearly points to their financial success.

With the commercial acceptance of this new technology for irrigation there are already attempts to refine the technology of irrigation even further. Lacking the means to monitor soil moisture and the anticipated demands of his crop, the farmer must, as a management decision, overirrigate. The economic consequences of insufficient moisture far outweigh the cost of the extra water. Under the pressure of impending water and energy shortages a number of studies are being conducted to develop more efficient strategies for husbanding water in the production of agricultural crops.

Our studies have determined that scheduling to take advantage of the water stored in the soil during the fall, winter and early spring and the holding back of irrigation to take account of occasional rains during the growing season make it possible to conserve anywhere from a third to a half of the water needed for sugar beets or corn. Engineers with the Department of Agriculture's Agricultural Research Service in Idaho and with the U.S. Bureau of Reclamation and the Agricultural Research Service in Colorado have developed computer programs that continuously calculate the amount of water that is lost to a field through evaporation from the soil and transpiration through the leaves of the crop. The amount for a specific crop such as alfalfa or corn is adjusted by an experimentally determined coefficient as the crop grows. Such programs make it possible to more precisely set the times of irrigation and the amounts of water required to refill the root zone. We have used this system of scheduling at our research station near Scottsbluff, Neb., to successfully predict when the next irrigation is needed and to determine which crops should be watered first.

Because of variations in soils and rainfall the soil moisture in individual fields must be monitored. That is not easy, but this year the Bureau of Reclamation plans to schedule the irrigation of some 300,000 acres by monitoring soil moisture and informing farmers of their water requirements. Scheduling for individual farms not on bureau projects is available to farmers in Nebraska and Arizona through a commercial service; well over 100,000 acres were scheduled in this way last summer.

Research workers in Texas have developed a "stress day" method for predicting irrigation demand. Studies in California and Texas indicate that the ultimate yield of a crop may be affected by an insufficiency of water more strongly at some stages of growth than at others. For example, the yield of corn is affected by an insufficiency of water more strongly at the stage of pollination than at other stages. Sorghum, on the other hand, seems not to be as susceptible in this regard as corn. On the basis of this information a computer-assisted strategy has been developed in Texas to determine the optimal scheduling for the application of water as a function of the stage of growth, the available quantity of water, the rate of evaporation and transpiration and the predicted weather probabilities. Such a strategy could be of profound economic importance to farmers. Knowing that they had a given amount of water available, they could follow the strategy to husband their supply for the optimal time of application.

In order to predict the response of corn and other crops to environmental stress more precisely, mathematical modeling methods are being developed at several universities. From data on light intensity, air temperature, day length and soil moisture one can estimate the leaf area, carbohydrate production, dry-matter weight and grain yield of corn by computer simulation. For example, we have found that corn grows about 20 percent per day in the early stages of its growth, and that this growth rate can be reduced to zero by moderate soil-moisture stresses. The plants do not wilt permanently, however, until the soil moisture drops to a much lower level. At the permanent-wilting point there is irreversible damage to the tissues of the plant.

At stresses in the moderate range the stomata, or pores, of the leaves close, blocking the further loss of water but also blocking the diffusion into the stomata of carbon dioxide, which is of course essential for photosynthesis. With the aid of contact auxanometers (devices that automatically monitor the diameter of a plant's stem) we can detect a change in growth rates within three to five minutes. Plants can be raised under precisely controlled conditions in growth cabinets, and the effects of various changes in their environment (for example changes in light intensity or temperature) can be quickly determined. This information then serves as a foundation on which mathematical models can be built.

Plants appear to have biological clocks that activate specific responses. These clocks operate on the basis of degree-days: the cumulative effect of certain temperatures multiplied by the time spent at those temperatures. Our measurements indicate, for example, that it takes about 1,000 degree-days for a corn plant to proceed from germination to forming a tassel. Utilizing that information, computer programs can now predict the date when tasseling will begin, and since this is the most vulnerable time for corn such programs can enable farmers to anticipate those times of most critical importance for irrigation.

Exact measurements of the amount of water being evaporated from the soil or transpired through the plant leaves are being conducted with lysimeters. These devices are bins containing a cubic meter of soil set in an open field. The bin is mounted on a scale with an electronic monitor capable of measuring a change in weight as small as 10 grams in a total bin weight of 5,000 kilograms.

Wind speed, solar-radiation intensity, relative humidity and temperature are also measured continuously with instruments. The data are processed by computer, enabling us to develop models with even more predictive accuracy. Our studies in Nebraska show that for the normally dry climate of the Great Plains the microclimate of a field being irrigated by a center-pivot system is quite similar to that of an oasis in a desert. Hot, dry air is carried to the field by winds, giving rise to more evaporation and transpiration than would be caused by solar radiation alone. Under these conditions as much as 1.25 centimeters of water can be lost over a 24-hour period.

Coupled with the growth models, more complete models of the system represented by the soil, the plant and the atmosphere are now under study. These models will enable us to determine the water requirements as the plant advances through the sequential stages of its growth. By programming alternative or anticipated weather sequences one would be able to evaluate the effects of various irrigation-management strategies and select the optimal system after considering energy cost, water availability and crop yield.

As I have mentioned, irrigation systems are visible from spacecraft in orbit around the earth. Studies of films made on the manned Skylab flights of 1973 and 1974 showed that the center-pivot systems could be clearly identified and that the development of the crop could be monitored from

OBLIQUE AERIAL VIEW of a quarter-section circular cornfield in Nebraska shows how in the center-pivot irrigation scheme water is applied automatically by means of a series of sprinklers of the impact type mounted on a radial pipe that is in turn supported by a slowly advancing row of mobile towers. The photograph was supplied by the Rain Bird Sprinkler Manufacturing Corporation of Glendora, Calif.

GROUND-LEVEL VIEW of a recently installed center-pivot system demonstrates its ability to accommodate to rolling terrain. The wheeled towers in this example are driven by electric power. The photograph was supplied by Valmont Industries, Inc., of Valley, Neb.

TELLTALE PATTERN of hexagonally close-packed circular fields of crops grown under center-pivot irrigation appears in this false-color composite image made by the multispectral-scanning system on board the LANDSAT earth-resources satellite. The site is near the Kufra oasis in the Libyan Desert some 550 miles southeast of Benghazi. The picture was made from an altitude of nearly 600 miles.

ANOTHER PERSPECTIVE on the complex of center-pivot systems at Kufra is provided by this oblique aerial photograph. Irrigated fields here, supplied with water from a vast underground reservoir, are somewhat larger than usual, measuring almost a mile in diameter.

space. The LANDSAT unmanned-satellite program has proved to be even more valuable, since the coverage extends over a considerable period of time.

The Skylab flights carried several sensor systems, including conventional cameras capable of making photographs in different regions of the spectrum, a special earth-terrain camera and a multispectral-scanning device. The LANDSAT satellites use the multispectral scanner exclusively. Since the LANDSAT satellites complete one revolution in a north-south orbit every 103 minutes, they are capable of observing any given location on the earth's surface once every 18 days as the earth rotates through the plane of the orbit.

Information from LANDSAT is beamed to tracking stations and stored on magnetic tape. The images are recovered by the same process in which television stills are obtained from tape recordings. LANDSAT relays information in four spectral bands: Band 4 (a visible green wavelength), serving primarily for the evaluation of atmospheric conditions; Band 5 (a longer-wavelength visible green) for vegetative growth; Band 6 (in the near infrared) for both vegetation and water, and Band 7 (a longer-wavelength near infrared) for water alone. Skylab was able to scan 13 bands in all.

The imagery for each band is in black and white, but false-color composites are most useful for interpreting the results. With this imagery our group has been able to identify center-pivot irrigation systems (as well as surface-irrigated systems) throughout Nebraska. As of last summer the center-pivot systems covered some 1.3 million acres. It is clear that the technology now exists for quickly conducting inventories of large areas in order to determine the patterns of development in irrigation systems. Such inventories will be of vital interest to government bodies in assessing the utilization of water resources and in projecting their future management.

The inventories have been further refined with the aid of flights by conventional aircraft and high-altitude RB-57 aircraft. Plant leaves that are under moisture stress have a higher temperature than leaves supplied with ample water. Precise monitoring of crop temperatures will make it possible to detect areas of inadequate irrigation. That will make possible more accurate crop forecasts, an increasingly important need of private and government planners.

The world population is expected to double in the next 35 years. Food production must be increased to meet the demand of this expanding population. Since the prime agricultural areas are for the most part already developed, the technology for increasing food production must rely on the more intensive and more efficient utilization of present resources. The center-pivot irrigation system, offering a level of control that cannot be attained in most other irrigation systems, will prove to be a major technological tool in this fateful effort.

Historical Supernovas

Early records indicate that seven of these huge stellar explosions were seen over a period of 1,500 years. At the recorded positions of the new stars remnants of the explosions can be observed today

by F. Richard Stephenson and David H. Clark

The vast majority of stars in a typical galaxy such as our own are extremely stable, emitting a remarkably steady output of radiation for millions of years. Occasionally, however, a star in an advanced stage of evolution will spontaneously explode, and for a few months it will be several hundred million times intrinsically more luminous than the sun. Such a star is a supernova, and at the time of its greatest brilliance it may emit as much energy as all the other stars in its galaxy combined.

Supernovas are enormously interesting, not only because they are the most spectacular of stellar events but also because the remnants and ejecta of such explosions are among the most unusual objects known to astrophysics today. It is believed that supernova explosions give rise to pulsars, black holes, high-energy cosmic rays, the heavy elements, certain expanding nebulas, extended sources of radio emission, high-velocity "runaway" stars hurtling through our galaxy at speeds approaching a million miles per hour, possibly gravitational radiation and probably most of the X-ray sources in the galaxy.

We know from observing supernovas in other galaxies that such catastrophic explosions are relatively frequent: about one per century in a galaxy such as our own. Supernova explosions tend, however, to occur near the central plane of a galaxy, and since the central plane of our own galaxy is filled with obscuring dust, few supernovas have been observed over the span of recorded history. The last supernova to be observed in our galaxy was in 1604, five years before Galileo trained his telescope on the heavens. Since we lack telescopic observations of the supernovas that have exploded within the galaxy, it is important that the historical records we do possess of those events be analyzed with care.

What are the historical records? What kind of information do we need? What has been learned from analyses of the records? How does the information gleaned from those records supplement present-day observations of the gaseous remnants of some of the supernovas?

Ideally what one would like to find in the historical records are accounts of new stars that describe in detail the position of the object, its changes in brightness, its color and how long it remained visible. In practice such completeness is seldom realized. It is nonetheless well to take full advantage of whatever information can be obtained from the early observations.

First one needs to be at least reasonably confident that the star in question was a supernova and not simply an ordinary nova. A nova resembles a supernova in many respects, except of course in scale. A typical nova outburst, in which a star throws off a very small fraction of its mass, lasts for weeks rather than months, and the energy released is less than a ten-thousandth of that released in a supernova explosion. The expanding shell of gas dissipates in a few decades, leaving a small, high-temperature, low-luminosity star as the only remnant. Novas are much commoner than supernovas, and if a nova is close to us in the galaxy, it can be for a short time one of the brightest stars in the sky. There have been six particularly bright novas in the 20th century alone; they appeared in 1901, 1918, 1925, 1934, 1942 and last year. All of them temporarily rivaled first-magnitude stars, and for a few days the nova of 1918 was almost as bright as Sirius, the brightest star in the sky.

If we concentrate on those stars in the early records that remained visible for a considerable period, say at least six months, we can be fairly confident that the record is one of a supernova rather than of a nova. What is the nature of the records? The two most recently recorded supernovas in our galaxy are comparatively well documented. The supernova of 1604 was carefully observed and described by Johannes Kepler and David Fabricius, and the supernova of 1572 was the subject of a lengthy memoir by Tycho Brahe. Both stars, particularly Kepler's supernova of 1604, were also carefully observed in the Far East.

For earlier supernova observations the kinds of historical sources that might come readily to mind are the Greek and Latin classics and the late Babylonian astronomical texts. Those sources turn out to be unproductive. There are no references to supernovas in the Greek and Latin classics. According to Pliny, the appearance of a new star in 134 B.C. inspired Hipparchus to compile a star catalogue, the earliest known, which gave the coordinates of more than 1,000 stars. The third-century historian Justin, however, makes it clear that the new "star" Hipparchus saw was actually a comet, a fact that is confirmed by Chinese annals for that year. The late Babylonian astronomical texts, which span the period from 700 to 50 B.C., have been analyzed by Abraham J. Sachs of Brown University. Although the texts contain a large number of records of eclipses of the sun, eclipses of the moon, occultations of stars by the moon and references to comets, they contain no references to novas or supernovas.

As far as we know the only useful records of historical supernovas are those that were kept in medieval Europe, in the Arab lands and in the Far East. John C. Brandt and his colleagues at the Goddard Space Flight Center of the National Aeronautics and Space Administration have suggested that certain cave paintings made by Indians of the southwestern U.S. represent observations of the supernova of A.D. 1054 (the supernova that gave rise to the well-known Crab Nebula in Taurus). Although the hypothesis is attractive, it has been opposed by some students of these Indian cultures, and even if it is correct, the pictures tell us almost nothing about the star.

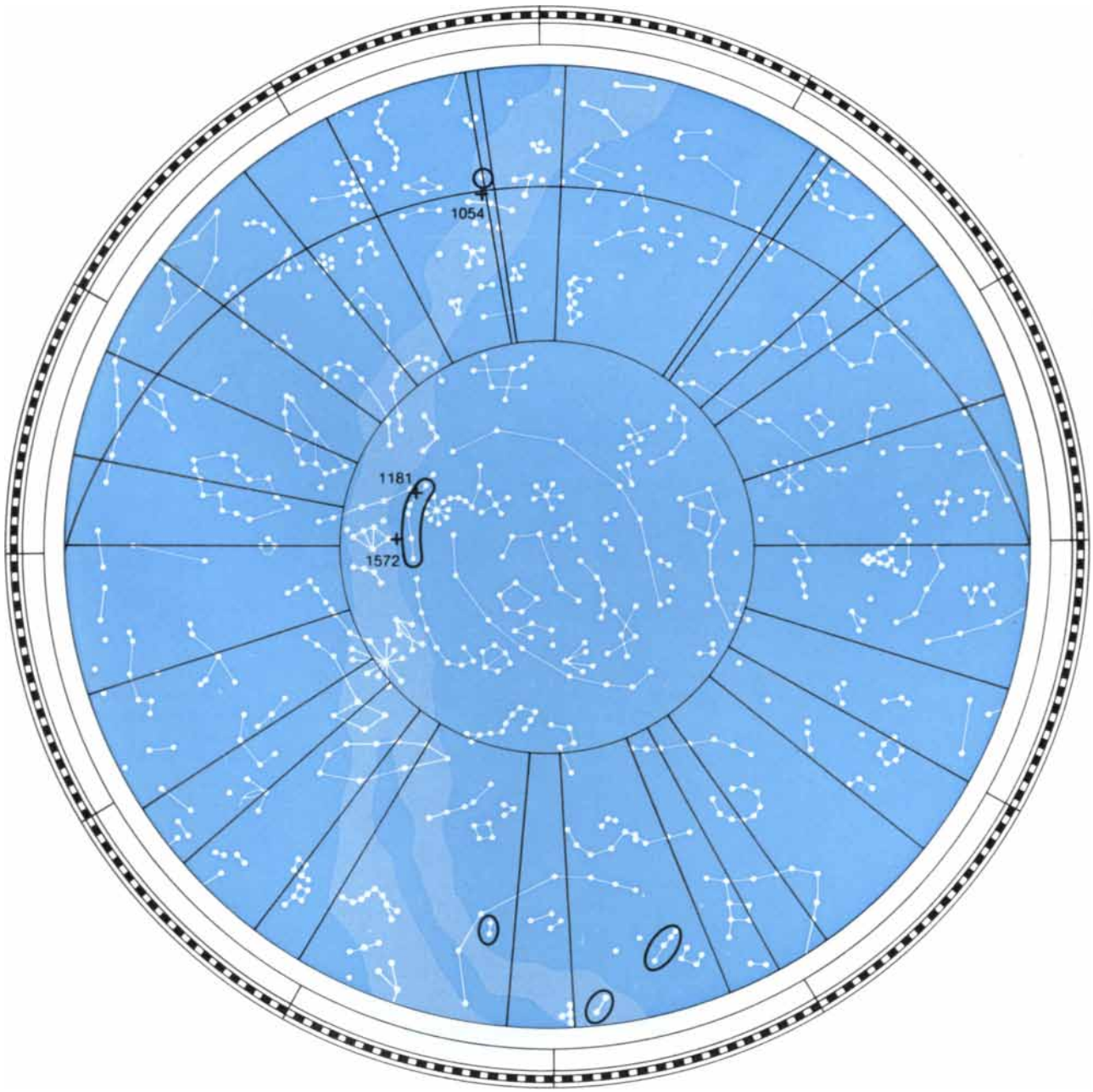
After A.D. 1000 there was a flowering of literary activity in the numerous monasteries scattered throughout Europe. Many of them began to keep chronicles of local events. Not surprisingly the chronicles included occasional references to natural phenomena such as earthquakes, eclipses and comets. The reports are often detailed and make fascinating reading. Many of the earlier chronicles were published a century ago, in two compilations titled *Monumenta Germaniae Historica* and *Rerum Italicarum Scriptores,* which are well known to students of medieval history. Many of the chronicles written after about 1200, however, are inaccessible to anyone other than the scholars who are familiar with them.

Although the medieval European chronicles might thus appear to be a promising source of information about supernovas, there are only two references to a supernova

in the better-known works. Both refer to a supernova in 1006. An intriguing later work is the continuation of the chronicle of Cosmas of Prague by a canon of the cathedral of Vyšehrad in Prague. The canon, whose work covers the period from 1126 to 1142, obviously had a keen interest in astronomy. Writing as an eyewitness, he recorded observations of solar and lunar eclipses, auroras, meteors and parhelions ("sundogs," or bright spots appearing on opposite sides of the sun). It is evident, however, that he had only a hazy knowledge of planetary motions and that he could not even identify the planet Venus with certainty. For example, in 1131 he reported seeing two strange stars, but he gave their locations only vaguely. "I do not believe," he wrote, "that there are men who know which among them was called Lucifer." (Lucifer was another name for Venus.) It seems probable that both objects were Venus.

A partial explanation for the lack of reports of new stars in the medieval European chronicles is that their authors may have had no interest in the stars or knowledge about them. Before one criticizes these authors too severely, however, one might ask how many educated people today who have no particular interest in astronomy can identify Venus. It seems likely that a solitary new star, unless it was extremely brilliant, would pass unnoticed by everyone but the most persistent sky watchers.

There is a less obvious explanation of why only occasional references to supernovas are found in the chronicles of medieval Europe. There is evidence, mainly

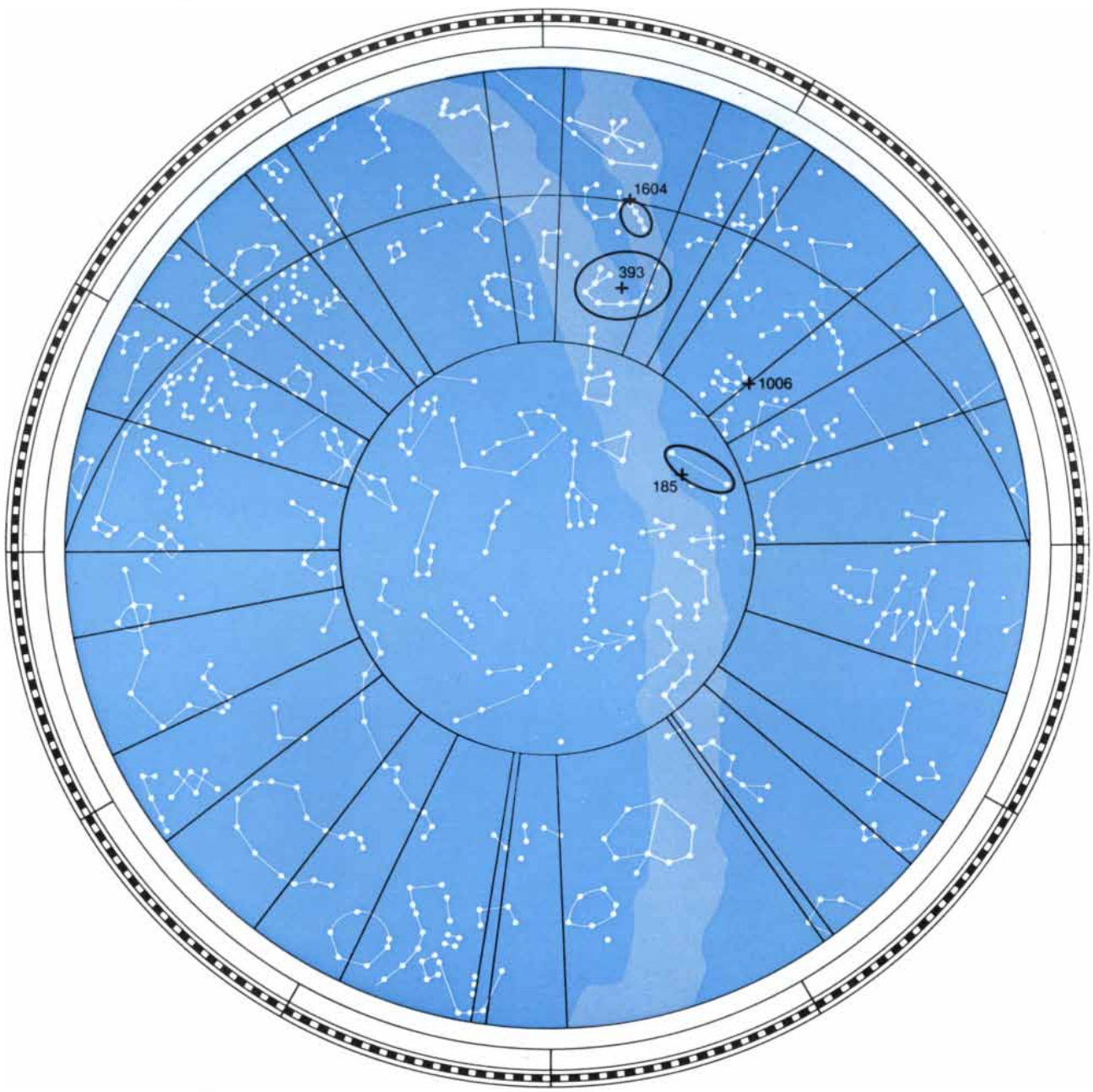


CHINESE STAR CHART, copied by Jesuits in Peking in the 18th century, shows the small Chinese constellations, or asterisms, in the northern celestial hemisphere. The black crosses give the position of three of the seven "guest stars" observed by Chinese court astronomers. Since the Chinese records indicate that these bright new stars persisted for more than six months, they were almost certainly supernovas rather than ordinary novas, which usually last for considerably less than six months. Number next to each cross is the year the new star was observed. Five asterisms with a line around them are some of those mentioned in text of this article. Counterclockwise from top they are *T'ien-kuan* (the star now designated Zeta Tauri), *Ch'uan-shê* (the Guest Houses), *Tsung-hsing* (the Ancestor Star), *Tsung-chêng* (the Supervisors of the Nobility) and *Huan-chê* (the Court Eunuchs). Light colored area running across the chart is the Milky Way.

from chronicles of the Far East, that two bright supernovas were visible in the 11th century. One was the new star that appeared in Taurus in 1054, which was far enough north to have been readily visible to European observers. The object does not seem to have been much brighter than Venus, however, and it may well have failed to attract attention.

The same cannot possibly be said of the new star of 1006, which at its maximum brilliance appears to have been as bright as the moon. This supernova appeared, however, in the southern constellation Lupus. Nevertheless, it was so bright that in spite of its southerly position it should have been prominent to observers south of the latitude 48 degrees north, so prominent that even a casual observer could hardly have failed to be impressed by it. Yet to our knowledge the star is recorded only in the chronicles of the monasteries of St. Gallen in Switzerland (at 47 degrees north latitude) and of Benevento in Italy (at 41 degrees north latitude). It seems that the Aristotelian concept of a perfect, changeless celestial vault was firmly rooted throughout Christendom until the Renaissance. The noted historian of science George Sarton concluded: "The failure of medieval Europeans and Arabs to recognize such phenomena was due not to any difficulty in seeing them but to prejudice and spiritual inertia connected with the groundless belief in celestial perfection."

