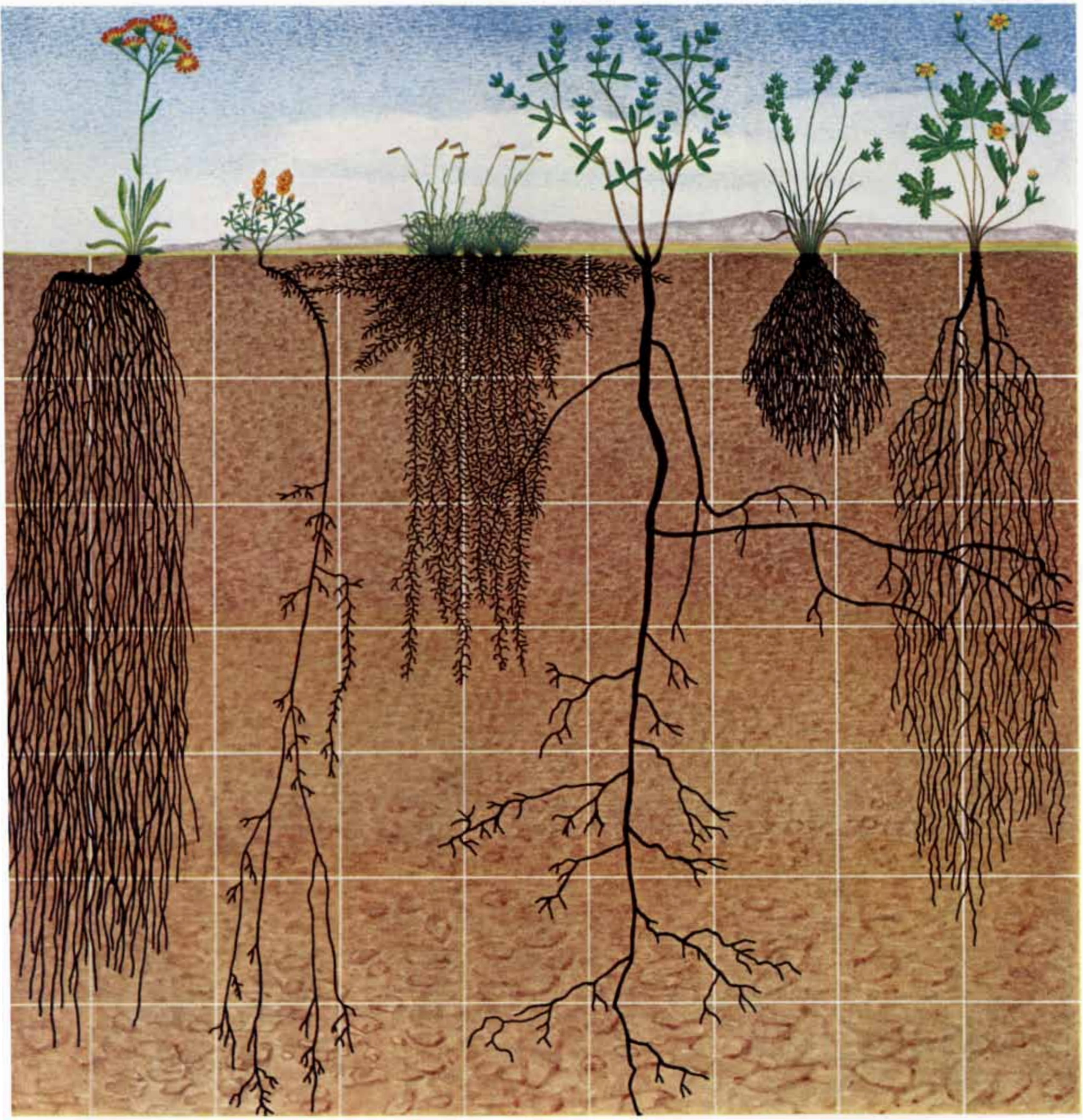
SCIENTIFIC AMERICAN



ROOTS

ONE DOLLAR

May 1973

"Hi. We're the Uniroyal Tire Thrill Drivers."

Uni: We leap off ramps through flaming hoops. We do hairpin turns, near collisions, 2-wheel driving. That's how we demonstrate the strength and traction of Uniroyal Steel Belted Radials. Watch for us on TV.

Roy: Daring we may be, but dumb we're not. If we couldn't count on Uniroyals in hairy situations, we wouldn't take their money. We're not looking for grief.

Al: So if you want to feel the same confidence on the road as we do on the track, even in rain or freeway traffic, get yourself some Uniroyals. We use 'em. And we don't like to push our luck.


straight to Australia
The straight line from London to Sydney runs across a European capital. Can you name it?
Did you say Brussels?
We can then tell you two things:
1. You have a remarkably accurate mental picture of the world map (or perhaps you looked at a map?).
2. You are mistaken.
The capital on the true straight line London–Sydney is Copenhagen – the gateway of the Trans-Asian Express to Bangkok and Singapore, which is far shorter than any other route.
Trans-Asian Express flights on Wednesdays (SAS) and Saturdays (THAI International) continue from Bangkok as Thai International services, with through-going aircraft from Copenhagen, to Sydney via Singapore.
This is not only the shortest way to Sydney. It stands out among alternative routes to Sydney in another way, too.
If you are one of those travellers who are somewhat less than enthusiastic about two consecutive nights in an aircraft seat, you can now travel to Sydney and spend only one night en route.
Leave Europe on a Wednesday – arrive in Sydney on Thursday. Or leave on a Saturday and arrive on Sunday. The shortest way there is.
You have two alternatives when you go to Australia:
FLY VIA COPENHAGEN-
OR FACE A DETOUR
SAS
SCANDINAVIAN AIRLINES
General Agent for Thai International

Established 1845 **SCIENTIFIC AMERICAN** May 1973 Volume 228 Number 5

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1973 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $10 PER YEAR, U.S., POSSESSIONS AND CANADA; $12.50 PER YEAR, ALL OTHER COUNTRIES. EUROPEAN SUBSCRIPTION COPIES MAILED C/O INTERNATIONAL DISTRIBUTORS, KOEPOORTBRUG 1, B 2000 ANTWERP, BELGIUM.

THE COVER

The painting on the cover shows the root systems of six plants of the North American prairie. From the left they are a hawkweed of the genus *Hieracium*, a scarlet globemallow of the genus *Malvastrum*, a blue grama of the genus *Bouteloua*, a slimflower scurfpea of the genus *Psoralea*, a fescue of the genus *Festuca* and a cinquefoil of the genus *Potentilla*. The shape of a plant's root system is determined in part by heredity and in part by soil and water conditions, but all root systems, regardless of shape, perform the same function, absorbing water and mineral nutrients from the soil and carrying them aboveground (see "Roots," by Emanuel Epstein, page 48).

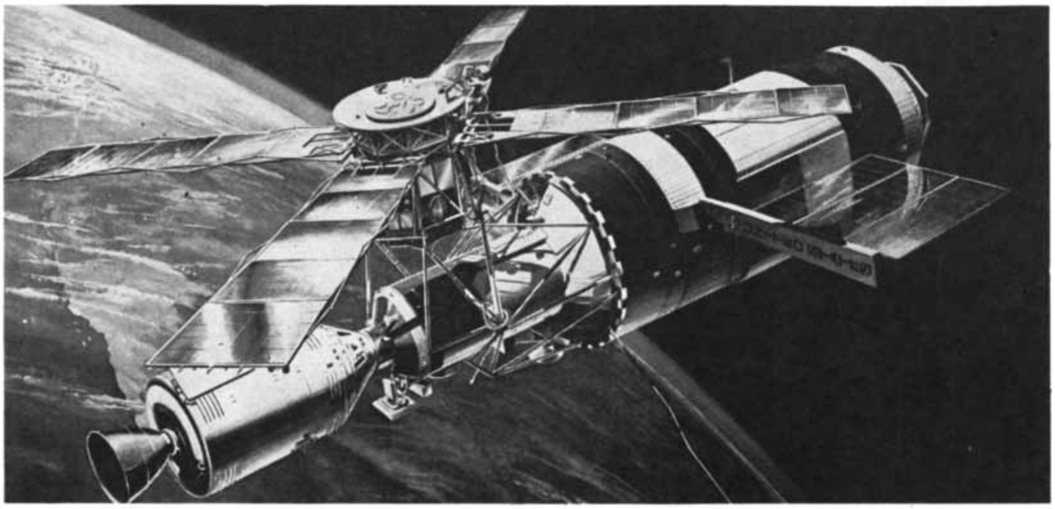
THE ILLUSTRATIONS

Cover painting by Eric Mose

Page	Source
16–23	Jerome Kuhl
25	European Art Color Slides
26	Globe Photos
27–29	Ilil Arbel
31–33	Dan Todd
34–35	L. L. Ban and W. M. Hess, Columbia Carbon Company
36–40	Dan Todd
49	Soil and Water Institute, Agricultural Research Service
50–53	Enid Kotschnig
54	Ilil Arbel
55–58	Enid Kotschnig
62	National Aeronautics and Space Administration
64–72	Allen Beechel
75–77	Jerome Snyder
78–79	Ilil Arbel
80	Jerome Snyder
81	Jerome Snyder (*top*), Ilil Arbel (*bottom*)
82	Ilil Arbel
83	Jerome Snyder
84	National Central Library, Florence
86–90	Ilil Arbel
95	Richard G. Zweifel (*top*), Jon Brenneis (*bottom*)
96	Barnes Engineering Company (*top*), Jon Brenneis (*bottom*)
97–98	Tom Prentiss
99	Richard M. Meszler, University of Maryland
100	R. Igor Gamow and John F. Harris
102–106	Ilil Arbel
108	Roger Hayward
109	Ben Rose (*top*), Roger Hayward (*bottom*)
110–112	Roger Hayward

Skylab, the world's loftiest photo studio.

Inside, Nikon cameras study invisible airglow, measure the sun's corona, and watch a spider at work.



Space photography has come a long way since its early days when just bringing back a picture seemed quite an achievement. The photographic tasks assigned to Project Skylab in particular are far more diverse and exacting than any that came before.

That means more sophisticated equipment. Versatile enough to handle the variety of assignments planned, yet easy enough to permit a non-expert to obtain the needed picture quality. In short, 35mm equipment. Which means Nikon, the most sophisticated 35mm system available.

Nikon cameras had already demonstrated their capabilities in previous missions, including Apollo 15, 16 and 17. These were essentially standard Nikon Photomic FTN models, modified to fit the conditions encountered in a vacuum. Special materials were used to prevent molecular changes (outgassing), new means of lubrications devised, some controls adapted for easier handling with bulky gloves.

Thus when NASA decided on 35mm equipment for Skylab, the Nikon system was able to take it in stride. It was quite a challenge, though, considering that subject matter ranges from the infinite to the infinitesimal.

Experiment SO 63

Purpose: To record ultra-violet emission from the airglow layer of the upper atmosphere. Equipment: Two motor-equipped Nikon cameras controlled by a specially designed Nikon interval timer. One camera is fitted with a multiple-layer coated Auto-Nikkor 55mm f1.2 lens for recording visible light, the other with a newly designed 55mm f2 Ultra-Violet Nikkor lens (the first of its kind and, like all Nikkor lenses, made from Nikon optical glass). The latter is mounted on a track enabling it to follow the curvature of the earth and providing automatic, simultaneous triggering of the conventional-lens camera at a predetermined point. Photos from both will be matched in evaluating the results of this experiment.

Experiment TO 25

Purpose: Coronagraph contamination study, monitoring the presence of particulate matter near the spacecraft and measuring the solar corona. Equipment: Nikon Photomic FTN camera with multiple-layer coated Auto-Nikkor 55mm f1.2 lens, in fixed position. (Originally, NASA had considered a larger camera for this assignment but found 35mm equipment more suitable). A special reticle was designed for this camera's finder screen, containing degrees, x and y axes, and digital computer markings. Absolute finder accuracy is vital because the sun's image must be kept within a central 1mm circle. Since the camera is fixed, the observing astronaut gives directions in computer language for computer-controlled changes in the spacecraft's position and attitude.

Other Nikon Skylab photography is very much of the kind you might do yourself. Included are closeup studies of a spider web spun in space as well as photos of various activities aboard the spacecraft.

In fact, the reasons Nikon works so well in outer space are the same reasons that have made Nikon the first name in 35mm photography on earth. Matchless versatility, repeatability and reliability. Ingenious design that keeps it ahead technologically while defying obsolescence. Ruggedness bordering on the incredible. Above all, picture quality that literally put 35mm photography on the map.

It's a camera worth looking into, for pleasure as well as for scientific pursuits. At Nikon dealers everywhere. (Ask also about the Nikon School of Photography.) Or, write for literature folio 19. Nikon Inc., Dept. SA, Garden City, N.Y. 11530. Subsid. of Ehrenreich Photo-Optical Industries, Inc. EPI (In Canada: Anglophoto Ltd., P.Q.)

The toughest thing about building a network is finding a place to start.

When you get into something as complicated as networks, it's easy to get in trouble.

What's difficult is knowing where to begin.

Start with us.

After all, you'll be using a lot of minis. And they'll probably be our PDP-8's and 11's. Simply because they're the minis almost everybody buys.

You'll be needing all the interfaces for those minis. For synchronous lines. Asynchronous lines. Memory to memory. We already have them all. All off the shelf. Ready to go. So you can hook up your network the way you want it, when you want it.

You'll want a host computer, a big timesharing computer, that fits right in with all our minis and works the way you do.

We make that, too. It's our DECsystem-10.

In fact, almost 30% of the computers used by independent timesharing services are DECsystem-10's. Networks are their business, so they have to go with the best computer available.

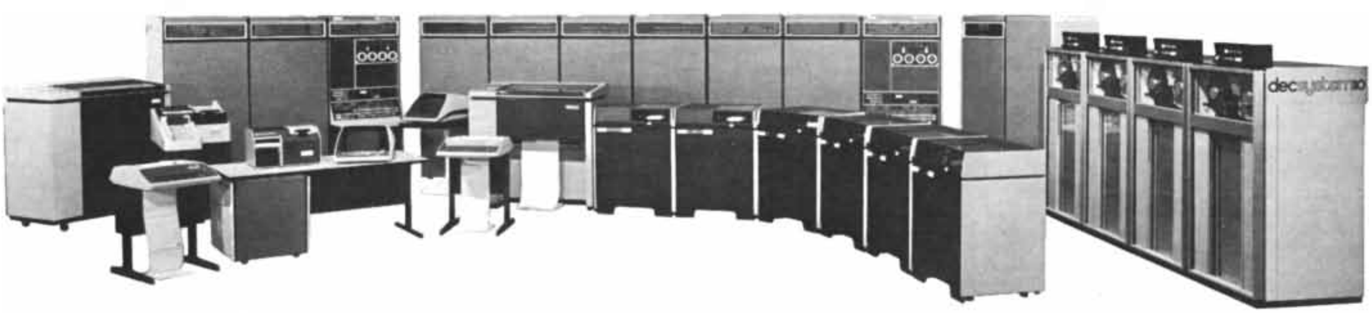
If you're interested in networks, we're the people to talk to. We have more timesharing installations than anyone else in the world.

You can come to us because of all our minis, or all our interfaces, or our DECsystem-10 host computer, or our timesharing experience.

But we expect you'll come to us because we're the only company that gives you all of them.

DECsystem-10 Group, Digital Equipment Corporation, 146 Main Street, Maynard, Mass. 01754. (617) 897-5111. European headquarters: 81, route de l'Aire, 1211 Geneva 26. Tel: 42 79 50. Digital Equipment of Canada Ltd., P.O. Box 11500, Ottawa, Ontario. K2H 8K8. (613) 592-5111.

digital



DECsystem-10: It'll make your network work.

LETTERS

Sirs:

Professor Wang is to be congratulated for his lucid presentation of some of the relations between Chinese writing and the Chinese languages [SCIENTIFIC AMERICAN, February]. His contribution is particularly appropriate at a time when intellectual and other exchanges with China appear to be on the verge of a quantum increase. I should like to comment on a statement in the article that, while not threatening to its exposition, strikes me as unfortunate: "The phenomenon of tones seems to be confined to Chinese and to some of the languages of Southeast Asia that have been heavily influenced by Chinese."

It is indeed the case that the languages of China and Southeast Asia are for the most part tonal. It is also the case, however, that an even larger number of tone languages exist in Africa west of Ethiopia and south of the Sahara. I refer to the Sudanic, Bantu, Bushman and Hottentot groups. Moreover, various tone languages are found in southwestern Mexico: Mixteco, Mazateco, Amuzgo, Chatino, Chinanteco, Chocho, Cuicateco, Otomi, Tlapaneco, Trique and Zapoteco. In the U.S., Navaho and Apache are probably the best-known examples.

Aside from abundant internal linguistic evidence for tone language in Africa and North America (not to mention some tone-partial languages in Europe), there are the ethnographically well-publicized cases of Luba (*et al.*) drum signaling and Mazateco "whistle speech." Both of these phenomena are derivable only from resources available in tone languages.

JAMES R. JAQUITH

Anthropology
Social Sciences Division
Southern Illinois University
Edwardsville

Sirs:

In response to various communications on my article the following points are made for clarification.

Concerning the matter of tones, it is true many African and Amerindian languages have been called "tonal," as have some Indo-European languages such as Lithuanian, Swedish, Serbian, etc. However, in all these languages pitch patterns are used in significantly different ways from those in Chinese. For example, in Swedish the patterns are predictable in terms of the morphology and the rest of the syllable, in the Bantu languages the distinctions are restricted to only pitch levels, etc. I have discussed these differences more precisely in a paper, "Phonological Features of Tone" (*International Journal of American Linguistics,* Volume 33, pages 93–105, 1967).

In the sentence quoted in Mr. Jaquith's letter my purpose was to highlight the unique way pitch patterns are used in Chinese. Here almost all the morphemes are one syllable long, accompanied by exactly one tone. Furthermore, these tones make significant use of contours in addition to levels; they are not morphologically derived, and they are mostly not predictable from the rest of the syllable. To my knowledge no other language group uses pitch patterns in this way; it was in this sense that I used the word "tone" in the quoted sentence.

As for the drum signals and whistle speech mentioned in Mr. Jaquith's last paragraph, I must point out that such "surrogate systems" need not be based on tones, and cannot be taken as evidence for tones. In the April 1957 issue of *Scientific American,* André Classe described a very interesting case of whistle speech in the Canary Islands, which is based on a version of Spanish.

Readers interested in the relative antiquity of writing systems will find I. J. Gelb's *A Study of Writing* (The University of Chicago Press, 1963) a particularly useful reference. In my article I wrote "Sumerian is the only language we know of that has extant written materials that antedate Chinese ones.... But Sumerian and its derivative orthographies died out long before the beginning of the Christian Era." Among the derivative orthographies are usually included the cuneiforms of Akkadian, Elamite, Hittite, etc., of which the Persian cuneiform was the latest in use (between the sixth and fourth centuries B.C.). Of the writing systems currently in use Chinese characters reach the farthest back in time in terms of extant samples, the oldest of which are the "shell-bone inscriptions" of the second millennium B.C.

An intriguing observation on Chinese characters has recently emerged in several studies on aphasia by Japanese researchers. Japanese is written with a combination of Chinese characters and a phonetic syllabary. It appears that these two components of the Japanese writing system involve separate processing mechanisms in the brain. Some patients have lost control of the phonetic syllabary but can continue to use the far more complex Chinese characters. Such results have important implications for studies relating to reading ability.

On page 57 of my article the inadvertent deletion of the word "not" has caused some confusion. In the paragraph on ethnic minorities of China the sentence should read: "The languages of some of these groups are *not* related genetically to Chinese but belong. . . ."

Finally, I wish to thank the many readers who wrote me to express enthusiasm at the rapidly improving relations between China and the U.S. I share this enthusiasm, of course; the two cultures have so much to contribute to each other. My article on the Chinese language was offered, in a sense, as a minor token of "intellectual exchange" through the pages of *Scientific American.*

WILLIAM S-Y. WANG

Department of Linguistics
University of California
Berkeley

Scientific American, May, 1973; Vol. 228, No. 5. 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 Harry T. Morris, 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 25 cents.

Change of address (or other subscription correspondence) should be addressed to Circulation Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Please notify us four weeks in advance of change. If possible, kindly furnish an address imprint from a recent issue. Send both old and new addresses or complete and mail Post Office Form 3578.

☐ Please enter change of address.

☐ Please enter subscription.
U.S. and Canada:
☐ 1 year, $10 ☐ 2 years, $18 ☐ 3 years, $25
All other countries:
☐ 1 year, $12.50 ☐ 2 years, $23 ☐ 3 years, $32

NAME

NEW ADDRESS

OLD ADDRESS

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MAY, 1923: "Now in a flash comes a host of new archaeological material that may fill out to an undreamed of extent the picture of the age of the Egyptian pyramids. Biban el-Moluk, the valley of the kings' tombs, is a wild, desolate region amid the hills behind the western plain of Thebes. Some 60 tombs in the valley were already known when the Earl of Carnarvon began his search for any that might still lie concealed. On November 5 of last year Mr. Carter, after years of fruitless effort, came on a step cut in the rock under the path leading to the tomb of Ramses VI. As soon as Lord Carnarvon had arrived from England in response to a hasty summons, the steps and passages of the *L*-shaped approach were cleared. The tomb was that of King Tutenkhamon, and it had escaped the destructive attacks of the XX Dynasty robbers who were brought to trial under Ramses IX. The antechamber that the discoverers first entered is the source of practically all the treasures that were removed this year. A walled-up doorway was first opened on February 16 of this year and revealed the hoped-for sepulcher. It is reasonable to assume that here for the first time an untouched royal Egyptian burial has been found."

"Sonic sounding is rendered possible by the fact that sound vibrations, passing through water and striking a solid surface, are returned as an echo to the source from which they originated. Working on this principle, Dr. H. C. Hayes has developed at the Engineering Experimental Station in Annapolis a method of determining ocean depths, which has been used by ships of our navy in some very successful and striking demonstrations. Of these the most notable was a series of soundings taken aboard the destroyer *Stewart* while she was en route from Newport, R.I., to Gibraltar from June 22 to June 29 of last year. In the nine days of the trip the *Stewart* took 900 soundings, the speed of the ship being 15 knots. Very interesting is the resulting view of the mountains and valleys of the Atlantic. The Navy is to be congratulated on the success of this new and valuable aid to navigation."

"Professor J. F. McClendon of the University of Minnesota recently discussed the world's supply of iodine in relation to the prevention of goitre. Practically all of the iodine of the earth's surface is in the sea, which contains about 60 billion metric tons of iodine in the form of soluble inorganic salts. Judging by the prevalence of goitre, there is often a deficiency of iodine in our food and drink. Omitting the details of the local distribution of goitre, there is a wide goitre belt across the nation that includes the mountainous and glaciated regions. Since the run-off from these regions has carried away so much of the soluble material, it seems likely that the goitre belt is a low-iodide belt. Since the sea contains the bulk of the iodine supply, the transfer of iodine from the sea to our food or drink should be increased. Salt might be made an important source of iodine in our dietary scheme. Salt could easily be prepared from sea water with the retention of the iodine compounds and at a cost not exceeding that of present day table salt."

SCIENTIFIC AMERICAN

MAY, 1873: "The increasing tide of emigration settling toward our Western territory, together with the rapid settlement of large tracts of fertile and arable land by classes devoted to agricultural labor as a means of subsistence, has given rise to the development of an industry that has assumed proportions of almost incredible magnitude. We allude to the manufacture of agricultural implements, without the aid of which the profitable working of large farms would be a physical impossibility. Of these great works the largest and most complete are located in Chicago. The most extensive is the manufactory of Messrs. C. H. McCormick & Brother, and it is engaged in the production of the celebrated McCormick reaper."

"The substance known as codeine is beginning to find use in medicine. Like morphine, narcotine and several other alkaloids, it occurs in opium in combination with certain vegetable acids, principally meconic acid. From 100 lbs. of opium only 6 or 8 ozs. of codeine is obtained. Codeine is more soluble in water than morphine; it is also soluble in alcohol, ether and ammonia. The physiological effects of codeine resemble those of morphine in many respects. According to Robiquet, a dose of .3 or .4 of a grain produces in 24 hours, particularly in an excitable person, a sensation of comfort and repose and a refreshing sleep, and a dose of from 1.8 to two grains produces heavy sleep with a feeling of intoxication after waking, sometimes also nausea and vomiting. More than three grains in 24 hours cannot be taken without danger of serious consequences."

"In Africa the results of the year have been the rescue of Dr. Livingstone by the *Herald* reporter Stanley and the knowledge of the explorations of Dr. Schweinfurth in the regions west of Khartoum and to within 3½ degrees of the Equator. The latter traveler has found a race of pigmies, or dwarfs, supposed to be the same as those described by Herodotus. Sir Bartle Frere has arrived at Zanzibar and has communicated with the Sultan relative to the suppression of the slave trade."

"The annual report of the President and Registrar of Cornell University furnishes a gratifying exhibit of the rapid and substantial growth of that institution. There are at present 500 students, and the faculty consists of 40 professors and instructors. We note the erection of a college of mechanical engineering, provided with a machine shop and all accessories through the munificence of Mr. Hiram Sibley of Rochester. The college library ranks third in size, and second in value, of those of its kind in the country. The university now comprises five large buildings of stone, three of brick and two of wood, constructed almost entirely with the aid of money, aggregating $1,400,000, donated by friends. The regulations of the institution state that its benefits are open to all, but only students resident in the State of New York receive free instruction."

"We regret to announce the death of John Stuart Mill, a writer and thinker of great celebrity whose works are known to the civilized world. He was the son of James Mill, the author of *A History of India* and a speculative philosopher of great reputation. It is as a logician of the highest order, whose reasonings led him to sympathize with the cause of freedom in all countries, that John Stuart Mill will be remembered. He died at his country house at Avignon, France, in the 67th year of his age."

THE AUTHORS

HORST FEISTEL ("Cryptography and Computer Privacy") has been a member of the staff of the Thomas J. Watson Research Center of the International Business Machines Corporation since 1968. He received his early education in Berlin and obtained his bachelor's and master's degrees in physics from the Massachusetts Institute of Technology and Harvard University respectively. During World War II he was a member of the staff of the M.I.T. Radiation Laboratory, and thereafter he worked for a number of years for the Department of Defense. He writes that his contact with defense problems gave rise to "a deep and lasting interest in cryptographic coding." Apart from his work he is interested in military technology ("particularly warships") and European history. He notes that he also practices and enjoys cooking.

LEE SALK ("The Role of the Heartbeat in the Relations between Mother and Infant") is clinical professor of psychology in pediatrics at the Cornell University Medical College and director of the division of pediatric psychology at New York Hospital–Cornell Medical Center. Born in New York (the younger brother of Jonas Salk), he obtained his degrees at the University of Michigan: his bachelor's degree in psychology in 1949, his master's degree in sociology in 1950 and his Ph.D. in psychology in 1954. He writes that one of his major concerns is teaching people how to be parents. "I have been conducting a first-year-of-life series for new mothers, fathers and babies that involves four meetings in the course of the first year. I have also embarked on an experimental course for junior high school students on the responsibilities of parenthood." Salk is the author of *How to Raise a Human Being* (with Rita Kramer) and *What Every Child Would Like His Parents to Know*.

J. G. DASH ("Two-dimensional Matter") is professor of physics at the University of Washington. After receiving his bachelor's degree from the City College of New York in 1944 and his master's degree and Ph.D. from Columbia University in 1949 and 1951 respectively, he worked at the Los Alamos Scientific Laboratory until he took up his present work in 1960. Dash writes that the recent publication of *A Life of One's Own* by his wife, Joan, is "the most exciting thing that has happened recently in our family." He adds: "We have three children: Michael and Elizabeth, who are students at the University of Washington, and Anthony, who is in high school. Michael is a physics major, Elizabeth is in journalism and Anthony is in weight lifting, Latin and Shakespeare. I don't think I have any major interests outside of physics, but I do a number of things occasionally: tennis, hiking and photography. Molecular films are so interesting that I avoid spreading myself too thin."

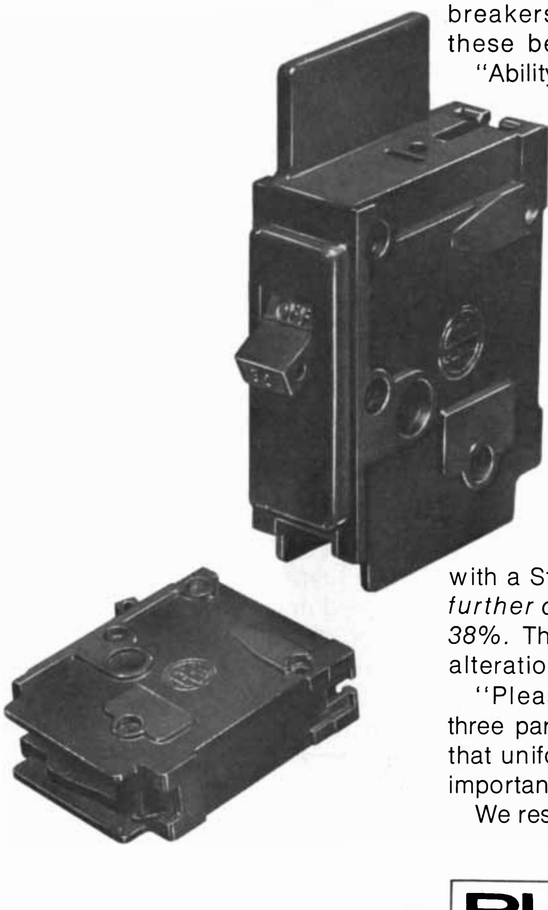
EMANUEL EPSTEIN ("Roots") is professor of plant nutrition at the University of California at Davis and plant physiologist at the university's Agricultural Experiment Station. Before he went to Davis he spent eight years as a plant physiologist with the U.S. Department of Agriculture. Epstein obtained all his degrees from the University of California system; his Ph.D., which he received in 1950 from the University of California at Berkeley, is in plant physiology. Epstein writes that his research is on the mineral nutrition of plants, particularly their responses to high levels of salt, and on the cellular mechanisms of ion transport that plants employ in regulating their mineral economy. He is also interested in genetic and ecological aspects of the subject.

D. P. McKENZIE and J. G. SCLATER ("The Evolution of the Indian Ocean") are respectively at the University of Cambridge and the Massachusetts Institute of Technology; McKenzie is senior assistant in research in the department of geodesy and geophysics and Sclater is associate professor of marine geophysics. They were contemporaries at Cambridge, where each received his Ph.D., and for a short time at the Scripps Institution of Oceanography. McKenzie has also worked at the University of California at San Diego, the California Institute of Technology, Columbia University and Princeton University. He has been involved in the development and application of ideas on plate tectonics. Sclater, who was a research geophysicist at Scripps from 1965 to 1972, writes that McKenzie "is the theoretician and I am the observationalist."

WESLEY C. SALMON ("Confirmation") is Norwood Russell Hanson Professor of the Philosophy of Science at Indiana University. In September he will become professor of philosophy at the University of Arizona. He was graduated from the University of Chicago in 1947 and obtained his master's degree there the same year; his Ph.D. (in philosophy) is from the University of California at Los Angeles. "I am much interested in many problems in the philosophy of physics," he writes. "They include space, time and causality." Salmon recently completed a term as president of the Philosophy of Science Association.

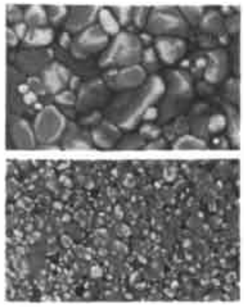
STILLMAN DRAKE ("Galileo's Discovery of the Law of Free Fall") is professor of history at the Institute for the History and Philosophy of Science and Technology at the University of Toronto. He writes: "I received a bachelor's degree in philosophy from the University of California at Berkeley in 1932. After some graduate work in mathematics I entered the investment field, specializing in bond issues for public works. In 1967 I left business to accept appointment here. Aside from my work my main interests are music (viola da gamba) and Sherlock Holmes; I was long Bodymaster of the Scowrers, San Francisco's scion society of the Baker Street Irregulars, and rejoice particularly in my reconstruction of the Camberwell Poisoning Case, a triumph of sheer logical deduction in which, from a single sentence, I was able to determine the name of the victim, the reason for the murder and the identity of the killer."

R. IGOR GAMOW and JOHN F. HARRIS ("The Infrared Receptors of Snakes") are at the University of Colorado; Gamow (the son of the late physicist George Gamow) is assistant professor in the department of aerospace engineering sciences and Harris is a graduate student in the department. Gamow did his undergraduate work at the university in traditional biology and his graduate work in molecular biology. After receiving his Ph.D. he went to the California Institute of Technology as a postdoctoral fellow; he worked with Max Delbrück, who, he writes, "introduced me first to the wonderful world of sensory physiology and second to the photoresponding fungus Phycomyces." He adds that since joining the engineering faculty at the University of Colorado he has found that many exciting problems "can best be solved by overlapping the fields of biology and engineering." Gamow says he enjoys "skiing in the winter and tennis all the time." Harris has spent the past academic year teaching underprivileged high school students in New York; he expects to return to his graduate studies in the fall. His interests include flying and photography.

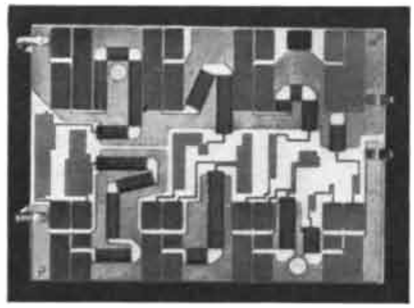


WESTERN ELECTRIC REPORTS

1500° C furnace was specially designed to fire these new substrates. The relatively low temperature results in smooth substrate surfaces for practically fault-free thin film bonding.



Electron micrographs show the great difference in grain size between new ceramic material (lower) and the previous material (upper).



Thin film integrated circuit shown here is part of a resistor network. It is one of many that benefit from the improved substrate. Metal leads on sides are bonded by thermocompression to tantalum nitride resistor film.

Smoothing the way for perfect thin film bonding.

Aluminum oxide, or alumina, is considered to have the best combination of properties for thin film circuit substrates. Until recently, however, the bonding of metal elements to gold-coated tantalum nitride resistor film on alumina was somewhat unpredictable.

Now, an advance at Western Electric has made it possible to get practically fault-free bonding of these materials.

This new perfection in bonding came through the development of finer grained alumina substrates.

The process has four basic steps: milling, casting, punching and firing.

During milling, alumina is combined with magnesium oxide, trichlorethylene, ethanol and a unique deflocculant. For 24 hours, this mixture is rotated in a ball mill. In a second 24-hour period, plasticizers and a binder are included.

The deflocculant plays a major role by dissipating the attraction forces that exist between the highly active alumina particles. This prevents thickening, which would ordinarily make an active alumina mixture unworkable.

The 48 hours of milling is followed by casting. When the material comes off the casting line, it is in the form of a flexible polymer/alumina tape, dry enough to be cut into easily handled sections.

After casting, a punch press cuts the material into the desired rectangles or other shapes. Holes can be punched at the same time.

Finally, because of the use of active alumina, the material is fired at an unusually low temperature which results in smooth substrate surfaces for reliable thin film bonding. The finished substrate is then ready for the various processes of thin film circuit production.

In developing this new process, engineers at Western Electric's Engineering Research Center worked together with engineers at the Allentown plant.

Conclusion: This new way to produce substrates is a truly significant contribution for thin film circuit production.

The ultimate gain from this smoother substrate is for communications itself. For through the achievement of nearly perfect bonding of metal leads to tantalum nitride, thin films can be produced with even greater reliability and economy.

Western Electric

We make things that bring people closer.

Established 1845 SCIENTIFIC AMERICAN May 1973 Volume 228 Number 5

Cryptography and Computer Privacy

Computer systems in general and personal "data banks" in particular need protection. This can be achieved by enciphering all material and authenticating the legitimate origin of any command to the computer

by Horst Feistel

There is growing concern that computers now constitute, or will soon constitute, a dangerous threat to individual privacy. Since many computers contain personal data and are accessible from distant terminals, they are viewed as an unexcelled means of assembling large amounts of information about an individual or a group. It is asserted that it will soon be feasible to compile dossiers in depth on an entire citizenry, where until recently the material for such dossiers was scattered in many separate locations under widely diverse jurisdictions. It will be argued here, however, that a computer system can be adapted to guard its contents from everyone but authorized individuals by enciphering the material in forms highly resistant to cipher-breaking.

Traditionally those most dependent on secrecy have been military men and diplomats. Their work often calls for the element of surprise, and surprise implies secrecy. Whatever need ordinary people have had for secrecy has remained essentially an individual problem and has rarely been of public concern; lovers and thieves have solved their requirements for communications privacy as best they could. This state of affairs changed little until about the middle of the 19th century. At about that time scientific methods and modes of thought were finally enlisted to improve the techniques of cryptography. Nevertheless, the techniques employed in secret communication remained largely pencil-and-paper operations until well into this century.

Cryptographic encipherment can be achieved in two essentially different ways: by ciphers or by codes. A helpful distinction between the two is as follows. A cipher always assigns substitute symbols to some given set of alphabet letters. Being alphabetic in character, the cipher enables one to say anything that one can print on a typewriter in the particular language that can be represented by its keyboard letters. This means that with a cipher one can say things that have never been said before or even been anticipated as needing to be said. A code, on the other hand, is intrinsically semantic in character. A code can convey only meanings thought of in advance and provided for in a secret list such as a code book.

One should perhaps mention that today the word "code" is frequently used in a sense that does not imply cryptography. The word then usually has a broad meaning embracing sets of symbols with special relations. Thus one speaks of error-detection and error-correction codes, data-compression codes, codes for telecommunications purposes, commercial codes and codes involving all kinds of highly intricate electrical signals and wave forms.

In tackling the privacy problem presented by modern computers at the Thomas J. Watson Research Center of the International Business Machines Corporation we have given the central role to cipher techniques. It will not be possible here to cover the entire subject of "data bank" confidentiality and the securing of computer operations. I do hope to show, however, that certain principles underlying data encipherment and authentication of sources are pertinent to these broad problems.

In modern machine-to-machine networks of the type needed for a data bank the notion of secrecy embraces more than mere message concealment. A data bank includes a network of terminal-to-computer communications. The communications lines connecting the terminals and the computer centers are wide open not only to tapping but also to deliberate alteration and corruption of traffic. In addition to the more obvious aspects of data secrecy, one must therefore provide adequate protection against operational deception of the system. Mere error-detection will not do. The system itself must make it extremely unlikely that an unauthorized but clever and sophisticated person can either enter, withdraw or alter commands or data in such a system.

That is vitally important, because the slightest amount of false data entered in a data-bank system, whether by accident or by intent, can make much of the system's operation worthless. Computers without adequate protection are easily deceived, particularly by operators who thoroughly understand their operation. Indeed, the terminals them-

CLEAR	CIPHER 1	CIPHER 2
A	F	○
B	E	□
C	K	#
D	J	△
E	N	♡
F	P	∃
G	O	✳
H	C	⧅
I	D	✠
J	Y	⊗
K	U	▽
L	W	☽
M	H	⦶
N	M	⊖
O	B	∞
P	Z	₥
Q	L	⬡
R	V	⧄
S	X	◎
T	G	⌂
U	T	Σ
V	R	⊠
W	S	+
X	I	⧈
Y	Q	☼
Z	A	×

CLEAR TEXT

S	E	N	D	M	O	N	E	Y

CIPHER 1

X	N	M	J	H	B	M	N	Q

CIPHER 2

◎	♡	⊖	△	⦶	∞	⊖	♡	☼

SUBSTITUTION, the basic operation in cryptography, is illustrated by two tables in which clear-text letters of the alphabet are replaced by cipher equivalents: other letters (*middle column*) or arbitrary symbols.

selves can be misused for cleverly designed data insertion. In order to preserve the purity of a data bank it will be necessary to authenticate the legitimate origin and character of any data received and to do so rapidly and with very high margins of safety.

Let us begin with the most elementary fact about ciphers: All cryptography amounts to substitution. In its simplest form a substitution can be defined by a tableau, or table. The left side lists the ordinary letters of the alphabet in "clear text"; the right side lists their cipher equivalents, the substituted values. The amateur is frequently impressed with the enormous number of possibilities of arranging such substitution, or permuting, alphabets. For the English alphabet with 26 letters there are $1 \times 2 \times 3 \ldots \times 26$ ways of writing down unique substitution alphabets. (Such a product is referred to as 26 "factorial," written 26!. Any number $n!$ is the product of all the integers from 1 to n.)

The $n!$ possible permutations of a tableau with n entries constitute the number of possible "keys." Now, 26! is a very large number, more than 4×10^{26}. Even so, any simple alphabetic substitution can be broken easily by frequency analysis. If the letter *Q* occurs more frequently than any other in a fairly long sample of cipher text, the analyst can be pretty sure that a cipher *Q* stands for *E*, the letter that occurs more frequently than any other in English clear text, and so on.

Nothing is gained, of course, by replacing the letters of a substitution alphabet with mysterious-looking symbols. An analyst could not care less how complicated the substitute symbols may look. He simply replaces them with a normal alphabet or number substitute of his own and proceeds with his frequency analysis. The mysterious-symbol approach does, however, demonstrate the flexibility with regard to substitute symbols. Since substitution tableaus, in spite of their well-known weaknesses, are of basic importance in cryptographic design, let us consider the possibility of introducing genuinely useful substitution symbols.

This can indeed be done. The binary number system, consisting of just two symbols, 0 and 1, is ideally suited for cryptographic processing by computers. With n binary digits one can generate 2^n distinct binary codes. Thus with a "block" of five binary digits one can generate 2^5, or 32, distinct combinations, or more than enough to encode the 26 letters of the alphabet [*see top illustration on opposite page*]. If we wish to name or label more items, we can increase our reservoir of distinct binary numbers merely by increasing the size of the digit block. Every time we increase the block size by one digit position we double the number of possible codes. Hence a six-digit code would provide 2^6, or 64, distinct codes, or enough to include numerals, punctuation marks and so on.

In addition to the ease with which binary digits can be represented in electronic circuitry (by on and off signals, for example) they have all the advantages of ordinary decimal numbers. Thus binary numbers can be added, subtracted, multiplied and so on, just as ordinary decimal numbers can. As we shall see, arithmetic manipulation plays a vital role in cryptographic techniques designed for computers.

Now we can introduce complexity in a fundamentally different way. Instead of using just one substitution table we can use several, in some disarranged but prearranged order. The ordering pattern constitutes a key. If we use only two tables, tagged 0 and 1, then a typical key might be 1101101. We are now dealing with multialphabetic substitution. With this new source of complexity available the question arises: Can one not now simplify the substitution tables, perhaps by making them smaller? The simplest binary substitution one can perform is on single message digits, in which case there are only two possible distinct substitution tables. Let us therefore set up two substitution tables, one for each of the two basic types of key [*see bottom illustration on opposite page*]. The table marked Key 1 simply interchanges the 0's and 1's; the table marked Key 0 preserves their identity. These are the only two possibilities. Now, it happens that the same effect can be obtained by the operation known as addition modulo 2: two digits of the same kind add up to 0, two digits of the opposite kind add up to 1. And so, in this special case, the pattern key can be called an addition key.

Before proceeding further let me explain that from here on we shall tacitly assume that the messages we wish to handle cryptographically have first been translated into a sequence of binary digits. Any kind of information, be it letters, musical sounds or television signals, can be represented by binary coding. We shall no longer worry about the original meaning of the messages. We shall deal only with their binary representation.

We are now ready to see what happens when we encipher a sequence of binary digits, let us say 0001010, into a new sequence using the two keys, Key 1 and Key 0, in some arbitrary sequence: 1101101. Remembering the rule that Key 1 interchanges 0's and 1's and that Key 0 leaves digits unchanged, we get the following:

Message:	0001010
Key:	+ 1101101
Cipher:	1100111

This represents addition modulo 2. It has the convenient property that subtraction is the same as addition, so that the message can be recovered simply by adding the sequence of digits in the key (which is known to the intended recipient) to the sequence in the cipher:

Cipher:	1100111
Key:	– 1101101
Message:	0001010

The question immediately arises: Is this simple cipher of any practical value? Since the cipher in effect employs only two substitution tables of minimal size it is perhaps obvious that we must switch from one to the other frequently and in random fashion, that is, add a random sequence of key digits. Suppose we do. The astonishing answer is that we then have a potentially undecipherable cipher. From the viewpoint of information theory, what this cipher does is to add to each bit of message information one bit of key information (misinformation!). This is enough to completely destroy any kind of structure the original message may have had, provided that the key digits are picked at random, say by flipping a coin, and that the key sequence has the same length as the message and does not ever repeat.

Why is this system genuinely undecipherable? Actually it is no "system" at all. The basic cryptographic transformation amounts to no more than random addition of single digits. It is as trivial as that. The method derives its strength solely from the fact that for each message digit we completely and randomly change the key. This is the only class of ciphers that one can prove to be undecipherable in the absolute sense.

Even if an opponent made a brute-force attempt to break the system, for

CLEAR	BINARY EQUIVALENT	BINARY CIPHER
A	00000	00100
B	00001	01001
C	00010	10001
D	00011	01010
E	00100	00011
F	00101	10011
⋮	⋮	⋮
X	10111	11101
Y	11000	01110
Z	11001	10110

TRANSLATION TO BINARY SYSTEM is required for computer cryptography. A binary substitution-table cipher is constructed, for example, by first translating each letter of the alphabet into a five-digit binary number and then enciphering each binary equivalent.

a

KEY 0

CLEAR	CIPHER
0	0
1	1

KEY 1

CLEAR	CIPHER
0	1
1	0

b

KEY	MESSAGE	
	0	1
0	0	1
1	1	0

c

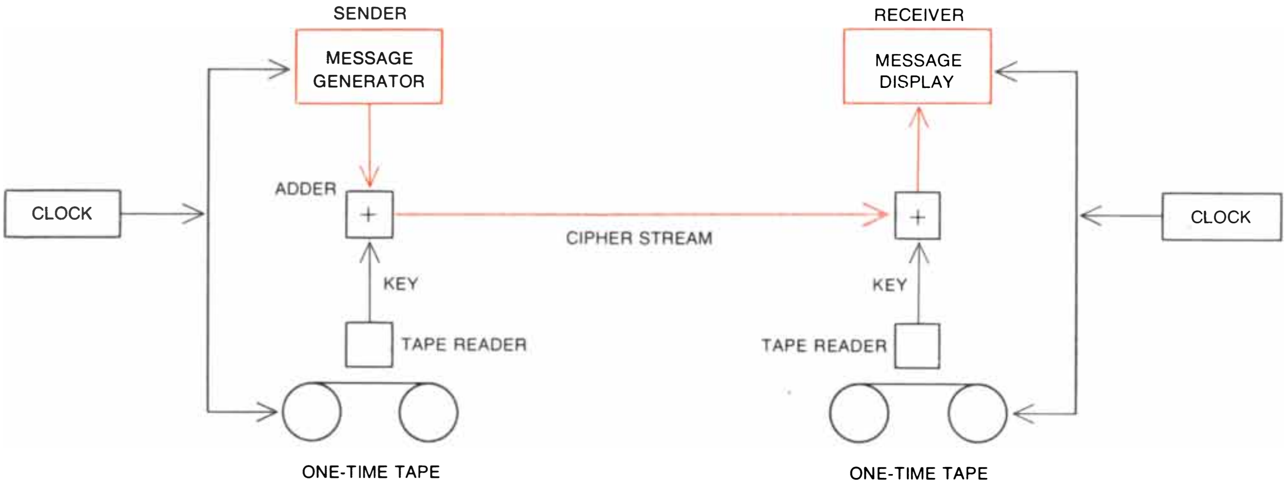
MESSAGE		KEY		CIPHER
0	+	0	=	0
0	+	1	=	1
	+	0	=	1
1	+	1	=	0

d

F	=	00101
KEY	=	10110
CIPHER	=	10011

CIPHER	=	10011		
KEY	=	10110		
		00101	=	F

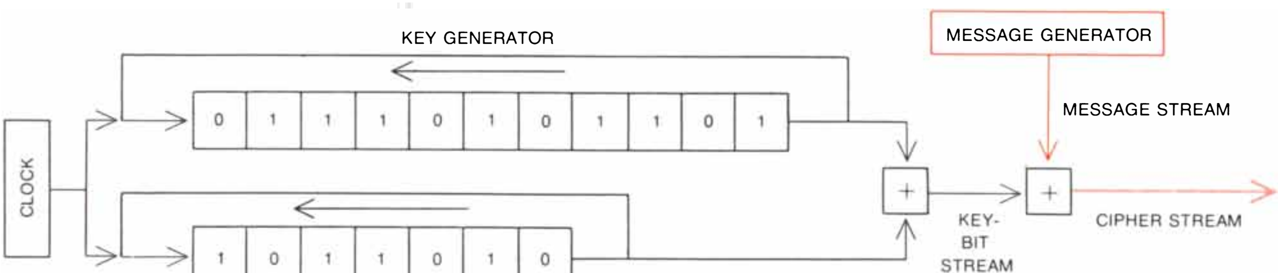
BINARY SUBSTITUTION in its simplest form allows only two possibilities (*a*): in one table (*Key 0*) there is no change from clear to cipher; in the other table (*Key 1*) the cipher is the opposite of the clear. In this case of binary substitution the two tables can be represented by a single addition table modulo 2 (*b*); in this case binary substitution is equivalent to binary addition (*c*). The cipher key for addition modulo 2 is an arbitrary sequence of 1's or 0's. To encipher the binary equivalent of a message letter, one adds a key digit to each message digit (*d*). To decipher, one subtracts the key (the same as adding modulo 2).



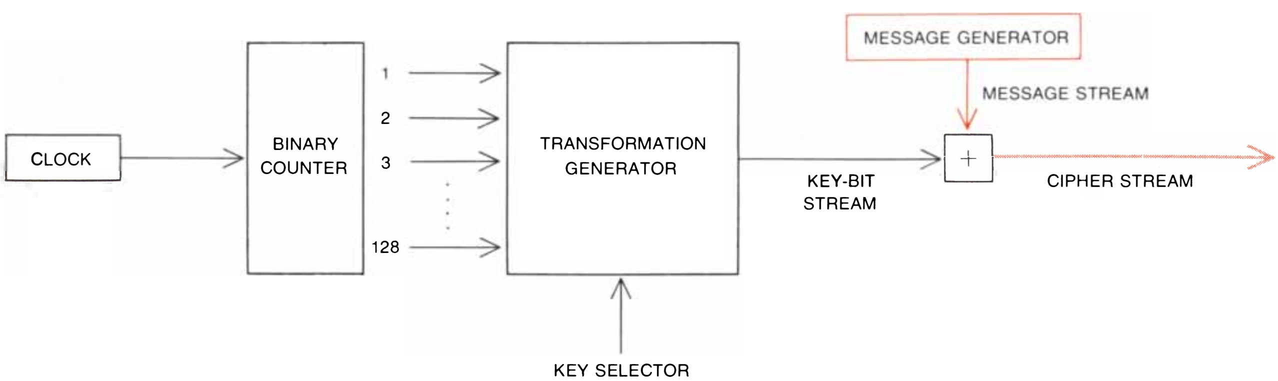
"ONE-TIME-TAPE" SYSTEM requires that the sender and the receiver have tapes carrying the same key, synchronized by clock devices. The digits of the key and of the message are added, the resulting cipher stream is transmitted and the key is subtracted (added modulo 2). For large traffic volumes a vast store of key digits must be conveyed securely to the receiver and filed there.

example by trying all the possible addition keys (2^6, or 64, in the case of our six-digit message), he would get all possible clear texts, including the one we had actually enciphered. Thus if we had enciphered the name "Tom" (which would actually take a minimum of 15 binary digits), the analyst would find among his trial decipherments all English three-letter names, such as Joe, Jim, Job and so on, including Tom, but no clue as to which name was correct. Not even a god or demon who could try all possible keys in an instant could ever establish any consistency. The system is well known and is in actual use by all major governments under various names, such as the Vernam system or "one-time pad."

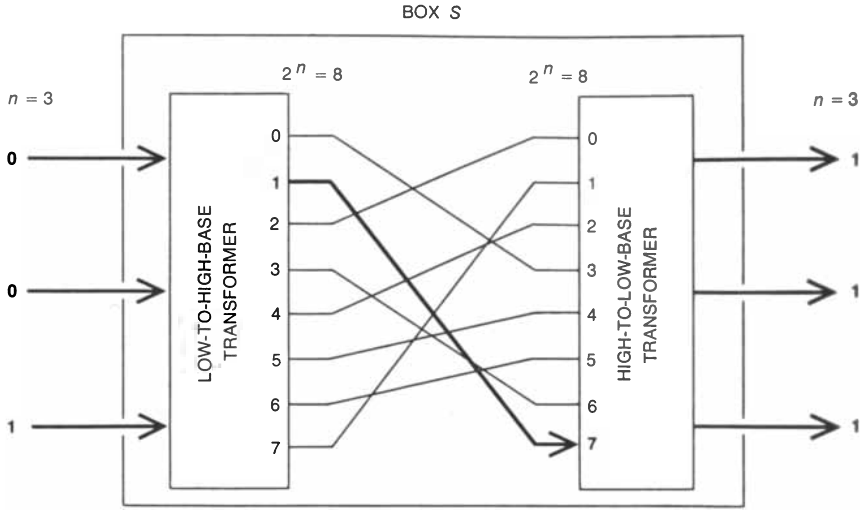
In a practical system two identical tapes of random key digits are prepared; the tapes can be of any type—printed, punched, magnetic or whatever. One tape is held by the sender and the other is transmitted by "noninterceptible means," say by guarded courier, to the legitimate receiver. When the sender wants to transmit a message, he converts it first to binary form and places it in an apparatus that adds a digit, modulo 2,



PSEUDORANDOM-TAPE SYSTEM incorporates two tape loops of key digits, or bits, that are added together to generate a "pseudorandom" key-bit stream that is in turn added to the message stream as in the one-time-tape method and is deciphered by an identically generated key at the receiver. Short tapes thus simulate long ones, but the resulting internal periodicities may be useful to an analyst.



PSEUDORANDOM STREAM can be generated in more complex forms. In this generalized representation a binary counter provides the input numbers for the transformation generator, which supplies the key-stream digits to be added to the message stream.