The Arabs were restricted in their outlook on the universe by the influence of Aristotle, but that influence apparently did not have quite the same sway over them that it had over the Europeans. The astronomy of the medieval Arabs was far superior to that of the medieval Europeans. Bernard R. Goldstein of the University of Pitts-



SECOND CHINESE STAR CHART is a companion to the one on the preceding page; it shows as much of the southern celestial hemisphere as was visible to Chinese observers. Position and date of four remaining historical supernovas are again indicated by a cross and a number. As can be seen on both maps, the supernovas tend to lie close to Milky Way. From top to bottom asterisms with a line around them are *T'ien-chiang* (the Celestial River), *Wei* (the tail of the Occidental constellation Scorpius) and *Nan-mên* (Alpha and Beta Centauri).

burgh, who has intensively searched Arab chronicles, astrological treatises and other documents, has found three detailed accounts of the supernova of 1006. It seems possible that there are also Arab records of the supernova of 1054, and so we have initiated a search of the literature of the period. So far, however, we can still only express the hope that useful results will emerge.

There remain the chronicles of the Far East. The astronomy of the Orient was virtually unrivaled until modern times, both in the scope of its concepts and in the extent and variety of its observations. By a very early period in China, possibly well before 1000 B.C., a complex and rigid astrological system had been devised. It remained virtually unchanged up to modern times. The night sky was subdivided into some 250 asterisms, or small constellations, each consisting of perhaps five or six stars. Each asterism represented a province, a personage or some other entity. For example, the asterism *Tsung-chêng* was the Supervisors of the Nobility, *Huan-chê* was the Court Eunuchs, *T'ien-shih* was the Celestial Market, *T'ien-chiang* was the Celestial River, *Tsung-hsing* was the Ancestor Star and *Ch'uan-shê* was the Guest Houses.

A celestial event such as the appearance of a planet, a comet or a new star in a particular asterism was regarded as a precursor of terrestrial events. There was a standard prognostication for every omen. For example, one passage in the *Chin Shu,* or *Book of the Chin Dynasty,* written in A.D. 635, has been translated by Ho Peng Yoke of Griffith University in Australia: "The two stars of *Tsung-chêng,* southeast of the 'Imperial Throne,' represent the officials in charge of the nobility. When a comet guards these stars, while their color also fades, they predict that these officials will be busily engaged. When a 'guest star' guards them, changes of rank and titles are to be anticipated."

The delay between the appearance of an omen and its acknowledged fulfillment tended to be comfortably long. In the *Hou Han Shu,* or *Book of the Later Han Dynasty,* written between A.D. 300 and 450, the solar eclipse of A.D. 120, which was almost total at the Chinese capital of Lo-yang, was regarded as boding ill for the dowager empress. It was not until two years and three months later, however, that the portent was fulfilled. (The dowager empress died.) A delay of that length was typical: it gave the prediction a good chance of being right.

To administer such a complex astronomical and astrological system civil servants were appointed to maintain a constant watch on the sky night and day. The office was hereditary, a fact that doubtless has much to do with the system's resistance to innovation. An almost identical system was later adopted by the Koreans and the Japanese, so that after about A.D. 700 independent observations were being made in all three countries.

YEAR	CONSTEL-LATION	MAGNITUDE	DURATION	GALACTIC LATITUDE	REMNANT RADIO SOURCE	RECORDS
185	CENTAURUS	−8	20 MONTHS	0° ± 2°	G 315.4 − 2.3	CHINESE
393	SCORPIUS	−1	8 MONTHS	0° ± 5°	G 348.5 + .1 OR G 348.7 + .3	CHINESE
1006	LUPUS	−8 TO −10	SEVERAL YEARS	+14.5°	G 327.6 + 14.5	CHINESE JAPANESE EUROPEAN ARABIC
1054	TAURUS	−5	22 MONTHS	−5.8°	G 184.6 − 5.8 (CRAB NEBULA)	CHINESE JAPANESE
1181	CASSIOPEIA	0	6 MONTHS	+2° ± 2°	G 130.7 + 3.1	CHINESE JAPANESE
1572	CASSIOPEIA	−4	18 MONTHS	+1.4°	G 120.1 + 1.4	CHINESE KOREAN EUROPEAN
1604	OPHIUCHUS	−2.5	12 MONTHS	+6.8	G 4.5 + 6.8	CHINESE KOREAN EUROPEAN

NEW STARS recorded over the past 2,000 years that remained visible for at least six months are listed in chronological order. For each star the Occidental constellation in which it appeared is given, together with the star's probable apparent magnitude when it was at its maximum brilliance, the duration of its visibility, its estimated latitude above or below the central plane of the galaxy and the celestial radio source identified as the remnant of the explosion. The scale of magnitudes is a logarithmic one in which negative numbers indicate a greater brightness than positive numbers, and in which each magnitude is some 2.5 times brighter than the one below it. Zero magnitude is the brightness of Vega and −4 magnitude is the brightness of Venus.

Since the prognostication depended on knowing the asterism in which the new star had appeared, the position of the star was usually carefully recorded. Furthermore, it was clearly recognized that certain "guest stars" were different from planets and comets in that they remained in one celestial position as they grew brighter and then gradually disappeared from sight. It is in that class of objects that the supernovas are to be found.

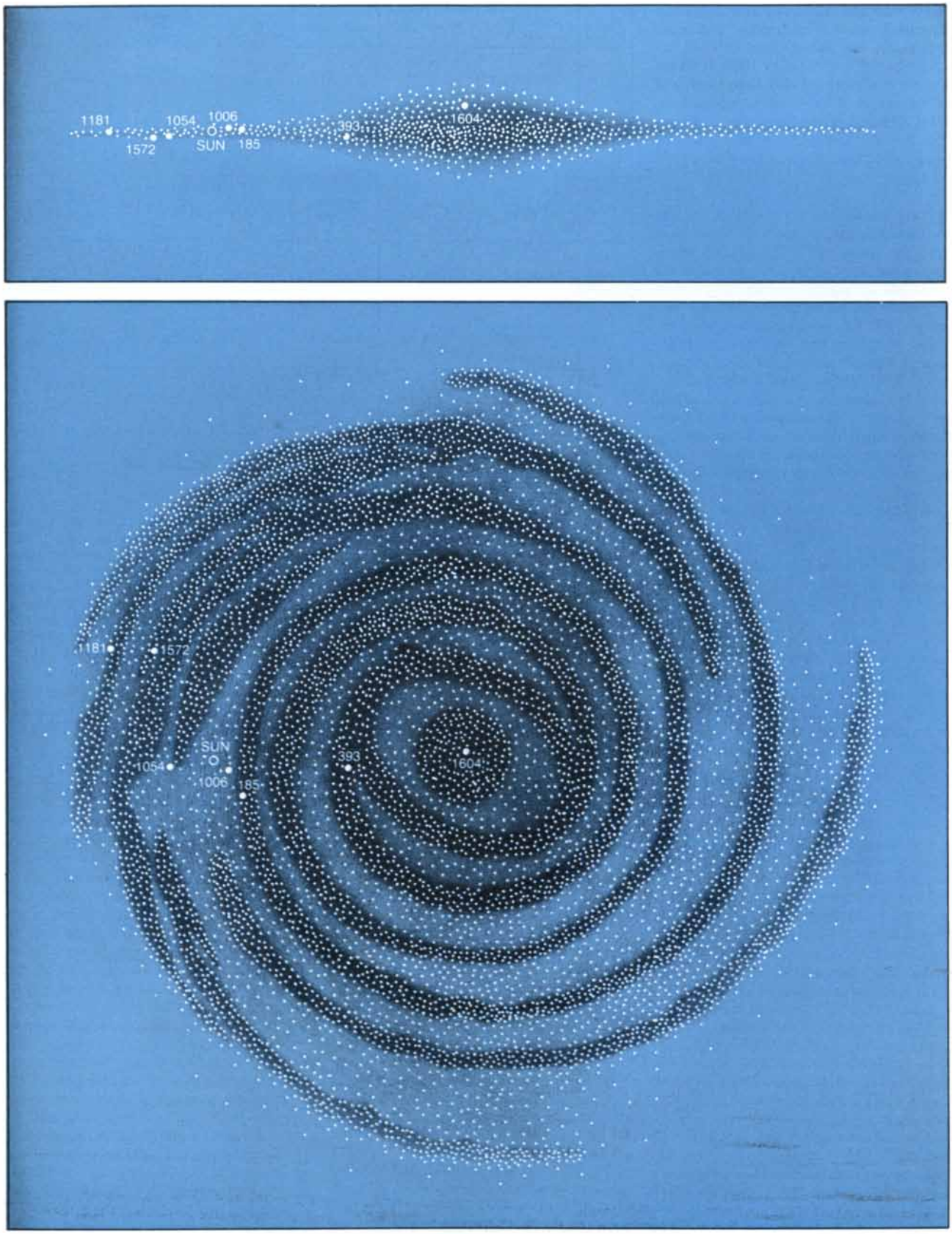
More than 50 guest stars are recorded in the Far Eastern chronicles. Most of them were undoubtedly ordinary novas that were close to the solar system. Where the new stars were observed for only four or five months, their positions are poorly recorded. Our criterion that the object must be visible for six months proved to be quite convenient for distinguishing the supernovas from the novas, because the observations of new stars that were visible for more than six months are of a superior quality. It is thus owing mainly to the diligence of the Oriental astronomers that seven probable supernovas are recorded in history.

The next step in identifying the historical supernovas is to search the recorded position today for some remnant of the explosion. Hundreds of supernova remnants are known; most of them have not been identified with any specific supernova. There are four principal kinds of remnant. One is a pulsar, a rapidly rotating neutron star believed to be the remains of the star that exploded. A second is a visible expanding gaseous nebula. A third is an extended source of radio emission, and a fourth is an extended source of X rays.

Of the objects within our galaxy that have been positively identified as the remnants of supernovas only the Crab Nebula and the remnant of the supernova in the constellation Vela are definitely known to display all four features. For the remainder the extended radio source is the most obvious remnant and appears to be common to all supernova explosions. It happens, however, that because of the obscuring dust in the central plane of our galaxy, emission at visible and X-ray wavelengths can be detected only from supernova remnants that are comparatively close to the solar system. The detection of a pulsar is subject to similar observational limitations. Moreover, it is by no means yet certain that all supernovas give rise to an observable pulsar.

The radio remnant of a supernova typically has the structure of a ring or a shell. It also has a nonthermal spectrum, a spectrum produced by charged particles spiraling in a magnetic field and emitting what is known as synchrotron radiation. Synchrotron radiation has two hallmarks: the intensity of the radiation is greater at the longer wavelengths and the radiation is strongly polarized, that is, the direction in which the electric component of the electromagnetic field oscillates is constant.

On the basis of these criteria of classifica-



POSITIONS OF HISTORICAL SUPERNOVAS (*large white dots*) **in our galaxy are shown in both a cross section of the galaxy** (*top*) **and a plan view** (*bottom*). **Supernovas tend to lie near the central plane of the galaxy. The galactic plane, however, is filled with interstellar dust** (*gray*) **that would dim the light of supernovas other than those comparatively close to the solar system. An exception is the supernova of 1604, which was high enough above the galactic plane for it to be visible even though it was nearly half the diameter of the galaxy away from solar system. Diameter of galaxy is about 100,000 light-years. The numbers again give the years the new stars were observed.**

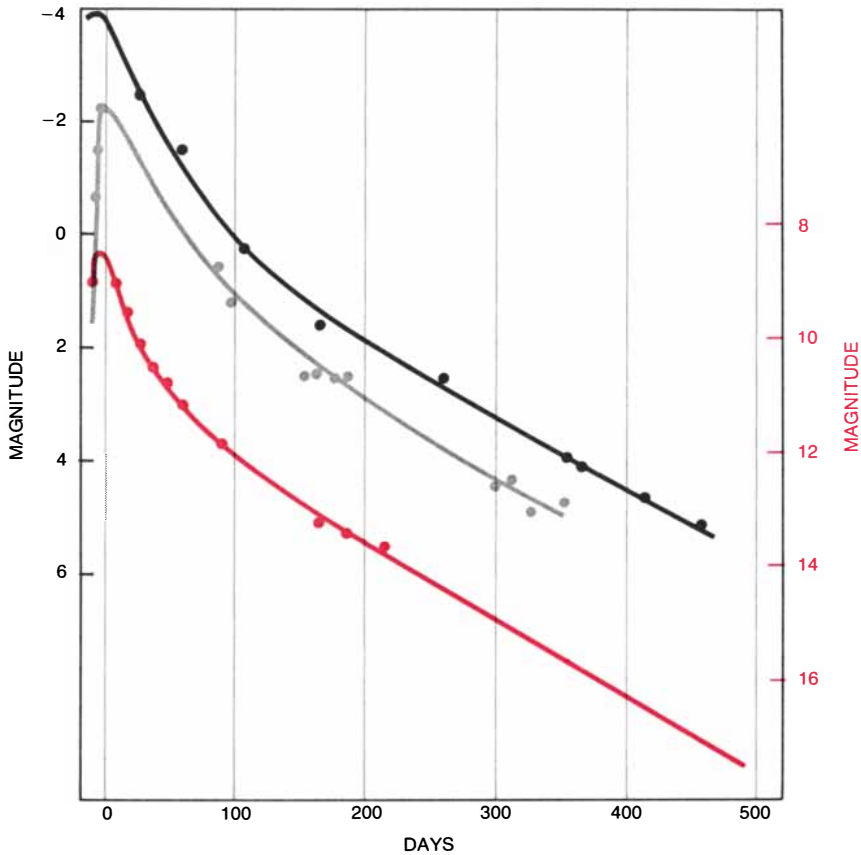
tion the first comprehensive catalogues of radio-emitting supernova remnants in our galaxy were compiled independently by Douglas K. Milne of the Australian Commonwealth Scientific and Industrial Research Organization (C.S.I.R.O.) and by Dennis Downes of the Max Planck Institute for Radio Astronomy in Bonn. A few of the sources in the catalogues have now been reclassified as clouds of ionized hydrogen not produced by supernovas. Many new supernova remnants have been identified, however, by James L. Caswell of the C.S.I.R.O., Anne Green of the Max Planck Institute and one of us (Clark) from a survey of the southern portion of the galactic plane made with instruments in Australia (at the Molonglo Radio Observatory of the University of Sydney and at the C.S.I.R.O. radio observatory at Parkes). The most recent catalogues list some 120 intragalactic radio sources that are believed to be supernova remnants, and it is in such catalogues that candidates for the remnants of historical supernovas are now being sought.

The new stars that remained visible for at least six months are presented in the illustration on page 103. We shall discuss the objects in reverse chronological order, concentrating on the supernova of A.D. 185 in somewhat greater detail than the others because it has particularly attracted our attention.

Kepler's supernova, which appeared in the constellation Ophiuchus in the year 1604, was carefully observed both by leading astronomers in Europe and by court astronomers in China and Korea. The position of the star was so accurately measured by Kepler and Fabricius that in 1943 Walter Baade of the Mount Wilson Observatory had no difficulty finding at that position a small patch of nebulosity that is clearly a remnant of the explosion. Furthermore, the European and Korean astronomers compared the brightness of the supernova with that of the neighboring planets and stars. Those comparisons make it possible to draw a remarkably precise curve of the increase and decrease of the supernova's light [*see illustration on this page*].

Two principal types of supernova can be identified on the basis of light curves and spectra. The shape of the light curve that Baade constructed from the European observations proves conclusively that Kepler's supernova of 1604 was of Type I, that is, a supernova that was originally a star with a mass roughly the same as that of the sun. (A Type II supernova is a star far more massive than the sun.) We are in the process of improving Baade's light curve by incorporating the Korean observations.

The position of Kepler's supernova, when it is corrected for the precession of the earth's axis over the intervening centuries, is identical with the position of a powerful radio source designated G4.5 + 6.8. That designation gives the source's galactic longitude measured along the galactic plane from a point in the direction of the galactic



LIGHT CURVES OF THREE SUPERNOVAS plotted by Walter Baade in 1945 show the apparent magnitude of each object with respect to the length of time it was visible. Both Tycho's supernova of 1572 (*black*) and Kepler's supernova of 1604 (*gray*) behaved like a supernova observed in 1937 (*color*) in the galaxy IC 4182. Magnitude scale at the left refers to the apparent brightness of the supernovas in our galaxy; the scale to the right refers to the brightness of the supernova in IC 4182. Duration of visibility is adjusted so that day zero is supernova's maximum. All three supernovas are Type I: explosions of a star slightly more massive than the sun.

center (in the constellation Sagittarius) followed by the source's galactic latitude measured north or south of the galactic plane. The radio source, which is the remnant of the supernova, has a diameter of about three minutes of arc. A high-resolution radio map has been constructed by Steven Gull of the University of Cambridge from observations made with the One-Mile Radio Telescope at Cambridge and with the radio interferometer at the Owens Valley Radio Observatory of the California Institute of Technology. The map shows that the object has the form indicative of an expanding shell of gas, a form that is characteristic of supernova remnants.

Tycho's supernova of 1572 was also observed both in Europe and in the Far East, but it is only Tycho's observations that provide useful information. The position he determined for the object allows us to identify it with the radio source G120.1 + 1.4, and once again the light curve indicates that it was a supernova of Type I. Tycho's supernova, which appeared in the constellation Cassiopeia, was considerably brighter than Kepler's; at its brightest it rivaled Venus. A high-resolution radio map of its remnant made with the Westerbork radio telescope in the northeastern Netherlands again shows a symmetrical shell at radio wavelengths. The position of the outer edge of the shell corresponds to the position of filaments observable at visible wavelengths.

The supernova of 1181, unlike Kepler's supernova and Tycho's, does not seem to have been particularly bright. At its brightest it was about magnitude zero, or approximately as bright as the star Vega. It was recorded only in China and Japan, and its location with respect to three neighboring asterisms in Cassiopeia was carefully described. The records convey little information about the supernova's light curve, so that its type is uncertain. Its position is in excellent agreement with the position of the nonthermal radio source G130.7 + 3.1, and it is highly probable that the two are the same. Unlike most other supernova remnants, however, the radio source is brightest at the center instead of at the edges.

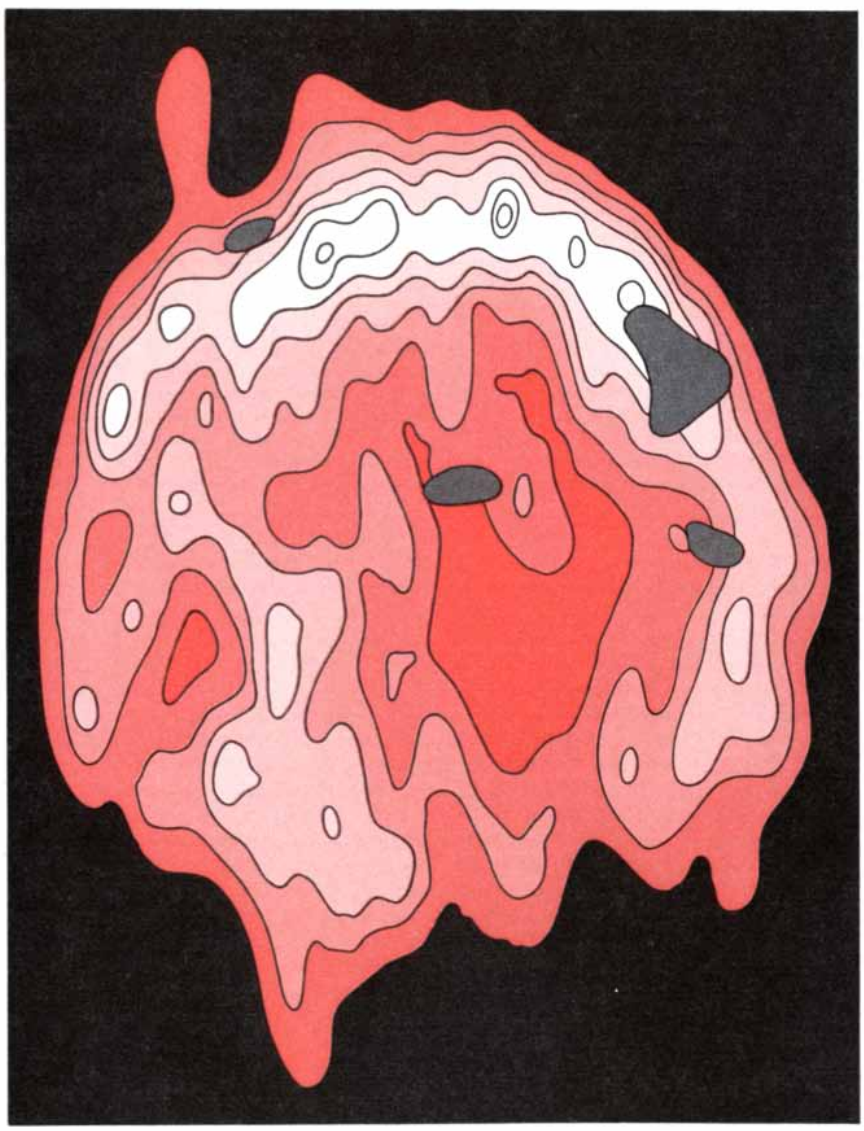
The supernova of 1054 was again recorded only by the Chinese and the Japanese. From the combined reports we learn that the star was visible in the daytime for 23 days and that it appeared close to the third-magnitude star Zeta Tauri. The possi-

bility that the supernova was associated with the Crab Nebula was first suggested in 1921 by Knut Lundmark of the University of Lund. The rate at which the visible filaments of the Crab Nebula are expanding indicates that it was created within about a century of 1054. Moreover, that conclusion seems to be confirmed by the rate at which the period of the pulsar embedded in the center of the nebula is lengthening. The pulsar is responsible for much of the radiation from the Crab Nebula. It is believed that a pulsar's rate of rotation decreases fairly regularly with the passage of time; the rate at which the rate of rotation decreases gives some measure of the pulsar's age. The pulsar in the Crab Nebula is the only one known to be associated with a supernova recorded in historical times.

Ho Peng Yoke and his colleagues have criticized the identification of the Crab Nebula with the supernova of 1054. They observe that the nebula lies northwest of Zeta Tauri, whereas in the historical records the new star is said to lie to the southeast of Zeta Tauri (called in the records *T'ien-kuan*). We have replied to their criticism in detail, pointing out that any observer can easily make a relatively small mistake in specifying directions in the sky. There is every reason to believe the Crab Nebula is one of the few known supernova remnants the age of which has been precisely determined.

The supernova that appeared in 1006 is the only one known to be recorded in both European and Arab records before the Renaissance. It was also carefully observed by astronomers in China and Japan. Almost every source comments on the extreme brilliance of the star. A Chinese record states: "It shone so brightly that objects could be seen by its light." One Arab record states: "Its rays on the Earth were like the rays of the Moon." Another comments: "Its light illuminated the horizon and... its brightness was a little more than a quarter of the brightness of the Moon." A Chinese record indicates that the star was observed for several years.

From those reports Goldstein has estimated that the apparent magnitude of the star at its brightest ranged between -8 and -10, indicating that it must have been very close to the solar system, perhaps only 4,000 light-years away. (The galaxy is some 100,000 light-years in diameter.) The Chinese and Japanese records are independently confirmed by the records from the monastery of St. Gallen in Switzerland. There the star must have been very low in the sky, and the southern horizon is quite mountainous. Thus by accident the St. Gallen account places a valuable southern limit on the galactic latitude of the star, and it supports the identification of G326.6 + 14.5 as the remnant of the supernova. The remnant, which is about half a degree in diameter, has a low surface brightness at radio wavelengths. The best radio maps available, which have been made by Milne with the 210-foot radio telescope at Parkes, suggest that it is a partial ring.