IN	OUT
0 0 0	0 1 1
0 0 1	**1 1 1**
0 1 0	0 0 0
0 1 1	1 1 0
1 0 0	0 1 0
1 0 1	1 0 0
1 1 0	1 0 1
1 1 1	0 0 1

SUBSTITUTION BOX, unlike the stream devices, is general, and includes both linear and nonlinear transformations: it does not merely add a 0 or 1 to input digits but can substitute for any input digit block any output digit block. It consists essentially of two switches. One converts a binary number of n digits into a one-digit number base 2^n; the other reconverts. The box therefore provides 2^n internal terminals that can be wired in $n!$ (n factorial, or $1 \times 2 \times 3 \ldots \times n$) different ways; in the case of this box with $n = 3$, that means 40,320 different wirings, or different tableaus like the one illustrated. A box of this kind with $n = 128$ would defeat analysis, but it is not possible to realize technologically.

from the key tape to each message digit [*see top illustration on opposite page*]. At the receiver the coded message is recorded and run through a machine similar to the encoder that adds (subtracts) a digit, modulo 2, from the key tape to each digit of the coded message, thereby yielding the clear text. Naturally the key tape must be stepped with absolute synchronization to duplicate the encoding operation.

The fundamental drawback to the Vernam system is that for every "bit" of information transmitted the recipient needs to have in his possession in advance one bit of key information. Moreover, these bits must be in a random sequence that cannot be used a second time. For large traffic volumes that is a severe restriction. Because of this requirement the Vernam system has usually been reserved for top-secret messages.

To get around the problem of supplying the recipient with large volumes of random key digits in advance, designers and inventors have devised many ingenious schemes for generating very long pseudorandom streams of digits from several short streams according to some algorithm. The recipient of a coded message can then be provided with a generator that operates exactly like the one used to add pseudorandom digits to the original message. An algorithm of course implies systematic routines, which in turn introduce regularities for which an analyst may search.

One basic method for building such a generator is to use two or more key-bit tapes that are added, bit by bit, to yield a compounded stream. For example, the simple one-time tape can be replaced with two tape loops of prime or relative-prime lengths. In this situation the lengths have no common factors and the compounded stream has the length of the product of the component streams: two tapes of lengths 1,000 and 1,001 respectively produce a compounded stream that will not repeat for 1,000 × 1,001, or 1,001,000, digits. The tapes are circulated through an adder that makes a modulo-2 addition of the digits sampled, one from each tape [*see middle illustration on opposite page*]. The output of the adder serves as the key digit for encoding the message. It is therefore important that the compounded stream exceed the length of all the messages one expects to send over a reasonable period of time.

Since any bit-by-bit adder is linear, it is inherently weak, but it can be cryptographically strengthened in any number of ways. One can pile complication on complication by introducing intricate feedback arrangements, perhaps linked in some fashion to the message, or by introducing such nonlinear mathematical operations as substitutions on digit blocks of suitable size. The nonsecret cryptographic literature contains many fascinating designs for pseudorandom-stream generators, all of which can in principle be reduced to one basic scheme [*see bottom illustration on opposite page*]. One way or another they generate a pseudorandom number by performing complex mathematical operations on some ordered sequence of input numbers, thereby transforming it in a way that is supposed to confound an analyst.

The reader will be surprised and disappointed to learn that the Vernam class of ciphers—the only class that can be proved to be undecipherable in the absolute sense—is not perfect at all in another sense: it does not in itself offer protection against clever deception on nonredundant traffic. Whether a message is encoded by the use of truly random digits or by pseudorandom-stream digits, in bit-by-bit encipherment a single error that arises during transmission is confined to a single digit position; the error does not "propagate," or spread, through the rest of the message. The cipher does not introduce intersymbol dependence. When the message itself is in a "natural" language such as English text, the normal redundancy context makes it easy for a human reader to spot an occasional error. Thus if some of the five binary digits representing the letter *E* were corrupted to turn it into the binary sequence for *F* (so that, for example, SECRET came out SECRFT), a human interpreter would detect the error immediately.

The situation is quite different in a

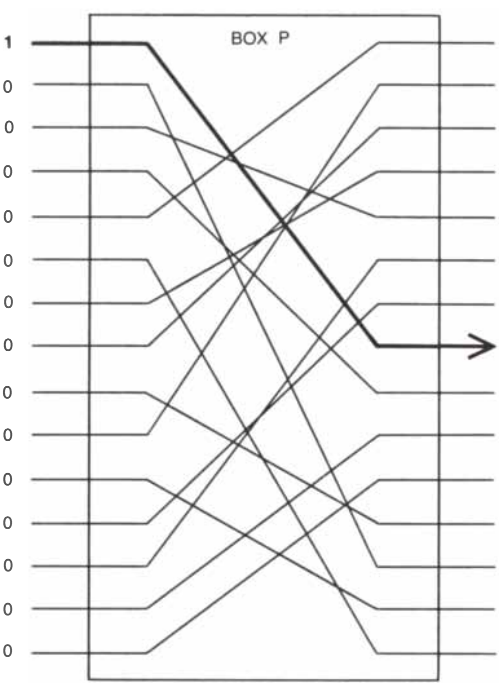
computer environment. Here the data transmitted may often be nonredundant if, for example, they are purely numerical, and an error in a single digit can cause an avalanche of computational errors. Study of the problem has shown that simple error-detecting codes are inadequate for guarding the integrity of computer data against possible tampering by a human expert. What is required is not mere error-detection but cryptographically protected authentication. Surprisingly, this is best achieved by relying on certain principles inherent in the cipher structure itself. Rather than trying to modify the stream concept, let us take a fresh look at the basis of all cryptography: substitution on blocks of message digits.

We shall refer to any cipher that converts n message digits into n cipher digits as a block cipher. For example, a block cipher would be one that turns 00000, standing for a clear-text *A*, into, say, 11001, the cipher equivalent of *A* according to some permutation key, exactly as a tableau does. To see how such a binary transformation is performed by an electronic device let us consider a substitution on only three binary digits [*see illustration on preceding page*].

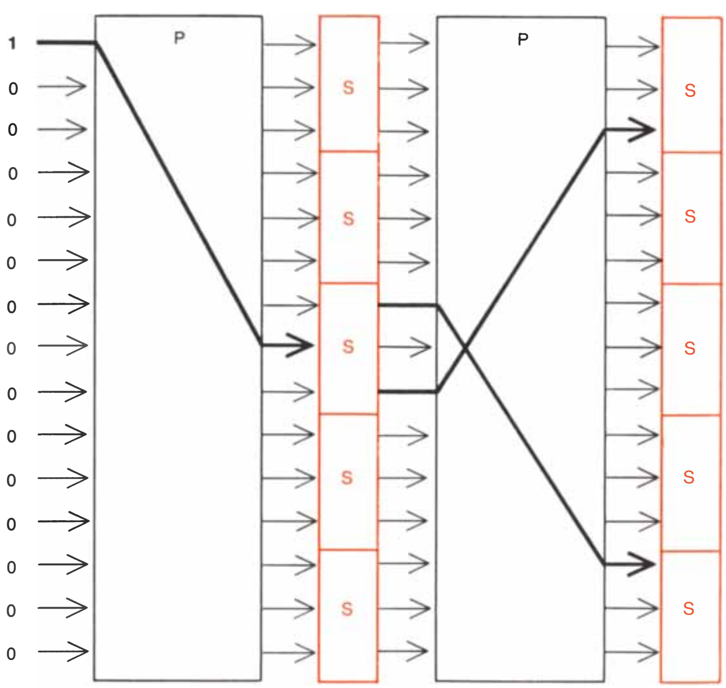
Three binary digits can represent eight items: 2^3 equals eight. The substitution device consists of two switches. The first converts a sequence of three binary digits into its corresponding value to the base eight, thereby energizing any one of eight output lines. These eight lines can be connected to the second switch in any one of 8!, or 40,320, ways. We are at liberty to decide which one of these 40,320 distinct connection patterns, or wire permutations, is to be made between the first switch and the second switch. The role of the second switch is to convert the input, presented as one digit to the base eight, back into a three-digit binary output.

If the substitution device were built to handle a five-digit binary input, it could be used to encipher an alphabet of 32 letters. The number of possible connection patterns between the two switches would then be 32!. That would seem to be an incredibly large number of keys, but the cipher produced must still be regarded as glaringly weak: it could not resist letter-frequency analysis. The weakness is not intrinsic; the device described is mathematically the most general possible. It includes, for any given input-output dimension, any possible reversible cipher that has been or ever could be invented; mathematicians would say it represents the full symmetric group. It is completely "nonsystematic": one permutation connection tells an opponent nothing at all about any other connection. The problem is not intrinsic, then, but is related to size. In spite of the large number of keys, the "catalogue" of possible inputs and outputs is too small: only 32. What is required is a catalogue so large that it is impractical for any opponent to record it. If we had a box with 128 inputs and outputs, for example, an analyst would have to cope with 2^{128} (or more than 10^{38}) possible digit blocks, a number so vast that frequency analysis would no longer be feasible. Unfortunately a substitution device with 128 inputs would also require 2^{128} internal terminals between the first and the second switch, a technological impossibility. This is a fundamental dilemma in cryptography. We know what would be ideal but we cannot achieve the ideal in practice.

Perhaps one could find a device that is easy to realize for a large number of inputs. One might, for example, build a box with, say, 128 input and 128 output terminals that are connected internally by ordinary wire crossings [*see illustration at left below*]. Such a "permutation box" with n! terminals would have n! possible wire crossings, each of which could be set by a different key. It could be built easily for $n = 128$. Although this provides a usefully large number of keys (128!), we are now faced with a new difficulty. By the use of special trick messages it is possible to read out the complete key to such a system in only $n - 1$ (in this case 127) trials. The trick



PERMUTATION BOX can handle very many terminals but it only shuffles positions of digits. An opponent can learn its wiring by feeding in inputs with single 1's and seeing where 1's come out.



PRODUCT-CIPHER SYSTEM combines *P* boxes and *S* boxes. The *P* boxes have a large number of inputs (represented by 15 in the illustration) and the *S* boxes a number that is manageable for such

is to introduce a series of messages containing a single 1 at $n-1$ positions; the position of the 1 in the output betrays the particular wire crossing used in the box. The flaw in the simple permutation box is again that it is a linear system.

We need a compromise that will at least approximate the features of the general system. We are led to the notion of a product cipher in which two or more ciphers are combined in such a way that the resulting system is stronger than either of the component systems alone. Even before World War I various cumbersome ciphers using several stages of encipherment were studied. The first genuinely successful example was probably the one devised by the Germans that was known as the ADFCVX system. We need only observe here that it coupled "fractionation" with "transposition." By that procedure a message was broken into segments and the segments were transposed. The important fact to note here is that the result of a product cipher is again a block cipher; the goal, of course, is that the cipher behave as much as possible as if it were a general substitution cipher.

Between World War I and World War II interest in product ciphers almost totally disappeared because of the successful development of rotor, or wired-wheel, machines, which belong to the general class of pseudorandom-stream generators. A typical rotor machine has a keyboard resembling that of a typewriter. Each letter is enciphered by the operation of several wheels in succession, the wheels being given a new alignment for each new letter according to an irregular and keyed stepping algorithm. The message is decoded by an identical machine with an identical key setting.

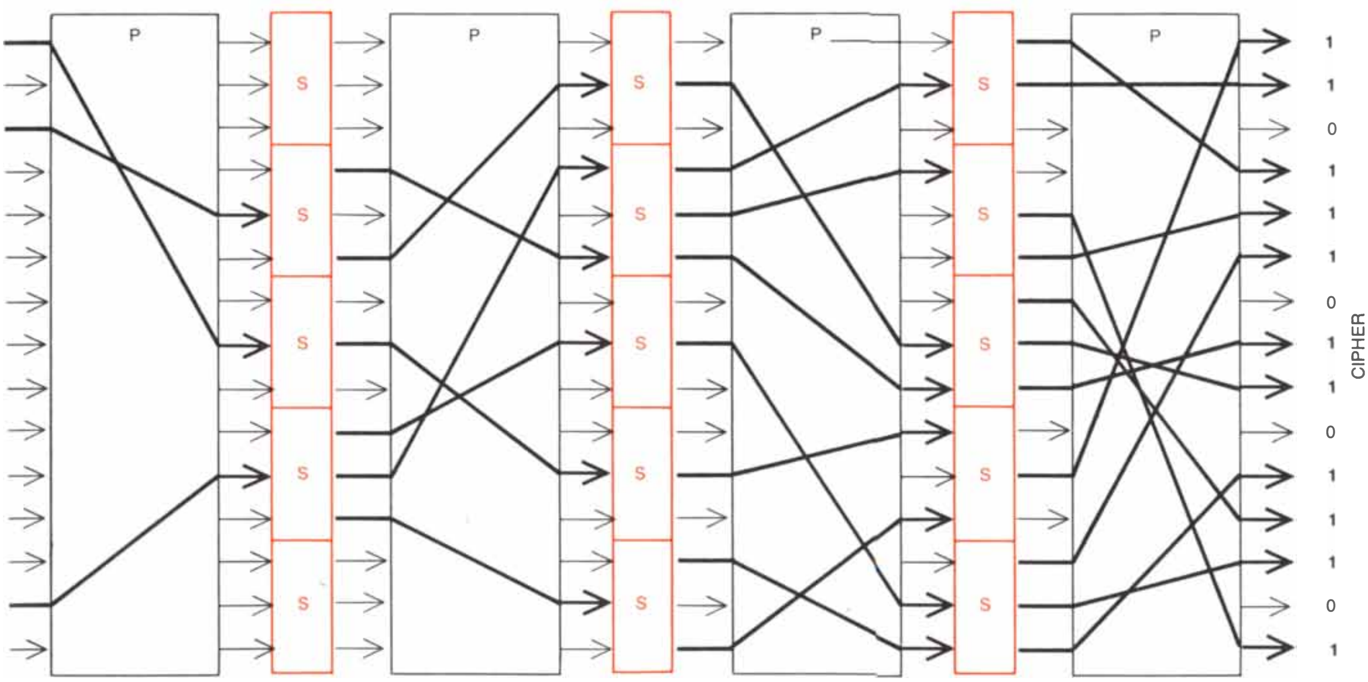
The modern interest in product systems was stimulated by a paper by Claude E. Shannon titled "Communication Theory of Secrecy Systems," published in the *Bell System Technical Journal* in 1949. In a section on practical cipher design Shannon introduced the notion of "mixing transformation," which involved a special way of using products of transformations. In addition to outlining intuitive guides that he believed would lead to strong ciphers, he introduced the concepts of "confusion" and "diffusion." The paper opened up almost unlimited possibilities to invention, design and research.

The manner in which the principles of confusion and diffusion interact to provide cryptographic strength can be described as follows. We have seen that general substitution cannot be realized for large values of n, say $n = 128$, and so we must settle for a substitution scheme of practical size. In the IBM system named Lucifer we have chosen $n = 4$ for the substitution box. Even though 4 may seem to be a small number, it can be quite effective if the substitution key, or wire-crossing pattern, is properly chosen. In Lucifer nonlinear substitution effectively provides the element of confusion.

We have also seen that a linear permutation box is easy to build even for $n = 128$. The number of input and output terminals is simply equal to n. Being a pure digit-shuffler, a device that merely moves digits around without altering the number of 1's in the data, the permutation box is a natural spreader of confusion, that is, it can provide optimal diffusion.

In the Lucifer system the input data pass through alternating layers of boxes that we can label *P* and *S*. *P* stands for permutation boxes in which n is a large number (64 or 128) and *S* stands for substitution boxes in which n is small (4). Whereas either *P* boxes alone or *S* boxes alone would make a weak system, their strength in combination is considerable.

One measure of strength is depicted in a device in which for simplicity the *P* boxes have $n = 15$ and the *S* boxes have $n = 3$ [*see illustration on these two pages*]. If we imagine this sandwich of boxes being "tickled" by addressing it with a specially selected input, which might consist of a number made up of



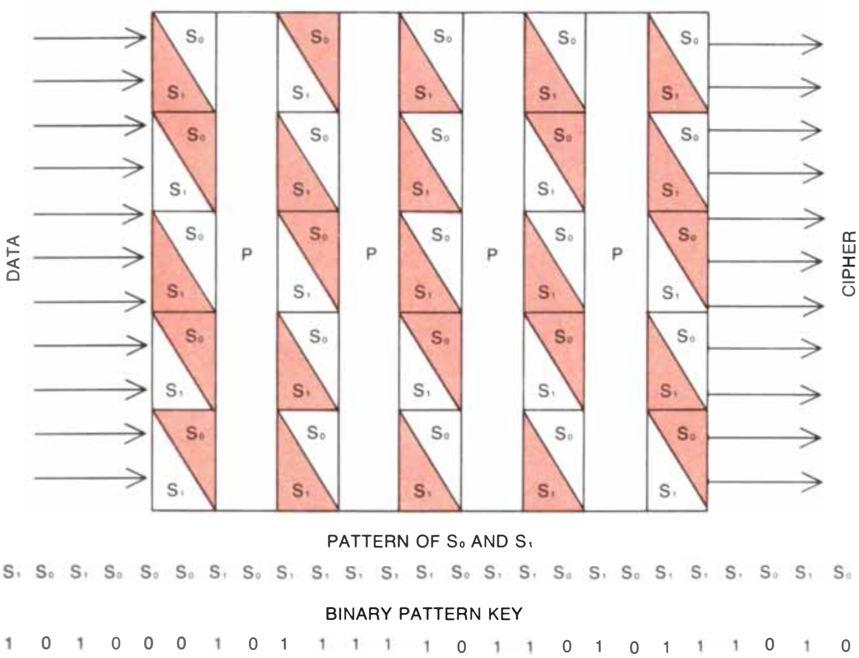
devices—three in this case. The *P* boxes shuffle the digits, providing "diffusion." The *S* boxes provide nonlinear substitution and thus "confusion." In this simplified example the input includes a single 1 and 14 0's. Because the *S* boxes are nonlinear, they can potentially increase the number of 1's; meanwhile the *P* boxes move the 1's around. The result can be an unpredictable avalanche of 1's.

14 0's and only a single 1, it is easy to see confusion and diffusion at work. The first box, *P*, moves the single 1 to some box *S*, which being a nonlinear device may convert the 1 to a three-digit output having as many as three 1's. In the diagram it actually generates two 1's. The next box, *P*, shuffles the two 1's and feeds them into two separate *S* boxes, which together have the potential of generating as many as six 1's. From here on the diagram is self-explanatory. As the input moves through successive layers the pattern of 1's generated is amplified and results in an unpredictable avalanche. In the end the final output will have, on the average, half 0's and half 1's, depending on the permutation keys chosen for the various *P* and *S* boxes.

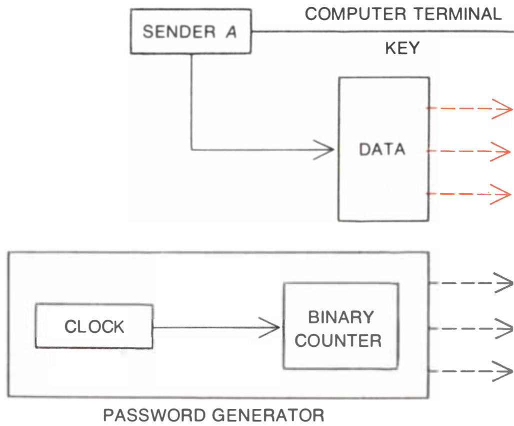
The important fact is that all output digits have potentially become very involved functions of all the input digits. Since all the boxes are independently keyed, the avalanche and its final effect cannot be predicted. The goal of the designer, of course, is to make it very difficult for an opponent to trace the pattern back and thus to reconstruct the permutation keys on *S* and *P*.

In the actual Lucifer system we found it advisable to introduce a few important changes. For example, box *S*, being perfectly general, could be accidentally keyed to behave exactly like a *P* box, in which case the entire system would be no better than a single layer of *P*, which could be exposed by tickling. To avoid this result both *S* and *P* boxes are permanently keyed with permutations that are considered strong; these permanent keys will be known to anyone owning the boxes. Hence we need another keying capability, preferably one that can be represented by a binary number.

This can be accomplished by building a sandwich in which each *S* box has two different permanent keys and can thus be present in two different possible states, S_0 and S_1. The sequence of these states in any one sandwich constitutes a keyed pattern unknown to a potential opponent. The pattern can be represented by a binary key that in effect says which of two substitution tables is to be used, just as in the case of the two-table substitution discussed earlier [*see illustration below*]. (In the diagram there are 25 *S* boxes, and so the key is 25 digits long; in an actual device there would be many more *S* boxes and a correspondingly longer key.) The key digits can be loaded into a key register in the cryptographic device and possibly recorded magnetically in the key card assigned to the authorized user of the system. When two states of the *S* box are used in this way, the resulting cryptogram exhibits a sensitive intersymbol dependence that makes all output digits complicated functions not only of all message digits but also of all the digits in the key. The



INDIVIDUAL KEYING CAPABILITY is required because the keys built into the *S* and *P* boxes are known to anyone with access to the same system. Individual keying capability can be provided as shown here. At each *S* position in the system there are two possible states, S_1 or S_0. The pattern in which they are intermixed is established by a binary key as shown.



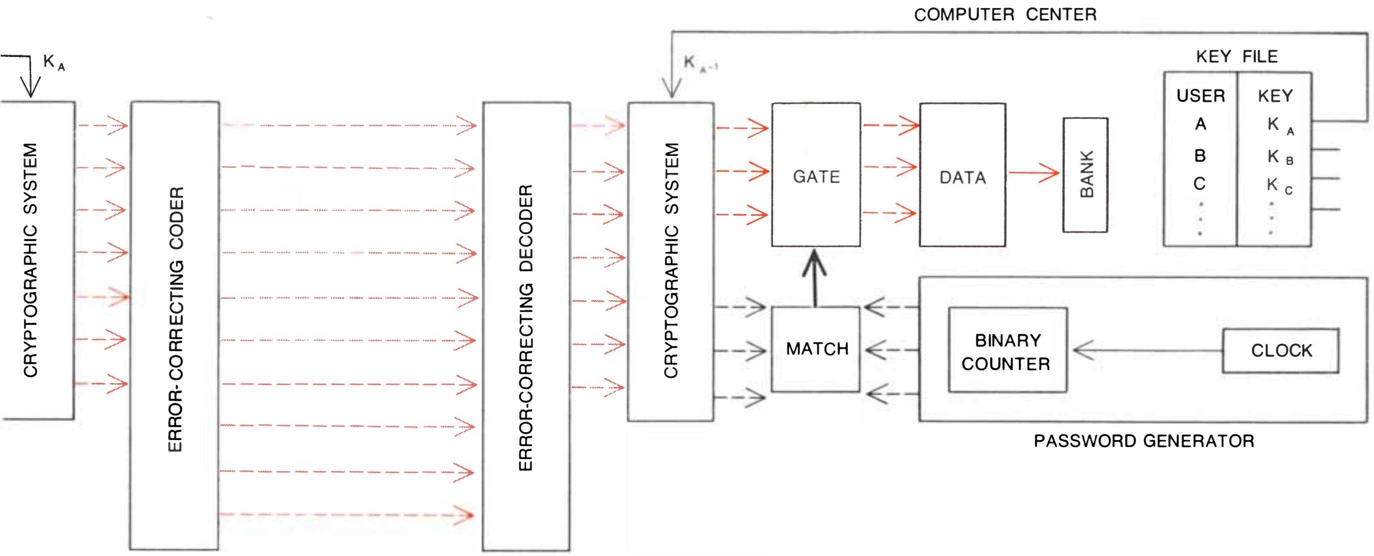
COMPLETE SYSTEM combines a password generator, a cryptographic product system such as the *P* and *S* sandwich in the preceding illustrations and error-correction. The password generator provides a fresh pass-

concept has so far proved to be resistant to penetration by mathematical analysis.

Although the strong intersymbol dependence is a necessary (but not sufficient) index of cryptographic strength, it has a dark side: it implies great sensitivity to noise or interference during transmission. The corruption of a single digit position can, through the inverse-avalanche effect, result in complete garble at the receiver. Modern communications techniques make the problem less serious, however, at least for nonmilitary uses.

Moreover, the strong interdependence among digits provides a surprising and unexpected windfall. Because the system is so sensitive and so violent in its response to changes, it is automatically an ideal agent for detecting such changes, whether generated by accident or by clever intent. And so we have now simultaneously combined high message secrecy with a tamper-proof alarm for error-detection.

To make use of this automatic feature we need only reserve space for a "password" within a given block of message digits. The password is a sequence of digits fed into the message stream automatically by the sending apparatus and does not concern the person using the system. The role of the password is to tell the receiving apparatus that the message has not been intentionally tampered with or seriously degraded by noise during transmission. The enciphering process keeps the opponent from knowing how message bits and password



word block (*black dashes*) for each data block. The sender, using his individual key, introduces his data (*colored dashes*). Password and data digits are no longer traceable as such after being enciphered according to the key. Error-correcting digits are introduced and removed after transmission. The cryptographic system at the computer center deciphers the transmission according to the inverse of the sender's key, which has been selected from a secure file in the center, and extracts the password digits. If these match the password generated in the computer, the gate opens and the input data are admitted to the bank as being of legitimate origin.

bits are reflected in the cryptogram. If the password digits cannot be cleanly recovered by the decoder at the receiving end, the message is rejected.

The decisive role in this scheme is played by the password generator, which is required at both the transmitter and the receiver [*see illustration above*]. The password generator is really nothing more than a binary clock or counter that indicates the time or order number of the message in binary notation and combines a sequence of time digits with each block of data digits transmitted. We assume that at some initial time, say 8:00 A.M., the clocks at the two ends of the transmission channel are synchronized to the same stepping frequency.

Now let us see how the "password authentication scheme" provides privacy for the members of a centralized data-bank community who have access to a large central computer. Each member has his own private key, perhaps in the form of a sequence of binary digits recorded magnetically on a card. The keys of all members are listed in a suitably protected form at the central computer. Let us suppose that a member with Key *A* wishes to send a message to the computer. He inserts the card bearing his key in a typewriterlike terminal sitting on his desk, waits a second or two for a signal that the line is free and begins typing his message.

The message is automatically broken up into blocks of data digits (say 64) and combined at each tick of the binary clock with the password (which might also be 64 digits) that corresponds to the output of the clock at that moment. The resulting block of 128 digits is enciphered by its passing through the sandwich of *S* and *P* boxes that thoroughly mix password digits and data digits.

Since the resulting cryptogram is sensitive to transmission errors, it is fortified with an appropriate error-correcting code, which might be matched to the noise characteristics of the telephone line being used. The addition of the code lengthens the password-plus-message block by a few digits. The resulting cipher block is preceded by the address of the sender in the clear and is transmitted to the central computer. When the message arrives, the key *A* belonging to Member *A* is looked up in the listing and the inverse of Key *A* is fed into the decoder to decipher the cryptogram.

The decisive test now is whether or not the cryptogram as received will yield the same password as the one generated locally by the binary clock at the receiver. In the absence of interference, and if the cryptogram was indeed enciphered with Key *A*, the output of the decoder will consist of the data digit block and the password digit block. This is taken as sufficient proof that the cryptogram did in fact originate with Member *A*. The data are accepted as being of legitimate origin.

What happens if there is interference? If it is nothing more than sporadic bursts of noise on the transmission line, the error-correcting code will eliminate it and the message will pass the authentication test. If, however, the errors are of such a nature that they are not eliminated by the error-correcting code, even one false digit will produce an avalanche effect in the decoder and will garble the output. The passwords will no longer match. The system regards the cryptogram as being of suspicious origin and rejects the communication.

The password test would function just as reliably if someone were to record an intercepted communication and were to retransmit it when the password was no longer valid. The use of a false key, of course, is an instant cause for rejection. The system appears to be proof against any imaginable attempt to deceive it. Each binary digit invested in the password provides one bit of authentication information. If the password is n digits long, an opponent has only one chance in 2^n (or one chance in 2^{64} if $n = 64$) of generating by any means he may select a cryptogram that on decipherment will yield a password that is accidentally correct. The number 2^{64} is equal to about 10^{19}. It is not possible to authenticate more efficiently.

There is no reason why the security system described for a single link could not be expanded to provide security for all users of a network. Other cryptographic approaches are still being studied in our laboratory and elsewhere. It would be surprising if cryptography, the traditional means of ensuring confidentiality in communication, could not provide privacy for a community of data-bank users.

The Role of the Heartbeat in the Relations between Mother and Infant

Most mothers exhibit a natural preference for holding their baby on the left side of their chest. There is reason to believe that the sound of the adult heart has a soothing effect on the infant

by Lee Salk

Some years ago I was stimulated by a series of random observations on poetry and anatomy to pursue what may seem to be an unlikely project for someone who works in the behavioral sciences. As a psychologist I had learned that the hypothalamus, a small region at the base of the brain, plays an important role in the expression of emotions. Yet in my leisure reading I was repeatedly impressed by the fact that, in referring to the wellspring of deep emotions such as love, poets and lyricists from all periods in history and from all parts of the world consistently chose another part of the anatomy: the heart. References to the heart in connection with the expression of love are of course found not only in poetry, literature and song but also in everyday language: "I love you from the bottom of my heart," "My heart longs for you" and "I am heartbroken."

When I tried to apply my professional knowledge to these expressions ("I love you from the bottom of my hypothalamus" or "My hypothalamus longs for you"), the results somehow lacked a true romantic ring. The outcome was no better when I envisioned receiving a St. Valentine's Day card with Cupid's arrow piercing a hypothalamus. I quickly abandoned these attempts, but I was still left with many unanswered questions about why the heart and love are so often linked. As I continued my haphazard observations, my respect for folk knowledge grew.

One day in New York's Central Park Zoo I reflected that a mother rhesus monkey was holding her newborn infant not close to her hypothalamus but close to her heart. Moreover, in 42 subsequent observations of the same monkey and infant, I found that she held the newborn on her left side 40 times and on her right side only twice. Could it be, I wondered, that "Close to a mother's heart" is more than just an expression? Could the phrase have a basis in behavior and represent a psychobiological process? A survey of the scientific literature revealed no studies concerned with the problem. I therefore undertook an investigation designed to answer the following question: What does a human mother do when she is presented with her newborn infant?

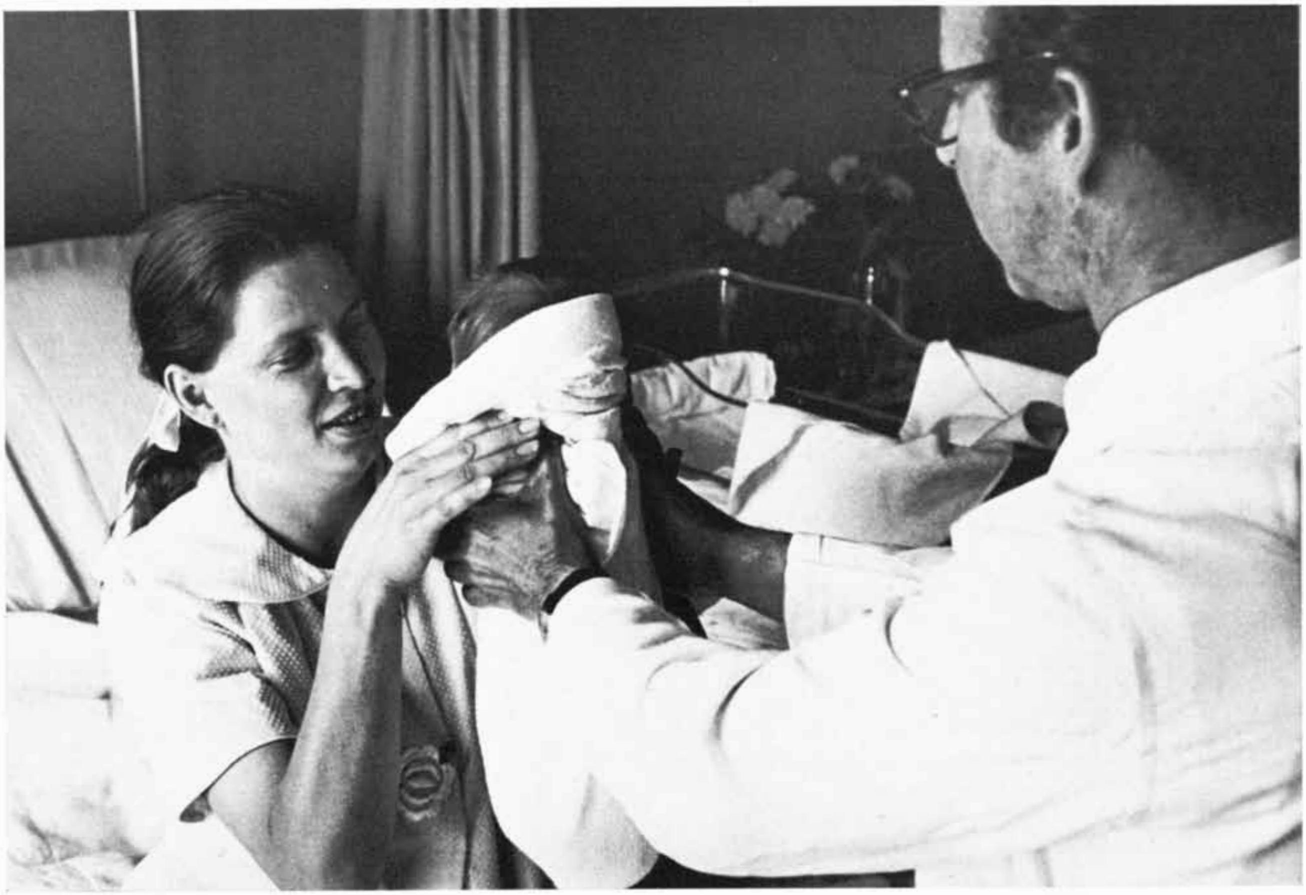
Of 255 right-handed mothers whom I observed during the first four days after they had given birth, 83 percent held their baby on the left side of their chest or on their left shoulder and 17 percent held the baby on the right side. I also observed 32 left-handed mothers: 78 percent held the baby on the left side and 22 percent on the right. All the deliveries had been normal, and the mothers had had access to their infants soon after birth.

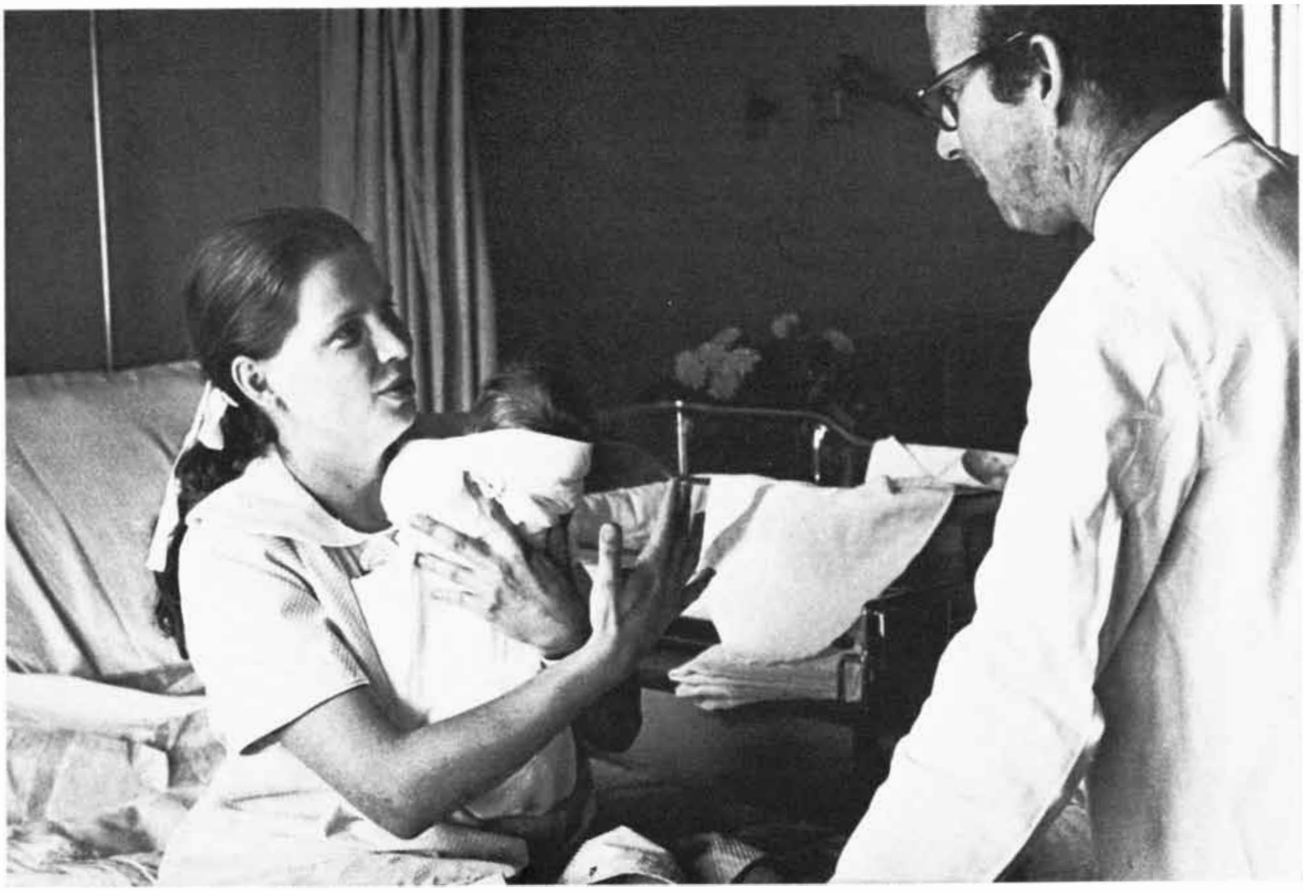
When the left-handed mothers were asked why they held their baby on the left side, they thought for a moment and replied: "I'm left-handed and can hold my baby better this way." When right-handed mothers were asked why they held their baby on the left side, they replied: "I'm right-handed, and when I hold my baby on the left side it frees my right hand to do other things." It seems unlikely that the right- and left-handed mothers were doing the same thing for different reasons. I felt they were giving me a rationalization for an automatic response that was not related to handedness.

In an independent study I. Hyman Weiland of the University of Southern California observed a large number of mothers holding their infant while they were at a well-baby clinic. He found a definite preference for holding the infant on the left side. As a control, he collected observations of shoppers carrying packages approximately the size of a baby. He noted on which side the shoppers held the package as they left a supermarket that had automatic doors, so that it was not necessary for the person to free one hand to open the door. Of the 438 adults he observed, exactly half held the package on the left and half on the right. No preference for one side or the other was shown by either sex. Weiland then postulated that holding a package does not arouse anxiety but holding a baby does.

To find out if there was a left-side preference for holding an object when a person was in a state of anxiety, Weiland and his colleague Zanwil Sperber tested patients undergoing dental therapy. Most people admit to being anxious when they go to the dentist, and Wei-

WORKS OF ART, from all parts of the world and from all historical periods, depicting an infant being held by a woman will, in about four cases out of five, show the infant being held on the woman's left side. This tendency is particularly noticeable in paintings of the Madonna and the Christ child, such as the one on the opposite page by Piero della Francesca, the 15th-century Italian painter of the Umbrian school. The painting, titled "The Virgin and Child with Two Angels," originally came from the Church of Santa Maria delle Grazie near Senigallia and is now at the Galleria Nazionale delle Marche in Urbino.





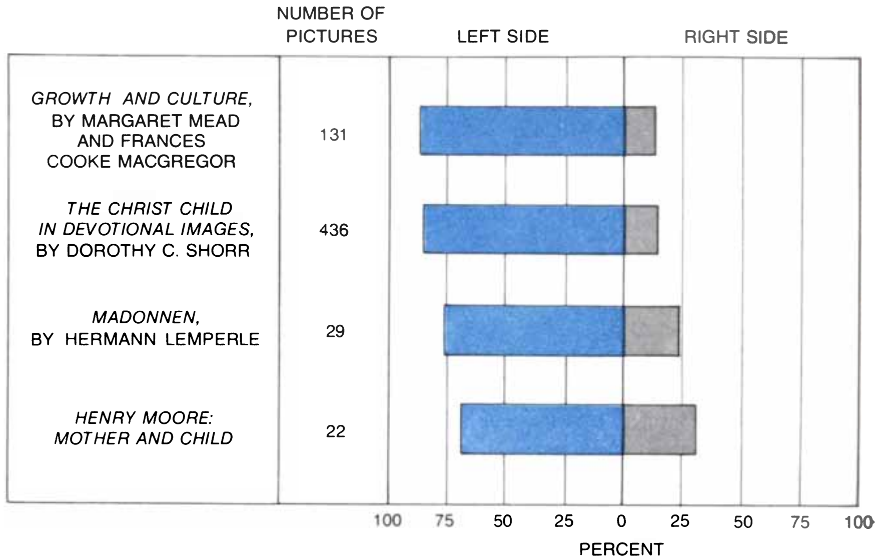
SIDE PREFERENCE for holding a baby is determined by presenting the infant directly to the midline of the mother's body (*top photograph*). The side on which the mother holds the baby is noted and she then is asked questions to obtain background data.

land and Sperber decided to see how a person holds a five-inch rubber ball while being treated. Each patient was told: "One is less aware of pain when he concentrates on some activity such as pressing an object between his forearm and chest. We are currently investigating a number of such objects to determine if size, texture, shape or consistency are of importance." The results showed that patients chose a left-side placement significantly more often than would be expected by chance. This was true even though the dentist was left-handed and had to work on the patient's left side, which made it slightly more awkward for the patient to hold the ball toward the left.

In another study Weiland and Sperber found that when adult women were asked to hold a small pillow against their chest, they showed no side preference. When they were told to imagine that they were holding an endangered infant, however, more women tended to hold the pillow on the left side. Additional studies, which focused on more specific elements of the subject's emotional state, led Weiland and Sperber to conclude that it was not anxiety alone that produced a left-side preference; after all, imagining that the object was a baby gave rise to similar behavior. Their results support the notion that anxiety affects such behavior and show a special emotional effect that a baby has on a human adult.

After I had begun my study of mothers and infants I also began to pay special attention to paintings and sculpture involving a child being held by an adult. Of 466 such works of art that I examined in museums and galleries, 373 (80 percent) depicted the child being held on the left side and 93 (20 percent) on the right side. I then made a study of photographs, paintings and other artistic representations from various cultures and again found that when a child is held by an adult a significantly larger number of pictures show the child being held on the left side rather than on the right [*see illustration on this page*]. It seems clear on the basis of these data that there is a significantly strong tendency of mothers to hold their offspring on the left side of their body.

When I was asked what my observations might mean, I was hard put for some time to claim anything more than that they were simply naturalistic observations not previously reported in the scientific literature. Then I noticed some maternal behavior that seemed to be at variance with the data I had accumulated. Of the mothers who brought their children to the follow-up clinic for premature infants at the New York Hospital–Cornell Medical Center, an unusually large number held their baby on the right side.



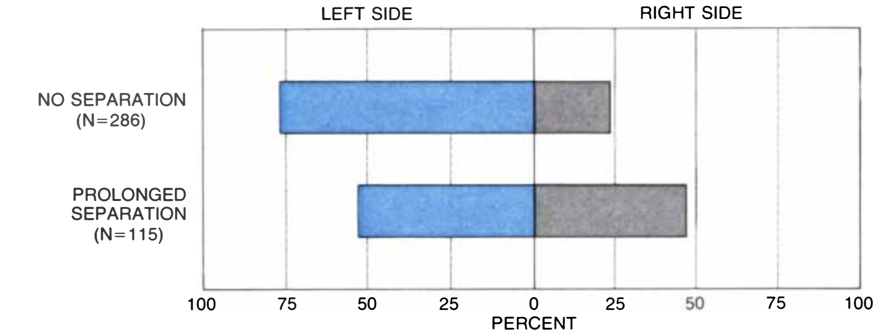
SURVEY of several books containing a large number of photographs and artistic representations of infants and adults was conducted by the author. Of the pictures in which a child is held by an adult, a significant majority show the child being held on the adult's left side.

Although many factors could cause mothers of a premature infant to behave differently from mothers of a full-term infant, one factor seemed to be the most likely cause: immediately after birth the mother of a premature infant is separated from the baby for a prolonged period, whereas the mother of a full-term infant has early contact with her baby. I undertook a large-scale study to find out if prolonged postpartum separation of the mother and her infant has an effect on the mother's response to the baby. Since we had already observed that mothers of a premature infant showed some departure from the tendency to hold the baby on the left side, it seemed that a logical index of the mother's response would be the side on which she held her child.

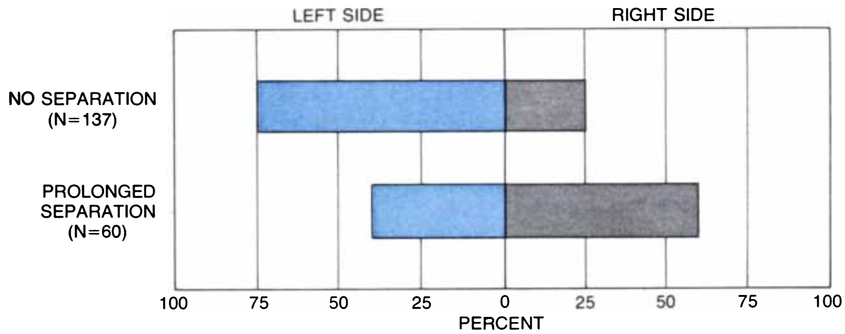
One group, the experimental group, consisted of 115 mothers who had experienced prolonged separation from their infant after birth. We arbitrarily set 24 hours or more as the definition of prolonged separation. During this period the mother did not hold her infant at all. Most of the mothers in the experimental group had had a premature baby (gestation period less than 38 weeks). Some infants in the experimental group, however, were full-term and had been separated from their mother for medical reasons. The control group of 286 mothers was selected at random from mothers who came to the well-baby clinic for regular baby checkups. None of these mothers had experienced a prolonged postpartum separation from their infant, and all had handled their infant during the first 24 hours after birth.

The procedure for all subjects was as follows. The experimenter approached the mother while she sat in the waiting room and ushered her and the baby into an examining room. The mother was told that a survey was being conducted and asked if she would cooperate. The experimenter then took the baby and presented the child directly to the midline of the mother's body. The experimenter noted on which side the mother held the baby and then asked a series of questions to obtain background data, including length of postpartum separation from the baby, number of siblings, length of postpartum separation from siblings and handedness of the mother.

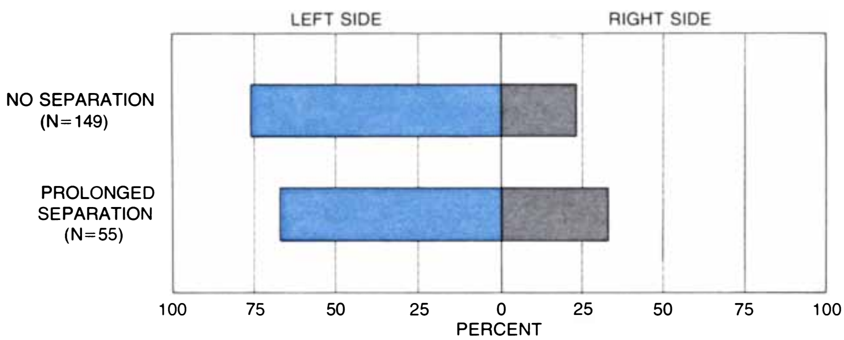
The data strikingly confirmed the earlier results. Mothers in the control group who had experienced no postpartum separation showed a marked preference for holding their baby on the left side (77 percent). In contrast mothers who had experienced prolonged postpartum separation did not show a side preference: 53 percent placed the baby on the left side and 47 percent on the right. From these results it appears that prolonged postpartum separation does alter the response of mothers to their baby. Moreover, the results suggest that the time immediately after birth is a critical period during which the stimulus of holding the infant releases a certain maternal response. This process may have



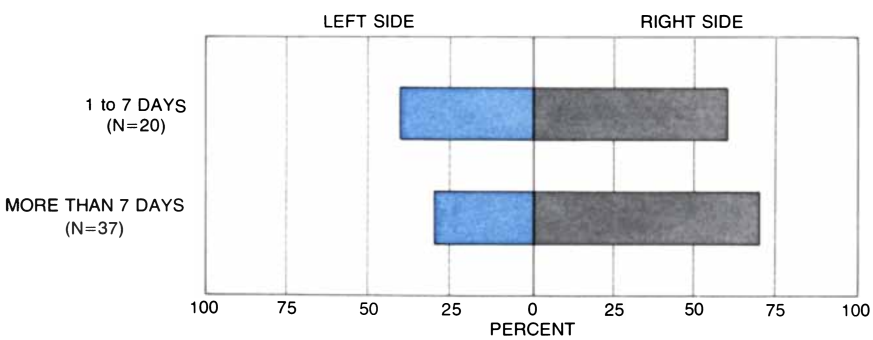
SEPARATION of mother and infant for 24 hours or more after delivery is associated with a change in side preference for holding the baby. Mothers who experienced no postpartum separation showed a marked preference for holding their babies on the left side; mothers who had a prolonged separation from their infant showed no particular side preference.



MOTHERS OF FIRSTBORNS who were separated from their infant at birth preferred holding the baby on the right side. Mothers who had early contact preferred the left side.



MOTHERS WITH OLDER CHILD from whom they were not separated during the postpartum period showed left-side preference for holding their current newborn baby even if the mother and newborn were separated for a prolonged period immediately after birth. This indicates that the holding response, once established, persists in subsequent childbirths.



POSTPARTUM SEPARATION of mother and baby for from one to seven days led to as much alteration in the response of mothers as did separation for a longer period. It seems, therefore, that the establishment of a left-side preference takes place shortly after birth.

some resemblance to the phenomenon of imprinting in birds and mammals.

If holding an infant releases the holding response, then among the mothers in the experimental group we should find differences in response between those who did not have any other children and those who had had at least one child from whom they had not been separated at birth. Our data included this information, and we found the expected difference. Of the mothers in the experimental group, those who had had another infant with whom there had been early contact tended to hold their new baby on the left side. The mothers of firstborns in the experimental group, however, actually showed a right-side preference [*see second illustration from top at left*].

We next looked at whether or not the length of postpartum separation affected the mother's response. We divided the mothers of firstborns into two groups: those who were separated from their infant from one to seven days and those who were separated for more than seven days. The results indicate that the length of separation does not greatly affect the maternal response [*see bottom illustration at left*]. It therefore seems evident that the critical period for the establishment of the left-side holding preference in mothers is during the first 24 hours after birth.

To find out if prematurity of birth played a role in determining response of the mothers in the experimental group we compared the data from mothers with a full-term infant and mothers with a premature infant. Both groups showed the same tendencies, indicating that it was postpartum separation and not prematurity that affected the mother's holding response. Perhaps some physiological mechanism puts the mother in a state of increased sensitivity shortly after childbirth, and the holding response is stimulated by contact with the infant.

When a mother holds an infant on her left side, she is placing him near her heart. I wondered whether this maternal behavior was in concert with a need on the part of the infant. The notion seemed increasingly more plausible as I realized that the heartbeat of the mother is the first and most prominent sound heard by the infant before birth. In the uterus the infant is constantly exposed to the beat of the mother's heart, which is transmitted by way of the aorta into the amniotic fluid. During this time the infant is automatically fed, supplied with oxygen, kept at a constant temperature and shielded from the kind of stress it encounters after birth. After birth the rhythmical sensations cease and are replaced with unfamiliar, dissonant, nonrhythmical sounds.

Perhaps during the period in the uterus the infant builds an association between the rhythmical heartbeat and the tension-free state, or perhaps there is an imprinting of the heartbeat rhythm so that similar sounds later in life have a functional connection with the original experience. Consider the rhythm of music in societies throughout the world: from the most primitive tribal drumbeats to the symphonies of Mozart and Beethoven there is a startling similarity to the rhythm of the human heart. The more primitive the culture, the more closely the rhythm approximates the actual sound of the heart. In short, the maternal heartbeat, with respect to the fetal nervous system, occupies a place of primacy and constancy and may be associated with a feeling of well-being. This prenatal sensory experience may have a place in explaining, from a psychobiological point of view, the universal appeal of music.

To find out if a mother's tendency to hold her baby near her heart is congruent with the baby's need I set up an experiment in which newborn infants were exposed to a recorded normal adult heartbeat sound. The infants, born in a hospital servicing one of the boroughs of metropolitan New York, were placed in a nursery immediately after birth and were kept there for four days except for normal feeding by their mothers every four hours. The sound of an adult's heartbeat (72 beats per minute at 85 decibels) was presented without interruption day and night in the nursery.

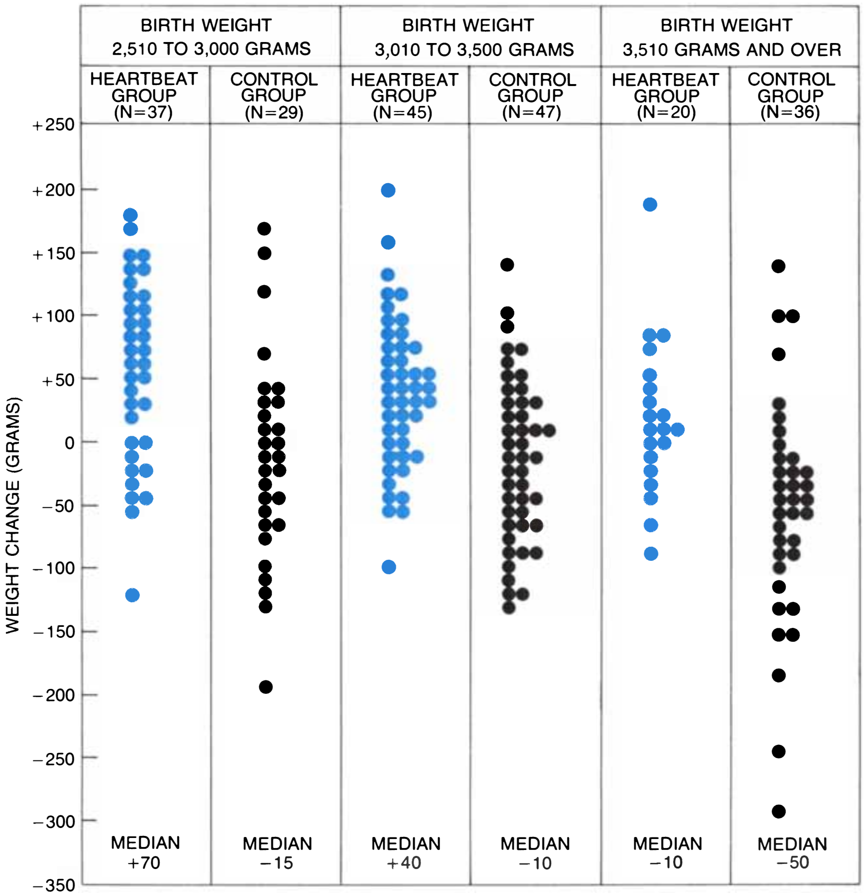
At the outset of the experiment we intended for control purposes to expose another group of newborn infants to rhythmical sounds other than a normal heartbeat. When these sounds (for example a heartbeat at 128 beats per minute) were played, however, there was an immediate increase in the crying and restlessness of the infants. We also observed these reactions when a constant hissing noise accidentally developed in conjunction with the normal heartbeat sound. Because of the critical influence of early infantile experience on later behavior, we discontinued the use of sounds other than a normal heartbeat. For a control group we monitored another group of infants in the same nursery without exposing them to the heartbeat sound. There were 102 infants in the experimental group and 112 in the control group.