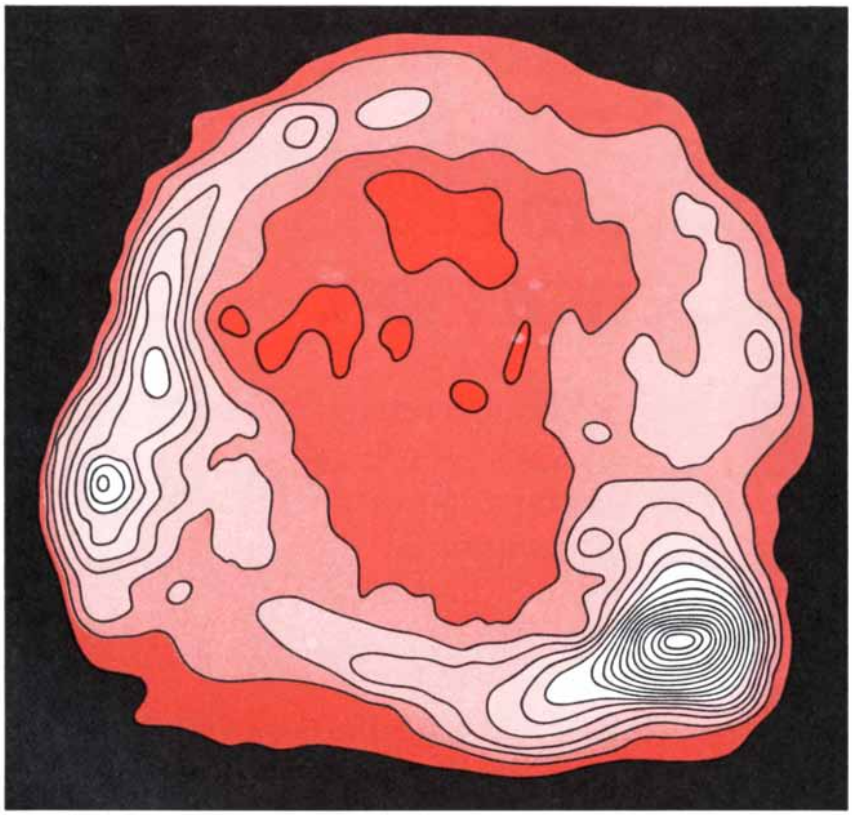


REMNANT OF SUPERNOVA OF 1604 was mapped by Steven Gull of the University of Cambridge with the Cambridge One-Mile Radio Telescope and the radio interferometer at the Owens Valley Radio Observatory of the California Institute of Technology. Lighter colors indicate more intense radio emissions; darker colors, less intense. Gray regions indicate position of visible filaments. Radio source has ringlike structure characteristic of supernova remnants.

The remaining two supernovas of which we have historical accounts were recorded only in China. We were the first modern investigators to call attention to the supernova of A.D. 393. There is no mention of the brightness of the star in the single record we have, but from the fact that the star was visible for a period of eight months we have inferred that it was probably brighter than magnitude -1, or close to the brightness of Sirius.

The new star was described as being within the asterism of *Wei,* which is the tail of the constellation Scorpius. The asterism is well defined by bright stars and forms a nearly closed curve lying on the galactic equator. We have found that there are three supernova remnants on the near side of the center of the galaxy (which would thus be free from the strong absorption of light by dust) that fit the observed location. The estimated age of two of them is close to 1,500 years. Further research may well establish which is the true remnant.

The last probable supernova on our list is the guest star of A.D. 185. The solitary Chinese record of it is in the astronomical treatise of the *Hou Han Shu,* where it is stated: "On the day *kuei-hai* in the tenth month of the second year of the Chung-p'ing reign period a guest star appeared within *Nanmên;* it was as big as half a mat; it was multicolored and it fluctuated. It gradually became smaller and disappeared in the sixth month of the year following the next year.



MAP OF RADIO SOURCE G315.4 − 2.3, the suggested remnant of the supernova recorded by the Chinese in A.D. 185, was made by the authors with the Mills Cross radio telescope at the Molonglo Radio Observatory of the University of Sydney. The visible gaseous filaments (*not shown*) are concentrated along the bright southwestern rim of radio remnant at lower right.

According to the standard prognostication this means insurrection." The text goes on to say that in the sixth year of the reign there was a rebellion that was ruthlessly suppressed.

According to the text, the supernova was first seen on December 7, 185, and disappeared between July 24 and August 21, 187. *Nan-mên,* which means Southern Gate, consists only of the two bright stars Alpha and Beta Centauri. The record implies that the guest star appeared between them. At Lo-yang, however, Alpha and Beta Centauri would have been no more than three degrees above the horizon. The absorption of light by the atmosphere at such a low altitude is between three and four magnitudes. For the new star to have remained visible for so long under such unfavorable circumstances suggests that it was very brilliant and hence close to the solar system.

There are four supernova remnants lying roughly between Alpha Centauri and Beta Centauri. Two of them are reliably estimated to be at a considerable distance from the solar system, on the order of 30,000 light-years. Moreover, their galactic latitude is low, so that their light would tend to be obscured by the dust of the galactic plane. It seems virtually impossible that outbursts at such a distance would be visible from the solar system.

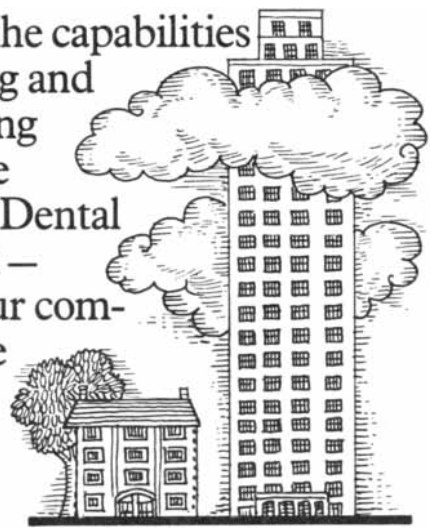
The other two remnants are the radio sources G315.4 − 2.3 and G315.4 − 0.3. Eric R. Hill of the C.S.I.R.O. has suggested that the first of these sources might be the remnant of the supernova. That source has a diameter of about 40 minutes of arc and shows the peripheral brightening characteristic of supernova-remnant shells. At visible wavelengths the remnant consists of bright filaments. It is estimated to be less than 6,000 light-years away and 2,000 years old. We are confident that Hill's suggestion that G315.4 − 2.3 is the remnant of the supernova of A.D. 185 is correct.

The study of historically recorded supernovas appears to have reached maturity. Little progress in the discovery of historical records of additional early supernovas seems likely, although it is possible that further information on the stars we have discussed may yet appear. It should be kept in mind that any new star seen for much less than six months was probably an ordinary nova.

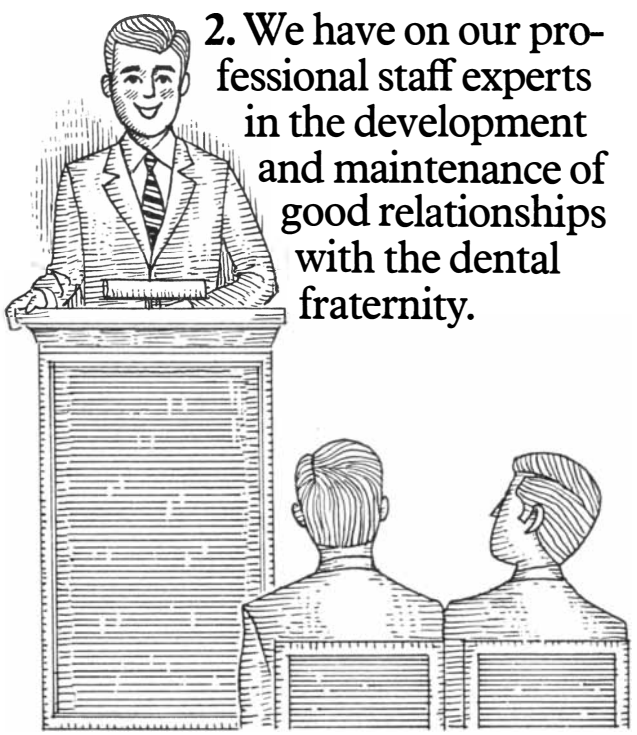
The Crab Nebula is undoubtedly the outstanding example of a supernova remnant. It may, however, have dominated the attention of astronomers for too long. A unified investigation of all seven of the historical supernovas we have discussed here might considerably advance understanding of how supernova remnants evolve.

5 good reasons why you shouldn't miss out on Metropolitan Life's Group Dental Plan.

1. We have the capabilities for designing and implementing precisely the *right* Group Dental plan for you — whether your company is large or small.

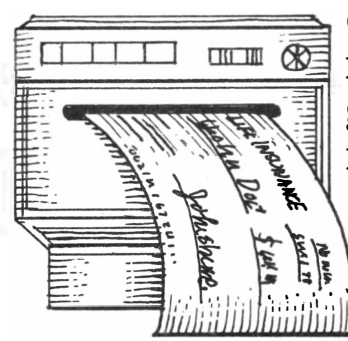


2. We have on our professional staff experts in the development and maintenance of good relationships with the dental fraternity.

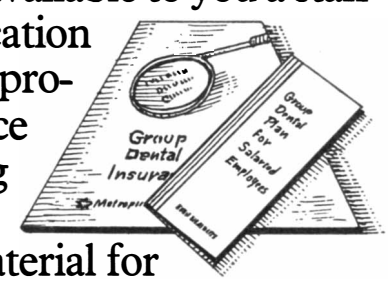


3. We also retain practicing dentists who serve as consultants to Metropolitan clients on a local basis.

4. Our computerized claims processing system functions with great efficiency from both a facility of payment and expense point of view.



5. We make available to you a staff of communication specialists to provide assistance in developing literature and other material for use in giving plan information to your employees.



For the full story on our Group Dental Plans, just call your Metropolitan representative.

Metropolitan
Where the future is now

Future Performance in Footracing

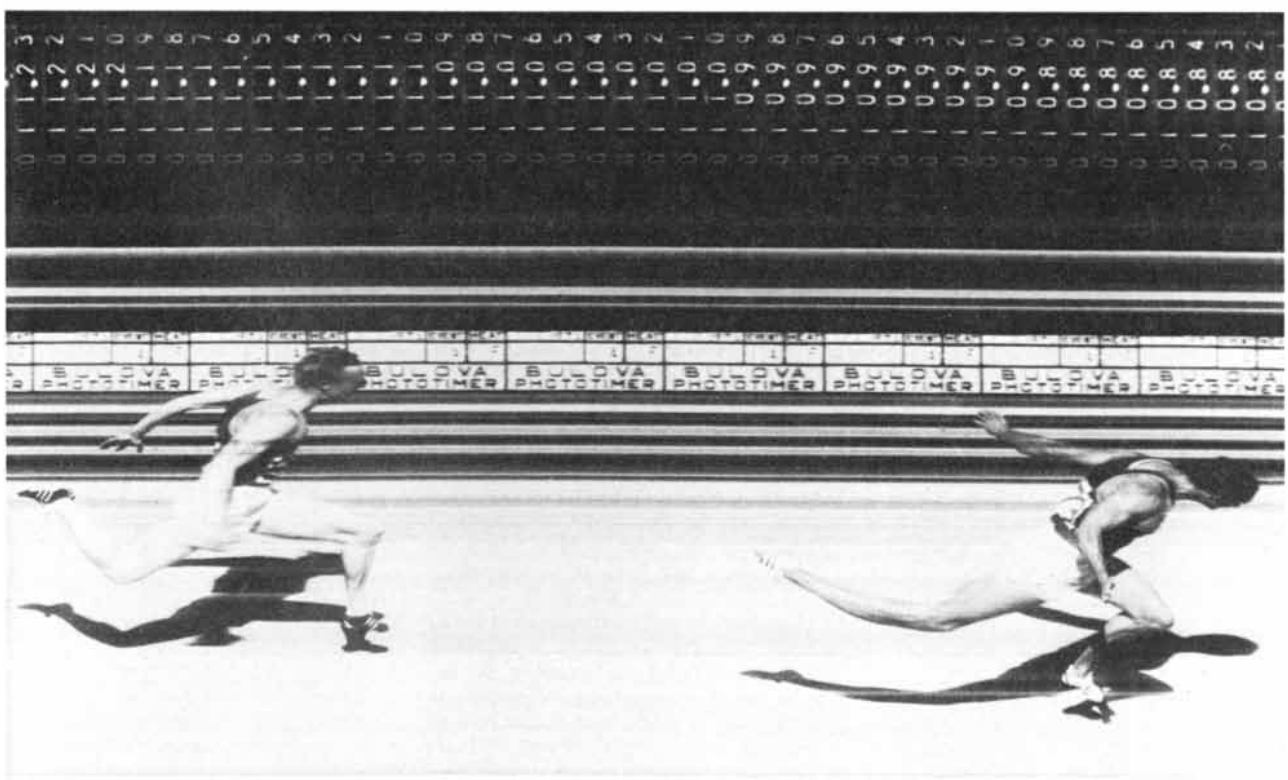
Running records are still far below human physiological limits. The restraints on performance are psychological: good runners do not work as hard once they have set a record or won a medal

by Henry W. Ryder, Harry Jay Carr and Paul Herget

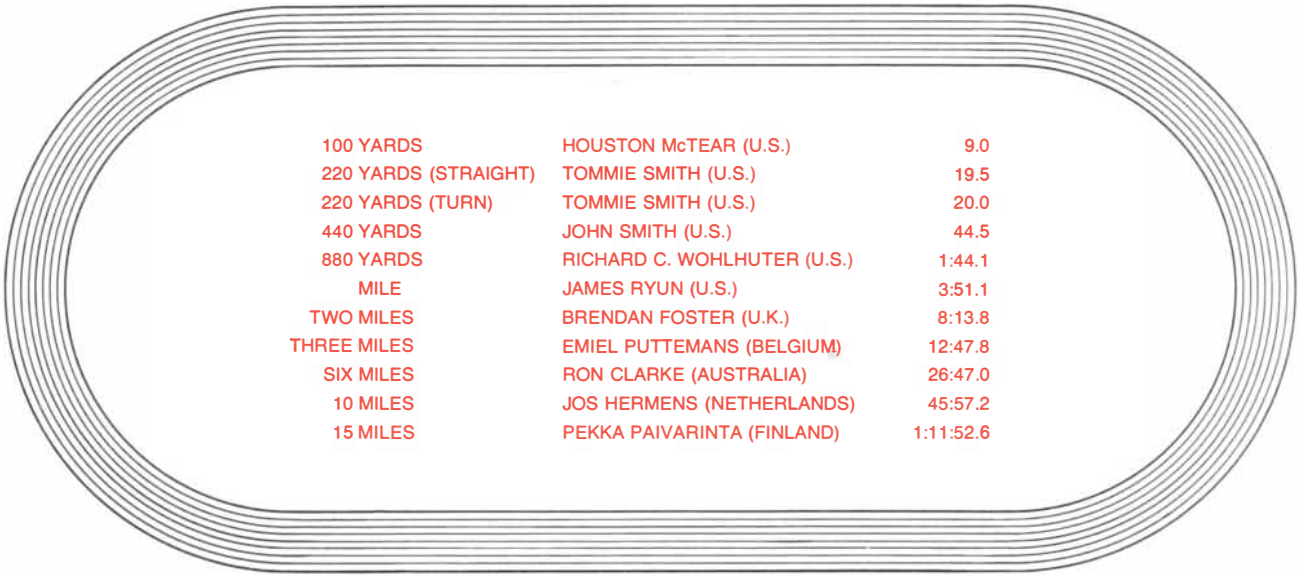
Roger Bannister, the first man to run the mile in less than four minutes (three minutes 59.4 seconds in 1954), has predicted that the first three-minute, 30-second mile will be run within 30 or 40 years. It is likely that several new records will be set in the running events of the 1976 Summer Olympic Games in Montreal next month. At every distance the running speed has increased steadily over the past 50 years. Why? There must be a physiological limit to human running speed, but it certainly has not yet materialized at any distance. Our investigations of footracing have led us to the conclusion that the barrier to be overcome by the runner who wants to be a champion is psychological: the last record set and the willingness of athletes to try to break it are the determining factors for the next record. Moreover, a champion stops not on achieving a given rate of speed but on winning a given medal.

By obtaining data from various sources on the performance of male amateur runners at distances of from 100 yards to 30 kilometers we have plotted the average speed in meters per minute of the runners who set world records over the past 50 years. The increase in speed is reflected in those plots [*see illustrations on pages 112 and 113*]. The rate of acceleration has averaged about .75 meter per minute per year. This increase in mean speed is a veritable crawl, so imperceptible at the shorter distances as to lead many coaches and athletes to feel that the limit of ultimate performance is being approached or may already have been reached.

In the 100-meter dash the mean speed improves by about .6 meter per minute per year. For a distance run in 10 seconds the improvement amounts to 10 centimeters



ELECTRONIC TIMING records the finish of a 100-meter race in the decathlon of a meet between the U.S. and the U.S.S.R. at the University of California at Berkeley in 1971. The winner was Russ Hodge of the U.S., and second place was taken by Boris Ivanov of the U.S.S.R. The same two men finished in the same order in the decathlon. What one sees in this electronically timed finish is not a simultaneous photograph of the two runners at the end of the race but a continuous photograph of the finish line showing each runner as he crossed it. His time for the race appears in seconds and hundredths of a second above him. In hand-timed races the timing is only to a tenth of a second. In the electronic timing apparatus employed in this race the film moves horizontally past a shutter that is always open, which accounts for the separate photographs of the runners and the distortion of their bodies. Present records are based mostly on hand timing.



100 YARDS	HOUSTON McTEAR (U.S.)	9.0
220 YARDS (STRAIGHT)	TOMMIE SMITH (U.S.)	19.5
220 YARDS (TURN)	TOMMIE SMITH (U.S.)	20.0
440 YARDS	JOHN SMITH (U.S.)	44.5
880 YARDS	RICHARD C. WOHLHUTER (U.S.)	1:44.1
MILE	JAMES RYUN (U.S.)	3:51.1
TWO MILES	BRENDAN FOSTER (U.K.)	8:13.8
THREE MILES	EMIEL PUTTEMANS (BELGIUM)	12:47.8
SIX MILES	RON CLARKE (AUSTRALIA)	26:47.0
10 MILES	JOS HERMENS (NETHERLANDS)	45:57.2
15 MILES	PEKKA PAIVARINTA (FINLAND)	1:11:52.6

PRESENT RECORDS in races at nonmetric distances are listed as recognized by the International Amateur Athletic Federation as of November 30, 1975. During 1975 two new records were set for the mile, by Filbert Bayi of Tanzania in May at 3:51 and by John Walker of New Zealand in August at 3:49.4. Houston McTear shares the record for the 100-yard dash with Ivory Crockett of the U.S., who in the spring of 1974 was the first man to run the race in nine seconds. McTear is listed here because he equaled the record most recently.

per year. The 10 centimeters are covered in .01 second. Since the 100-meter dash is timed at present only to the nearest .1 second, a new world record is expected in this event only once in 10 years.

For the marathon (26 miles 385 yards) the increase of about .9 meter per minute per year is about 116 meters for a distance that is 422 times longer than the 100-meter race. The marathon runner covers the 116 meters in about 21 seconds at the present record rate of 128.5 minutes for the event. Therefore the marathoner has a sporting chance at setting a new record every time out, whereas the 10-year span for a new record at the short distances is several times the sprinter's expected racing lifetime at peak speed. Skiing and swimming races are timed now to .001 second. With present technology the same could be done for the dash men, giving them a better chance at setting a record.

Although the general trend of improvement for all the distances is linear, significant deviations appear. Some of them seem random, some systematic. Much of the systematic deviation is to be expected, since any new record is a function of the present one. The successive points are highly correlated. The better a record is, the less likely it is to be beaten in any particular year.

Another prominent systematic deviation at most of the racing distances is a paucity of new records in the late 1940's and early 1950's, with a clustering of new records near 1955. For example, after Gunder Haegg of Sweden set the record for the mile at 4:01.4 in 1945, nine years passed before Bannister broke it. One often hears the barren period from 1945 to 1955 explained in terms of "barriers" that had to be surmounted. Bannister is widely said to have "cracked the four-minute barrier." Four minutes is a conveniently round figure, but in racing it applies uniquely only to the mile. The similar lags that appeared between records in many of the other distances cannot be explained by any such convenient "barrier." We regard the lag as resulting from the disruptive effect of World War II on the training of the athletes who might have set records and from the fact that years of training were required after the war for athletes to reach the levels of performance at which they could set new records.

In 1926 the British physiologist A. V. Hill, who was interested in analyzing the causes of the fatigue that limited the performance of racers, introduced the concept of the racing equation. "If one plots the average speed at which the record is made," he wrote, "against the length of the race (compressed in some manner—as by taking its logarithm—to bring all points onto the same diagram), then a curve of almost perfect smoothness is obtained, in which a few points only lie just below the curve.... These latter are those races in which the athletes have not been so concerned to break the record as in the rest."

At any arbitrary date the equation can be plotted as a series of points, representing mean speeds at defined distances. We call the result the mean racing equation. In making such a plot, based on the records existing at the end of 1924 and 1972, we found that the performances at the nonmetric distances were slightly poorer than the performances at the metric distances. (The 23 points of such a plot group themselves into four families. One consists of 12 races at metric distances, ranging from 100 meters to 30 kilometers. Another is made up of eight races at nonmetric distances, ranging from 100 yards to 15 miles. The third is a doublet consisting of the one-mile race and the metric distance closest to it, the 1.5-kilometer race, and the fourth is the marathon.) The differences increase slightly with distance.

These differences appear to be systematic, since the probability is small that all eight of a series of nonmetric races would be run slower than a series of metric races at comparable distances if the differences in performance were only random. We consider a relative lack of concern over records, as suggested by Hill, to be a reasonable explanation of these differences. In other words, the metric races are apparently more competitive than the nonmetric ones. We have no data on the frequency with which the two types of distance are run. It is true, however, that the metric distances are more popular on the European continent and the nonmetric ones are more popular in the U.S. and the United Kingdom. Therefore the metric races are being run in more countries than the nonmetric races, which could give rise to a more competitive environment for the metric races.

For some years the faster sprinters have on the whole been Americans and the faster distance runners Europeans, regardless of whether the distances run were metric or nonmetric. The differences in performance between the metric and nonmetric races cannot be laid to the fact that the races are often run by different men, since the men who compete in both types of race usually do better at the metric distances than at the nonmetric ones. These systematic familial differences extend back over the past 50 years.

The marathon stands alone in excellence of performance. The current record (set in

1969 by Derek Clayton of Australia at two hours eight minutes 33.6 seconds) indicates an average speed for the 42.2 kilometers of 328.2 meters per minute. This rate is higher than the rate (327.8 meters per minute) at which the more recent record for the 30-kilometer race was set. Before 1973 at least 10 athletes had run the marathon at an average speed of 320.3 meters per minute, which was the rate that would have been expected in 1973 according to the projection of the long-term trend. The record expected in 1973 from consistent performance at the nonmetric distances was 304 meters per minute. Some marathon runners have beaten this time of two hours 18 minutes the first time out. As far as we know, however, no novice has ever set a world record at any distance.

The inference is that a highly competitive situation brings out in the finest of athletes a level of performance of which even they are incapable under less challenging circumstances. This concept of responding to challenge is hardly new. Nevertheless, the strength of the effect seems to have been overlooked by many of the people who have employed the mean racing equations to predict future performance. The times in certain races are consistently better than those in others, but bringing the slower times up to the standard of performance found in the races run at the faster times is not simply a matter of encouraging better runners to run at those distances. What is needed is not better runners but bigger challenges.

With few exceptions world records are set by runners whose speed is constant within a few percent after the initial buildup. We call this phenomenon the Aesop principle after the fable of the tortoise and the hare, in which the moral is: Slow and steady wins the race. Because of this constraint elapsed time and distance can be interchanged in a mean racing equation with no loss of information and with two distinct advantages. First, the effect of each racing distance is isolated and not confounded with the peculiarities of the other distances, since the elapsed times decrease over the years but the distance of any race always remains the same. Second, if endurance is defined as the elapsed time during which a man can remain subject to a constant stress, we can substitute instantaneous velocity (rate of work) for stress and substitute distance for time; racing endurance can then be expressed as the distance a man can cover at a constant speed. In this way the mean racing equation can be considered as isolating the two components of racing performance—speed and endurance—and as making it possible to assign quantities to them.

These two components cannot be compared with each other directly in terms of how they account for the success of any particular race, since they represent different attributes. Each one can be compared, however, with its own previous value. For example, by comparing two races run at the same or nearly the same constant speed and establishing the ratio between the distance covered in the first race and the distance covered in the second race one can arrive at a quantity we call specific endurance. In the same way the increase in mean speed of all the races over the past 50 years can be expressed as specific speed.

A study of specific endurance reveals some extraordinary relations. Clayton ran the marathon in 1969 at the mean speed of 328.2 meters per minute; in 1864 Charles Lawes of Great Britain set the record for the mile (4:56) running at a mean speed of 326.2 meters per minute. Clayton showed a specific endurance more than 26.2 times higher than Lawes's, which is to say that working at the same rate as Lawes did, Clayton was able to do more than 26.2 times as much work.

The differences between the attributes of speed and endurance can be shown more clearly by plotting specific speed and specific endurance for all the racing distances at two well-separated times. We have made such a plot for the Olympic years 1924 and 1972. Any such plot emphasizes that a remarkable increase in endurance rather than speed is mainly responsible for the improvement in racing performance.

The dominant effect of the improvement in endurance is hidden by the custom of timing a race at a constant distance. If it were customary to measure how far runners can go in a given period of time, one would be measuring (and admiring) the distance they cover at a certain speed. In other words, the measurement would be recording their endurance and would neglect their comparatively small improvement in speed. Nonetheless, to the runner who wants to set a record it makes no difference which type of race he is running. He follows the Aesop principle all the way in either fixed-distance running or fixed-time running.

Physiologists and track coaches generally agree that a successful footrace is mostly run on the Aesop principle. The physiological argument is that for the total consumption of energy to be at a minimum the rate at which energy is used must be constant. Since the rate of energy consumption depends mainly on the instantaneous velocity, the speed at which the race is run must be rather constant in order for the runner to get the most out of his available energy. Hill has inferred that for a maximum effort the required constant pace must be such that the available energy is exhausted just as the race ends.

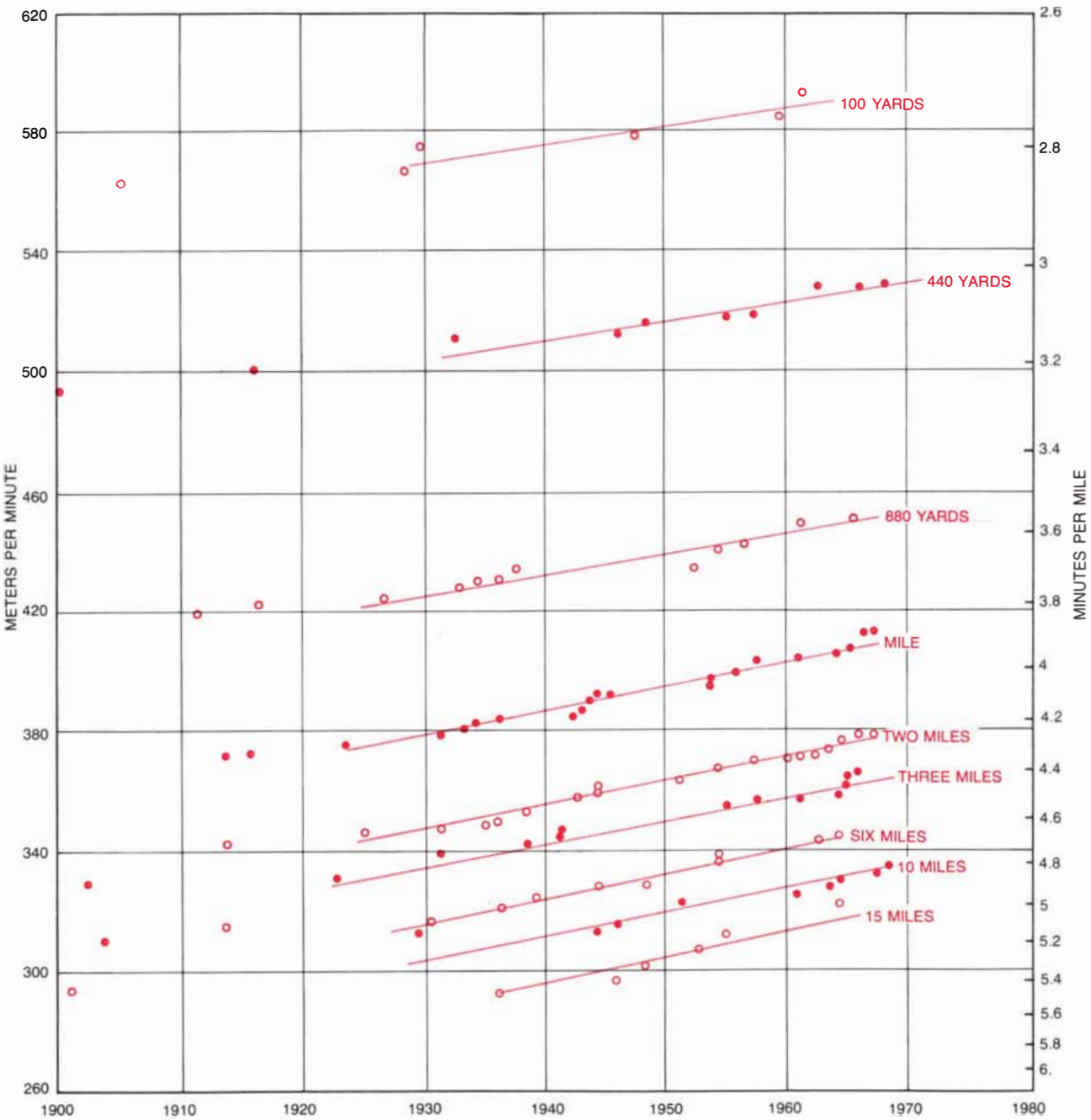
The availability of energy can be limited

100 METERS	STEVEN WILLIAMS (U.S.)	9.9
200 METERS (STRAIGHT)	TOMMIE SMITH (U.S.)	19.5
200 METERS (TURN)	DONALD QUARRIE (JAMAICA)	19.8
400 METERS	LEE EVANS (U.S.)	43.9
800 METERS	MARCELLO FIASCONARO (ITALY)	1:43.7
1,000 METERS	RICHARD C. WOHLHUTER (U.S.)	2:13.9
1,500 METERS	FILBERT BAYI (TANZANIA)	3:32.2
2,000 METERS	MICHEL JAZY (FRANCE)	4:56.2
3,000 METERS	BRENDAN FOSTER (U.K.)	7:35.2
5,000 METERS	EMIEL PUTTEMANS (BELGIUM)	13:13.0
10,000 METERS	DAVID BEDFORD (U.K.)	27:30.8
20,000 METERS	JOS HERMENS (NETHERLANDS)	57:31.6
25,000 METERS	PEKKA PAIVARINTA (FINLAND)	1:14:16.8
30,000 METERS	JIM ADLER (U.K.)	1:31:30.4

METRIC RECORDS are listed as recognized by the International Amateur Athletic Federation as of November 30, 1975. Steven Williams is one of six men to have run the 100 meters in 9.9 seconds; he is listed here because he did it most recently (1974). The record was first established by James Hines of the U.S. in 1968. The records held by Donald Quarrie and Lee Evans are based on electronic timing.

in ways other than exhausting the entire supply. The rate of oxidation for muscular effort is most efficient when it is aerobic, that is, when the rate of utilization of energy is equal to the rate at which molecular oxygen is supplied to the muscles. For maximal work when the demand for oxygen has exceeded the supply, the muscles can obtain a small amount of reserve energy by reducing pyruvic acid to lactic acid, although the process is a dead end in the metabolism of carbohydrate in muscle. It is done at the cost of accumulating lactic acid, which cannot be metabolized at all in muscle cells and has to be carried to the liver or the heart for oxidation or further reduction. Its accumulation in the muscle and the associated depletion of the necessary hydrogen donor, which at any time is present in an exceedingly limited amount, will quickly immobilize the runner. Since the energy available for either aerobic or anaerobic metabolism is thus limited by the rate at which it is consumed, this extra energy is available in an appreciable amount for only a limited time, measured in seconds. According to Sid Robinson of Indiana University, it should be used up as quickly as possible by a sprint at the end of the race.

The foregoing considerations bring up the further question of whether the principle of consuming energy at a constant rate should be restricted to aerobic metabolism, which is the more efficient process. The answer is that the runner at every stage of the race is already using both methods of energy conversion, with one or the other predominating at any particular stage. Early in a race of any distance it would be quite possible for him to run somewhat faster than the Aesop pace and still stay within his limits of balanced consumption of energy and oxygen, since at that early stage the enzymatic systems involved in the metabolic processes that develop energy have



IMPROVEMENT IN SPEED in footracing at nonmetric distances is charted for a period of approximately 50 years. Each circle represents a new record at each distance, and the sloping lines reflect the increase in mean speed in meters per minute and minutes per mile.

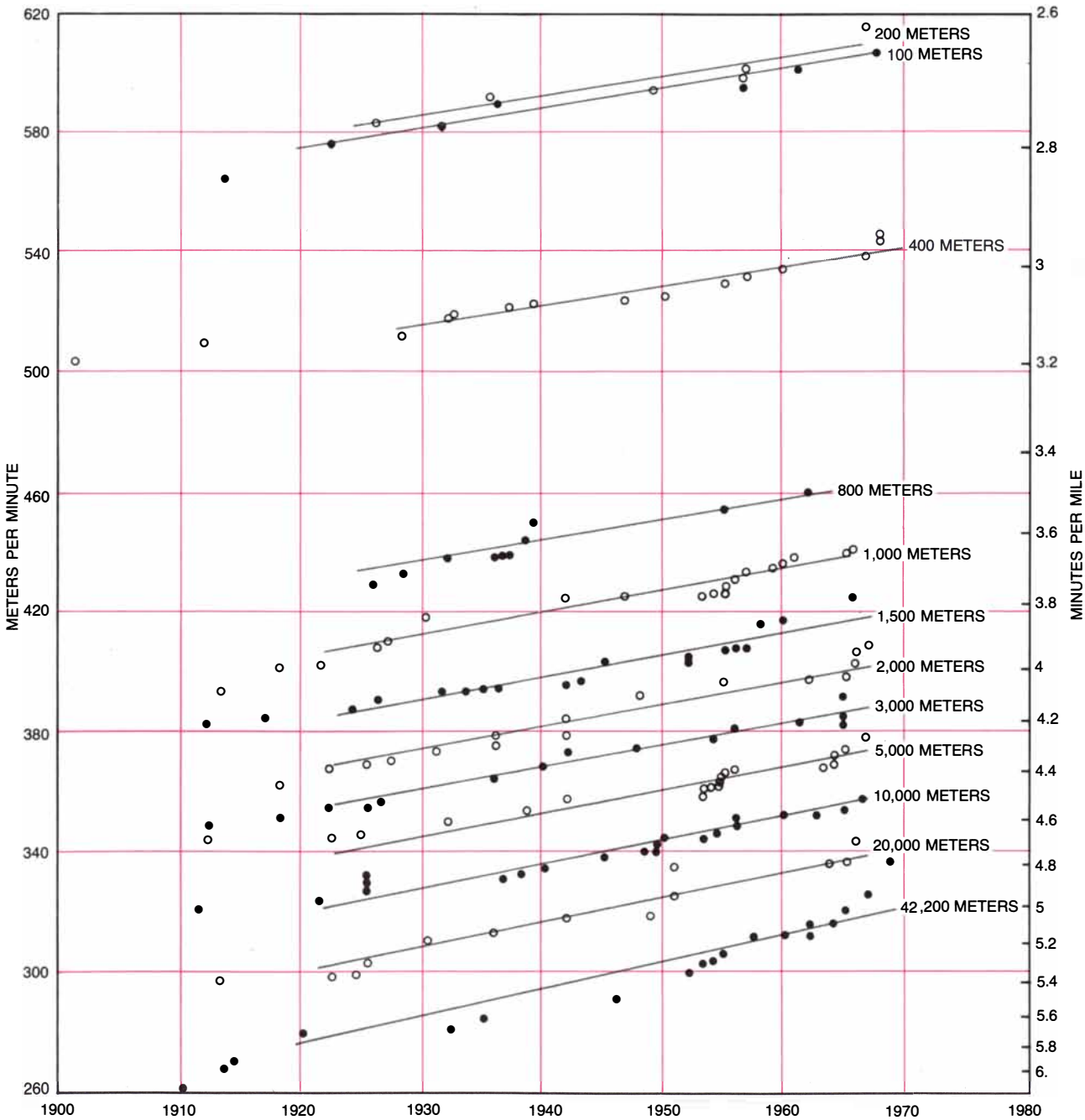
not been degraded by the accumulation of lactic acid.

It is rather surprising to learn how close the runner is at all times in a championship race to the limits imposed by the anaerobic use of energy. It is not unusual for the winner of a race to remark at the end that he had come close to quitting before the race was over. Certain runners, such as Vladimir Kuts of the U.S.S.R. (the holder of eight world records), have been noted for starting their races at a fast pace or for speeding up during the race, so that they win by exhausting their opponents. The technique succeeds by making other runners degrade their enzyme systems by working at a pace the winner has trained himself to tolerate longer. Kuts has said, however, that the technique does not serve for establishing records, only for winning races.

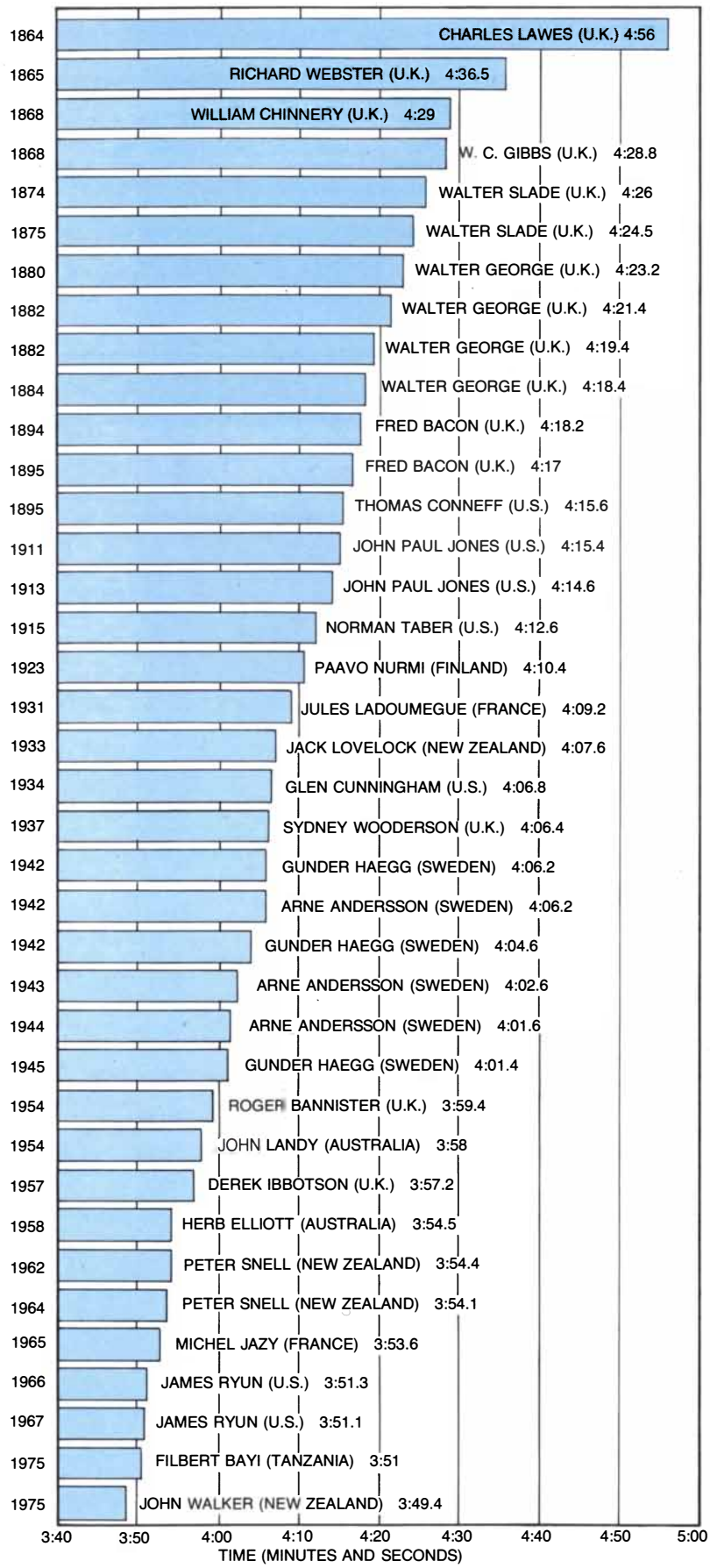
It is our impression that the principle of the steady conversion of energy stores to energy throughout the race may be an overriding consideration for the maximal use of the available energy. Therefore it is the key to maximal endurance and performance, even though it may not always be the way to win. The anaerobic runner, whether he is a sprinter or a marathoner, can ignore the Aesop principle only at the cost of failing to get the most out of his store of energy. Moreover, his supply of energy can be limited by the very act of spending it too quickly.

Although tactical variations in running a race may be of decisive importance in winning, they cannot account for the dramatic improvement in winning times over the past 50 years. There has simply not been any change in tactics that could account for the steady and remarkable increase in speed and endurance over that period of time.

It is of considerable significance that two-thirds of the men who establish a world record never set another one. About 20 per-



METRIC DISTANCES show a similar improvement. The performances in metric races are slightly better than in nonmetric races at comparable distances, apparently because on a worldwide basis metric races are run more frequently and are therefore more competitive.



cent of runners better their own record and 25 percent establish another record at a different distance. The man who establishes more than one record proves himself by that act to be in a superior class: unlike most of the record-setters, he has an even chance of setting yet another record. Even the most versatile athletes, however, set records only in a narrow range of distances, which is to say within a narrow range of speeds.

Whatever the amount by which a new record exceeds an old one, the magnitude of the improvement is overwhelmed by the change in the record with the passage of time. The greatest individual improvement in the record for a footrace was achieved by Emil Zátopek of Czechoslovakia, who lowered the time for the 20-kilometer race by 4.47 percent in two races two weeks apart in 1951. Yet since 1913 the record for that distance (set then by Hannes Kolehmainen of Finland at one hour seven minutes 40.2 seconds) has been lowered by 14.99 percent. In 1966 Tommie Smith of the U.S. lowered the record for the 200-meter race by 2.50 percent (from 20 to 19.5 seconds), but since Bernie Wefers of the U.S. set the record at 21.2 seconds in 1896 it has been lowered by 8.02 percent. Hundreds of athletes have run the mile in less than four minutes since Bannister first accomplished it; every year dozens of marathoners better the time of Zátopek in 1952, when he set the Olympic record. In 50 years the improvement in mean speed in every race has ranged from 5 to 15 percent. The real puzzles are why records are beaten so often, why, over a period of years, they are beaten so badly and why it is usually a matter of years between one record and the next.

Each performer has a racing pace and distance that are best for him. The greatest runners show some versatility in that they are unbeatable near their own best distance as well. Whether this superiority is due more to natural talent than to training is not important in the present context, because in time all records fall. The champions stop not at a given speed but when they set a record. Succeeding champions do the same. They telescope in their relatively short racing lives all the achievements of the great runners of the past and then stop with a gold medal, just as their predecessors did. Since it is the medal and not the speed that stops them, the speeds they reach cannot be considered in any way the ultimate physiological limit.

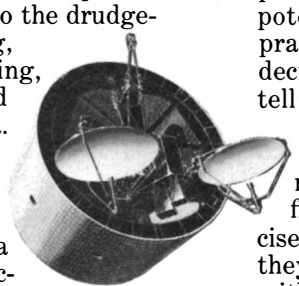
By continuing the plotted lines that reflect the increase in average speed at various distances over the past few decades, one can project the improvement that will be recorded in the future. In this way, for exam-

RECORD FOR THE MILE is traced from 1864 to the present. The steady increase in speed reflected in this chart has also been manifested at every other racing distance, metric and nonmetric. It appears likely that within 50 years the record will be down to 3:30.

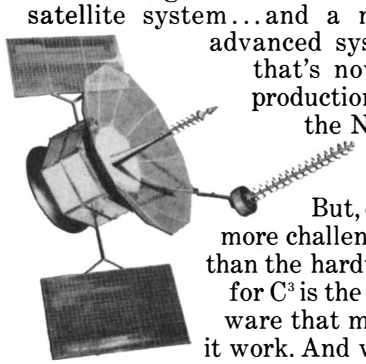
Conversation Pieces

C^3-Key to crisis management.

Every time there's a flare-up in Xenobia or Whereistan, staff officers have to work overtime. And, nowadays, crises seem to be erupting almost continuously. Tidal waves of data come in from all over the world, but what senior commanders need is *significant* information for making decisions. It's the age-old problem of command, control, and communications, or C^3 for short, but it's much more complicated than it used to be. So, we rely on sophisticated electronics to do the drudgework of sorting, storing, retrieving, correlating, and displaying data.



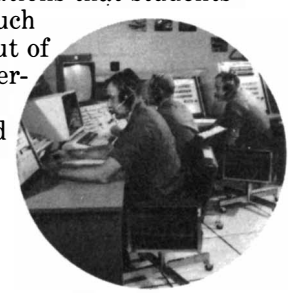
TRW builds a lot of these electronic systems; we also build the Air Force's global communication satellite system... and a more advanced system that's now in production for the Navy.

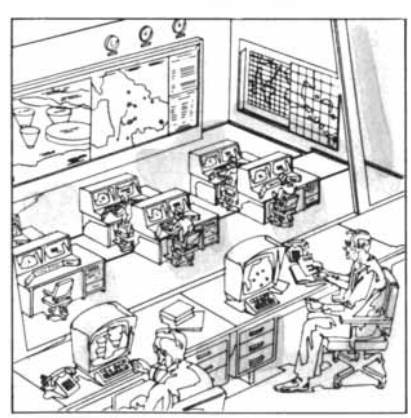


But, even more challenging than the hardware for C^3 is the software that makes it work. And we're using the term software, here, to mean more than just computer programming. It includes an enormous amount of front-end analysis and systems engineering. We emphasize this because we've found it's the only way to deliver systems that work properly, and do it without delays or over-runs.

For example, we've developed ASSIST, the Army System for Standardized Intelligence Support Terminals. It will eventually centralize the data from intelligence-gathering units and make it readily accessible.

We've also built a combined Arms Tactical Training Simulator for the Army. CATTS is a computer-based system that gives potential commanders low-cost practice in making battlefield decisions. Users at Fort Benning tell us it provides such realistic simulations that students get much more out of field exercises than they would without such training.





Then there are the Fleet Command Centers we're developing with the Navy. These systems help FLTCINCs and the CNO to manage any level of crisis; they also interface with WWMCCS, the Worldwide Military Command and Control System.

At the uppermost levels of command and control, we are supporting the Defense Communications Agency's development of a master plan for MEECN, the Minimum Essential Emergency Communications Network. Our modeling of advanced systems and concepts helps to give decision-makers a quantitative basis for achieving an optimum balance between adding to network survivability and meeting other vital defense objectives.

If you are interested in putting this broad-gage C^3 capability to work in your area of the national defense program, we invite you to write and let us know your specific needs.

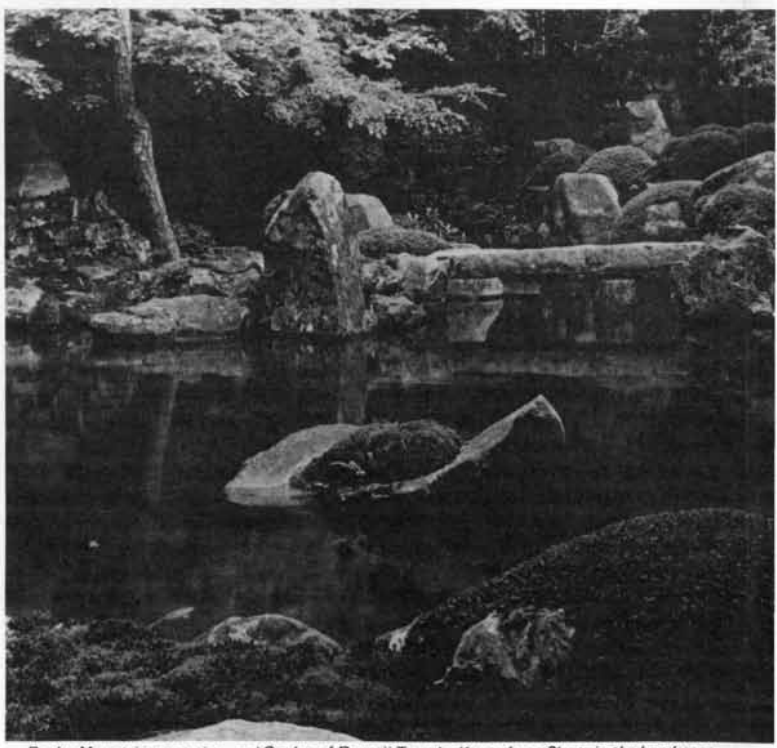
TRW®
SYSTEMS GROUP

Attention:
Marketing Communications,
E2/9043
One Space Park,
Redondo Beach,
California 90278

Freeman Books on Land and Landscape

Examine them free for 15 days.