We focused on weight change, food intake and amount of crying. A tape recorder with a microphone was placed in the nursery. A timing device turned on the recorder for 30 seconds every seven minutes: this provided us with a sample of nursery sounds throughout the experimental and control phases of the experiment.

Because weight change after birth is affected by the procedures used in delivery, weight change in the infants was measured from the day following birth to the fourth day. We found that 70 percent of the infants exposed to the heartbeat sound gained weight and that only 33 percent of the control group infants did so. In terms of actual weight gain the heartbeat group showed a median gain of 40 grams; the control group showed a median loss of 20 grams. There was no significant difference in food intake between the two groups.

The crying measure consisted of the percentage of time during which one infant or more could be heard crying on the tape recordings. There was crying 38 percent of the time in the heartbeat group of infants; in the control group one infant or more cried 60 percent of the time. On the average there were nine infants in the nursery at any one time. These results suggest that newborn infants are soothed by the sound of the normal adult heartbeat. Since there was no difference in food intake between the two groups, it is likely that the weight gain for the heartbeat group was due to a decrease in crying.

It is not in the nature of nature to provide living organisms with biological tendencies unless such tendencies have survival value. We often find in nature that the interaction between two organisms involves mutual satisfaction. In this connection when a baby is held on the mother's left side, not only does the baby receive soothing sensations from the mother's heartbeat but also the mother has the sensation of her heartbeat being reflected back from the baby.



WEIGHT GAIN from the day after birth to the fourth day after birth of infants exposed to the sound of a normal adult heartbeat in a hospital nursery was greater than the weight gain of infants in the same nursery who did not have the heartbeat sound played to them. Infants exposed to the heartbeat sound also cried much less than infants not exposed to the recorded heartbeat. There was no significant difference between the two groups of infants in the amount of food they ate. This suggests that the difference in weight gain was probably due to the soothing effect of the heartbeat sound, which led to a decrease in crying.

TWO-DIMENSIONAL MATTER

Films only a molecule thick vapor-deposited on atomically clean, highly uniform crystal surfaces serve as physical analogues of theoretically ideal two-dimensional systems

by J. C. Dash

We are imprisoned in a three-dimensional world but can imagine universes of higher or lower dimensionality. Edwin Abbott, in his book *Flatland,* explores a two-dimensional society of creatures who, having length and width but no thickness, are constrained to live and move on a plane surface. Severely restricted in form and motion by their loss of one dimension, this superficial society is held up to us as a caricature of our own fuller but still limited lives. Abbott's charming essay is based on simple geometry, but beyond the geometrical effects it is known that a reduction to two dimensions can cause more subtle and complex changes in the nature of things. The physical properties of lower-dimensional matter have interested theoretical physicists for many years. Among their results is the sweeping conclusion that in two-dimensional matter there can be no perfectly ordered states or structures; in other words, two-dimensional crystals, magnets, superconductors and superfluids cannot exist.

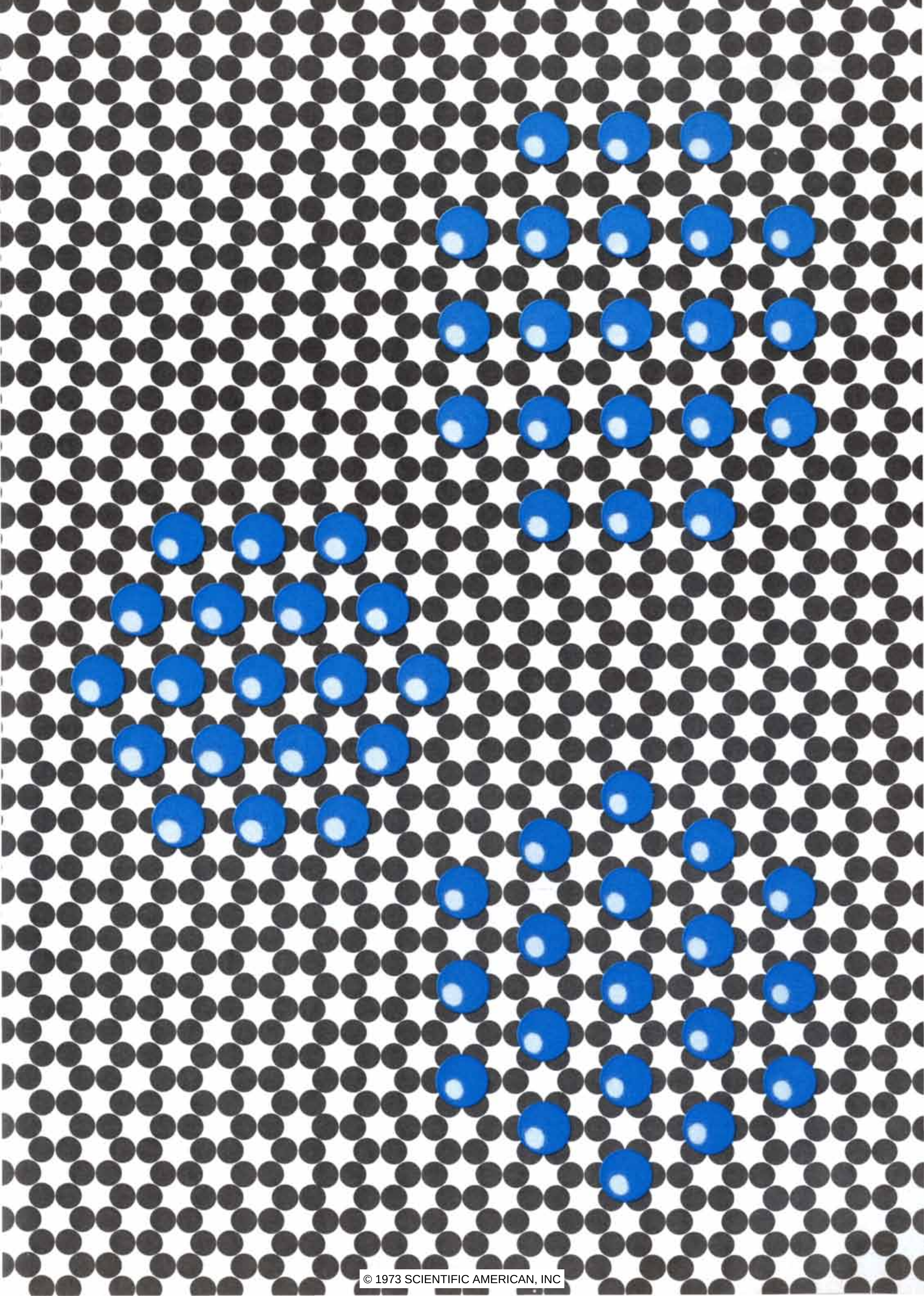
Some 40 years ago R. E. Peierls proved that if atoms could move and arrange themselves only in two dimensions, they would be unable to form a perfect crystal at any temperature above absolute zero. Yet when atoms form three-dimensional arrangements, as they do in natural bulk solids, the crystal structure remains regular and ordered over a range of temperatures. What disorder there is in either case arises from the atomic vibrations that represent one kind of heat motion in all solids, regardless of their dimensionality. All substances have a spectrum of vibrational frequencies ranging from zero to some 10 trillion vibrations per second. The spectrum for a given material is a pattern characteristic of its atomic forces, atomic masses and atomic structure.

There is a consistent distinction, however, between substances of different dimensionalities. Two-dimensional solids are much richer in low-frequency, easily excited vibrations. In the ideal state at absolute zero there are no thermal vibrations; the equilibrium positions of the atoms form a "space lattice," a regularly repeating structure that extends to the farthest reaches of the solid. As the crystal is heated the atoms vibrate more vigorously around their lattice locations, so that the atomic positions become more uncertain. Nevertheless, in a three-dimensional crystal the atomic positions will on the average continue to repeat with regularity, so that one can still describe the solid by means of a perfect lattice. In contrast, if a two-dimensional crystal were to be heated to any finite temperature, its abundant low-frequency vibrations would cause uncertainties in the atomic positions that would steadily increase with distance; at remote locations it would become impossible to distinguish whether a particular atom is the nth or the $(n+1)$th.

It has been shown by Felix Bloch, L. D. Landau, Pierre C. Hohenberg and others that the same principles that underlie the distinction between two-dimensional and three-dimensional crystals—the particular richness of weak modes of excitation in lower-dimensional systems—also apply to other forms of long-range order. Magnetic order, for example, means that some of the elementary magnetic units (the electrons) share a common direction in space. One can visualize them as magnetized needles, with a preferred orientation extending throughout the substance. Heating makes the needles wave about, causing the magnet to become weaker as its temperature is raised. In two-dimensional magnets there are many more low-frequency wavelike motions, and these are so easily excited that there would not be any preferred direction at finite temperatures.

The long-range order exhibited by a superconducting material, on the other hand, has to do with the fact that the superconductor has no electrical resistance, a fraction of its electrons being able to flow through the material with no loss of momentum. The superconducting electrons are all in a single quantum state of momentum, hence there is long-range momentum order. In a superfluid (liquid helium is the only known example) long-range momentum order corresponds to a flow of the liquid without viscous resistance. Long-range momentum order is progressively weakened as the material is heated to higher states of excitation; in both two-dimensional superconductors and two-dimensional

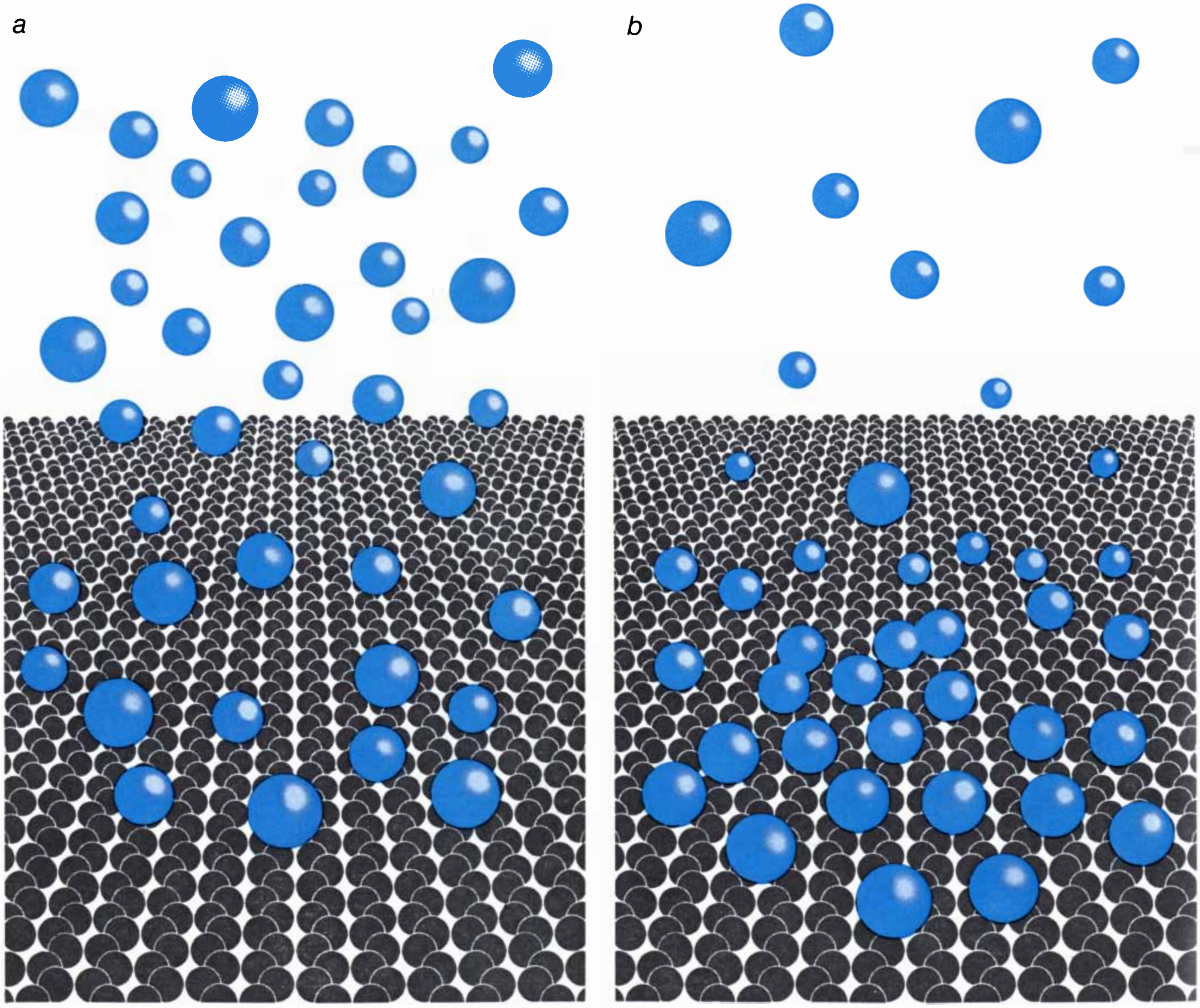
TWO-DIMENSIONAL CRYSTALS, more properly referred to as ordered lattice gases, are depicted in the highly idealized diagram on the opposite page. Each of the three different regular patterns shown represents a single layer of helium atoms (*colored balls*) vapor-deposited on a flat, uniform graphite substrate, which in turn consists of a geometrically perfect hexagonal array of carbon atoms (*black balls*). The structure at middle left was actually observed in an experiment involving xenon monolayers by James J. Lander and John A. Morrison of Bell Laboratories, using a low-energy electron-diffraction technique. Evidence has been found that the same structure occurs in the case of helium monolayers below three degrees Kelvin for a narrow range of atomic densities. The other two structures may occur in monolayers of helium at lower densities, but identification is much less certain.



superfluids the large number of low-energy excited states causes the destruction of long-range momentum order at any nonzero temperature.

It has recently become possible to put some of these theoretical predictions to experimental test, using molecular surface films as physical analogues of the ideal two-dimensional systems. The idea that films could be used as models of two-dimensional matter has occurred to many people. There is a class of molecules (typified by the noble gases helium, neon, argon, krypton and xenon) that are attracted to most other substances more strongly than to one another. When such a gas is cooled, it tends to adsorb as a thin film on the walls of the vessel rather than condense into droplets or crystals. The thickness of the film depends on the temperature and pressure of the gas; by controlling them one can make films only a few molecules thick, or of just a single molecular layer. When the temperature is quite low, virtually all the molecules in the gas become adsorbed, their thermal energies being too small to allow escape from the surface forces. The molecules may continue to move along the surface, however, provided that it is very smooth, even at such low temperatures. It is as if one dimension had been frozen out of the motion, changing the behavior of the system from three-dimensional to two-dimensional.

This oversimplified picture, however, ignores the atomic structure of the substrate. On an atomic scale the surface of a solid is not smooth. It is composed of atoms, which give the surface a regular texture; on a highly magnified scale this texture might be visualized as a cobblestone street. A gas molecule can be imagined as a ball of about the same size as a cobblestone. At high temperatures the ball is very energetic, bouncing vigorously and spending little time near the surface. At lower temperatures it is not bouncing along the street but only rolling, and although the surface is not perfectly smooth, the ball rolls so quickly that it skims over the top of the cobblestones, missing the cracks. As the temperature is lowered further the motion slows, and eventually the ball becomes captured in a crevice; now its only mo-

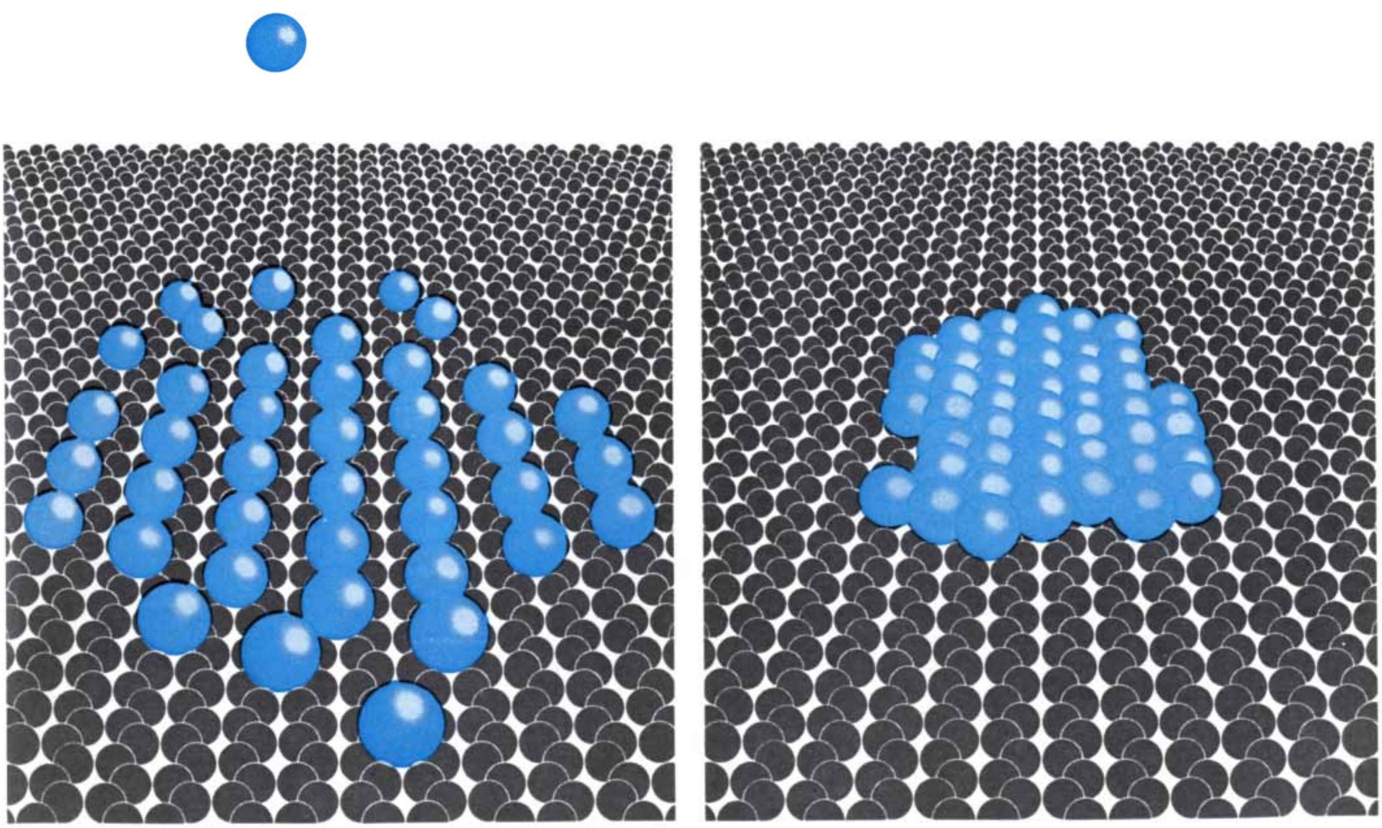


ADSORPTION of gas atoms (*color*) onto a solid substrate (*black*) is represented schematically in the four perspective views on these two pages. At a high temperature (*a*) the kinetic energy of the gas atoms is much greater than the force attracting the atoms to the solid surface, and only a few atoms are adsorbed. At successively lower temperatures (*b, c, d*) more and more atoms are adtion might be to rock back and forth within the crevice. Thus it is predicted that at sufficiently low temperatures the adsorbed molecules lose their surface mobility and become localized at definite adsorption sites.

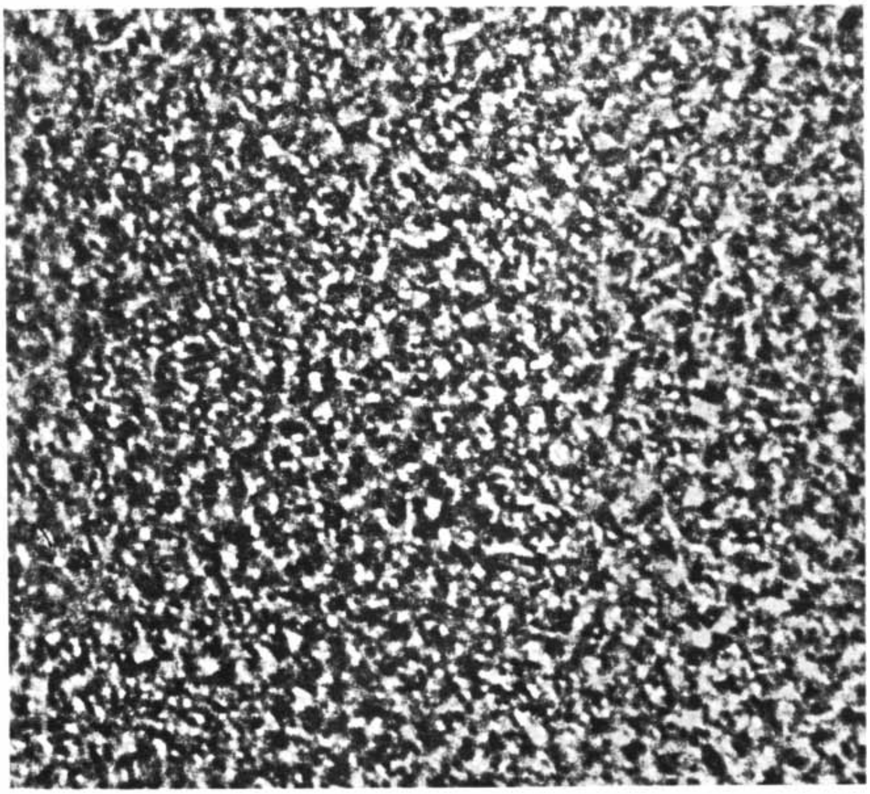
If the atoms are localized, the film phases will not correspond to any states of ordinary bulk matter. In ordinary substances the molecules can arrange themselves in any pattern in response to the forces from neighboring molecules and external fields. If conditions are changed slightly, the molecules can readjust by slight shifts in position. In a localized film, in contrast, small shifts are not possible; molecules can only make discrete jumps from one site to another. Their space has become quantized to a lattice of sites on the surface, and one speaks of such a film as a lattice gas.

A lattice gas can condense into lattice-liquid and lattice-solid phases. These transitions are easier to analyze theoretically than the condensation and melting of ordinary matter and therefore are extremely useful models of common phase transitions. The analysis of phase transitions presents such difficult theoretical problems that even the lattice-gas models can be understood only in an approximate way. There is one notable exception, however. In 1944 Lars Onsager presented an exact and complete solution of a two-dimensional lattice gas at the critical density (the density at which there are exactly half as many atoms as there are sites). Onsager showed that there is a sharp phase transition from the high-temperature disordered state, where the atoms are randomly arranged at sites, to a low-temperature ordered arrangement. This transition is qualitatively different from condensation and melting, which are classed as first-order phase changes. The transition of the lattice-gas model is a second-order phase change, of the same type as the liquid-gas critical point and many magnetic transitions. In fact, the lattice-gas model is formally identical with a particular idealization of magnetic systems known as the Ising model, and the solution of the problem is applicable to both types of system. Because the two-dimensional lattice-gas transition is the only phase change that has been solved exactly, it occupies a

c *d*



sorbed, until only a few remain in the gas phase. The one-atom-thick film that forms on the substrate can have different phases, corresponding to two-dimensional gases, liquids and solids; ordered arrays in register with the substrate can also occur (*c*). In the case of more tightly packed arrays (*d*) the substrate structure is less important and in some cases may act as a flat attracting surface.



CARBON BLACK, formed by heating polyethylene to a temperature of approximately 600 degrees Celsius, is one of the forms of carbon frequently used as substrates for thin-film studies. Heating removes all the other molecules, leaving a pure but disordered carbon structure.

position of great importance in statistical physics and serves as a paradigm of second-order transitions in general.

Attempts to test the theories by experiments on very thin films have been made in many laboratories for more than 30 years. The fact that some success has been achieved only within the past three years is attributable primarily to the difficulties in producing clean and uniform surfaces; real surfaces are never as visualized in the theoretical models. Molecular films of varying thickness coat all solid surfaces exposed to the air, even if they have been most scrupulously cleaned. An atomically clean surface freshly prepared in a vacuum or cleaved from a crystal and then exposed to the atmosphere soon adsorbs a variety of contaminants, ranging from greases and oils to light organic and inorganic molecules. The more lightly bound contaminants can be driven off by heating, but temperatures high enough to remove the rest cause some impurities to diffuse from within the solid to the surface. Even if these impurities are removed, a typical surface is still heterogeneous: a jumble of microcrystalline facets, laced with cracks and steps.