Sierra Nevada, photograph by Shelton from *Geology Illustrated*



Rocky Mountain vegetation, and Garden of Rengeji Temple, Kyoto from *Plants in the Landscape*

GEOLOGY ILLUSTRATED

John S. Shelton, Drawings by Hal Shelton

Geologist/pilot Dr. John Shelton has been taking aerial photographs since 1939, and this is his master work. *Geology Illustrated* is a breathtaking view of our home planet: its rocks, minerals, volcanoes, glaciers, lakes, rivers, earthquake faults, mountain ranges, and other features. The nearly 400 photographs and crystal-clear text provide an enjoyable way to understand geology. Shelton addresses the book to "thoughtful people who enjoy the outdoors"—to hikers, rockhounds, pilots (and passengers), environmentalists, and many others. "An artist of aerial photography, Shelton uses some 400 of his finest photos to illuminate a discussion of the whole-earth system. Not a traditional textbook, but a fascinating exploration of problems posed by asking, 'How did that come about?' Worth buying for the photos and book design alone, but you'll probably find yourself becoming interested in geology regardless of your original intentions."
—*Whole Earth Catalog,* Summer 1971

1966, 434 pages, 382 illustrations, $14.50

PLANTS IN THE LANDSCAPE

by Philip L. Carpenter, Theodore D. Walker, and Frederick O. Lanphear, all of Purdue University

Plants in the Landscape is a unique introduction to the principles and practices of ornamental horticulture in landscape architecture. No other single volume covers such a broad range of topics. Here is a wealth of information on subjects ranging from the landscape practices of ancient civilizations to no-nonsense, dollars-and-cents descriptions of the present-day landscaping industry. The focus throughout is on landscape plants: how to select and maintain them for healthy growth in a wide range of settings.

Plants in the Landscape is a beautiful book. The text is closely integrated with more than 400 illustrations (about 300 photographs). Whether you are building a new home or remodeling an old one, whether you are a professional landscaper or a weekend gardener, you'll find *Plants in the Landscape* a valuable addition to your library.

1975, 481 pages, 534 illustrations, 32 tables, $16.00

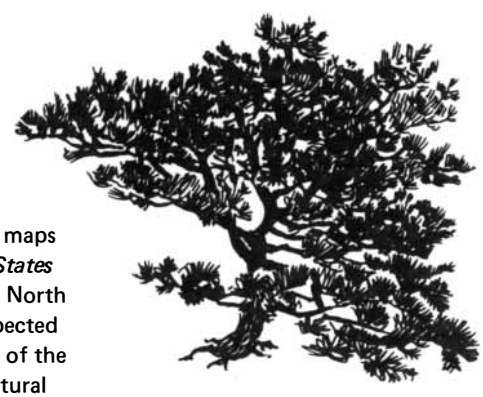
NATURAL REGIONS OF THE UNITED STATES AND CANADA

Charles B. Hunt, The Johns Hopkins University

With its generous use of photographs, line drawings, and maps (558 illustrations in all), *Natural Regions of the United States and Canada* provides a fascinating environmental tour of North America. Charles B. Hunt, a widely experienced and respected scholar, describes and explains the physical environment of the North American continent, Puerto Rico, and Hawaii. Natural regions are discussed in terms of their topography, geology, climate, soil, water, plants, and the interrelationship among all these factors.

"Just open this book to an area you already know. Within pages, the landforms, soils, colors, plants, and importance of your watershed will intensify. An encyclopedia of physiography written in academic style. Lots of maps. Lots of geologic history. The best book to practice mapping North America into your brain."—*Whole Earth Epilog,* 1974

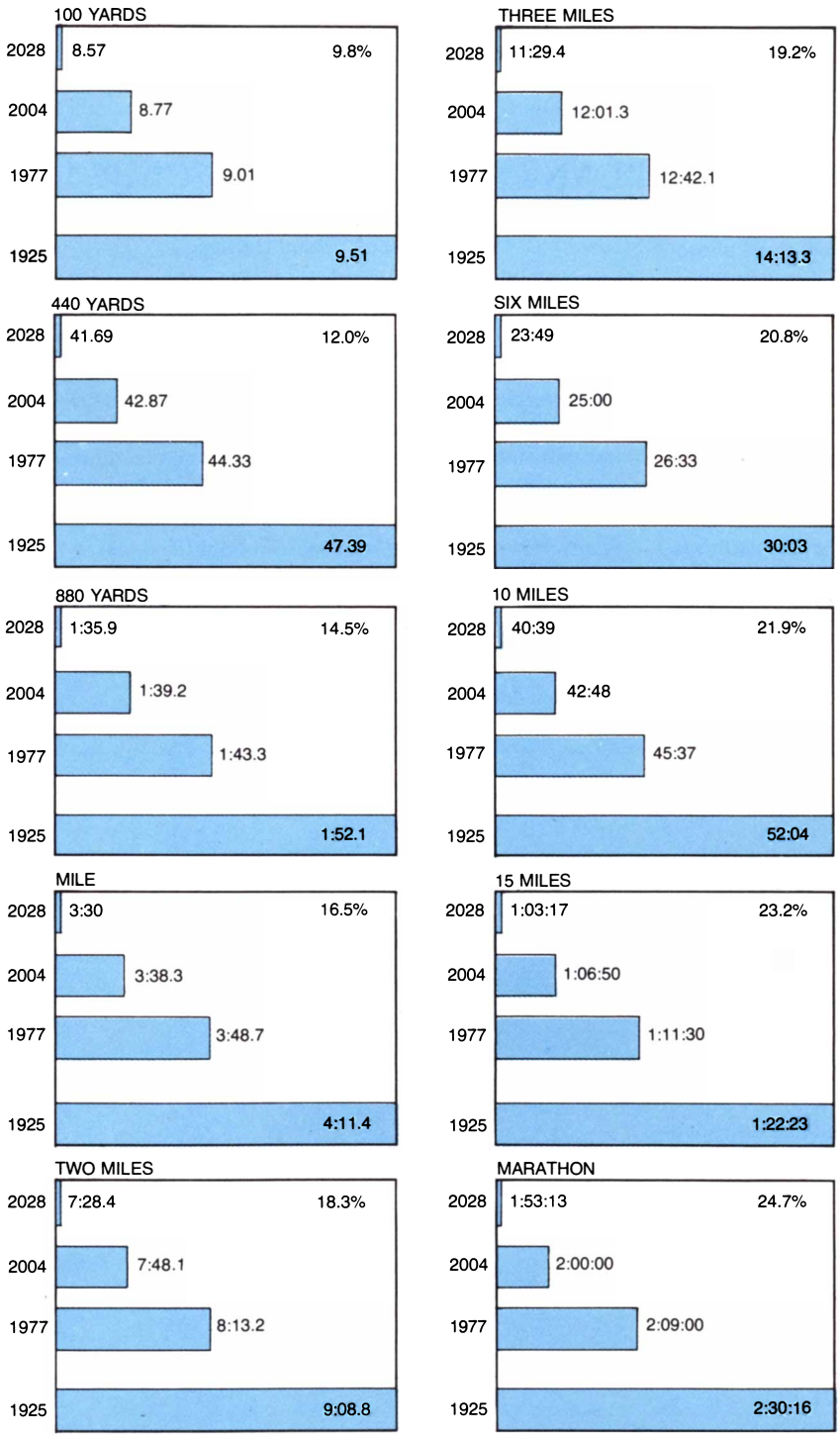
1974, 725 pages, 558 illustrations, 22 tables, $15.95



TO ORDER, use the adjacent card. If card is missing, send order with payment to:

W. H. FREEMAN AND COMPANY
Dept. SA, 660 Market Street, San Francisco, CA 94104

ple, the mile is projected to be run in 3:30 in the year 2028. (We should emphasize that this is a projection, not a prediction.) The mean speed would be the same as the mean speed at which the present record for the half mile was set. Similarly, it is projected that the quarter mile (440 yards) will be run at the rate achieved by Charles Paddock of the U.S. in setting the record for the 220-yard dash in 1921 and that the half mile will be run at the rate achieved by J. E. Meredith of the U.S. in the quarter mile in 1912. The mean speeds of the projected records seem to be well within present capabilities. The projections, however, imply that endurance will improve by a factor of two, and it is uncertain whether such an improvement in endurance will ever be within the limits of human performance at those relatively short distances.



FUTURE PERFORMANCE in footracing at a number of nonmetric distances is projected on the basis of the improvement in mean speed for each event since 1925. The percentage figure associated with each distance reflects the improvement in time from the record set in 1925 to the record projected for the year 2028. Hours, minutes and seconds are separated by colons; the numerals following each decimal point represent tenths or hundredths of a second. In the shorter distances (100 and 440 yards) the times shown are seconds and fractions of a second.

Many factors influence speed and endurance in one way or another. They include the date of the race, the challenge the race presents, age, talent, health, nutrition, drugs, hormones, body build, running economy, aerobic capacity, weight, body fat, training methods, coaching, tactics, self-discipline, the condition of the track, the state of the weather and so on. These variables are by no means mutually exclusive. Although we have no direct information on their effect, we believe a somewhat more detailed analysis of the physical meaning of the speed and date distributions reflected in the accompanying illustrations [*pages 112 and 113*] will shed light on how the variables influence performance.

The form of the speed-date distribution for each distance is the same as the simple proportional relation between stress and strain that is seen when a piece of metal is bent. With runners one can consider the stress as the speed required to beat the record and the strain as the effort (training) needed to run that fast and that far. We have no measure of the intensity of the training effort other than the reported distances and speeds at which athletes have trained over the years. It is a rather gross estimate at best. We cannot do much more than say that the aspiring runner has no more information than this about how long and how hard to work. He learns of the performances and training methods of the runners who have set records at his distance and makes what we assume to be a proportionately greater training effort than they did in terms of speed and endurance.

The specific training schedules that runners have followed show a great variety. Many of them have led to records. There must be a factor common to all of them that explains their success. This key factor seems to be related to the spectacular improvement in specific endurance: the ability to perform sustained work at a high input of energy.

One can regard the runner as having two adaptive needs: (1) to induce changes that will enable him to better meet the demands of his external environment and (2) to maintain homeostasis, the essential constancy of his internal environment. Whatever the mechanisms of his adaptation may be, his first consideration is to obtain maximal exposure to the conditions under which he will run. He must approach as a limit (a distant one) the condition of spending his life running at race pace.

With the repeated application of this training stress he specifically adapts by making the respiratory, circulatory, enzymatic, musculoskeletal, endocrinal and psychological adjustments necessary for him to perform most efficiently under that stress. One sees here an example of a principle

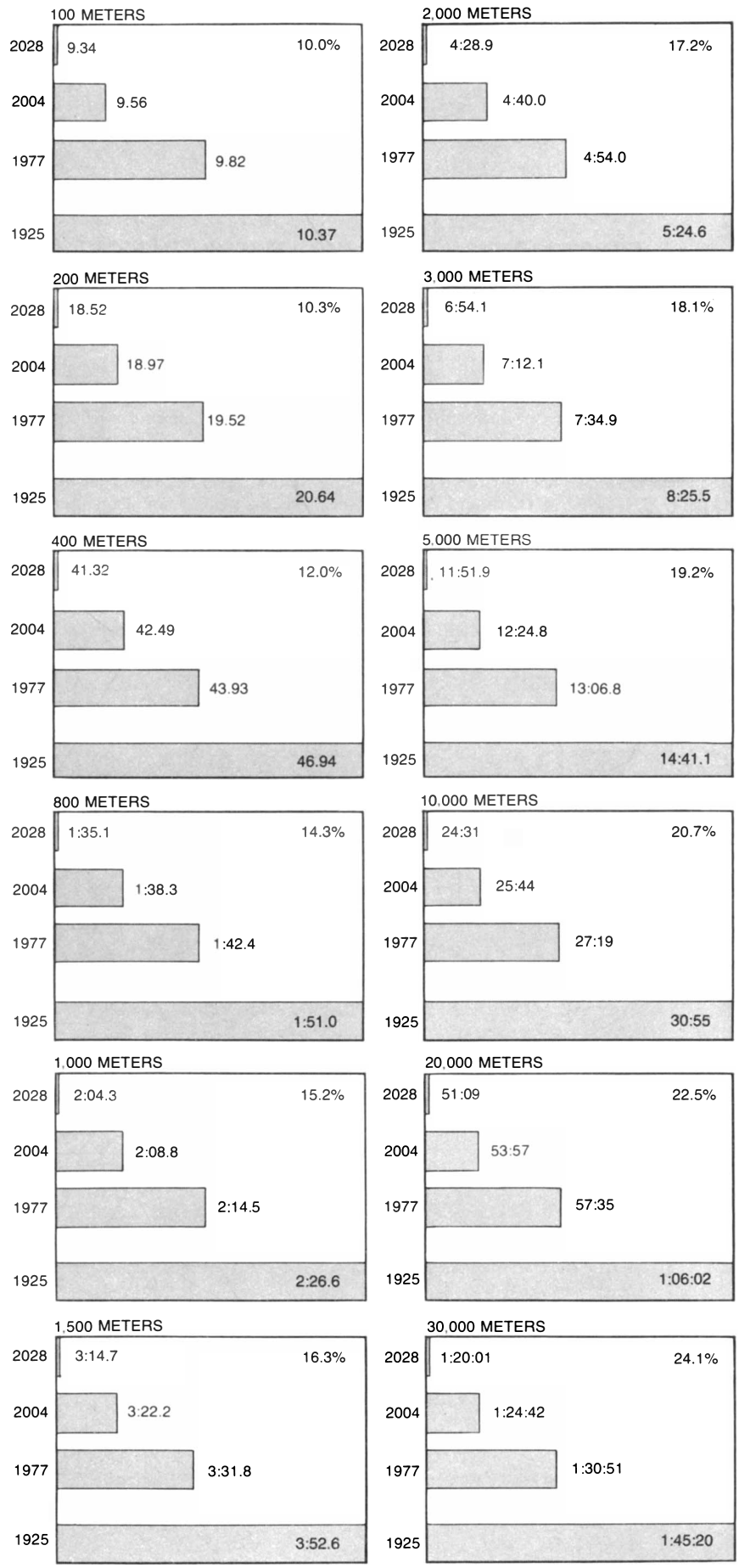
put forward by the French chemist Henry Louis Le Châtelier: If one of the factors of a system in equilibrium is changed, as by a stress, another factor or more than one will adjust to reduce the stress. As far as cellular metabolism in muscle alone is concerned, endurance training has been found to induce adaptive changes in respiration by tissues, in aerobic and anaerobic enzyme systems and in systems that store energy.

Because of such adaptations, new champions can run aerobically at speeds that forced their predecessors to run anaerobically. If the linear plots of speed and date are regarded as stress-strain curves, or curves reflecting rate of work and proportionate response, the pragmatic substitution of a date for strain or effort is seen as an artifact. In speculating about future performance in footracing on the basis of the improvement in the past one tends to focus on dates instead of the factor correlated with the passage of time, namely training effort, which is the level of strain in the past that has led to record performance.

The trained runner has satisfied both of his adaptive needs. Thanks to Le Châtelier's principle he has shifted his metabolic equilibrium so that he can run aerobically at an intense rate of work. Indeed, the literature of physiology contains no example of such adaptability that is better documented. The runner has achieved this adaptation while maintaining homeostasis, so that the body is employing its most efficient mechanisms for the conversion of stored reserves into energy.

If the challenge the runner has to master is not the race but the daily training, the factors that limit his training (not his racing) are what he must overcome in order to set a new record. The physiological, pathological and psychological reactions that limit proportionate increases in the intensity of training are already distressingly familiar to coaches and sports physicians. At present the factor limiting record performance may be pathological or psychological, but it is not physiological.

The physiological limits will appear when challengers can no longer be protected from breaking down while making a proportionately greater training effort or when such an effort is no longer followed by an increased response. Although these two considerations are dependent on time to a certain extent, neither is a function of date, so that projections based on past performance should not hold up in the future. With the provision of consummate challenges, skillful coaching and specialized medical supervision, running at every distance could improve dramatically. Indeed, the improvement is so firmly based on the psychological, physiological and biochemical nature of man that it is likely to continue for an indefinite number of years.



METRIC RECORDS are projected on the same basis as in the illustration on the opposite page. Times given for the first distances are in seconds and fractions of a second.

MATHEMATICAL GAMES

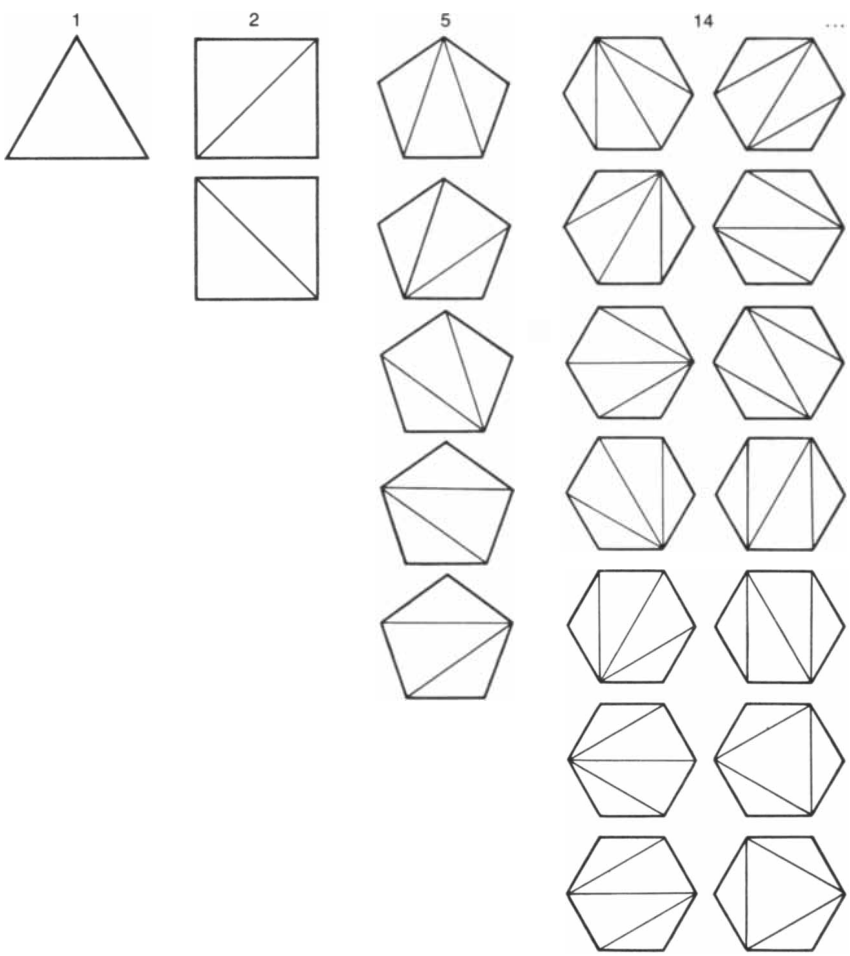
Catalan numbers: an integer sequence that materializes in unexpected places

by Martin Gardner

Any mathematician or scientist is highly likely to encounter infinite sequences of positive integers. If the sequence is simple, such as the doubling series (1, 2, 4, 8, 16, ...) or the squares (1, 4, 9, 16, 25, ...), it is immediately recognizable. And few mathematicians would fail to recognize the Fibonacci numbers (1, 1, 2, 3, 5, 8, ...) or the triangular numbers (1, 3, 6, 10, 15, 21, ...). If the sequence is unfamiliar, however, an enormous amount of time can be wasted searching for a recursive or nonrecursive procedure that generates the sequence. (A procedure is recursive if calculating a next term calls for knowledge of the preceding terms; a nonrecursive formula gives the *n*th term without such knowledge.)

It is hard to believe, but it was not until 1973 that *A Handbook of Integer Sequences* was published. This invaluable tool, compiled by N. J. A. Sloane of Bell Laboratories, lists more than 2,300 integer sequences in numerical order. A mathematician who encounters a puzzling sequence no longer needs to spend hours trying to find its generating formula. He simply looks for the sequence in Sloane's book. The chances are excellent that it is there, followed by a list of references where the reader can check on the nature of the beast.

Our topic this month is the *Handbook*'s sequence 577: 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796,.... The components of this sequence are called Catalan numbers. They are not as well known as Fibonacci numbers, but they have the same delightful propensity for popping up unexpectedly, particularly in combinatorial problems. In 1971 Henry W. Gould, a mathematician at West Virginia University (Morgantown, W.Va. 26506), issued a bibliography of 243 references on Catalan numbers; in many cases the authors were not even aware they were dealing with a sequence known for more than two centuries. This year Gould increased the number of references to 450. (Interested readers can get the entire bibliography postpaid from Gould for $2.) Indeed, the Catalan sequence is probably the most frequently encountered sequence that is still obscure enough to cause mathematicians lacking access to Sloane's *Handbook* to expend inordinate amounts of energy rediscovering formulas that were worked out long ago.



Leonhard Euler's polygon-triangulation problem

It was Leonhard Euler who first discovered the Catalan numbers after asking himself: In how many ways can a fixed convex polygon be divided into triangles by drawing diagonals that do not intersect? An example can be provided with triangles, quadrilaterals, pentagons and hexagons [*see illustration at left*]. Note that in every case, regardless of how the *n*-gon is triangulated, the number of diagonals is always $n - 3$ and the number of triangles is $n - 2$. It is easy to prove that this relation holds in general. The possible triangulations for each of these four polygons are the first four terms of the Catalan sequence.