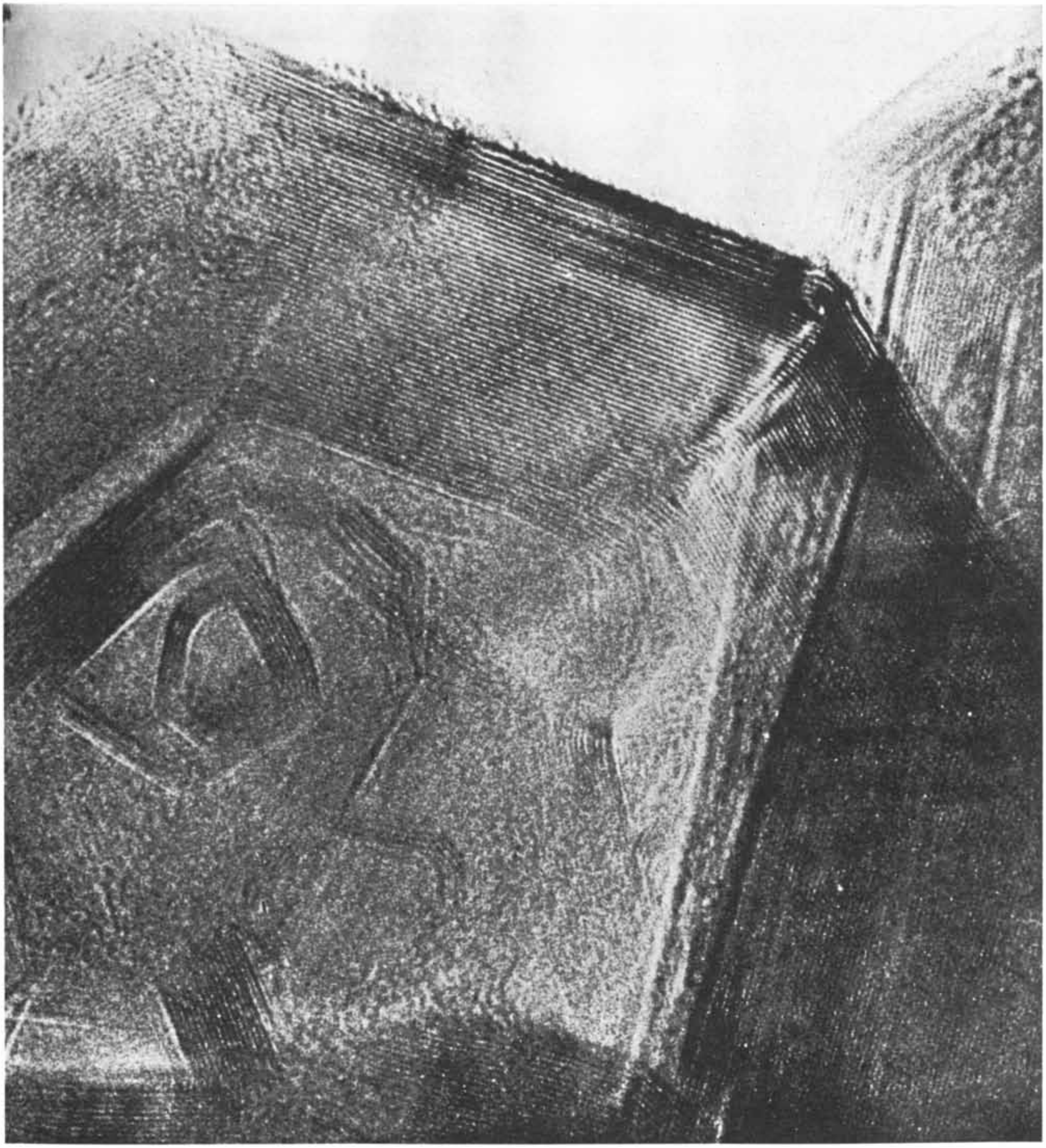
A monolayer film on such a surface will have a distribution of properties corresponding to each type of microscopic region, and this heterogeneity will obscure the fundamental properties of primary interest. What is wanted is a virtually ideal experimental surface: a single type of crystal facet, essentially free of strains, cracks and foreign chemicals. Fortunately a limited number of reasonably uniform surfaces can now be reliably prepared. Some of them have been used in recent years as substrates for monolayer films, and a rich variety of phases and phase transitions has begun to emerge.

The new knowledge of crystal surfaces comes from several groups of workers using different techniques. A number of important studies have been done by low-energy electron diffraction. In 1967 James J. Lander and John A. Morrison of Bell Laboratories examined monolayers of 26 different types of molecule adsorbed on a single crystal of graphite. They could identify several two-dimensional phases: dense solids, amorphous structures, disordered lattice gases and ordered arrays of several patterns [*see illustration on page 31*]. There have been several more recent electron-diffraction studies of physisorption with other gases and other substrates, generally showing the existence of both ordered and disordered patterns, depending on the density and the temperature of the film.

The low-energy electron-diffraction technique is extremely important, but it cannot answer all our questions about monolayers. For the effects of dimensionality that concern us here other methods are more suitable. Unfortunately they require uniform substrates that have a larger area than those that are used for the electron-diffraction work have, but a few suitable materials have been discovered. The best known is graphitized carbon black, which is produced by heating ordinary carbon-black powder to high temperatures. The heat treatment causes the graphite crystal structure to form on the outer surfaces of the particles, all with the same orientation of crystal planes with respect to the surface [*see illustration on opposite page*]. The adsorption characteristics of graphitized carbon black are different from those of ordinary carbon black and most other materials. On ordinary materials an increasing pressure causes the amount of gas adsorbed to increase smoothly, but with graphitized carbon black there are distinct steps in the curve, corresponding to the completion of molecular layers [*see illustration on page 36*]. Such steps correspond to the behavior one expects on a uniform surface; their absence on typical nonuniform surfaces results from layers being completed at different pressures on different portions of the sample.

The vapor-pressure curves of gases adsorbed on exfoliated graphite show even sharper steps than those of gases adsorbed on graphitized carbon black, an indication of still greater uniformity of surface. Exfoliated graphite can be produced from graphite crystals by a chemical process. Graphite has a layered structure, with strong forces between the atoms in each layer and relatively weak forces between layers. The weak interlayer forces and large separations allow certain foreign molecules to fit between the graphite planes. When such a layered compound is heated suddenly, the foreign chemical tends to turn into a gas, causing an internal pressure that tears the graphite layers apart. The gas can be pumped away, leaving a foam of graphite layers all chemically clean and with the same type of crystal face exposed for adsorption. About five years ago André Thomy and Xavier Duval, working at the Center of Physical and Chemical Kinetics in France, showed that the vapor-pressure curves of gases adsorbed on exfoliated graphite have considerable structure during the formation of the first layer, suggesting the existence of two-dimensional gas, liquid and solid phases [*see illustration on page 37*]. Their measurements yield surface phase diagrams remarkably like the phase diagrams of bulk matter.

A limitation of the vapor-pressure technique is that questions of long-range order are not answered. The vapor pressure is primarily controlled by the attraction between the atoms and the substrate, whereas we are more concerned with the motions and arrangements of the atoms already on the surface. For example, it is impossible to tell from the vapor-pressure results whether the adsorbed atoms are mobile along the surface or are localized at sites. This is crucial for questions of crystalline order: although it is predicted that there can be no long-range order in a solid monolayer having a structure determined only



GRAPHITIZED CARBON BLACK, produced by heating ordinary carbon-black powder to temperatures near 3,000 degrees C., serves as a highly uniform substrate for the deposition of thin molecular films. The heat treatment causes the carbon to crystallize, forming the flat crystal structure characteristic of graphite on all the exposed surfaces of the carbon particles. In this high-resolution electron micrograph, made by L. L. Ban and W. M. Hess of the Columbia Carbon Company, the graphite sheets are seen edge on.

by its interatomic forces, a film that takes on some of the structural pattern of the substrate can be ordered for as great a range as the underlying material is.

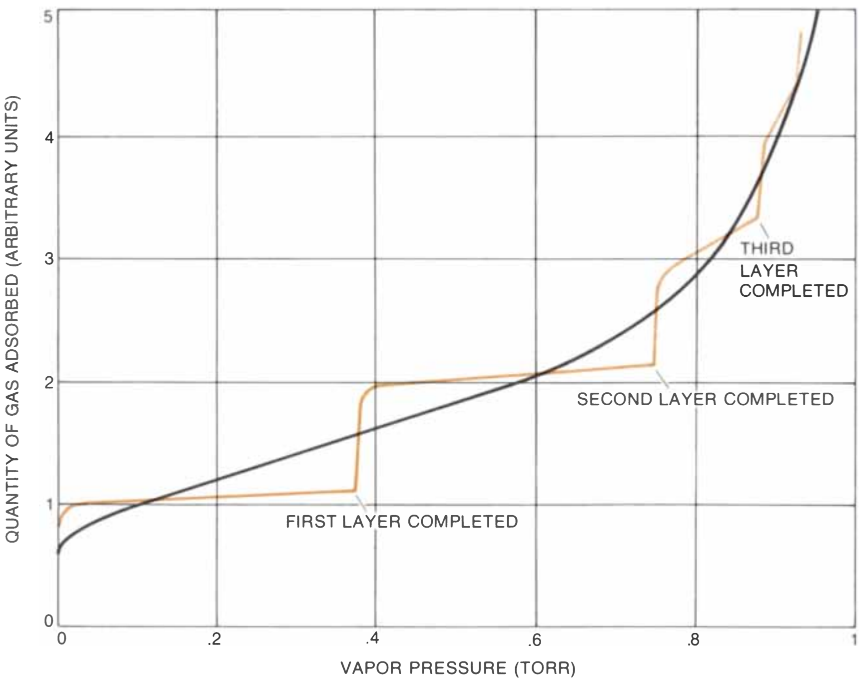
A more sensitive indicator of monolayer phenomena is the heat capacity. This traditional probe of bulk matter consists simply in measuring the temperature rises caused by adding definite quantities of heat to a substance. Heat-capacity studies have contributed significantly to our current understanding of three-dimensional materials. Applied to monolayers, they should measure just those changes of primary interest: the excitation of low-lying energy states of the film. Such experiments have been performed in many laboratories for several decades on a number of combinations of gases and substrates. Particular emphasis has been placed on films of the heavy helium isotope helium 4, primarily to determine how the superfluidity of liquid helium depends on film thickness. In the pioneer work of H. P. R. Frederikse at the University of Leiden in 1949 it was found that the superfluid transition, which occurs at 2.17 degrees Kelvin in the bulk liquid, becomes progressively attenuated and is shifted to lower temperatures for films thinner than about 10 layers of atoms. The films Frederikse studied were adsorbed on particles of jeweler's rouge. Subsequent work with glass substrates at Leiden, the University of Sussex, Cornell University and the University of California at Los Angeles traced the transition down as far as two or three layers of atoms and temperatures near .1 degree K. These results are consistent with the prediction that superfluidity is destroyed by a decrease from three to two dimensions, but closer examination reveals some complications.

When the thickness of the film is reduced to a monolayer or less, the specific heats are found to resemble those of a two-dimensional solid rather than liquid helium. It is as if the first layer were drawn so tightly to the surface that it could pack to the high densities characteristic of bulk solid helium. If one calculates the effective pressure in the first layer attributable to the adhesion forces, it is entirely reasonable to suppose that such tight packing would occur. The difficulty arises with submonolayer films, where the density is made less than the maximum coverage of a complete first layer. Although one would expect that the reduced density should cause the solid to give way to two-dimensional liquid and gaseous phases, the solid-like behavior persists. In a series of studies at the University of Washington we explored submonolayer films on copper and on rare-gas-plated copper and found solid-like behavior in all.

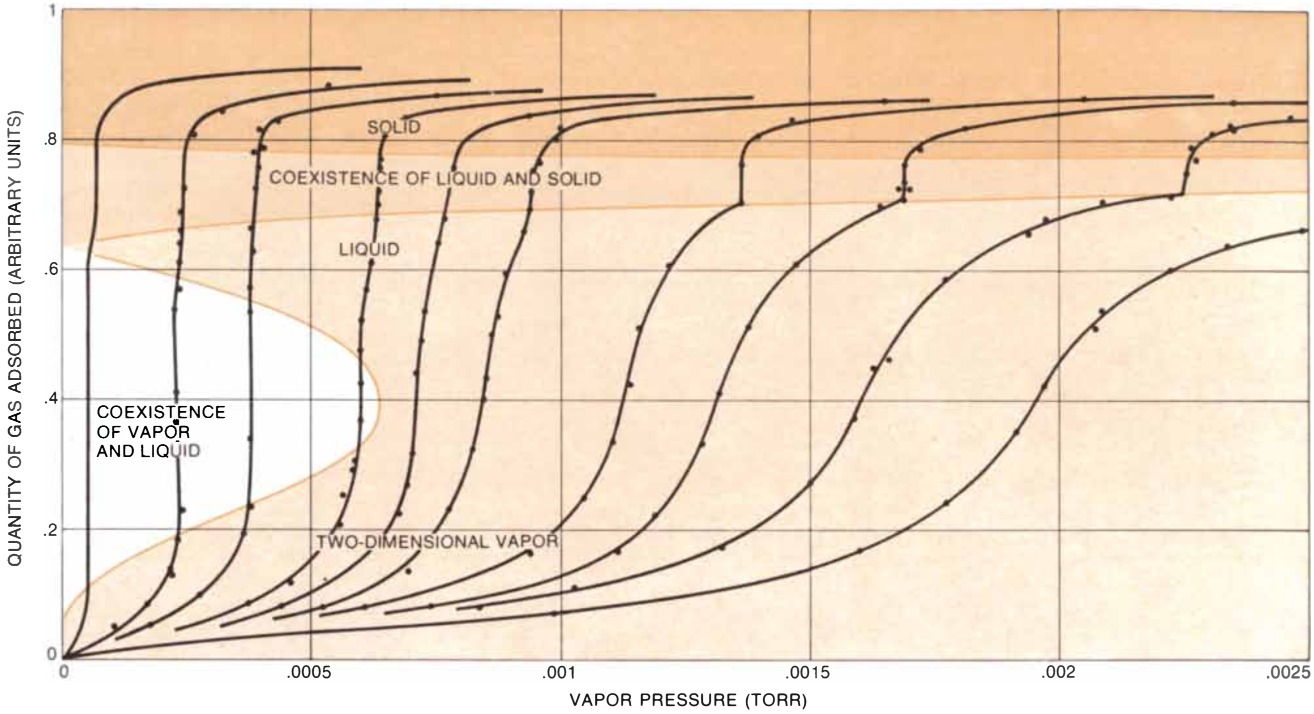
These results could be taken as satisfactory evidence that the properties we were observing were inherent qualities of the films themselves, as if the substrates were acting simply as flat attracting surfaces. But the great discrepancies between the results and our estimates of the forces between the helium atoms that would be needed to cause the solid to be so stable forced us to look for other causes. It was only then that we took more seriously the suggestions of our colleagues George D. Halsey, Jr., and William Steele, who from their experience in vapor-pressure work with a variety of systems believed that the experimental surfaces were far less uniform than we had hoped, and that their heterogeneity could be responsible in some way for the persistence of the two-dimensional solid. We therefore embarked on a search for other substrates that might be more uniform and also suitable for low-temperature heat measurements. Michael Bretz, who was then beginning his doctoral research, succeeded with a commercial form of graphite.

Bretz discovered, in virtually his first experiment with a low-density helium-4 film, the missing two-dimensional gas phase that had eluded us for so long. In the three years since then he and others in our laboratory have gone on to explore a rich complexity of phenomena in monolayers of helium 4, helium 3, neon and hydrogen [*see illustration on page 38*]. Among their results are confirmation of some of the predictions mentioned at the beginning of this discussion, and a number of surprises as well.

The two-dimensional gas phases of helium 4 and helium 3 occur at low densities, and at temperatures from one to four degrees K. Under these conditions the atoms are found to be highly mobile, traveling about in apparent defiance of the atomic texture of the surface. This mobility can be understood only as a quantum phenomenon. When fairly heavy, tightly adsorbed atoms are cooled to low temperatures, they eventually settle into definite sites on the surface, where they remain for times much longer than the duration of a typical experiment. Light, weakly bound helium atoms, on the other hand, cannot be localized for more than a very brief in-



VAPOR-PRESSURE CURVES are obtained by admitting a quantity of gas into a vessel containing a large-area adsorbent and then measuring the extent to which the adsorption of the gas causes the pressure to be lower than if no adsorption had occurred. The colored curve, obtained by André Thomy and Xavier Duval in France, represents this relationship for krypton gas adsorbed on a substrate of exfoliated graphite; the sharply defined steps in the curve correspond to the completion of individual molecular layers. For comparison, the black curve shows the vapor-pressure relationship for adsorption on a typical nonuniform surface; the absence of structure here indicates a heterogeneous surface, suggesting that individual molecular layers are completed at different pressures on various parts of the surface.



PHASE DIAGRAM OF TWO-DIMENSIONAL MATTER, remarkably similar in general appearance to the phase diagrams of simple bulk substances, was deduced by Thomy and Duval by measuring the vapor pressure of krypton adsorbed on exfoliated graphite at various temperatures during the formation of the first molecular layer. The curves were recorded at successively higher temperatures, ranging from 77 to 97 degrees K. The shapes of the curves suggest the existence of two-dimensional gas, liquid and solid phases.

terval—a few vibration periods—before they jump to other sites, no matter how low their temperature is. This rapid quantum diffusion gives the low-density monolayer the character of a mobile two-dimensional gas. From the temperature range over which the diffusion occurs one can obtain a measure of the depth of the atomic sites on the surface. Our experimental measurements of the mobility of helium atoms on graphite are consistent with theoretical estimates made recently by Donald Hagen, Anthony Novaco and Frederick J. Milford of the Battelle Memorial Institute.

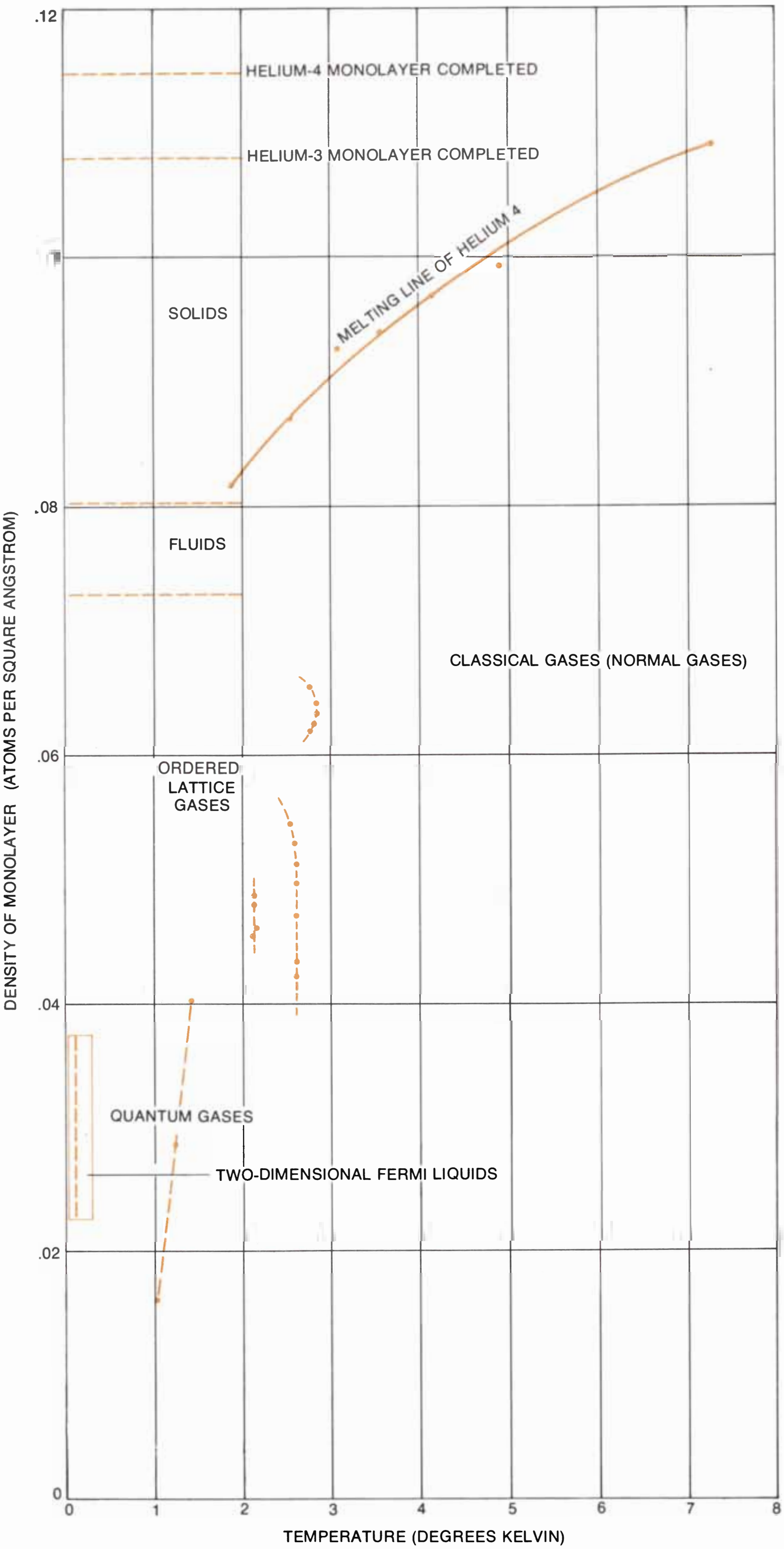
When the monolayer densities are increased, the effects of interactions between the adsorbed atoms become more important, producing changes of the kind that occur in ordinary gases before they condense to the liquid state. In two dimensions, however, there is far less tendency to condense, since a molecule can have fewer neighbors. This enables one to use films to explore the effects of interactions over a much wider range of densities and temperatures than is possible in three-dimensional gases. Very recently Robert L. Siddon and Michael Schick of our group have been able to explain several characteristics of the monolayer gas phases in terms of the scattering of helium atoms moving freely on a flat surface.

At still higher densities an entirely new set of features appears. As Bretz and I found for helium 4 and Donald C. Hickernell, Eugene O. McLean and Oscar E. Vilches found for helium 3, the heat capacity becomes sharply peaked near three degrees K. for monolayer densities within a narrow "critical" range. Analysis shows that these peaks are signals of second-order phase transitions to regular atomic arrangements, including one of the structures observed in low-energy electron-diffraction work on xenon. The shapes of the peaks are in striking agreement with Onsager's theory of two-dimensional lattice gases [*see illustration on page 39*]. An exact theory for the particular structure deduced for the films has not yet been worked out, but the close correspondence between the results and the structures that have been solved suggests a wider applicability of the theory than had been believed.

There are many experimental examples of second-order phase transitions in which the heat capacity has a peak resembling the theoretical shape, which rises logarithmically as the critical temperature is approached. The peaks occurring in helium films also have such a logarithmic shape. What is unique about the monolayer transitions is that, unlike other such peaks, these peaks have widths (that is, coefficients of the logarithmic function) that are very nearly equal to those of the exact theories.

The agreement seems remarkable in view of the differences in structure and atomic forces between the experimental systems and the models. There is a still more important complication in the real films: whereas the models assume that the atoms are localized at definite sites, we have seen that the individual atoms are highly mobile. The actual transitions must therefore involve a fundamental change in mobility when spatial ordering occurs. This change can be accomplished only by the cooperative action of both intermolecular and surface forces.

Returning to the analogy of the cobblestone street, the change can be visualized as follows. As long as there are few balls on the surface, the jumping from site to site proceeds with only infrequent encounters between balls. If their number is increased, there is greater chance that a jump is prevented, either because a site is occupied or because the forces between the nearby atoms, which tend to produce a definite pattern of occupied positions, are too strong compared with the thermal energy. Nevertheless, if the ratio of balls to crevices is small, some jump choices remain and mobility continues. When the critical density is reached, no easy jumps



HELIUM MONOLAYERS were found by the author and his colleagues at the University of Washington to exhibit a rich complexity of phase relationships. Some of the phases represented in this diagram are two-dimensional equivalents of the states of three-dimensional matter, whereas others also involve the atomic structure of the substrate. The identification of the various regimes was made by a detailed analysis of heat-capacity measurements.

are possible; hence the low-temperature state is both ordered and immobile.

Our conclusion that the transitions in the helium films must also involve the question of mobility was originally based on our measurements of the heat capacities. This view has since been confirmed by Robert J. Rollefson of our group using the technique of nuclear magnetic resonance on films of helium 3 adsorbed on graphite. His results show clearly that the atomic diffusion is effectively quenched at just the temperatures and densities of the sharp heat-capacity peaks. The structural transitions in films related to mobility are particularly interesting because they seem to involve the same principles as the metal-insulator transition in bulk matter. In metals the electrons tend to be associated with the individual atoms of the solid. Some of the outer electrons of each atom (the conduction electrons) can readily jump to another atom, provided that in so doing they do not move too close to another electron. This means that in metals the electrical conductivity involves a collective motion of many electrons, each moving so as to stay out of the way of all the others. It will perhaps be possible to understand more of this interplay by studying the analogous transitions in helium monolayers.

A new set of phenomena appears when the density is increased further. When more atoms are present than can be fitted into the regular array of sites, the entire orderly structure is destroyed and the film goes into a state primarily determined by the interactions of the helium atoms among themselves, ignoring the substrate texture. The frequent collisions apparently drive atoms out of their sites so that they spend larger fractions of time in migration, and the film takes on semblances of a two-dimensional liquid.

At still higher densities the monolayers freeze. The freezing of helium films is quite unlike that of bulk liquids. When a three-dimensional substance in equilibrium with its vapor freezes or melts, it remains at a fixed temperature while a definite amount of heat is extracted or added. This causes its heat capacity to become infinite at that temperature. If instead of allowing the substance to remain in contact with its vapor a typical solid is tightly sealed into a strong container with no vapor space, melting causes a rise in pressure that shifts the melting temperature to progressively higher values. This results in a heat-capacity curve shaped like a mesa with vertical sides [*see illustration on page*

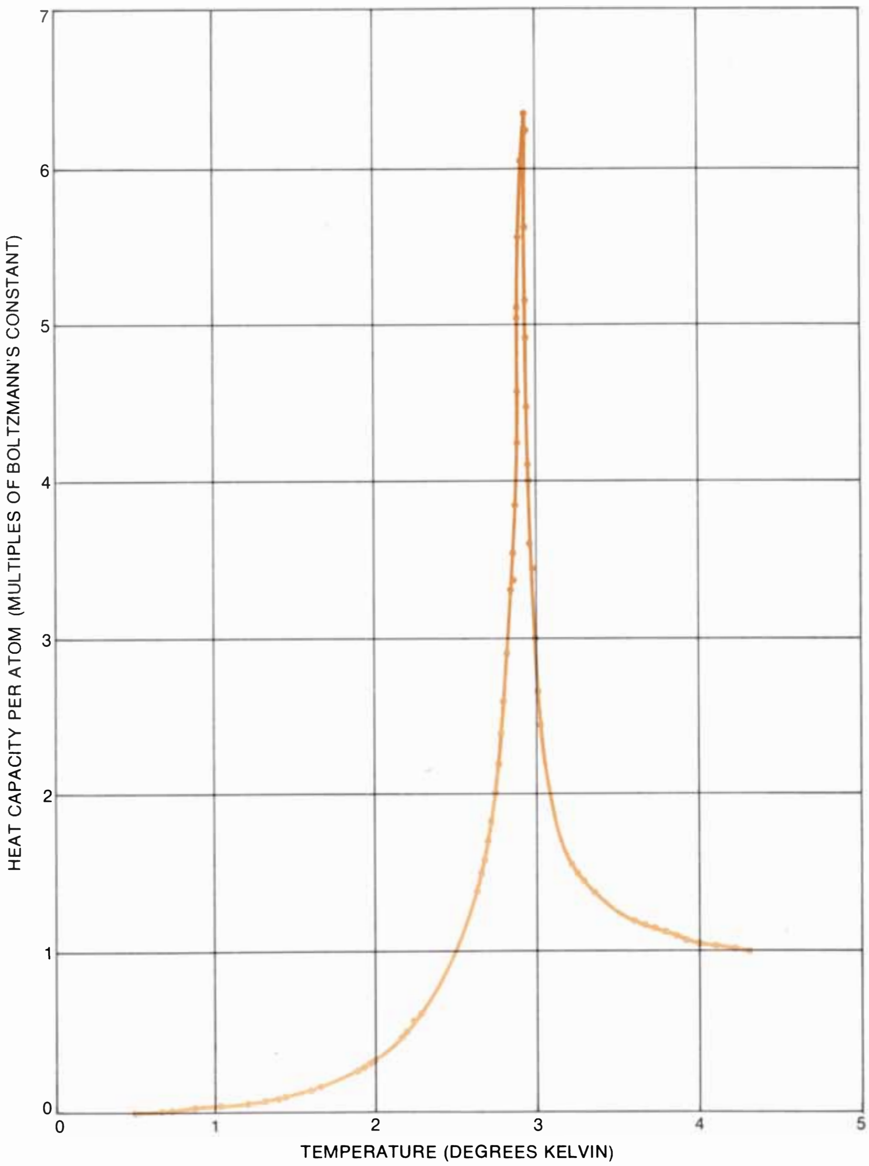
40]. Such experimental shapes have been obtained in tests on bulk helium 3 and helium 4. In contrast, we found that the liquid-solid transformations in helium monolayers produce nearly symmetric peaks with no vertical sides. This indicates that their melting does not begin suddenly at a well-defined temperature but rather occurs gradually as the film is heated. Thus there is no simple distinction between its liquid and solid phases; instead of a "melting line" of temperatures and pressures at which both phases can be present simultaneously (as for instance the region where ice and liquid water can coexist), there is only a single phase, which has more solid-like or liquid-like properties, depending on the temperature. The continuous nature of the transition implies that there is no sudden destruction of long-range order as the film is heated. Instead the perfection of the crystal structure in the film has a finite range that steadily decreases as the temperature increases. That is precisely what can be attributed to the theoretically predicted effect of reduced dimensionality on long-range crystalline order.