Applying an induction process that he described as "quite laborious," Euler obtained the following recursive formula:

$$\frac{2 \times 6 \times 10 \times \ldots (4n - 10)}{(n - 1)!}.$$

Numbers above the line have the form $(4n - 10)$, where n is a positive integer greater than 2. The exclamation mark is of course the factorial sign. It stands for the product of all positive integers from 1 through the preceding expression. For example, if $n = 6$ (the sides of a hexagon), the formula becomes

$$\frac{2 \times 6 \times 10 \times 14}{5!} = 14.$$

Unusually simple recursive formulas are obtained by putting another 1 in front of the series: 1, 1, 2, 5, 14,.... Let k be the last

number of a partial sequence and n the position of the next number. The next number is then

$$\frac{k(4n-6)}{n}.$$

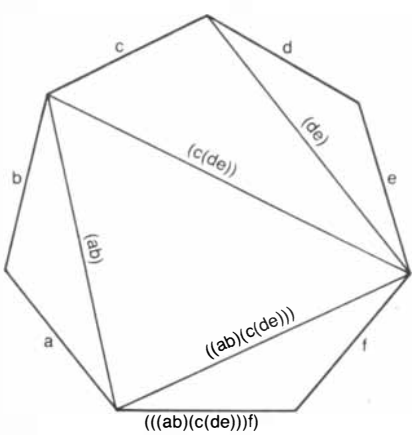
Johann Andreas von Segner, Euler's 18th-century contemporary, found a whimsical recursive procedure for the same form of the Catalan sequence. Write the partial sequence forward, then put below it the same numbers in backward order. Multiply each top number by the one below it and add all the products; the result is the next number of the sequence. For example,

$$\begin{array}{r} 1 \quad 1 \quad 2 \quad 5 \quad 14 \\ \times \underline{14 \quad 5 \quad 2 \quad 1 \quad 1} \\ 14 + 5 + 4 + 5 + 14 = 42\,. \end{array}$$

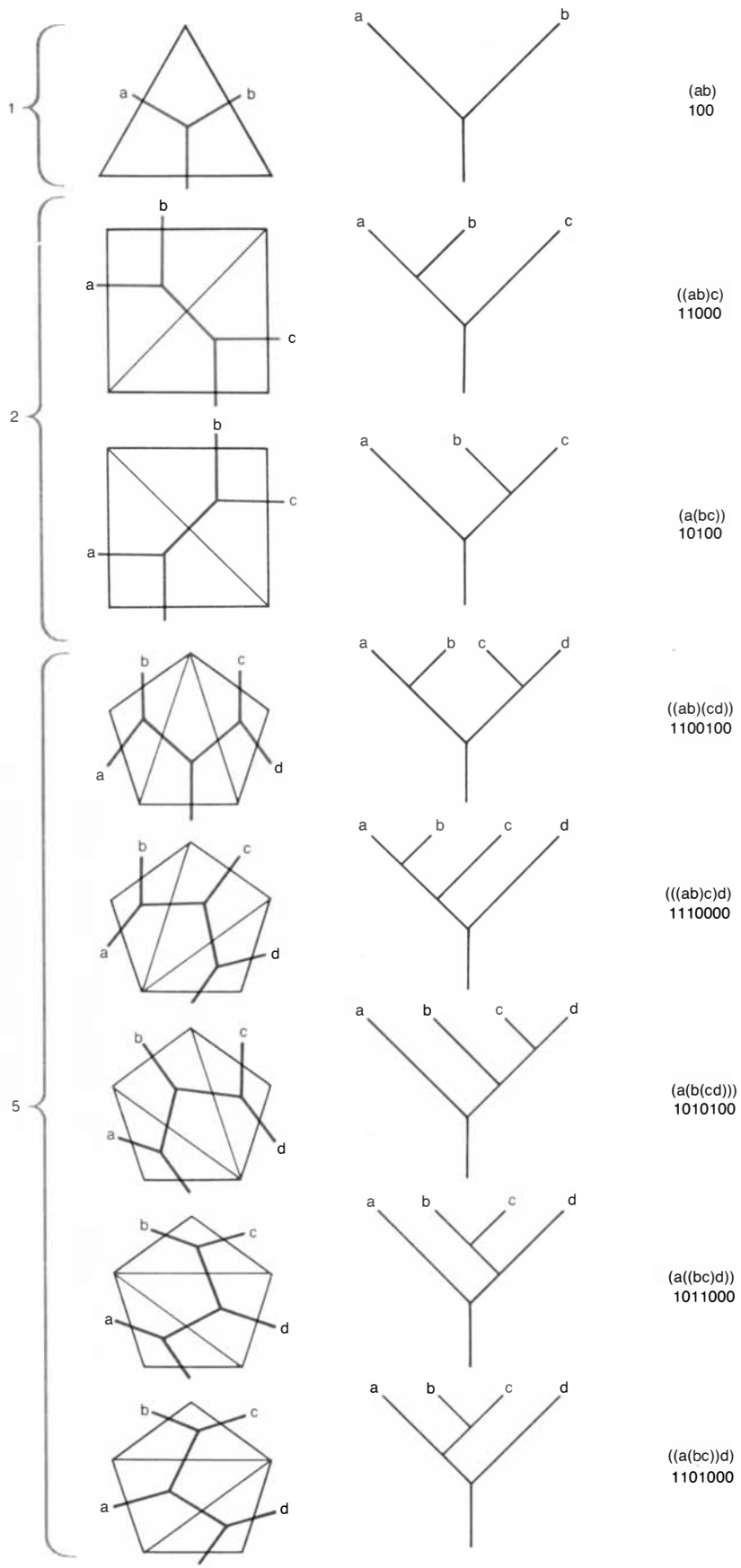
Euler's polygon triangulation is isomorphic with some seemingly unrelated problems. It was Eugene Charles Catalan, the Belgian mathematician for whom the sequence is named, who in 1838 solved the following problem. We have a chain of n letters in a fixed order. We want to add $n-1$ pairs of parentheses so that inside each pair of left and right parentheses there are two "terms." These paired terms can be any two adjacent letters, or a letter and an adjacent parenthetical grouping, or adjacent groupings. In how many ways can the chain be parenthesized?

For two letters, *ab,* there is only one way: (*ab*). For three letters there are two ways: ((*ab*)*c*) and (*a*(*bc*)). For four letters there are five ways: ((*ab*)(*cd*)), (((*ab*)*c*)*d*), (*a*(*b*(*cd*))), (*a*((*bc*)*d*)) and ((*a*(*bc*))*d*). The numbers of these ways, 1, 2 and 5, are the first three Catalans, and the Catalan sequence enumerates the ways of parenthesizing all longer chains.

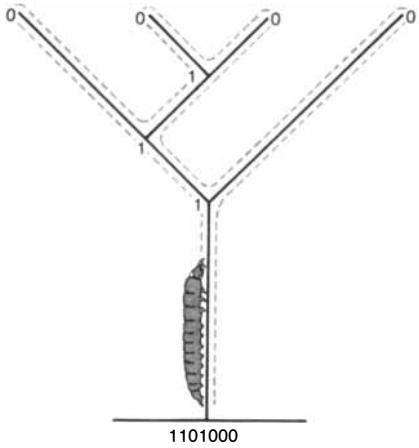
H. G. Forder, writing on Catalan numbers in 1961, showed a simple way to establish one-to-one correspondence between the triangulated polygons and the parenthesized expressions. An example is a triangulated heptagon [*see illustration below*]. Label its sides (excluding the base) *a* through



Parenthesized triangulation of a heptagon



Triangulation and planted trivalent trees



A worm generating a tree's binary number

f. Every diagonal spanning adjacent sides is labeled with the letters of those sides in parentheses. Each remaining diagonal is then lettered in similar fashion by combining the labels on the other two sides of the triangle. The base is lettered last. The expression for the base is uniquely determined by the dissection. If you apply this technique to the polygons portrayed in the illustration on page 120, you will obtain the parenthesized expressions shown at the right in the illustration at the right on the preceding page.

The British mathematician Arthur Cayley proved that Catalan numbers count the number of trees that are planar, trivalent and planted. A "tree" is a connected graph (vertexes joined by edges) that has no circuits. "Planar" means that it is drawn on the oriented plane without intersections. "Planted" means that it has one "trunk," the end of which is called the "root." The graph can thus be drawn to simulate a tree growing up from the ground. "Trivalent" means that at each vertex (except at the root and at the ends of branches) the tree forks to create a point where three edges meet.

The illustration of this structure on the preceding page is almost self-explanatory. The colored lines show how each triangulation corresponds to a planted trivalent tree. Next to the polygons corresponding trees are drawn in conventional form. It is easy to see how the grouping of a tree's branches corresponds to its parenthesized expression. Below each expression we convert it to a binary number by replacing every left-hand parenthesis with 1 and every letter with 0, ignoring all right-hand parentheses. These binary numbers are convenient shorthand ways of designating the polygon dissections and their trees. Right-hand parentheses are not needed because, given the left-hand ones and the method of grouping letters, the right-hand parentheses can always be added in a unique manner.

The Polish mathematician Jan Lukasiewicz found a pleasant way to obtain each tree's binary number [*see top illustration on this page*]. Picture a tree with four top ends. They are labeled 0 and the trivalent vertexes are labeled 1. Imagine a worm crawling up the trunk and around the entire tree along the broken path in the illustration. At each vertex the worm calls out the label. Once a vertex is called it is not called again. In this example the worm calls out 1101000, which proves to be the very binary number we obtained from the tree's parenthesized expression.

In 1964 it was discovered that normal planted trees are also counted by Catalan numbers. They are planted trees of n vertexes, including the ends but not the root. They can also be described as planted trees of n edges. A vertex in such a tree can have any valence.

Many ways have been found for showing a one-to-one correspondence between trees of this kind and the planted trivalent trees. The simplest one was pointed out by Frank Bernhart [*see bottom illustration on this page*]. The trivalent trees are drawn so that at each vertex of valence 3 the edges point up, down and to the right. Imagine that each horizontal edge [*shown in color*] shrinks to a point and disappears. If there is a trivalent vertex at the right end of the edge, it is carried to the left to merge with the vertex at the left. All the vertical edges remain distinct. This simple transformation changes all planted trivalent trees of n ends into all planted trees of n edges.

A worm crawling up and around any tree in this new set (the trees at the right in the illustration) will call out the same binary number as the tree's partner if it alters its procedure as follows. The worm starts at the bottom vertex rather than at the root. Each time it crawls up an edge it calls 1 and each time it crawls down an edge it calls 0.

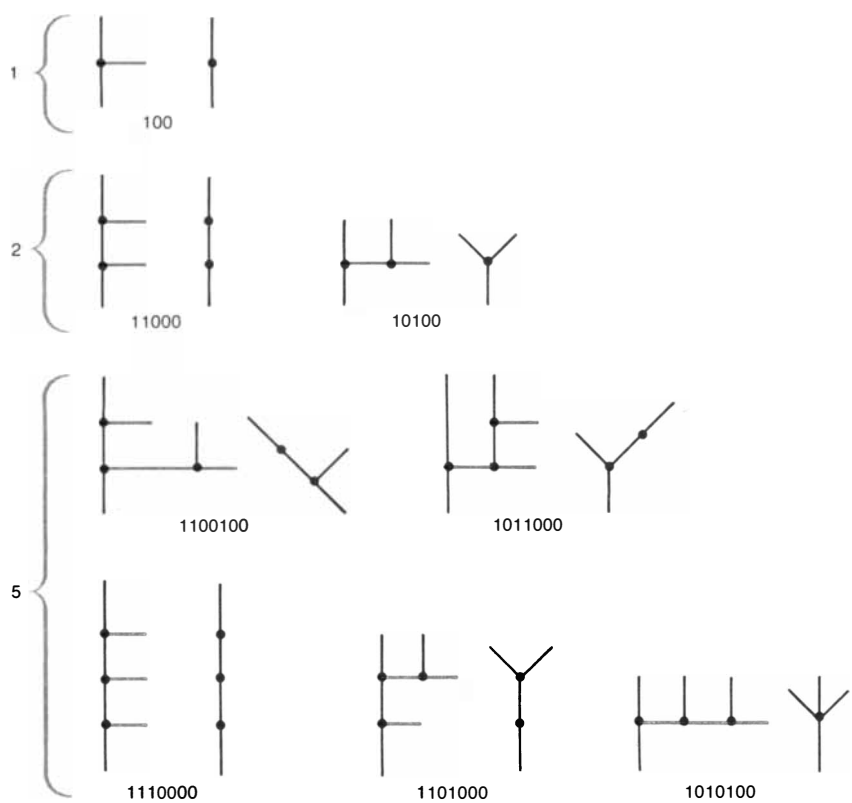
Consider chessboards of sides 2, 3, 4,.... All squares north and west of a main diagonal are shaded [*see illustration on page 124*]. We are to move the rook from the lower left corner to the upper right corner. It cannot enter a shaded cell, and its only allowed movements are north or east. For a board of side n how many different paths can the rook take?

Once more the Catalans give the answer. Below each board of side n write the binary number for the planted trivalent tree of n ends. Taking the binary digits from left to right, move the rook one square to the right for each 1 and one square up for each 0. (The final digit is ignored.) This pattern generates a path, and in this way all the rook paths are obtained.

Here are seven more recreational problems solved by the Catalans. For the first five I shall indicate how the corresponding binary numbers (ignoring final digits) solve the problem.

1. Two men, A and B, are running for office. Each man gets n votes. In how many ways can the $2n$ votes be counted so that at no time is A behind B? (1 = vote for A, 0 = vote for B.)

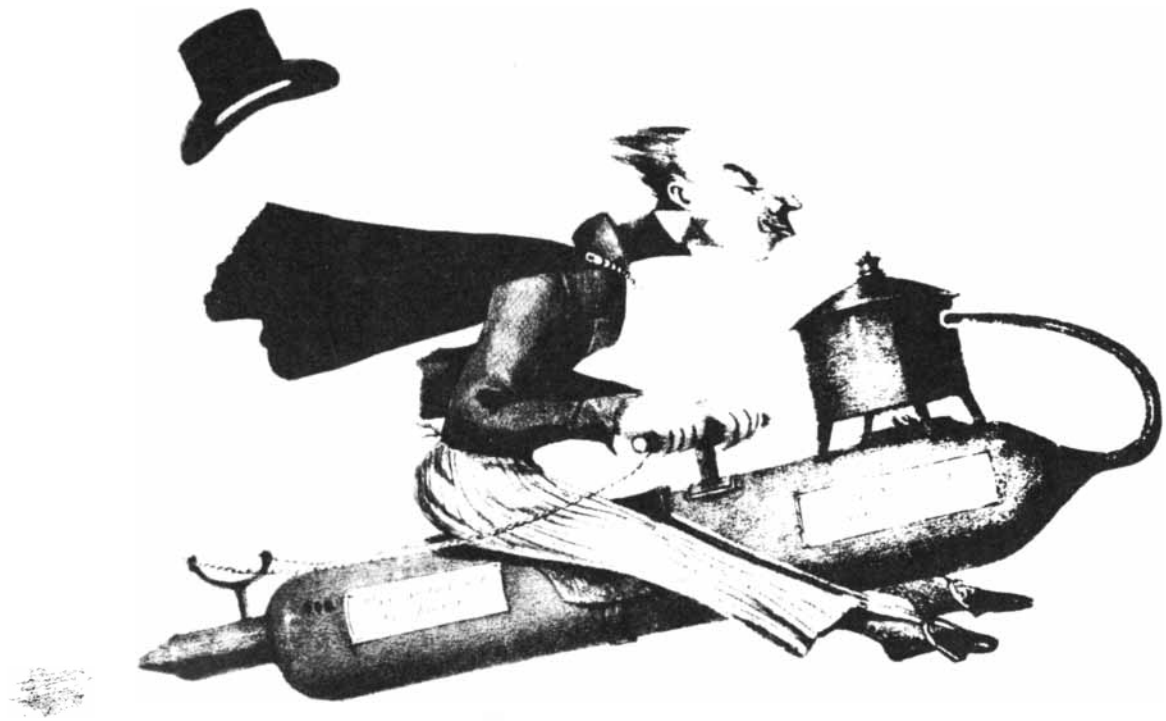
2. Place a penny, a nickel and a dime in a row. On the penny put a stack of n face-up playing cards with values in consecutive order from bottom to top. The cards are moved one at a time from the penny to the nickel or from the nickel to the dime. (No other moves are allowed.) By mixing these two types of moves you will end, after $2n$ moves, with all the cards on the dime. Giv-



Transformation of planted trivalent trees to planted normal trees

THESE BOOKS SOLVE PROBLEMS

See for Yourself: Examine Them Free for 15 Days



CONCEPTUAL BLOCKBUSTING

A Guide to Better Ideas

James L. Adams, Stanford University

Conceptual blocks are those frustrating mental walls that get between a problem and its solution. *Conceptual Blockbusting* is a witty, practical, and delightfully illustrated guide to breaking down these walls and improving your creative abilities.

Engineer-artist James L. Adams provides many helpful suggestions for identifying and eliminating the most vexing creative blocks:

*How to avoid the physical, cultural, and emotional restraints that so often stymy problem solvers

*How to "see" the true dimensions of a problem (and how to tell when a problem really doesn't exist)

*How to choose the correct "thinking language."

Conceptual Blockbusting also introduces readers to the classical theoretical writings on creativity—by Freud, Jung, Maslow, and others—and to the modern techniques of applied creativity such as brainstorming and Synectics. No matter what your background, you'll find that reading this book is a fine way to start limbering up your "mental muscle." "The best guide to creative thinking I have yet encountered. . . . An amusing, pictorial, and insightful journey through the mind."—*Human Behavior,* May 1975

1974, 137 pp., 42 illus., softcover $3.95

HOW TO SOLVE PROBLEMS

Elements of a Theory of Problems and Problem Solving

Wayne A. Wickelgren, University of Oregon

How to Solve Problems analyzes and systematizes some of the basic methods of solving mathematical problems. Each method—inference, classification of action sequences, state evaluation and hill climbing, subgoals, contradiction, working backward, and relations between problems—is considered in a separate chapter. The methods are described in terms of a modern theory derived from research in artificial intelligence and computer simulation of thinking. Examples include chess problems, logical puzzles, and railroad switching problems as well as problems frequently encountered in science and engineering. The author provides hints to the solutions, allowing the reader to arrive at them independently. Complete solutions are provided after the hints. With the exception of material in the final chapter, the book can be understood by readers with no more mathematical background than high school algebra and plane geometry. Whether your interest in problem solving is professional, recreational, or both, you will find this a helpful book.

1974, 262 pp., 73 illus., softcover $6.00

TO ORDER, use the adjacent card. If card is missing, send order with payment to:

W. H. FREEMAN AND COMPANY
Dept. SA, 660 Market Street, San Francisco, CA 94104

en n cards, how many different permutations can you achieve on the dime? (1 = move from penny to nickel, 0 = move from nickel to dime.)

3. An inebriated man leaves the door of a bar and staggers straight ahead. His steps are equal, but before each step he has a random choice of going forward or backward. How many ways can he take $2n$ steps that will return him to the door? (1 = step forward, 0 = step back.)

This random walk can be given other forms. A king starts on the first row of a chessboard and moves one square forward or back along the file to end on the starting square after $2n$ moves. Draw a space-time diagram of the moves, with time measured along the horizontal base line. The zigzag path can be viewed as the profile of a mountain with peaks an integral number of miles high and no peak higher than n miles. The paths depict all mountain ranges of this type.

4. An even number ($2n$) of soldiers, no two the same height, line up in two equal rows, *A* and *B*. How many ways can they do it so that from left to right in each row the heights are in ascending order and each soldier in row *B* is taller than his counterpart in row *A*? (Number the soldiers 1, 2, 3,... according to increasing height, and number the digits of the binary numbers from left to right. The 1 digits give the numbers for row *A* and the 0 digits give the numbers for row *B*. The problem is easily modeled with playing cards.)

5. Tickets are 50 cents, and $2n$ customers stand in a queue at the ticket window. Half of them have $1 each and the others have 50 cents each. The cashier starts with no money. How many arrangements of the queue are possible with the proviso that the cashier always be able to make change? (1 = 50 cents, 0 = $1.)

6. Hexaflexagons are curious toys made by folding straight or crooked strips of paper into a hexagonal structure that alters its "faces" when it is flexed. (They are described in the first chapter of my *Scientific American Book of Mathematical Puzzles & Diversions*.) A regular hexaflexagon of a specified type passes through different states as it is flexed. The total number of states, for all varieties of regular hexaflexagons of n faces, is a Catalan number. For example, a hexahexaflexagon (six faces) can be made in three ways. The total number of states is the Catalan number 42.

If we ignore the states and ask in how many essentially different ways a regular hexaflexagon of n faces can be made, the answer is provided by a sequence that counts the triangulations of convex polygons when rotations and reflections are excluded. This remarkable sequence (No. 942 in Sloane's *Handbook*) is 1, 3, 4, 12, 27, 82, 228, 733, 2282, 7528,....

In unpublished papers Bernhart and other flexagation addicts describe ways of mapping the changes of states, as a flexagon of n faces is flexed, by tracing paths around the lines of a triangulated polygon of $n + 1$ sides.

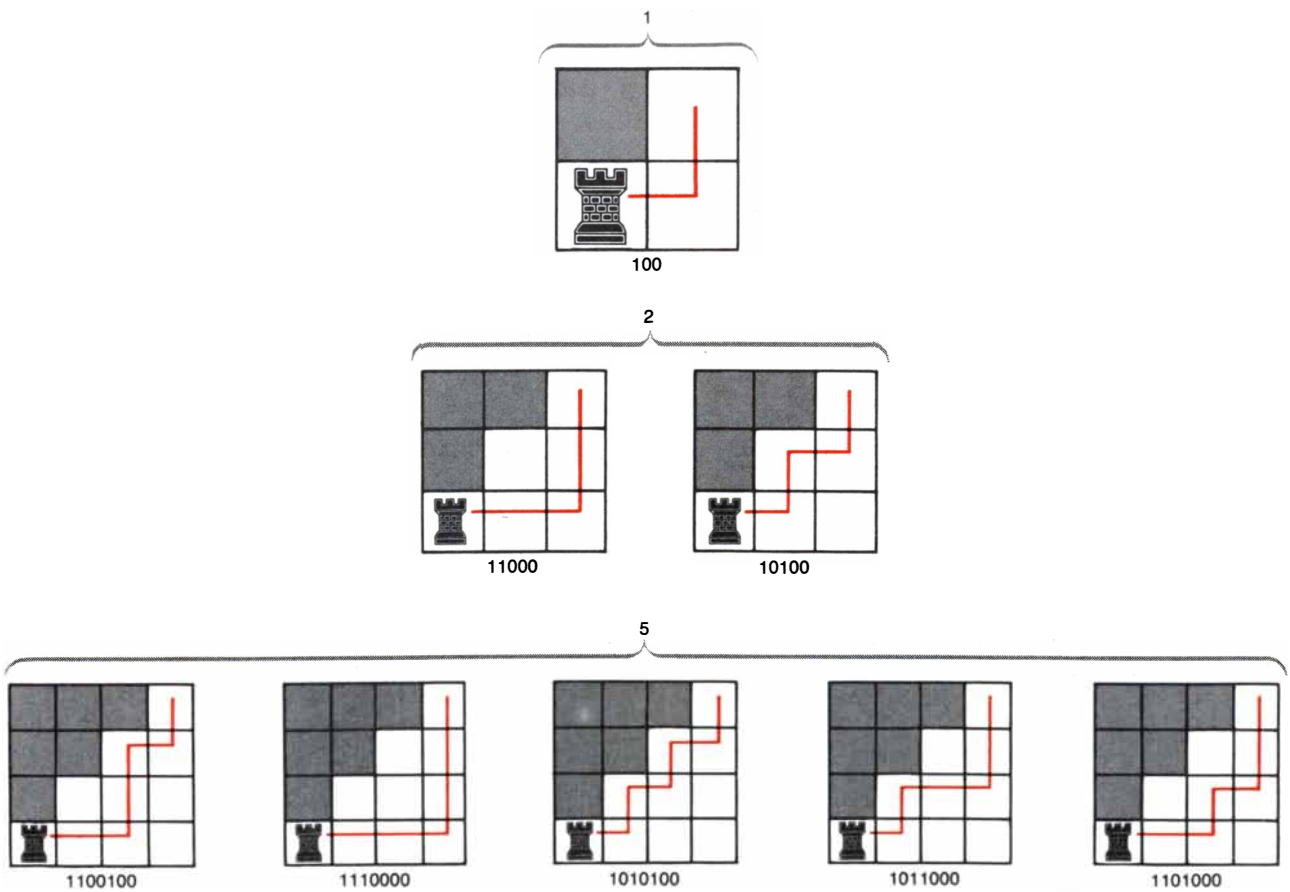
7. An even number of people are seated around a circular table. Each extends one arm and they clasp hands in pairs, but in such a way that no pair of joined arms crosses another. Given the number of pairs, in how many ways can this be done? More precisely, place $2n$ spots in fixed positions on the circumference of a circle and then find all the ways they can be paired by drawing nonintersecting chords.

Can you find a simple geometric way to establish a one-to-one correspondence between this problem and any of the above problems? Next month I shall give a pretty way of doing it that I obtained from Bernhart.

A nonrecursive formula for the nth Catalan has different forms depending on how the positions of the Catalans are numbered. The formula is simplest if the sequence begins 1, 2, 5,.... In this numbering the nth Catalan is

$$\frac{(2n)!}{n!(n+1)!}.$$

If the series begins 1, 1, 2, 5,..., it turns out that odd Catalan numbers greater than 1 appear at all positions, and only at those



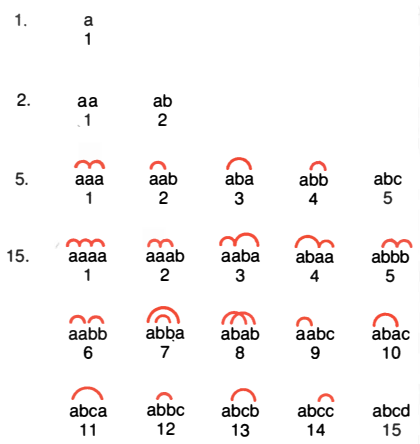
How Catalan numbers count a rook's paths

positions, that are powers of 2. Thus the fourth, eighth, 16th and so on Catalans are odd. This is only one of many unusual properties of the sequence that have been discovered.

A word of caution: When one works on combinatorial problems, it is easy to confuse the Catalan sequence with a closely related one: 1, 2, 5, 15, 52, 203, 877,.... As Gould points out in notes on his bibliography (which also includes a separate listing of references on the above series), when structures are complicated, it is easy to miss a 15th structure (when $n = 4$) and be tricked into supposing you have encountered a Catalan sequence. The numbers are called Bell numbers after Eric Temple Bell, who published a lot about them. They count the partitions of n elements. For example, the number of rhyme schemes for a stanza of n lines is a Bell number. A quatrain has 15 possible rhyme schemes. A 14-line sonnet, if convention is thrown to the winds, can have 190,899,322 (the 14th Bell) distinct rhyme schemes. But, you may object, who would write a sonnet with a rhyme scheme such as *aaaaaaaa aaaaaa*? Allowing a word to rhyme with itself, James Branch Cabell conceals just such a sonnet (each line ending with "love") in Chapter 14 of *Jurgen*.

The illustration on this page shows how Bell numbers count the rhyme schemes for stanzas of one line through four lines. Lines that rhyme are joined by curves. Note that not until we get to quatrains does a pattern (No. 8) require an intersection. Joanne Growney, who worked this arrangement out in 1970 for her doctoral thesis, calls the schemes that do not force an intersection of curves "planar rhyme schemes." Bell numbers count all rhyme schemes. Catalan numbers are a sub-sequence that counts planar rhyme schemes.

The Bell sequence is No. 585 in Sloane's *Handbook*. But the Bells chime another story that we must postpone until a future column.



How Bell numbers count rhyme schemes

Many readers quite properly chided me for carelessness when I described (in March) Colin R. Blyth's paradox of the man and the three pies. It was I (not Blyth) who said that the man's decision was to maximize his "expectation of satisfaction." What he is maximizing is, in Blyth's words, "his best chance" of getting the most satisfying pie. It is a subtle but important difference. Both the dining statistician and the doctor have a choice between two intents: maximizing their average of satisfaction in the long run or maximizing their chance of getting the best pie or drug on a particular occasion.

To put it another way, Blyth's pie eater is minimizing his regret: the probability that he will see a better pie on the next table. His doctor counterpart, as Paul Chernick suggested, could be trying to avoid a malpractice suit that might result if a dissatisfied patient went to another doctor and got more effective treatment. "Is the case of a scientist closer to that of a player in the spinner game," asked George Mavrodes, "or is it closer to that of the statistical pie eater?... I do not know the answer to that question."

John F. Hamilton, Jr., revised the dialogue between the waitress and the statistician as follows:

Waitress: "Which pie will be better tonight, *A* or *B*?"

Statistician: "The odds are on *A*."

Waitress: "What about *A* and *C*?"

Statistician: "Again, *A* will probably win."

Waitress: "I see, you mean *A* will probably be the best of all."

Statistician: "No, actually *C* has the greatest chance of being the best."

Waitress: "Okay, cut the funny stuff. Which pie do you want to order, *A* or *C*?"

Statistician: "Neither. I'll have a slice of *B*, please."

The paradoxes of confirmation are not, of course, paradoxes in the sense of being contradictions but paradoxes in the wider sense of being counterintuitive results that make nonsense of earlier naïve attempts, by John Stuart Mill and others, to define the meaning of "confirming instance." Philosophers who discuss the paradoxes are not ignorant of statistical theory. It is precisely because statistical theory demands so many careful distinctions that the task of formulating an inductive logic is so difficult.

Since I wrote February's column on packing cubes with $1 \times 2 \times 4$ bricks considerable progress has been made toward solving the general problem of proving impossibility for all odd-order cubes. The second version of my impossibility proof for order 5 (using planar sections) was found by David A. Klarner to generalize in a manner that proves impossibility for all cubes with sides equal to plus or minus 3 (mod 8), thus settling half of the odd orders. Klarner also found an impossibility proof for order 7, and similar proofs were later found by Frank Barnes, Jeffrey C. Lagarias, Dermott O. Murphy and Maupou Patrice. It remains to be determined whether or not any of these proofs can be extended to cover all odd-order cubes.

BOOKS

Wounds, mites, Eadweard Muybridge, machine disasters and ancient Greece viewed from the air

by Philip Morrison

THE HEALING HAND: MAN AND WOUND IN THE ANCIENT WORLD, by Guido Majno. A Commonwealth Fund Book. Harvard University Press ($25). Only one book of Pliny remains to us. It is a monument to obsessive data-collecting. The author wrote only by night, had books read to him during rubdowns and at table and scolded his friends, "Your interruption has cost us ten lines!" This tireless Roman army officer and civil official collected 20,000 facts from 2,000 volumes by 100 authors in the course of two years, just a few years before his death in A.D. 79 (probably of a heart attack as he approached Vesuvius in eruption). His reputation today is low; he is remembered mainly for the mistakes he repeated. Not here, however. The cheerful scholar, skeptic and medical scientist who has produced this extraordinary volume admires the old fellow greatly and cites almost eight pages of his remarks, grouped by topic, with genuine pleasure. (For example: "Physicians acquire their knowledge from our dangers, making experiments at the cost of our lives" and "There is a marvelous neatness in the titles of Greek books, but when you get inside them, good heavens, what a void!")

Professor Majno even found an unnoticed pearl in Pliny. There is a plain account of the juice of a certain plant that can stanch bleeding and cure cough; it was named *ephedron*. We have known the drug, an analogue of adrenalin, only since its efficacy was reported 50 years ago by a young American at Peking Union Medical College who had tested popular Chinese herbs. (It was the one find of 2,000 samples he examined.) The Mediterranean remedy seems to have been independent of the very old Chinese tradition. Strangely enough, the drug had been extracted and purified in the 1880's, but it was ignored because at full strength it poisoned the experimental dogs. "Asthmatics went on coughing and wheezing—and all the while it lay written in the books... that a gentler, impure decoction of *Ephedra* would have brought them instant relief."

Professor Majno is a pathologist "primarily interested in sick people," a specialist in inflammation. He began to write a historical preface to a monograph, he says, and "the preface took over." We are grateful, if not quite persuaded. There is not much about wounds in the citations from Pliny; throughout the book what is opened up is a lighthearted and yet compelling path toward the pleasures and depths of classical learning. The author is a courageous and catholic amateur of history who delights in inference, an experimenter, a questioner and a cultivated writer of imagination, humor and compassion. He has entered the past over a daunting span of space and time with the best of all preparation: real knowledge of the discipline whose history he teases out. He follows the wound (there is no problem with its diagnosis) from before history through Sumer, Egypt, Greece, China, India, Alexandria and Rome—with no narrow view of what he is seeking. He is as interested in how cuneiform is construed as he is in just what some learned editor concludes about contested phrases; he fears no digression and he nourishes the eye with text facsimiles, diagrams, sketches of his own, experimental results and archaeological and artistic finds—a treasure made coherent out of diversity. Although he does not normally seek primary historical novelty (Pliny's drug aside), his work is above all a delectable lesson in how to know the past. The bibliography is critical and extensive. He has written to or talked with the contending scholars; he explains his own tentative best surmises, and he draws the reader along on a journey over the millenniums, through clinical laboratories and into the libraries at one and the same time.

The volume is as much a delight to view as to reflect on, with 15 color plates, hundreds of figures and such elegancies as marginal sketches to identify the chapters. It would be of little use to summarize this rich summary-history; before the reader sets out for the bookstore or the library it will suffice to sample almost telegraphically a half-dozen points. Most of these remarks stand each for an entire well-knit chapter.

The apes administer first aid along with their more obvious grooming. "Nature does most of the job, while the chimps perform a few helpful gestures... *in the pursuit of pleasure*.... But to turn to my surgical colleagues: would anyone deny that cleaning up an untidy wound still is, deep down, a pleasure?"

There is a photograph of a hand impression in the late Paleolithic Gargas cave, showing fingers that appear to be chopped short. Missing-joint counts suggest that what is represented is a sign language of bent fingers, not ritual amputation. And yet the sacrifice of fingers among hunting peoples is known.

Powdered malachite and chrysocolla, two lovely blue-green copper minerals, were typically used for Egyptian eye makeup and paint. Here is an ancient carved head so decorated; with it we see a Petridish test that makes plain the antibiotic effect of the two cosmetic powders. There is a serious function for eye paint; Dioscorides confirmed it. Not for nothing does a fashion endure over five millenniums.

Gaping wounds close to some extent by natural contraction. (One basic mechanism was discovered in the author's Harvard laboratory while he was completing this book: "a good omen.") The contractile tissue draws the closure much more easily when a wound has corners than when it is round; the healing time in rabbits is shortened by a week. Hippocrates was willing to cut to change a circular wound into an oval one; round wounds can heal only by changing shape, and so they are slow to heal and therefore dangerous.

The balm of Gilead was a treatment par excellence for wounds, with "an overtone of salvation." Myrrh is soluble and is a clear-cut bacteriostatic agent. "I conclude that the ancient use of myrrh as a wound drug was fully justified." How did it come to be adopted? First, the resin fills gashes in plants; perhaps that was a clue, conscious or not. Second, wounds used to smell, and so the fragrant resin was welcome. Third, resins never decay in nature; the analogy is real. And best of all, the resins actually helped.

Professor Majno faces up to China (depending of course on scholarly translations) in a brief chapter full of novelty. Readers who recall the Hemingway short stories will be surprised to learn that the Chinese surgeons who made eunuchs, like the unhappy Kansas City kid in the story, took away all the external genitalia. The citation is grim; it describes 1929 practice that cannot have been much easier back in the Han.

Here too is the dread white-hooded cobra and the scheme of Ayurvedic medicine for treating snakebite. The affair is described as the concrete living experience of an ancient doctor on a busy day, not in the abstract codified style of the Sanskrit texts. "Treatments that really treat are so rare in ancient medicine" that the ligature tied above the bite deserves and receives special mention.

The mandibles of a soldier ant are said to be efficacious for clamping a wound; they may have been the first sutures. Majno tried them; see the photograph in color. Indeed, one formicologist (the late T. C. Schneirla) reported that his clothes had been sutured to his skin by the appropriate species. ("I have ceased to doubt about the ancient Hindu sutures.")

There is a loving page about papyrology, that specialty of epigraphy. Police reports and accounts of an enema doctor make intimate the chapter on Alexandria. It is Galen who ends the book. The chapter on him begins with a photograph of the formidable

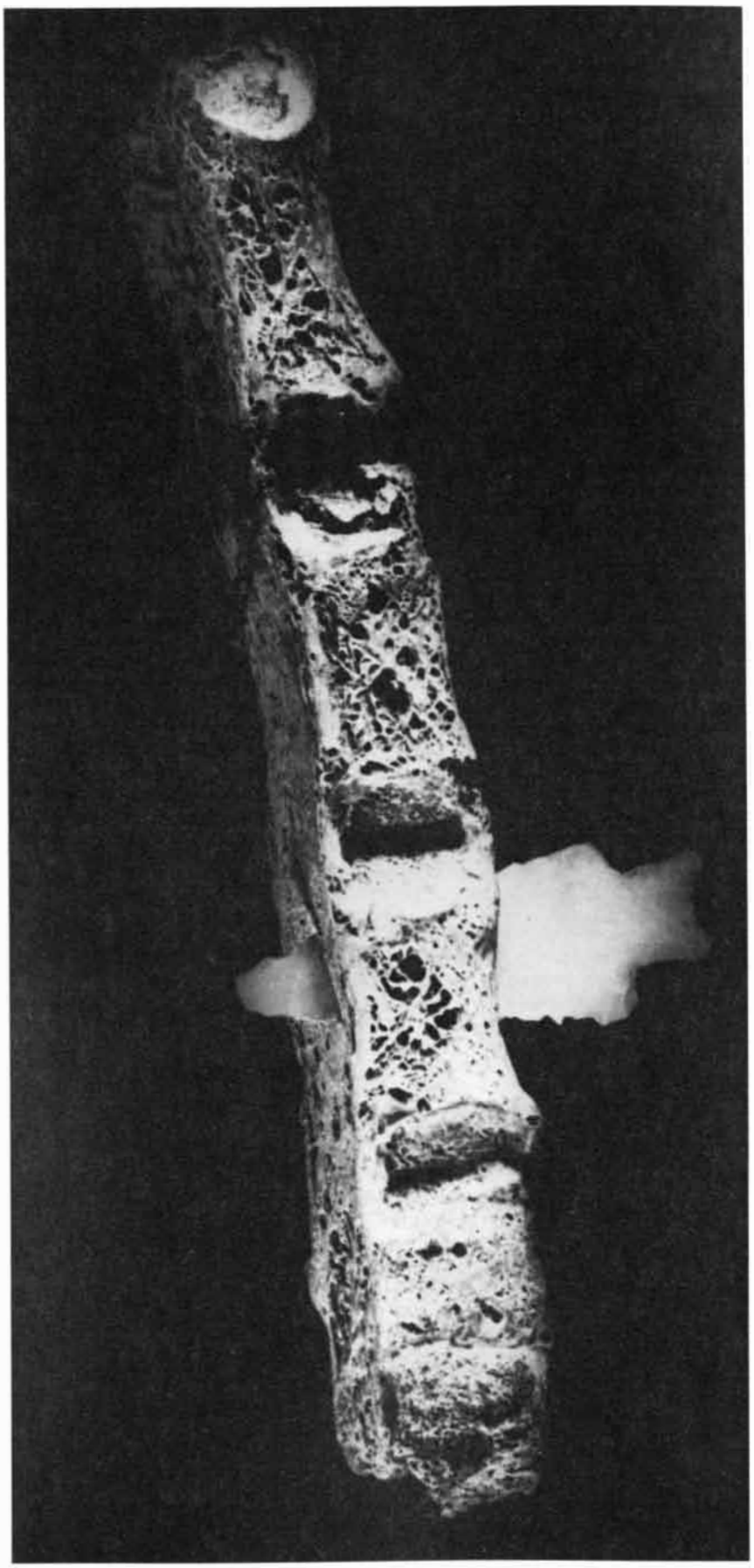
shelf of his works in the only available edition, from the 1830's (admittedly a bit overplayed, since it includes a complete translation from the Greek to Latin). "I must confess that when it came to Galen I broke my own rule: I did *not* read the twenty-two volumes." It is fascinating that medical professors routinely teach that this Greek of the four humors produced a fifth sign of inflammation, that he added to *rubor, tumor, calor* and *dolor* (redness, swelling, heat and pain) another sign: disturbed function. "The concept of a fifth sign is definitely catchy.... It sounds like progress." It is not, however, progressive in Galen. He did not even repeat the four cardinal signs of the Roman Celsus: Galen "consumed gallons of Roman ink, but took no notice of anything Roman." It was the great Rudolf Virchow who enunciated the fifth sign in 1858; soon the handbooks and textbooks included the fifth sign as general wisdom, and in 1919 someone ascribed it to Galen! "After Galen, the history of the wound grinds to a halt for at least one thousand years."

A review can only clumsily transmit the feel of this remarkable book. Few indeed are works that so intimately connect past learning with present experience; the volume is itself a kind of healing, uniting in mutual benefit the classicist and the experimenter, the scholar and the general reader.

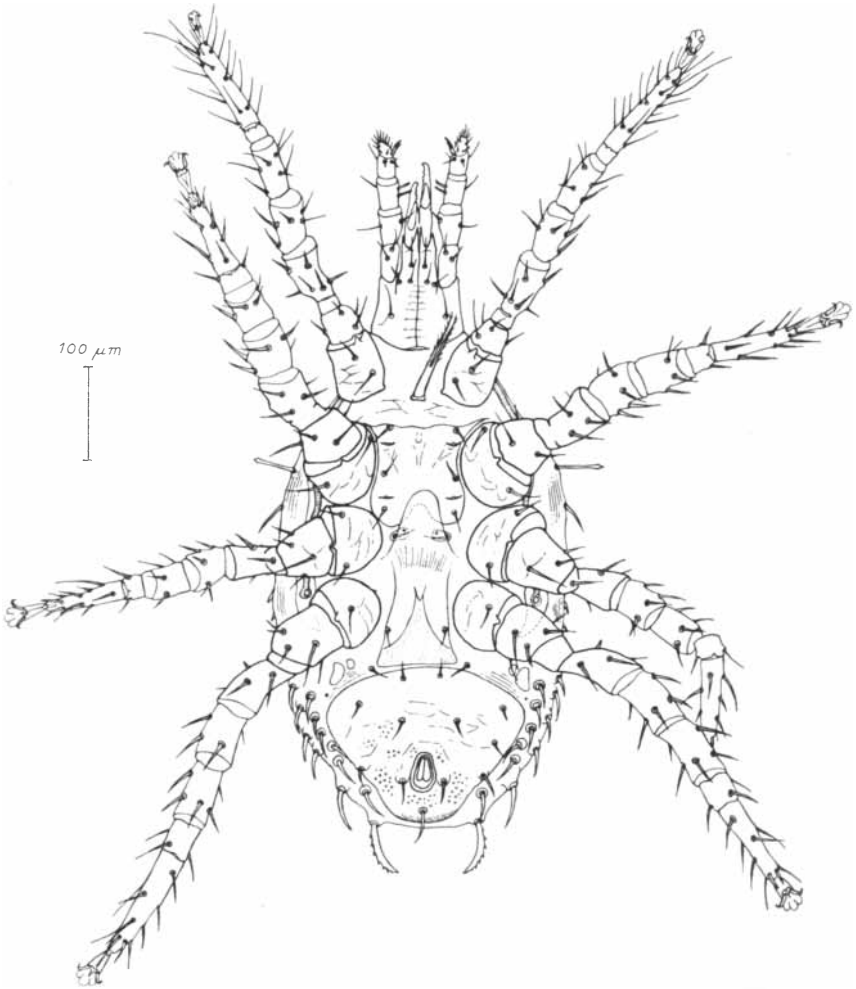
MITES OF MOTHS AND BUTTERFLIES, by Asher E. Treat. Cornell University Press ($35). The golden age of taxonomy harks back to young Darwin and Wallace. Lands newly found by biologists were rich in new forms of plants and animals, and the delight of discovery might lie behind the next bush. Amateurs can still enjoy that experience in their more intimate way, but the pros nowadays are the intrepid people still farther afield, on the Amazon or in Antarctica, or are careful scholars at home aiming for encyclopedic detail summed up in a quasi-legal or, nowadays, mathematical style; the herbariums and the drawers of skins and skulls have long been packed with the type specimens.

And yet the hunt for new species and genera and even for new families (logically as broad as the cat family) is far from over. In this book, at once meticulous in detail and warmly personal, we sniff a little still the air of the golden age. All one had to do was drop his view to the submillimeter level, where little creatures abound that are still new to science. The 1,400 listed acarologists of the world distinguish some 30,000 species of mites—more than there are of spiders—and the count is growing fast. One need not travel far; a footnote tells us that "the summer of 1975... was notable in [Massachusetts] for an extraordinary abundance of the moth ear mite." That remarkable species was discovered 20 years ago by Professor Treat, a City University of New York biologist; he has found quite a few other new species since.

In this volume the author discusses his small, lively corner of the world of mites:



A flint arrowhead in a human sternum, from **The Healing Hand**



Ventral view of the mite **Lasioseius lacunosus,** *from* **Mites of Moths and Butterflies**

the nearly 100 species so far found on moths and butterflies as parasites, hitchhikers or scavengers as the case may be. "Every insect, looked at closely enough . . . is an exotic island whose bays and coves may harbor a lurking pirate." The moth may be living, caught at a night light and held gently under the microscope, or it can be quite dead, netted 100 years ago, its dead mite population unseen all these years. The yield is not small; one common species of night moth in Professor Treat's backyard averages over the years about 25 percent infestation (with a nice Poisson-like distribution ranging up to one unhappy moth that was host to 17 half-millimeter mites, mostly found burrowed in the scales at the base of the hind wings, feeding). The same moth species examined in various museum moth collections yielded an 8 percent infestation rate to "a somewhat cursory examination." The great collections of the showy Lepidoptera "remain almost entirely untapped," also unwitting collections of mites.

The most commonly seen mite of moth or butterfly, recognized by collectors for 150 years, is a "little red or orange-colored object attached by its beak to a leg, an eye, a wing vein." These are the larvae of one mite group that grow up to leave and live free after having engorged themselves as minute vampires "until their distended bodies look like tiny toy balloons." The host seems none the worse for them. Another form is responsible for a dermatitis known as "grain itch" or "hay itch," sometimes a severe affliction of people who come in close contact with grassy material containing various insect larvae (often, but not only, moth species) on which the mite is parasitic. There were a couple of thousand cases at the Indiana State Fair in 1950 and 1951.

The remarkable reproduction of these mites is illustrated in one of many simple line drawings of the strange, unprepossessing beasts. The gravid female becomes "a huge turgid sphere" two millimeters in diameter, five or 10 times her original overall length. First the eggs and then up to some 250 adults appear within that sack, without passing through clear larval stages. About one in 20 of the progeny is a male, and males are the firstborn. They facilitate the birth of their sisters by obstetrical intervention, pulling them out through the sack opening. The males are said to live briefly as ecoparasites with their mother; they cannot survive removal. The young females are generally impregnated on the spot, and they leave home and mother promptly to seek their livelihood on a new host.

Most mites are not very specific to the host, but some are, and the most remarkable story is one the author has been a leader in unfolding over the past 10 or 15 years. This is the story of the moth ear mite. The color frontispiece shows a snug colony of a dozen blind mites in the "sculptured caverns of the insect's ear." The moth's ear is a sensitive passive sonar, rigged to detect the deadly ultrasonic locator pulses of the moth's night-flying enemy, the bat. The ear mite sets up housekeeping, punctures the membranes to feed and lays her eggs, and "within a few days the offspring of even a single brood female can crowd the ear with mites and eggs." The moth is rendered deaf—but only in one ear. The other ear is almost never attacked, even when "both sides of the collar and neck are swarming with young females ready to seek a new home." This unilateral attack is pretty surely an exquisitely adaptive form of parasitism; if the bat gets the moth, the mites go too, but one functioning ear is enough to enable the moth to evade most bats.

There are other mite species that place their eggs in moth ears, but their larvae do not feed there, so that the ears remain essentially undamaged; these other mites, the "ear-sparers," are normally found in both ears. It is not that moth ears are somehow unsymmetrical mite homes; the ear mites can be led experimentally to grow in the uninfested ear all right. When the first mite boards a moth, however, she makes a hesitant choice at the midpoint between the ears and goes to the left or the right. The chosen ear becomes home to a mite brood, and as other mites come aboard they follow with little hesitation, lodging "in the same inn, never visiting the empty tavern across the street." There is probably a chemical trail to the ear of choice. The individual mites are not left- or right-handed; a mite that has chosen a right ear may well opt for the left after an experimental transfer to another moth.

There is a lot to learn about mites and moths, both as pure biology and for economic benefit. This book has the style and air of a piece of delightful natural history, but it is rich in technical details and mite legs drawn bristle by bristle; it is an expensive and expert work, meant for the specialist, in which the general reader can find much good. It is clear that taxonomical fervor is undiminished under the microscope: a certain African mite species might be confused with another, but your informed observer will not make that error, since "they have only twelve instead of sixteen rows of deutosternal denticles."

MUYBRIDGE: MAN IN MOTION, by Robert Bartlett Haas. University of California Press ($18.50). The view from the giant Ferris wheel at the World's Columbian Exposition at Chicago in 1893 shows the dome of the Moorish Palace and the mile beyond it, lined with hootchy-kootchy dancers, wax museums and a variety of other noisy attractions. Right at hand, in a special white rectangular hall with 4,000 square feet of floor area, was "the first

commercial motion-picture theater in the world." Its offerings were serious, billed as lectures on "Animal Locomotion, especially in relation to Design in Art," and it bore the unlikely name of Zoopraxographical Hall.

It was for a few not very successful months the venture of Eadweard Muybridge, born in 1830 in Kingston upon Thames as plain Edward James Muggeridge and the hero of this true romance of California. Muybridge's claim to the first motion picture is a precise one, urged here with discrimination. The device Muybridge exhibited he had first displayed in 1879. It consisted in essence of a glass disk around the edge of which marched one or two dozen photographic images made in quick sequence by as many cameras, triggered by trip wires or electrical signals. The disk was spun in front of a projection lantern along with a counterrotating slotted disk of metal. The effect was a "semblance of actual life" in a brief sequence: "the consecutive phases of movements by Men, Women, Children, Horses, Dogs, Cats, Wild Animals, and Birds... running, jumping, boxing, dancing... or engaged in other muscular exercises... projected by the electric light on a large screen."