More recently we have examined some features of neon monolayers. We find that their properties contrast sharply with those of the helium films. For example, Bradley G. Huff has found that neon on graphite has extremely strong and narrow heat-capacity peaks, approaching the spiked shapes of first-order melting in bulk crystals at their vapor pressure. This seems to contradict the theoretical prediction in that it implies the sudden dissolution of long-range order. We interpret the results as indicating that neon films consist of two-dimensional solid clusters that have impressed on them some of the regularity of the substrate. Analysis indicates that the impressed regularity is only partial, with the primary structure being dictated by the interactions of the neon atoms, but that at relatively long and periodic distances the atoms fit into the same pattern with respect to the substrate. This is therefore a much more complicated situation than the one in the helium films, which apparently ignore the substrate structure. The reasons for the distinction are not obvious, and we hope to explore it further by studying other gases.

The new observations of films on graphite raise many questions about the detailed nature of the several phases and of adsorption in general. Recent studies by Robert L. Elgin, David L. Goodstein and G. Alexander Stewart of the California Institute of Technology have contributed greatly to our understanding of the phases of helium 4 on graphite. The adsorption of helium and neon on graphite are also being explored by John G. Daunt and Eugenio Lerner at the Stevens Institute of Technology.

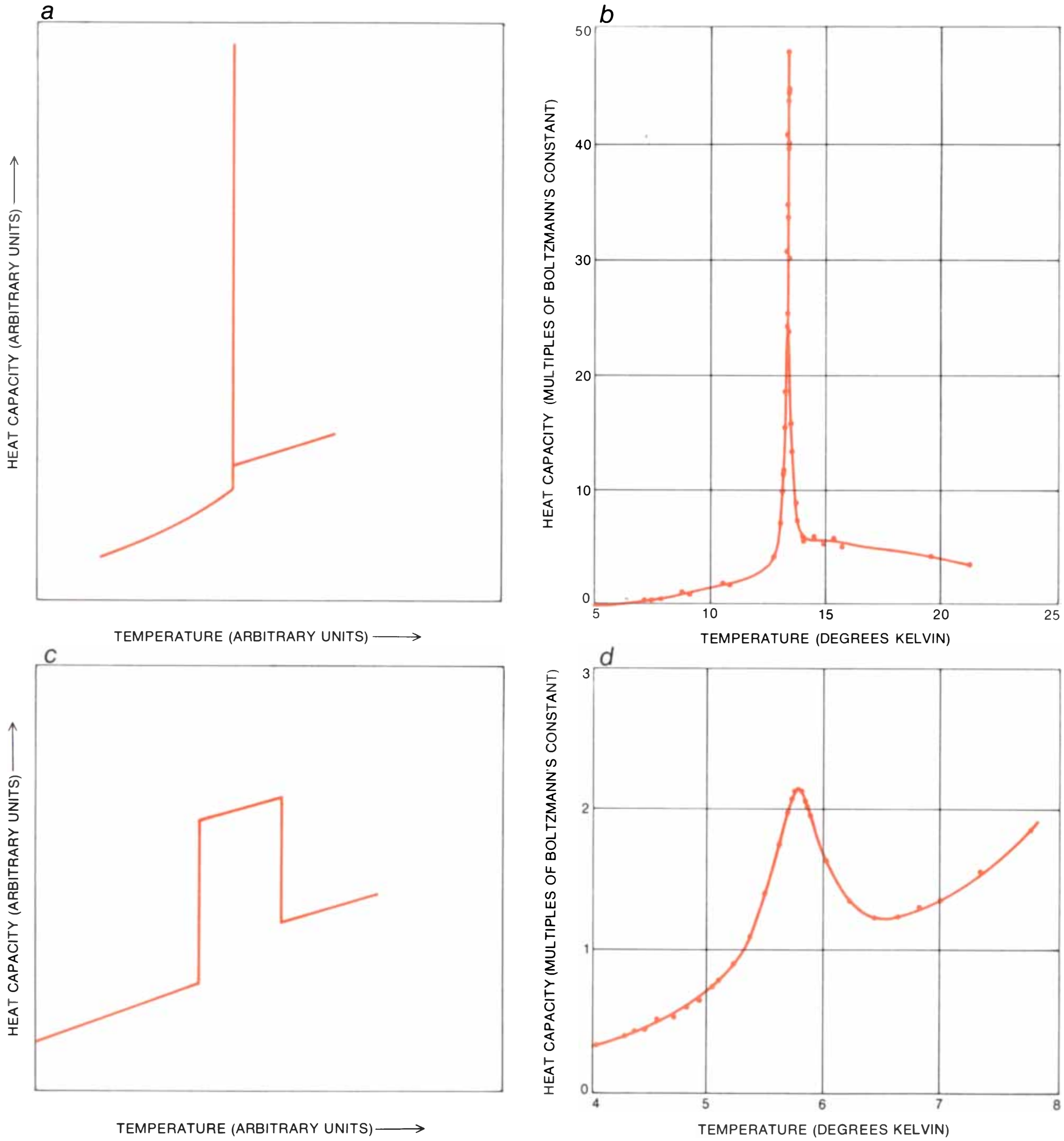
If these experiments seem to confirm the predicted absence of long-range crystalline order, they have not been as conclusive with respect to momentum order. It is true that in neither the gaseous nor the liquid phases of helium-4 monolayers has any evidence of a superfluid transition appeared. Other factors, such as substrate structure, however, may be interfering with its onset. Moreover, the heat capacity might not be a sensitive probe of superfluidity in thin films. Although in bulk liquid helium superfluidity begins at the temperature of the "lambda point" heat-capacity peak, there is no known reason why this should also be true in films. Therefore other techniques are being explored.

Within the past year we have begun



HEAT CAPACITY of the heavy helium isotope helium 4 adsorbed on graphite was measured for a monolayer containing one-third as many atoms as adsorption sites. (The heat capacity of a substance is simply a measure of the temperature rises caused by adding definite quantities of heat to the substance.) The peak indicates a second-order phase transition from a two-dimensional fluid state (at temperatures above three degrees K.) to an immobile and localized state (at temperatures below three degrees K.); the low-temperature state is believed to have a regular structure similar to the two-dimensional crystalline pattern portrayed at middle left in the illustration on page 31. The shape of the heat-capacity peak is in close agreement with theoretical predictions of two-dimensional lattice-gas transitions.

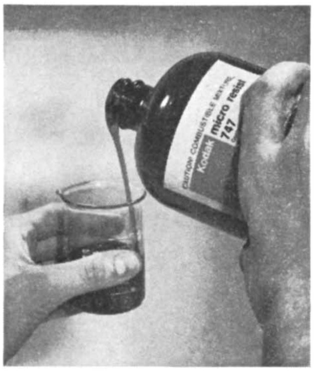
to extend our studies to multilayer films, repeating with graphite substrates the types of measurement previously done with other surfaces. Several new features have shown up in helium, hydrogen and neon films, indicating that the nonuniformities of typical substrates can cause a blurring of film properties that extends through several molecular layers. The new measurements suggest that under some circumstances a thick film can behave as a set of separate layers, each retaining its individuality and its two-dimensional character. These results pose an intriguing and still unanswered question: What conditions control whether a multilayer film acts as a set of separate layers or as a thin slice of bulk liquid? Or, in a physical sense, how does a film evolve from two-dimensional to three-dimensional matter?



HEAT-CAPACITY CURVES of three-dimensional and two-dimensional samples are compared. The heat capacity of a typical three-dimensional crystal in equilibrium with its vapor (*a*) becomes infinite at its melting point, since the substance remains at a fixed temperature while a finite amount of heat is supplied. In neon monolayers (*b*) the heat-capacity curve is sharply spiked. This behavior indicates that the neon freezes into "islands" of two-dimensional solid and that the solid has impressed on it some of the long-range crystalline order of the underlying graphite. For a typical three-dimensional crystal enclosed in a rigid container with no vapor (*c*), melting causes a rise in pressure that shifts the melting temperature; the heat-capacity curve has vertical sides corresponding to the initial appearance of the liquid and the final disappearance of the solid. In helium monolayers (*d*) melting occurs at higher temperatures for films of higher density, but the heat-capacity peaks do not have vertical portions; these characteristics indicate a continuous melting process, consistent with the theoretical prediction that two-dimensional solids cannot have long-range crystalline order.

We want to be useful...and even interesting

Sauce



This type of liquid first put men on the moon. Now it has made popular gift items out of watches without moving parts and pocket electronic computers.

In the years after Napoleon, when Europe was settling down to peaceful pursuits, word spread among a certain type of industrious artist about treated pitches and glues that varied their solubility according to light exposure. It was true. Materials and methods were found that allowed light to lighten the labor of delineation. Eventually wood cuts and steel engravings virtually vanished from the printed page. Resins that resisted etchants after hardening by light made photoengraving possible. Photoengraving permitted dissemination of photographers' work by printing press.

A century after the invention of photography, Kodak chemists, paid to comb troublesome variables out of useful phenomena, were learning how to build photopolymers from scratch. Light-sensitive resists for spraying on metal plates were offered. The graphic arts industry generally declined our offer, but after a while the electronics industry seized on our product for carving up semiconductive materials in clever patterns. The patterns rapidly grew cleverer and cleverer and smaller and smaller, and in more and more layers of substances that affect electron mobility in more and more ways, and today electronics people talk less of vacuum pumps than of adhesion, resolution, resistance (different from the kind that comes in ohms), image stability, pinhole propensity, photosensitivity, wet and dry coating thickness, and device compatibility.

The thought occurs that a few curious souls might like to trace one of the roots of the miraculous watches, computers, and journeys to the moon to its ultimate hairs.

Here is a transformation matrix between user's and maker's concerns about the sauce:

	Polymer characteristics	Viscosity	Solids content	Haze	Metal ion content	Ash	Moisture	Gel and particulates	Sensitizer concentration	Solvent composition
Adhesion										
Resolution										
Resistance										
Image stability										
Pinhole propensity										
Photosensitivity										
Wet coating thickness										
Dry coating thickness										
Device compatibility										

Through experience, bitter and sweet, each maker learns values to assign to the boxes and what to do where concern is high. Take the last column. Take xylene, a common solvent, a distillate of petroleum. Its isomer composition and tendency to leave residuals that affect resistance and pinholing depend not only on distillation parameters but on whether the crude was Middle East, Texan, or North Sea.

As for the photopolymer itself, let's say $\left[\underset{R_1}{CH}-CH_2-\underset{R_2}{CH}-CH_2 \right]_x$, we had better mind our statistics. The distribution of x, R_1, and R_2 for each batch must each be in the right range and breadth. Too many macromolecules of low x, image loss during development and image distortion during post-bake. Too many with high x, incomplete development and scum formation. Too many high in R_1 and low in R_2, too much adhesion and too little sensitivity. Vice versa, vice versa.

And that's electronics manufacturing today.

Not by exegesis alone but by performance must we vie for favor. Should more immediately useful facts about Kodak photo resists be wanted, D. W. Frey, Kodak, Rochester, N.Y. 14650 can provide them.

Saucer

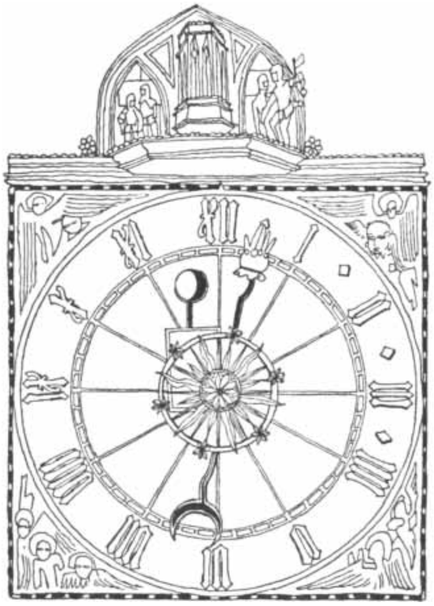
Wow. Wow!

Alternative hypothesis: Foreign particle picked up by film adheres when wet by developer. No development occurs under it. In fixer, particle loosens around edges. Partial fixation over outer portion of undeveloped area. Full load of halide remains where fixer didn't reach at all. During printing, halide modulates light like silver.

The trouble with Kodak's Film Technical Services Division in Rochester is attitude. Simple chemistry. Simple physics. On occasion, a little psychophysics. No interest in the occult. (Particularly not during working hours.) Same goes for Irving Goetz of Hamburg, N.Y., apparently. Pointed his home-built telescope at the moon, processed the 35mm film in a small tank, obtained the above result, then of his own free will gave us a look at the negative. We have refrained from dunking it back in hypo. Would probably have detracted from its unusual character.

Kodak

Enslaved by rationality



SCIENCE AND THE CITIZEN

Fat Protons

In physical experiments interactions typically reach some limiting value with the intensification of one or more variables: heat, pressure, kinetic energy or whatever. One objective in building accelerators capable of imparting more and more energy to subatomic particles has been to see if and when such limiting values will manifest themselves in a variety of particle-particle interactions. In general to construct a satisfactory theory one must have confidence that a limiting value exists along some crucial dimension. Theorists are obliged to reexamine their premises when an experiment demonstrates that the limiting value assumed by an existing theory turns out to be nonexistent on being put to the test.

Such a crisis has recently been precipitated by experiments at CERN (the European Organization for Nuclear Research) showing that the proton, instead of approaching a limiting size as the energy imparted to it is increased, reaches a minimum size and then begins to grow sharply again. The "size" of a proton or any other subatomic particle is defined by its area of influence over another particle. If the size of the proton is measured by letting two protons collide at relatively low energies, roughly between one billion and 10 billion electron volts (GeV), subsidiary effects become important; for example, the colliding particles may make each other resonate. Because of these low-energy effects each proton acts as a bigger target for the other proton than it does at somewhat higher energies. The area of influence, or target size, is termed the particle's cross section, and it is usually expressed in millibarns. (A millibarn is 10^{-27} square centimeter.)

It has been known for many years that the proton reaches a maximum size of slightly more than 47 millibarns at about 2 GeV. At 10 GeV the size has fallen below 40 millibarns and at 70 GeV the size appears to have leveled off at 38 millibarns. Measurements between 15 and 60 GeV were made with the 70-GeV accelerator at Serpukhov in the U.S.S.R., the world's most powerful before the completion last year of the 300-GeV accelerator at the National Accelerator Laboratory in Batavia, Ill. Measurements made at Batavia between 100 and 300 GeV appeared to confirm that the proton's size was leveling off as expected at 38 or 39 millibarns.

Energies greater than 300 GeV have now been produced at CERN with the aid of the facility known as the Intersecting Storage Rings. Protons accelerated to 31 GeV are injected into two large intersecting rings, where they initially circulate in opposite directions without colliding. After the number of particles in each circulating beam has been built up to a sufficient size, one of the beams is magnetically diverted so that it collides with the other. The collision produces energies equivalent to up to 2,000 GeV, or about seven times the energy that has been achieved at Batavia.

Recently physicists from CERN, from institutions in Pisa and Rome and from the State University of New York at Stony Brook have reported measuring the size of the proton at energies between 280 and 1,500 GeV. The unexpected result is that over this energy range the size of the proton climbs in virtually a straight line from 38 to 43 millibarns.

The result is surprising not only in its own right but also because it places in doubt an important theory developed by the Russian physicist I. Ya. Pomeranchuk stating that the size of an antiproton when measured in collision with a proton should always be greater than when the colliding particles are both protons. At energies up to about 60 GeV the antiproton-proton cross section remains significantly higher than the proton-proton cross section but the antiproton-proton curve is pointed so sharply down at 60 GeV (where it has a value of 43 millibarns) as to raise the possibility that it might intersect and fall below the proton-proton curve at energies above 100 GeV. If it were to fall below, it would completely undermine the present theoretical understanding of antiparticles and of antimatter in general. It should soon be possible to measure the antiproton-proton cross section in the critical region between 100 and 300 GeV with the Batavia accelerator and with a comparable machine now under construction at CERN.

Bombers Away

One of the more obvious shortcomings of the agreements reached last year as a result of the first round of the strategic-arms-limitation talks (SALT I) between the U.S. and the U.S.S.R. was the omission of any restrictions on the development, construction or deployment of modern long-range bombing aircraft. This omission, it was pointed out by some critics of the SALT I accords, could lead in time to a major revival of the strategic-arms race in this particular area of nuclear weaponry.

The announced plans of the U.S. Air Force for an entirely new strategic bomber, to be called the B-1, are now viewed by these same critics as a confirmation of their fears. So far the Air Force has received $1.3 billion for research and development on the B-1. If all the funds asked of Congress for the construction of the B-1 are approved, however, the Air Force forecasts a total expenditure of $11 billion over the next 10 years for 244 of the new bombers. The proposed B-1 force, if it is built, could carry as many as 5,280 nuclear weapons, 1,140 more than could be carried by a similarly projected fleet of modernized B-52 bombers. The additional 1,140 nuclear weapons would increase the total number of deliverable U.S. nuclear warheads in the 1980's to some 18,600 (6,500 in submarine-based missiles, 2,500 in land-based missiles and 9,600 in aircraft-delivered bombs). This projected total would nearly triple the number of deliverable nuclear weapons now possessed by the U.S.

According to a recent research study

of the proposed B-1 bomber program by the Center for Defense Information, an independent group of non-Government military analysts, "the basic question is: 'Does the U.S. need 18,600 strategic nuclear weapons in the 1980's?'" A summary of the study group's findings appears in a recent edition of the CDI newsletter, *The Defense Monitor*.

To begin with, the study points out, bombers are no longer the primary element of U.S. strategic forces, owing to improvements in missile technology in recent years. Furthermore, the SALT I treaty limiting anti-ballistic-missile (ABM) defenses makes the bomber less important as a "hedge" against the failure of both land-based and submarine-based missile systems.

The projected modernization of the 300 existing B-52 Series G/H aircraft would, in the view of the CDI analysts, make them "as useful as the B-1 through the 1980's." These late-model B-52 bombers are generally considered capable of penetrating the present Russian air defenses. With the improvements now under way, the B-52 should "nearly equal the proposed B-1's penetration capability." Present programs also call for doubling the nuclear-weapon-carrying capacity of the B-52 by the 1980's.

Although the proposed B-1 would fly faster and perform better than the B-52, the study continues, "this improvement in performance is not significant because both bombers would carry the same stand-off weapons." In and of themselves, therefore, the performance advantages of the B-1 over the B-52 Series G/H would appear to be "not compelling reasons for production of the new aircraft."

The Air Force's plans for the future use of the B-52 fleet are, however, in doubt. Although replacing all the B-52's with B-1's would add only 1,140 nuclear weapons to the U.S. arsenal by the 1980's, the B-1 force may simply be added to the B-52 G/H force, thereby greatly increasing the number of strategic bombers available to the U.S. in the 1980's.

As far as cost is concerned, the proposed supersonic B-1 bomber, which is designed to accomplish the same strategic mission as the subsonic B-52 bomber, would cost more than five times as much: $45 million for each B-1 as against $8 million for each B-52. As a final consideration the study adds that the U.S. currently has 531 long-range strategic bombers, whereas the U.S.S.R. has only 140.

All in all, the CDI analysts conclude, "the Air Force has not proved a need for the B-1, nor has it proved the urgency of the 1980 deployment."

The Organic Superconductor

Superconductivity, the absence of electrical resistance in certain substances at temperatures close to absolute zero, has been confined to metals, alloys and intermetallic compounds. Now, however, there is experimental evidence that organic solids may be able to act as superconductors. Moreover, superconductivity has been observed in such solids at temperatures three times higher than the highest superconducting temperature in any other substance.

A group at the University of Pennsylvania led by Alan J. Heeger and Anthony F. Garito reported to the American Physical Society in March that they have observed sporadic superconducting behavior in organic molecular crystals. Through the combination of compounds that met the requirements for acting like metallic superconductors, Heeger and Garito were led to the organic salt dimethyltetrathiofulvalene-tetracyano-quinodimethan. They have not yet been able to stabilize the superconducting state in the salt, but they feel that it may be possible to do so by "fine-tuning" the chemical structure of the compounds in it.

The temperature at which the organic salt has exhibited superconductivity is 60 degrees Kelvin. The highest superconducting temperature previously known, 20.4 degrees K., was for the niobium-aluminum-germanium alloy $Nb_3(Al_{.75}Ge_{.25})$. The present organic compounds become insulators at temperatures lower than 50 degrees K., because of changes induced in their crystal structure. Although Heeger and Garito have so far studied only crystalline organic substances, they believe the knowledge and techniques resulting from their work can be applied to noncrystalline organic and polymeric solids as well.

The Polyunsaturated Steak

It has recently been discovered that cattle fed "protected" polyunsaturated vegetable oils produce polyunsaturated body fat and milk fat. Now, according to the results of a study conducted in Australia, it appears that a diet of polyunsaturated beef and milk products is effective in reducing the cholesterol level in the blood of human subjects.

Under normal circumstances unsaturated oils are hydrogenated to more saturated forms in the rumen of the bovine digestive tract. It has been found that this hydrogenation can be prevented by mixing such oils with the milk protein casein, drying the mixture and treating a powder of it with formaldehyde. The unsaturated-fat content of the beef and milk is raised in proportion to the amount of protected unsaturated oil in the feed.

In the Australian study six healthy people were fed a carefully controlled menu. The only sources of fat were beef, drippings, cheese, butter, milk and cream. In the control, or saturated-fat, diet the products were obtained from cattle that had been fed grass and oats. Products for the unsaturated-fat diet came from cattle that had been fed supplements of protected vegetable oil. The proportion of fat in both diets was kept between 45 and 50 percent of the total calories.

The study was conducted by Paul J. Nestel and his colleagues at the Australian National University at Canberra and at the Commonwealth Scientific Industrial Research Organization. They report in *The New England Journal of Medicine* that five of the six subjects showed a significant reduction (a mean of 10 percent) in blood cholesterol after three to four weeks on the diet of unsaturated beef and dairy products. They conclude that the diet was "just as effective as vegetable oils with similar ratios of polyunsaturated to saturated fatty acids."

Bicycle Technology (Cont.)

For nearly 100 years the triangulated tubular-steel frame has been a central element in the design of the bicycle (see "Bicycle Technology," by S. S. Wilson; Scientific American, March). An effort to improve on this time-tested strong, lightweight structure has now been announced. The new contender is a plastic bicycle, described by its manufacturers, the Original Plastic Bike Co. of Garden City, N.Y., as "the first production plastic bike in the world." The "virtually all-plastic" design features a one-piece plastic frame, plastic wheel forks, plastic handlebars, a plastic free-wheel mechanism, plastic front and rear hubs, plastic sprocket wheels, a plastic chain and a plastic 10-speed gear system of the *dérailleur* type. The hand brakes (of the caliper type) are also plastic, although the shoes and cables are made of conventional materials.

The new machine, unveiled at the recent International Cycle Show in New York, weighs less than 17 pounds. Its plastic parts are fabricated of a high-strength "structural resin" developed by the General Electric Company. According to spokesmen for the Original Plastic Bike Co., the plastic bicycle is "impervious to rust and the elements. The bikes can be molded—without the need for weld joints, priming, brazing and sanding. The color is throughout the thickness of the bike.... Because of the self-lubrication qualities of this patented plastic, maintenance and breakdown are virtually eliminated. It requires no grease or oil."

In addition to the 10-speed version exhibited at the cycle show, which will retail for "under $100," the firm expects to offer five- and three-speed models and also a 26-inch folding bicycle weighing only 16 pounds.

Slaughter in the New World

When the last continental ice sheet retreated in North America, there were at least 31 kinds of large mammals on the continent, among them mammoths, mastodons, horses, bears, saber-toothed cats, camels, goats, antelopes, tapirs, peccaries, capybaras, four genera of ground sloths and a giant beaver that weighed upward of 400 pounds. All became extinct at about the time man began arriving in the New World from Asia. Could man have been responsible? The question has been difficult to answer on the basis of archaeological evidence. Only seven unmistakable sites where men killed these extinct mammals have been found, and in all seven instances the animals were mammoths.

Paul S. Martin of the University of Arizona, writing in *Science*, has taken another approach to show how a small group of men could have given rise to the great extinction. Martin believes the slaughter began as soon as big-game hunters from Asia invaded the New World. Starting with the assumption that the large mammals of North America were probably as numerous as the large mammals in African game parks are today, he estimates that the ice-free parts of the continent supported some 50,000 pounds of big game per square mile. As the initial event in his hypothetical reconstruction he postulates that an Asiatic hunting band of 100 managed to reach the vicinity of Edmonton, Alberta, 11,500 years ago.

The maximum human population that good hunting ground can support is one person per square mile. Martin visualizes his band strung out in an arc with this average spacing, moving south of Edmonton across the prairie at a net rate of less than a mile per month, preying on the big game as it advanced. Within nine generations the hunters would have reached a point some 400 miles south of Edmonton. At an initial population growth rate of 3.4 percent per year, the number of people in the arc, which would by then have extended from the West Coast to Lake Superior, would have reached 61,000 and the advancing wave, with each person on the average occupying a square mile, would have been 100 miles deep. Within 250 years the arc would have reached the Gulf of Mexico and the total number of people in the hunting bands would have been 300,000.

If during this 250-year advance only one out of every four of these people had on the average killed the equivalent of one 1,000-pound animal per week, the big-game population in the zone of advance would have been reduced by 26 percent per year. Considering the low rate of reproduction among large mammals, an annual loss of this magnitude would have ensured the extinction of the large-mammal population within a decade. A century later the advancing arc would have reached the Isthmus of Panama and most of the big game in North America would have been extinct for more than 100 years.

Thereafter the sweep of the hunters could have been repeated across South America. The nomadic bands would have reached Tierra del Fuego some five centuries later, leaving that continent also stripped of its large mammals. In Martin's view the slaughter went so quickly that "perhaps the only remarkable aspect of New World archaeology is that *any* kill sites have been found."