The history is intricate. An Austrian officer had first used such a disk in 1853 to project moving pictures, but his were drawings. A Philadelphian, Henry Heyl, had in 1870 projected photographs of a waltzing couple in motion, but the pictures were stationary poses in various stages of dance. It was Muybridge who first analyzed real motion by making successive photographs and then reprojected them in quick succession, synthesizing the motion again from the actual sequential record. He demonstrated this first in the Italianate villa of Leland Stanford, on the site of the present-day campus in Palo Alto; those first shows presented Stanford's favorite horses at a trot and a gallop. Muybridge exhibited the Zoöpraxiscope with great acclaim to distinguished audiences in Paris and London over the next few years.

This was not yet the movies complete; it could handle only a few seconds of action. Thomas Edison opened his Kinetoscope to public admission in the spring of 1894, but it was a coin-in-the-slot peepshow device, too dim for projection, offering 15 seconds of "wonderfully lifelike movement." Edison had a critical new idea: the flexible strip of film. The brothers Lumière opened the first real movie in the Grand Café on the Boulevard des Capucines after Christmas in 1895. Theirs was motion in projection, from real photographs of motion on a strip of indefinite length. For one franc you could watch "The Arrival of a Train" and selected short features. Within the week they were playing to audiences of 2,000 a night.

This devoted biography, presented in a handsome design with plenty of period photographs, follows Muybridge through his strange and spectacular life. He came to San Francisco from a business career in New York in 1855, about the time of the Vigilance Committees. He set up first as a bookseller and publisher, with success. He survived a terrible accident in a stagecoach in Texas and eventually returned to California as a photographer, with a new name, new skills and renewed energy. In a crowded decade of work he had carried his wet plates from Alaska to Panama (his largest plates were of mammoth size, 20 by 24 inches). His photographs of Yosemite, the Farallons and the living rooms of the very rich are here, witness to his abundant talent and precious documents of a California that was already hyperbolic. It is a test of perception to find the Doyle family amidst the bric-a-brac, palmettos and statuary of their Menlo Park home. There is reproduced an extraordinary panorama of San Francisco from Mark Hopkins' house in 1877, and a fine crystalline stack of cannonballs in the ordnance yard at Alcatraz.

Muybridge married a woman half his age, a young photograph retoucher. The dramatic critic for the San Francisco *Post,* a handsome English man of the world, became a family friend; an affair began, a baby was born, and Muybridge came to suspect that the child was not his own. Muybridge went by train and team straight to the mine where Harry Larkyns was working, called him by name to the door and shot him dead "out of the darkness." After a sensational trial and an orotund defense the jury found Muybridge not guilty.

During those years Muybridge formed the relationship with Leland Stanford that molded the rest of his life. As a photographer of the elite, he entered and pictured the Stanford home. Then, when Stanford became interested in the problem of how a horse trots, he worked with Muybridge to settle the question by means of fast photography. The details are closely examined; it appears that neither man alone deserves the credit for the pioneer sequence photography Muybridge carried out under Stanford's patronage.

The two fell out in the early 1880's; Muybridge went on, at the University of Pennsylvania, to produce his remarkable series of sequence photographs of moving men, women, children, horses, dogs, tigers and the rest. That was his second achievement:

Before his famous motion studies Eadweard Muybridge made this photograph of a coffee plantation, from **Muybridge: Man in Motion**

analytical fast photography around which entire new fields of science and technology have since nucleated. The third achievement was his pioneering in commercial motion pictures; those few-second sequences exhibited from Palo Alto to Chicago were the first of all movies.

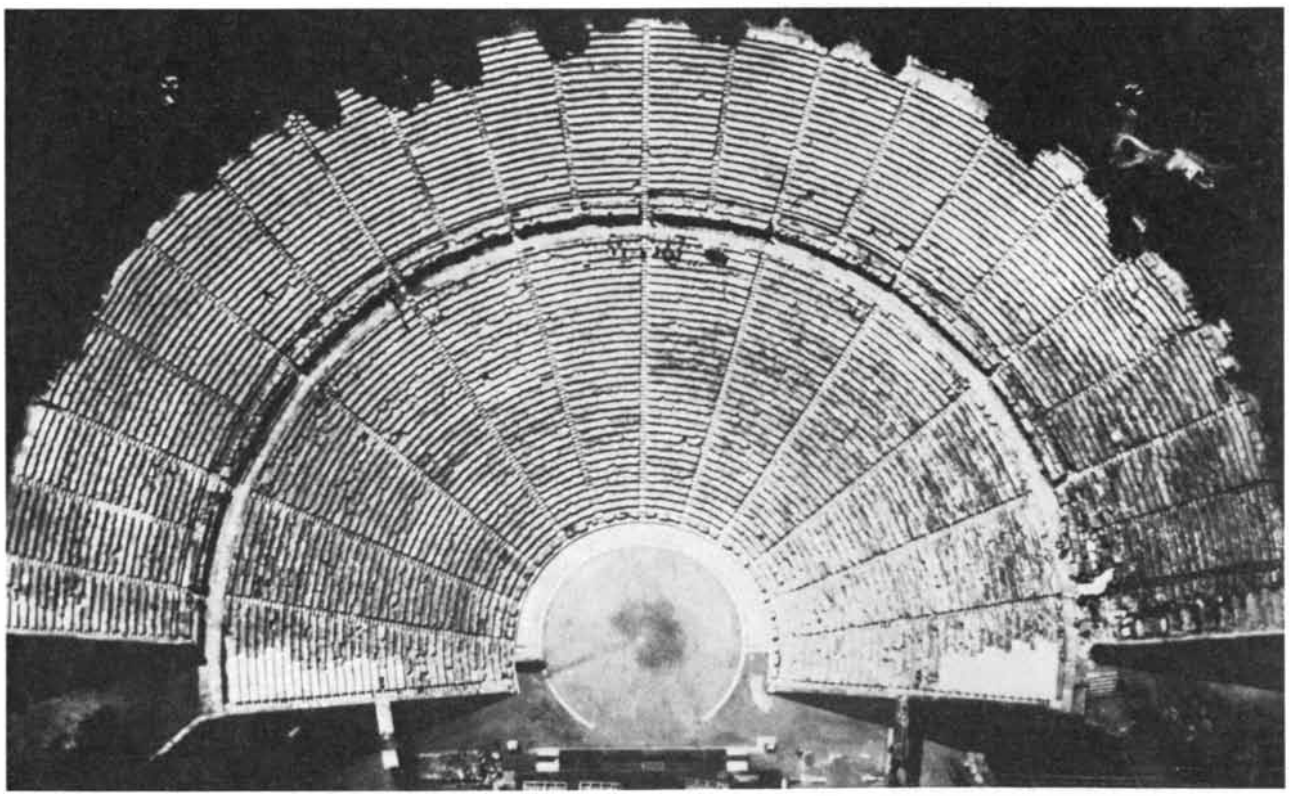
Muybridge came in his old age to resemble a white-bearded patriarch from the less angry engravings of William Blake; the frontispiece shows him at work with a spade, quite nude, a model for his own sequence photographs. Eight full pages of pictures from the Philadelphia period adorn this book. (There are now in print excellent book-length selections from Muybridge's bulky Philadelphia output, two in inexpensive versions and one in a deluxe version.) Robert Haas, a scholar at the University of California at Los Angeles, has spent some 25 years on and off in search of Muybridge and has taken full part in the revival of interest in the man that marked the centennial year of his first trotting-horse sequences.

ENGINEERING PROGRESS THROUGH TROUBLE, selected and edited by R. R. Whyte. The Institution of Mechanical Engineers ($15). Machines are born unto trouble as steam floats upward. A senior British turbine and diesel engineer has selected a couple of dozen published cases of such trouble, mainly from the proceedings of the British Institution of Mechanical Engineers, edited them to brevity and a modest technical level and presented them in a cautionary, instructive and readable account of the "exhilaration, excitement and agony which goes with engineering creation."

Corrosion, fatigue, resonance and plain error are the names of the chief enemies of proper function. They are nothing new; the immense literature on steam-boiler explosions includes data back to the 1850's. The number of reported explosions in Britain has dropped by a factor of 10 since that time and their toll of death and injuries by bigger factors still, even though the number and capacity of steam boilers, and their working pressures, have grown a hundredfold. A diagram compares a big boiler of 1859 with one of 1959; the early one would be lost, smokestack and all, in one corner of the huge modern steam-raiser. The greatest single cause of failures—one case in four—in a recent period was lack of water. Casey Jones looked at his water gauge, but some operators either forget to look or they trust a gauge they do not know how to test. Automatic boilers are a little more prone to certain accidents because "long periods of trouble-free working may lull the vigilance of the attendants." A low-pressure boiler of a marine type exploded in a New York office building in 1962 in the most disastrous of recent boiler accidents. A pressure switch and two safety valves all failed to operate, and a second pressure switch was on the closed side of a stop valve; 23 dead.

That powerful reagent seawater is one irrepressible source of trouble—which should come as no surprise. Land-based companies "light-heartedly diversifying into ocean engineering" particularly need to learn at the dear school of blue-water experience; 40 percent of all oceanographic equipment tested at one U.S. center fails catastrophically, and "100 percent suffers minor failures." Costs expand. One cannot withhold all sympathy from the luckless designers, since many of the errors are subtle. The aluminum hull of a fast patrol boat developed a rash of perforations along the waterline on one side. The cause: the boat had struck a glancing blow against a wooden jetty built with copper nails, a little copper rubbed off and the galvanic currents began to flow. Titanium alloys are "highly resistant to all forms of marine corrosion," but one must use titanium seawater tubing cautiously; those enduring and expensive tubes are prone to clog up, overgrown with mussels and barnacles. The growth is prevented by good old toxic copper.

Lord Rayleigh's great book on vibrations and sound "to this day contains the most important facts and theories bearing on all vibration phenomena in mechanical engineering. It was written at a time when engines ran at 100 rev/min." The diesel-driven German submarines of World War I suffered an epidemic of shaft failures, and for 50 years all ship diesel shafts have been precalculated against torsional vibrations. That did not help the Liberty ships of World War II. They were powered by triple-expansion steam engines "from drawings of a successful design made in 1916" and were not afterward checked for torsionals. They lasted out the war, but soon after it they began to drop their propellers into the ocean. Their competitive peacetime operators, free from the discipline of convoy,



Vertical aerial photograph of the theater at Epidauros, from **Wings over Hellas: Ancient Greece from the Air**

had increased the speed a little; the shafts crossed the critical speed, and engine-driven vibrations fatigued the main shaft.

The embarrassment of the proud liner *Queen Elizabeth II,* whose engines failed on her sea trials, was also the result of vibration. Turbine rotor blades stripped, both port and starboard: fatigue cracks through the base of the blades. A very careful analysis was made and published—something "most unusual, perhaps unique." The blades had been shaped according to an up-to-date standard pattern, but no full-scale tests of the pattern had been conducted, since no other turbines as powerful as those of the *QE*2 were contemplated. (Nowadays there is a need for them in big container ships.) Seaborne trials—safe enough, if distasteful—were the cheapest method "when things went right," although when they went wrong they could be very expensive. (The *QE*1 had thrown turbine blades too, on her maiden voyage in the 1930's.) The blades turned out to be resonant to high-frequency excitation—nearly five kilocycles at full speed—by variations in the steam wake left by the injection nozzles; the match of harmonics is inescapable at some speed. The shape of the root of the blades was changed, and a group of blades were bound by brazed wire so as to vibrate in unison; the fix was complete, and fully successful.

Here too is the story of a magnetite deposit: microparticles that formed when a little salt water contaminated the hydraulic servo-control system of a big new power-station turbogenerator. The tiny particles were polished to a fine layer of hard material by the close-fitting pistons of the governor and the emergency control system; once that layer grew thick enough the main piston seized, the emergency piston probably did the same and the turbine spun itself to destruction, hurling heavy wreckage several hundred yards.

There are many more stories—a few tragic, all rueful. The moral seems to be that the largest devices cannot be realistically tested, so that "progress and trouble go hand in hand." Overall, design clearly improves. Yet the flood of knowledge often conceals data that might have helped, and it becomes clear why cost overrun in development is commonplace, even by an order of magnitude. "A successful design engineer has a value beyond price."

WINGS OVER HELLAS: ANCIENT GREECE FROM THE AIR, by Raymond V. Schoder, S.J. Oxford University Press ($17.50). The aged DC-3 of the Hellenic Air Force banks and wheels slowly a few times at rather low altitude, about 1,000 feet, over some hill or foreshore. The side of the fuselage has a big opening; just inside, tethered by a six-foot strap, the indomitable archaeologist author of this remarkable work stands or kneels. With a 35-millimeter camera in his hands, four lenses ready on the floor for a quick change to suit each pass and the shutter speed set at 1/500th of a second on a sunny day, he signals the pilot for just the right approach, the right angle, the right height. Day after day—for close to 100 hours in the air—they flew all over Greece and the islands. The harvest was magnificent; these hand-held photographs are both beautiful and penetrating. They show site by site the look of the major places of ancient Greece—the ruins as they are, at exactly the scale that makes clearest the natural context of palace, temple, city wall and battlefield. About 80 sites are covered, each with one photograph or more in color and with identifying sketch maps carefully keyed to the pictures. A brief and knowing essay on the significance of the site accompanies the images, and for each site a half dozen or so references to primary archaeological papers and monographs are listed.

Of course there is Athens: a general view and the Agora, the Acropolis, the Stadium and more, in closeup oblique aerial shots. Here the Piraeus catches the eye most freshly. Although it is now the chief port of Greece, with "most of its ancient remains engulfed or long since gone," one still sees two circular coves in the shore, Zea and Munychia, where once 250 triremes lay in their sheds like the spokes of a wheel; now we see moored instead the pleasure flotilla of the sporting Athenians. Then there is Eleusis, no longer secret within its walls. The rock-cut steps and the column foundations are all that remain of the "dark Telesterion hall" where 3,000 initiates stood to witness the ceremonies, the climax in "a thrilling burst of light through an opening" in one of the largest roofed-over spaces in the Greek world.

There is the stunning theater at Epidauros, seen (as an elegant cup in one oblique shot and as a richly patterned fan from close above) in its "unrivaled harmonious symmetry and beauty," gray in the green wood. Blessed is a land whose theater is more striking than its fortresses. Here too, however, are the battlefields and the waters of the sea fights, Marathon and Salamis, Thermopylae and the bay of Navarino. There is the citadel of Tiryns, where Hercules labored. The island Thera is shown from far enough away to make plain the six-mile ruined ring where the sea replaced the volcano, in the explosion that generated the tsunami that is the kernel of truth in the ancient tale of the drowned splendors of Atlantis; the ruin of the classical city is also shown, in a closer vertical view. Thebes is here, the extent of the modern town delineating the line of the old wall, where the seven gates are marked but hardly traceable. Finally there is towering Olympus, with the snowy Throne of Zeus clearly seen, looking like a great glacial cirque cut into the limestone flank of the mountain.

Students, travelers, all who love Hellas and any who cannot resist handsome, clarifying images of famous places will find delight in this volume. Little has escaped the author, who is a professor at Loyola University in Chicago, but one would have welcomed a glimpse of the very shore where the fatal sea monster frightened the racing horses of Hippolytus, whose memorial sanctuary at Troizen we do see.

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

THE CHOICE OF VOTING SYSTEMS

THE THEORY OF COMMITTEES AND ELECTIONS. Duncan Black. Cambridge University Press, 1958.

SOCIAL CHOICE AND INDIVIDUAL VALUES (second edition). Kenneth J. Arrow. John Wiley & Sons, Inc., 1963.

THEORY OF VOTING. Robin Farquharson. Yale University Press, 1969.

COLLECTIVE CHOICE AND SOCIAL WELFARE. Amartya K. Sen. Holden-Day, Inc., 1970.

THE THEORY OF SOCIAL CHOICE. Peter C. Fishburn. Princeton University Press, 1973.

AN ELECTRON-HOLE LIQUID

CONDENSATION OF NON-EQUILIBRIUM CHARGE CARRIERS IN SEMICONDUCTORS. Ya. Pokrovskii in *Physica Status Solidi (A): Applied Research,* Vol. 11, No. 2, pages 385–410; 1972.

TEMPERATURE-DEPENDENT LUMINESCENCE FROM THE ELECTRON-HOLE LIQUID IN GE. G. A. Thomas, T. G. Phillips, T. M. Rice and J. C. Hensel in *Physical Review Letters,* Vol. 31, No. 6, pages 386–389; August 6, 1973.

TWELFTH INTERNATIONAL CONFERENCE ON PHYSICS OF SEMICONDUCTORS, STUTTGART. Edited by M. H. Pilkuhn. B. G. Teubner, 1974.

LIQUID-GAS PHASE DIAGRAM OF AN ELECTRON-HOLE FLUID. Gordon A. Thomas, T. M. Rice and J. C. Hensel in *Physical Review Letters,* Vol. 33, No. 4, pages 219–222; July 22, 1974.

ELECTRON-HOLE CONDENSATION IN SEMICONDUCTORS. Carson D. Jeffries in *Science,* Vol. 189, No. 4207, pages 955–964; September 19, 1975.

THE PURPLE MEMBRANE OF SALT-LOVING BACTERIA

CONSERVATION AND TRANSPORT OF ENERGY BY BACTERIAL MEMBRANES. F. M. Harold in *Bacteriological Reviews,* Vol. 36, No. 2, pages 172–230; June, 1972.

FUNCTIONS OF A NEW PHOTORECEPTOR MEMBRANE. Dieter Oesterhelt and Walther Stoeckenius in *The Proceedings of the National Academy of Sciences of the United States of America,* Vol. 70, No. 10, pages 2853–2857; October, 1973.

PHOTOPHOSPHORYLATION IN *HALOBACTERIUM HALOBIUM.* Arlette Danon and Walther Stoeckenius in *The Proceedings of the National Academy of Sciences of the United States of America,* Vol. 71, No. 4, pages 1234–1238; April, 1974.

THE STRUCTURE OF AN ION PUMP. Walther Stoeckenius in *The Structural Basis of Membrane Function: Proceedings of the International Symposium, Tehran, May 5–7, 1975,* edited by Youssef Hatefi and Lisa Djavadi-Ohaniance. Academic Press, Inc., 1976.

NAVIGATION BETWEEN THE PLANETS

RADAR OBSERVATIONS OF THE PLANETS. Irwin I. Shapiro in *Scientific American,* Vol. 219, No. 1, pages 28–37; July, 1968.

PLANETARY NAVIGATION: NEW CHALLENGES in *Astronautics and Aeronautics,* Vol. 8, No. 5; May, 1970.

MARINER 9 NAVIGATION. W. J. Neil, J. F. Jordan, J. W. Zielenbach, S. K. Wong, R. T. Mitchell, W. A. Webb, P. E. Koskela et al. in *Jet Propulsion Laboratory Technical Report,* JPL-TR-32–1586; November 13, 1973.

17,000 YEARS OF GREEK PREHISTORY

THE FRANCHTHI CAVE: A STONE AGE SITE IN SOUTHERN GREECE. Thomas W. Jacobsen in *Archaeology,* Vol. 22, No. 1, pages 4–9; January, 1969.

THE SOURCE OF THE FRANCHTHI OBSIDIANS. J. E. Dixon and C. Renfrew in *Hesperia,* Vol. 42, No. 1, pages 82–85; January–March, 1973.

EXCAVATIONS IN THE FRANCHTHI CAVE, 1969–1971: PART II. Thomas W. Jacobsen in *Hesperia,* Vol. 42, No. 3, pages 253–283; July–September, 1973.

FAUNAL CHANGE AT FRANCHTHI CAVE FROM 20,000 TO 3000 B.C. S. Payne in *Archaeozoological Studies: Papers of the Archaeozoological Conference 1974, Held at the Biologisch-Archaeologisch Instituut of the State University of Groningen,* edited by A. T. Clason. Elsevier Scientific Publishing Company, 1975.

PREHISTORIC MEN (eighth edition). Robert J. Braidwood. Scott, Foresman and Company, 1975.

CENTER-PIVOT IRRIGATION

WATER CONSUMPTION BY AGRICULTURAL PLANTS. M. E. Jensen in *Water Deficits and Plant Growth, Vol. II: Plant Water Consumption and Response,* edited by T. T. Kozlowski. Academic Press, Inc., 1968.

REMOTE SENSING TECHNIQUES FOR EVALUATION OF SOIL WATER CONDITIONS. Fred A. Schmer and Hal D. Werner in *Transactions of the ASAE,* Vol. 17, No. 2, pages 310–314; March–April, 1974.

MODELLING OF PLANT GROWTH FOR YIELD PREDICTION. W. E. Splinter in *Plant Modification for More Efficient Water Use,* edited by John F. Stone. Elsevier Scientific Publishing Company, 1975.

WATER INTAKE UNDER CENTER-PIVOT FROM TIME-VARYING APPLICATION RATES. John W. Addink, Donald L. Miles and Gaylord V. Skogerboe in *Transactions of the ASAE,* Vol. 18, No. 3, pages 523–525; May–June, 1975.

HISTORICAL SUPERNOVAS

ASTRONOMY. Joseph Needham in *Science and Civilisation in China, Vol. III: Mathematics and Sciences of the Heavens and the Earth.* Cambridge University Press, 1959.

ANCIENT AND MEDIAEVAL OBSERVATIONS OF COMETS AND NOVAE IN CHINESE SOURCES. Ho Peng Yoke in *Vistas in Astronomy: Vol. V,* edited by Arthur Beer. Pergamon Press, 1962.

NONTHERMAL GALACTIC RADIO SOURCES. R. Minkowski in *Nebulae and Interstellar Matter,* edited by Barbara M. Middlehurst and Lawrence H. Aller, in *Stars and Stellar Systems: Compendium of Astronomy and Astrophysics: Vol. VII,* edited by Gerard A. Kuiper. The University of Chicago Press, 1968.

SUPERNOVAE. I. S. Shklovsky. Wiley-Interscience, 1968.

FUTURE PERFORMANCE IN FOOTRACING

PERSPECTIVES OF ADAPTATION: HISTORICAL BACKGROUNDS. Chauncey D. Leake in *Handbook of Physiology, Section 4: Adaptation to the Environment,* edited by D. B. Bill. American Physiological Society, 1964.

PHYSIOLOGICAL FUNDAMENTALS OF TRAINING METHODS. H. Roskamm, H. Reindell and J. Keul in *Run Run Run,* edited by Fred Wilt. Tafnews Press, 1964.

THE NATURE OF ADAPTIVE REACTIONS AND THEIR INITIATION. Chandler McC. Brooks in *Physiology and Pathology of Adaptation Mechanisms: Neural–Neuroendocrine–Humoral,* edited by Eörs Bajusz. Pergamon Press, 1969.

MATHEMATICAL GAMES

SOME PROBLEMS IN COMBINATORICS. H. G. Forder in *The Mathematical Gazette,* Vol. 45, No. 353, pages 199–201; October, 1961.

EULER'S PROBLEM OF POLYGON DIVISION. Heinrich Dörrie, translated by David Antin in *One Hundred Great Problems of Elementary Mathematics.* Dover Publications, Inc., 1965.

A NOTE ON PLANE TREES. N. G. de Bruijn and B. J. M. Morselt in *Journal of Combinatorial Theory,* Vol. 2, No. 1, pages 27–34; January, 1967.

PRIME AND PRIME POWER DIVISIBILITY OF CATALAN NUMBERS. Ronald Alter and K. K. Kubota in *Journal of Combinatorial Theory: Series A,* Vol. 15, No. 3, pages 243–256; November, 1973.

12 YEARS OLD WORLDWIDE · BLENDED SCOTCH WHISKY · 86 PROOF · GENERAL WINE & SPIRITS CO., NEW YORK, N.Y.

Less work for father.

Admittedly, our personalized, deluxe, half-gallon pouring stand won't make it easier for a father to part with his Chivas Regal.

But it will make it easier for him to pour it.

So if you'd like one, just send a check for $6.95, together with the name you'd like imprinted on the name plate, to Chivas Regal Cradle, P.O. Box 5061, Smithtown, N.Y. 11787.

And do it as soon as you can.

After all, any father who deserves a half-gallon of Chivas shouldn't have to lift the bottle by himself.

Offer good in U.S.A. only, except where prohibited or restricted by law. Allow four to six weeks for delivery. Offer may be withdrawn without notice.