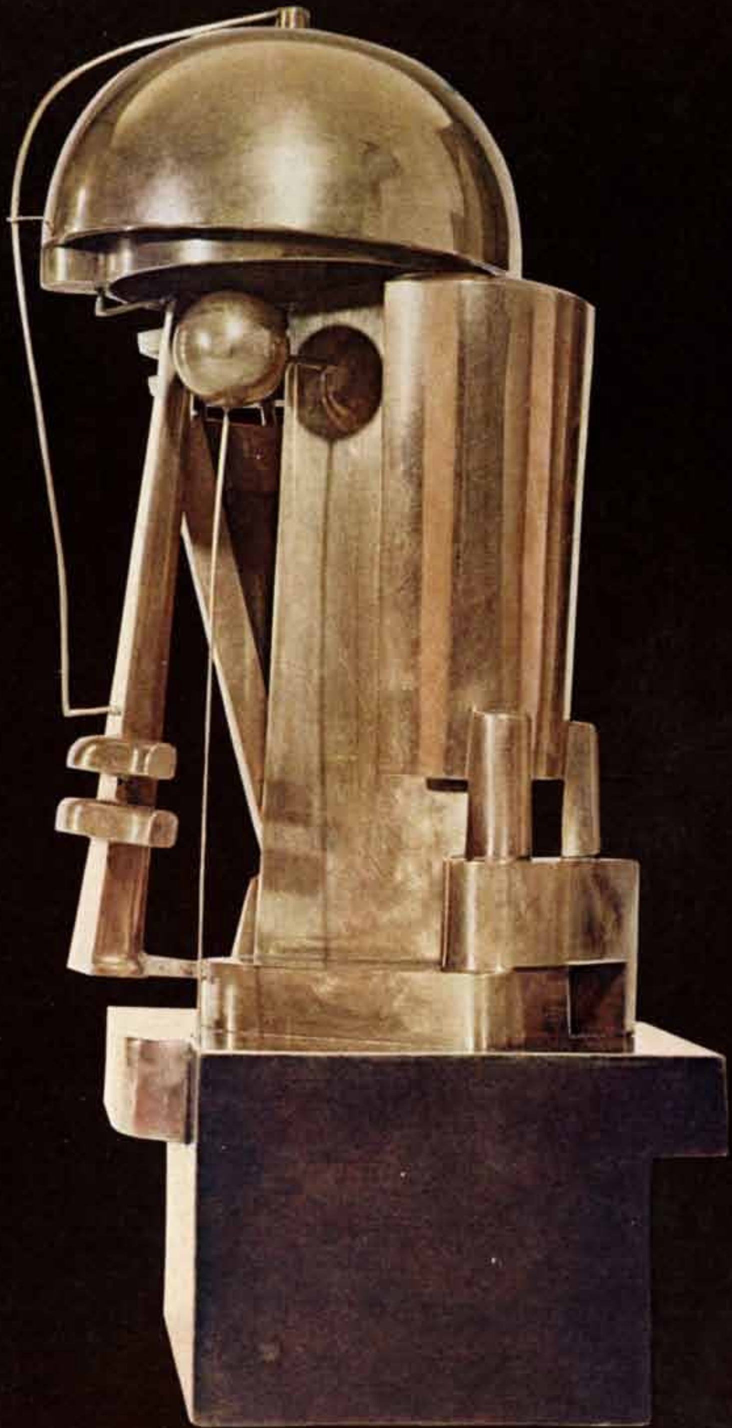
Home to Roost

The interdependence of all living things, the delicate balance among organisms in related biological niches, the inevitability of Nature's revenge when Man interferes with her subtle harmonies—these are among the clichés of popular ecology. They are ruefully illustrated by an epidemiological case history: a report on an outbreak of histoplasmosis that occurred at a junior high school in Ohio in 1970. The story is told in *The American Journal of Medicine* by Alan L. Brodsky, Michael B. Gregg, Matthew S. Loewenstein, Leo Kaufman and George F. Mallison of the U.S. Center for Disease Control in Atlanta.

On Thursday, May 7, more than twice as many students as usual were absent from the Willis Intermediate School in Delaware, Ohio—144 out of a student body of 872. The next day the absentee rate nearly doubled, and during the morning another 100 students went home; school was adjourned and health officials were notified. The "Willis flu," as it came to be called, was characterized primarily by fever, headache and chest pain, and to a lesser extent by a cough and sore throat. Cultures designed to identify viral or bacterial causative agents were negative. Most of the victims were back in school within a week; only five were hospitalized.

Investigation soon made it clear that the source was limited to the Willis School. The source was not cafeteria food (children who brought sandwiches from home got sick too) or the school water supply, which was found to be uncontaminated. This suggested that the disease was airborne and perhaps related to some common event. Investigators looked at the school calendar for the three weeks preceding the outbreak. There had been no assemblies, picnics or dances—but there had been Earth Day, April 22. The sixth-graders had gone out to clean up five local parks. Most of the seventh- and eighth-graders had merely cleaned out their lockers, but some special-education students in the eighth grade had volunteered to rake and sweep up debris around the building. In particular they had tackled an interior courtyard that had not been cleaned for some 20 years. It was a favorite roosting place for starlings and other birds and was encrusted with droppings. The children had worked there vigorously between noon and 1:15 on Earth Day and again the next day. Two of the air intakes for the school's forced-air ventilation system are in the courtyard. The kitchen and cafeteria adjoin the courtyard, and it developed that early in April one of the kitchen's two exhaust fans had been turned around so that it blew courtyard air directly into the kitchen and thence into the cafeteria.

Three weeks after the outbreak blood-serum samples obtained from 200 students selected at random were screened for antibodies to a number of respiratory diseases, including influenzas and other viral infections and histoplasmosis, which is caused by a fungus, *Histoplasma capsulatum*. The tests showed consistently high antibody levels only against *H. capsulatum*. Soil samples from the courtyard were found to be positive for *H. capsulatum*. Contaminated starling roosts are known to be a major source of histoplasmosis. As the authors of the epidemiological report suggest, "the result of a well-meant attempt at cleaning up the environment" was the cause of "the largest number of clinical cases of histoplasmosis ever to be reported in a single epidemic."



artist: belling museum of modern art


Looking for a wagon? The
Dodge has 58 ways for you to get the roominess of a wagon. More ways than any other manufacturer. And when you choose Dodge, you're choosing an engineering leader. Take the models here. With the exception of Colt, they're all equipped with Electronic Ignition which is virtually maintenance free. It also helps save you time and money by eliminating the troublesome points and condenser. That's why Dodge recommends changing spark plugs every 18,000 miles under normal driving conditions—compared to the 6,000 to 12,000 miles recommended by some of our competitors.
DODGE VARIATIONS ON A CLASSIC THEME... THE STATION WAGON.
Here's the wagon lovers' wagon. In two sizes. The spacious big Dodges—Polara and Monaco—and the family-sized Dodge Coronets. These models will carry a 4′ x 8′ sheet of plywood. Flat. Even with the tailgate closed. That's something some competitive wagon manufacturers seem to forget. And isn't room one of the reasons you're considering a wagon?
DART SPORT. THIS SLEEK LITTLE COUPE HIDES A WAGON IN DISGUISE.
Surprise! Dart Sport is like three cars in one. Count them. First, a five-passenger car. With the optional sun roof, you've got car two—a sun roof convertible. And with the optional fold-down rear seat, it's car three—a little wagon. With over 6½ feet of carpeted cargo area. Dart Sport Convertriple. Three cars in one.
CHRYSLER
MOTORS CORPORATION

biggest choice in the world is at your Dodge Dealer's.

SPORTSMAN. THE LARGEST SELLING COMPACT WAGON BUILT IN AMERICA.

The more you look at a Sportsman wagon, the more sense it makes. Take seating. There's room for five to fifteen people. Depending on the model. And you don't have to rough it inside. Sportsman wagons offer a full line of easy driving options such as power steering, automatic transmission, automatic speed control, and air conditioning. To name a few.

COLT WAGON. FROM JAPAN TO AMERICA C/O DODGE.

Why buy a little Dodge Colt wagon? Because it's got more room than a little car but costs less than a big wagon. And Colt is sold and serviced and backed by a network of Dodge Colt Dealers that stretches from one side of the U.S. to the other.

FACTORY-ENGINEERED TRAILER-TOWING PACKAGE.

The Dodge trailer-towing package has been specified and tested by Chrysler engineers. It is available on all models except Colt. On a big Dodge, this package can handle trailers weighing up to 7,000 pounds. And it's covered under your new car warranty. See your Dodge Dealer for full details.

Dodge

Dodge Trucks

ROOTS

The leaves of a plant gather the sunlight and carbon dioxide needed for photosynthesis; the roots gather the water and also the minerals that are essential not only to the plant but also to all animal life

by Emanuel Epstein

Of the 90-odd chemical elements found naturally on the earth only 24 play functional roles in living organisms. Most of these nutrients are mineral elements, and for life on land the source of the large majority of them is the soil. They enter the terrestrial biosphere through the roots of plants. Roots are therefore the agents for one of the key processes in the chemical economy of nature: the process that unlocks from the soil its store of inorganic nutrients, without which there can be no life as we know it.

Animals can supplement their diet by the direct use of salt, and grazing animals can do so by ingesting some soil along with herbage, but most of the essential nutrient elements present in the soil are no more available to animals than is the carbon present in the atmosphere as carbon dioxide. It is only after the leaves of green plants transform atmospheric carbon into carbon compounds by means of the process called photosynthesis that animals have access to the element. And it is only after plant roots absorb the mineral nutrients in the soil by means of the process called active transport that animals can obtain them. The substances acquired by both the leaves and the roots of plants are transferred into the food chains of the animal kingdom when the plants are eaten.

Roots engage in many physiological and metabolic activities that are common to all plant tissues. They grow and develop; they respire; they synthesize and degrade various compounds. Here, however, we shall be mainly concerned with their unique function of "mining" the soil for mineral nutrients.

Perhaps 5 percent of a plant's dry weight is minerals. Eight elements account for the bulk of this amount; they are potassium, calcium, magnesium, nitrogen, phosphorus, sulfur, chlorine and silicon. All are normally present in the "soil solution," as the water of the soil is called, at very low concentrations. Like leaves, roots have to absorb nutrient elements that are available only in very limited quantities. For example, the carbon dioxide assimilated by leaves is present in the atmosphere at a concentration of .03 percent, and the phosphorus taken up by the roots is typically present in the soil solution at a concentration of about one part per million. The nature of the extractive task has profoundly influenced the evolution of roots as organs of absorption and has given rise to similarities between roots and leaves.

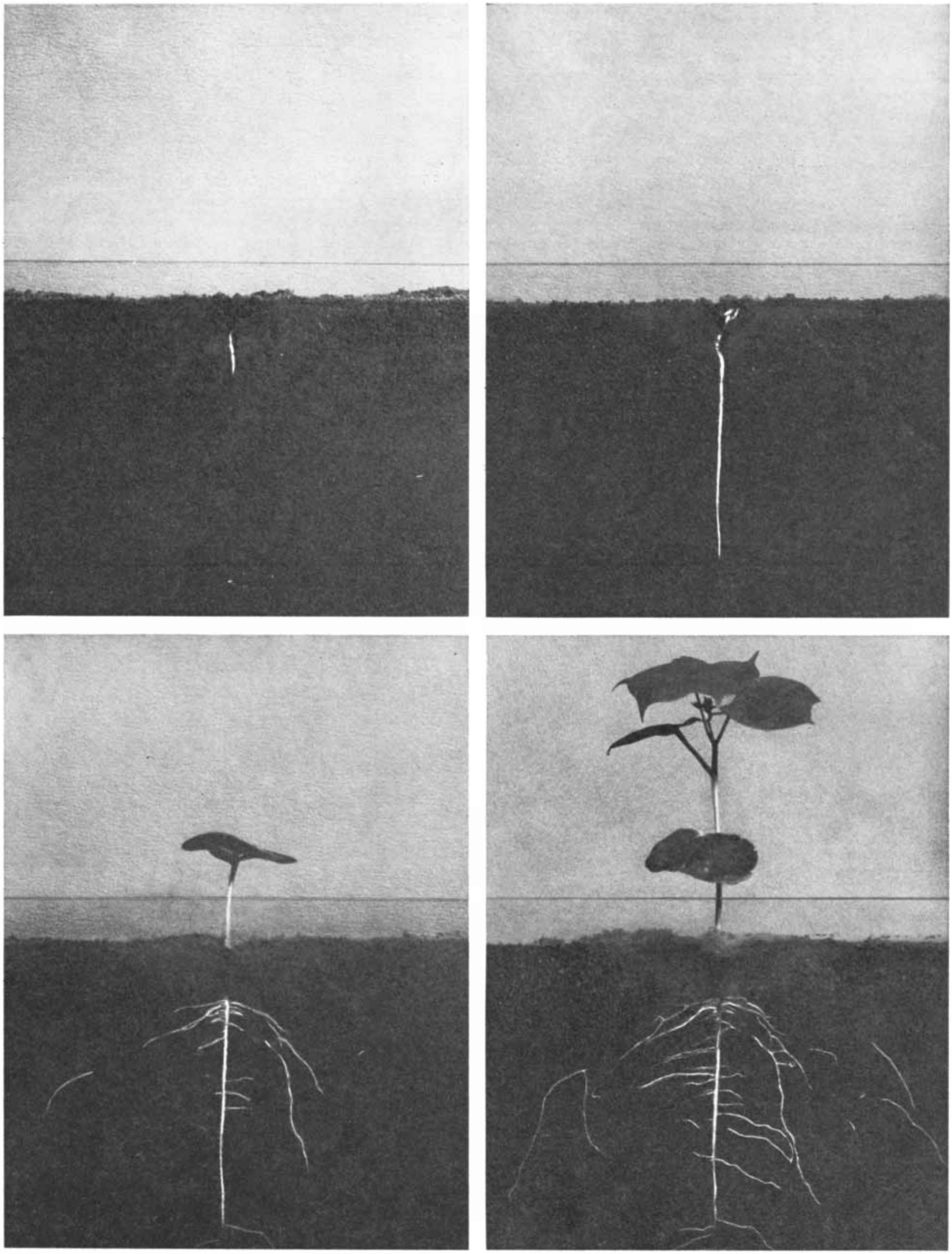
Root Geometry and Anatomy

Consider the physical shape of some higher animal, a rabbit, say, or a man. The evolutionary tendency in such animals has been to minimize surface area in relation to mass to the extent consistent with functional requirements such as locomotion and sensory perception. The trend to minimal area not only optimizes the animal's thermodynamic efficiency but also exposes the least possible surface to injury. Just the opposite is true of a plant. Plants consist of extensive systems of surfaces both above the ground and below it. The total area of a tree's leaf surfaces, for example, is several multiples of the ground area covered by the tree's crown. Such a large surface area is, so to speak, an engineering necessity for an array of organs with a dual function. The leaves must not only acquire from the ambient air a gas that is present at a very low concentration but also intercept another diffuse resource: the radiant energy of the sun. The larger the leaf area, the larger too the interface between the plant and these dilute inputs.

The same considerations apply to the plant's root system with even greater force. The roots abut on the ambient soil solution, absorbing from it the nutrients that are present in the form of inorganic ions at low concentrations. Since the soil solution is essentially unstirred, it can become depleted of nutrient ions in the zone of root contact because the diffusion of replacement ions from elsewhere in the solution is a slow process. In contrast to the leaf, which is rarely in prolonged contact with the same collection of air molecules, the root must work with substances that are relatively immobile in the soil.

In response to these features of the soil environment—extreme dilution and immobility of nutrients—evolution has given rise to root systems that, by means of progressive branching and ramification, bring an astonishingly large surface area into contact with the soil. The area of contact is further multiplied by the growth of "root hairs": fingerlike epidermal cells that extend into the soil roughly at right angles to the root axis. Some grasp of the dimensions attainable by root systems can be gained from an experiment conducted in the 1930's by Howard J. Dittmer at the State University of Iowa. He placed a single winter rye plant (*Secale cereale*) in a box of soil 12 inches square and 22 inches deep and let the plant grow for four months. At the end of that time he carefully liberated the root system from the soil and measured it. The total length, root hairs excluded, was 387 miles; the total surface area was 2,554 square feet. When the length and surface area of the root hairs were added in, the totals for the entire system were nearly 7,000 miles and 7,000 square feet.

The close contact between soil and roots, dictated by the function of the roots, is the reason plants are stationary. On land a mobile organism cannot se-

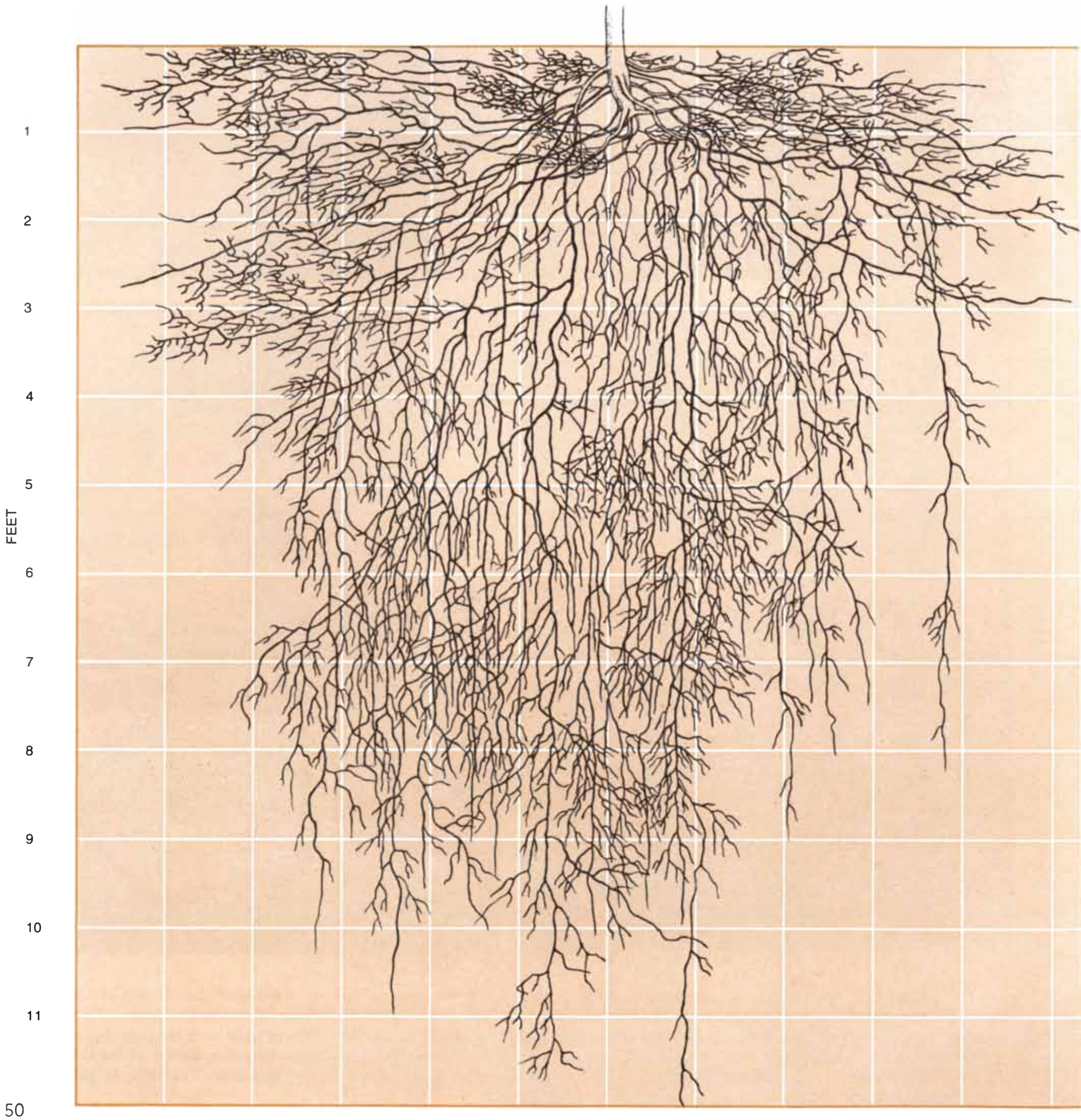


COTTON SEEDLING shows progressive root-system development over a 27-day period. The seedling is seen on the third day after planting (*top left*), on the fourth day (*top right*), on the ninth day (*bottom left*) and on the 27th day (*bottom right*). The taproot tip had grown out of camera range by the sixth day, and the taproot was 100 centimeters long by the ninth day. Within 12 days the taproot had reached the bottom of the compartment, 150 centimeters below the soil surface. Photographs were made at Soil and Water Institute of the Agricultural Research Service of the U.S. Department of Agriculture, located at Auburn University in Auburn, Ala.

cure nutrients from the soil. To do so it would have to be rooted, and an organism cannot be both rooted and capable of locomotion. A stationary existence, in turn, provides no evolutionary pressure of selection for an elaboration of sophisticated organs of sense and coordinated body movements or the central nervous system such organs and movements require. It may be predicted with some assurance that, if intelligent life on solid land is ever discovered elsewhere in the universe, there as on the earth the world of life will consist of two kingdoms: plant and animal. Only mobile organisms will develop intelligence and only stationary ones can secure from a solid substrate nutrient elements essential for the functioning of living matter.

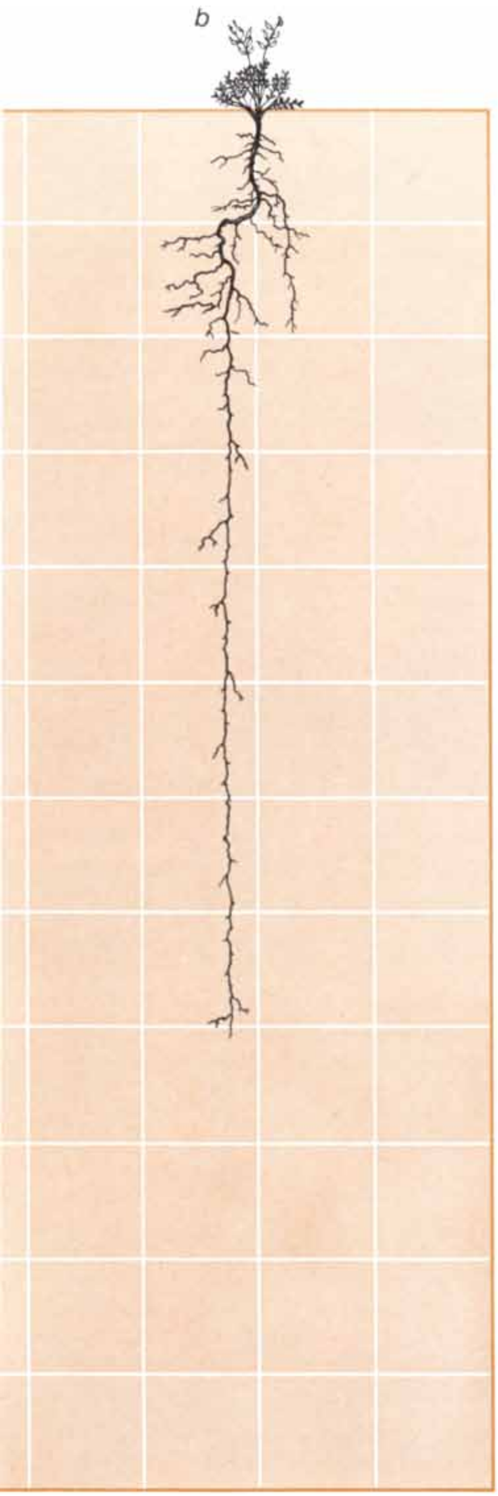
Three of the functions that a root must perform are clearly reflected in its anatomy. One is the absorption of nutrients and water from the soil solution. That is accomplished by the cells at or near the root surface: the outermost epidermal cells and a more abundant array of cells in the adjacent cortex. The cortex, in turn, ends in a single layer of cells that makes up the endodermis. Together these three arrays of cells surround the central stele of the root like a three-ply sleeve. Two further root functions consist in the upward transport of water and nutrients that are carried throughout the plant in the transpiration stream, and the downward movement of sugars and other products of photosynthesis to the parts of the plant that do not engage in photosynthesis. Among these parts are

a



the roots themselves, which are shielded from sunlight.

Cells within the central stele of the root perform both of these functions. The upward transport occurs in special "vessels": long, hollow conduits in the part of the stele called the xylem. The vessels are composed of cells, connected end to end, that have lost their cell contents and, being empty, can act as pipes. Water and minerals in solution ascend in these pipes as a result of the pull exerted when water evaporates from the plant's leaf surfaces. This upward flow of water and mineral nutrients is called the transpiration stream. The downward transport occurs in "sieve tubes": living cells in another part of the stele known as the phloem. The sieve-tube cells too meet end to end; the products of photosynthesis flow downward through the sieve-tube-cell cytoplasm, passing from one cell to the next through numerous pores located at the points where the cells meet.



Transport Studies

We know that mineral nutrients move from the soil solution into the epidermis of the plant root and onward from there until they are at last swept upward through a xylem vessel by the transpiration stream. By what paths do they travel? Experiments aimed at answering this question are seldom conducted in the field, because adequate control of the chemical composition of the soil solution is impossible. Instead the experimenter makes use of plants raised by the "solution culture" method. The plants are kept in containers filled with a nutrient solution. The solution is aerated to provide the oxygen the plant roots need and is also supplied with all the elements essential for plant growth in the form of dissolved inorganic salts. Plants as large as trees can readily be raised to maturity in this way.

For experimental purposes the culture solution is often specially prepared so that the particular mineral nutrient of interest to the investigator is absent from the solution altogether or present only at a very low concentration. This stratagem keeps the cells of the plant's roots fairly "empty" of the mineral. After a period of growth under these conditions the roots are cut off and immersed in other containers where the mineral under study is present in the solution in radioactively labeled form. The time the root tissue spends in the labeled solution is brief; in our laboratory at the University of California at Davis immersion for 10 to 20 minutes is typical.

At the end of the interval the roots are removed from the solution and a counter is used to measure the amount of labeled mineral the roots have absorbed. What is being measured in short-term experiments such as these is absorption by the cells of the epidermis and cortex only; in such brief periods the transfer of the nutrient mineral into the conducting tissues of the stele is negligible.

Let us consider what happens during such an experiment. As the root is immersed in the labeled solution, positively and negatively charged ions of the various dissolved nutrient salts diffuse quickly into the porous cell-wall spaces of the epidermis and cortex. The movement is a purely physical one, no different in kind from diffusion in soil or in any other porous matrix exposed to a solution.

At first there might seem to be no reason why the nutrient ions could not simply continue to percolate inward through the cell-wall spaces until they entered the conducting xylem vessels in the stele. In actuality this pathway is blocked at the point where the single row of endodermal cells separates the cortex from the stele. Embedded in the cell walls of the endodermis is a structure known as the Casparian strip (named for a 19th-century German botanist, R. Caspary, who first described it). The Casparian strip consists of waxlike substances that are impermeable to water and to charged particles and prevents any further diffusion of the ions.

Here is a paradox. Clearly the Casparian strip presents an impenetrable barrier to the further movement of mineral nutrients within the plant. Yet we know that in fact the nutrients do leave plant roots and travel up the stems. What other pathway is available? One possibility is that the ions migrate inward from cortex to stele by traveling from cell to cell along the strands of cytoplasm that connect them. In order to reach the cytoplasm of epidermal and cortical cells, however, the ions would first have to cross the plasma membrane that separates the cytoplasm from the cell wall. This too appears to be an impossibility; like the Casparian strip the fatty plasma membrane is also a permeability barrier that does not allow charged particles such as mineral ions to diffuse through it.

The Inside Track

Fortunately there is a solution to this paradox. We not only know empirically that the ions pass from root to stem; we can demonstrate that they do so by somehow penetrating the plasma-membrane barrier. For example, let us im-

DIFFERENT PATTERNS of root growth are exemplified by an apple tree (*a*) and a locoweed (*b*). The tree roots grow both laterally and vertically; the extent of the growth in each direction is determined not only by heredity but also by such environmental factors as the structure of the soil and its water content. The locoweed, a plant adapted to dry environments, has a characteristically deep taproot with little lateral growth.

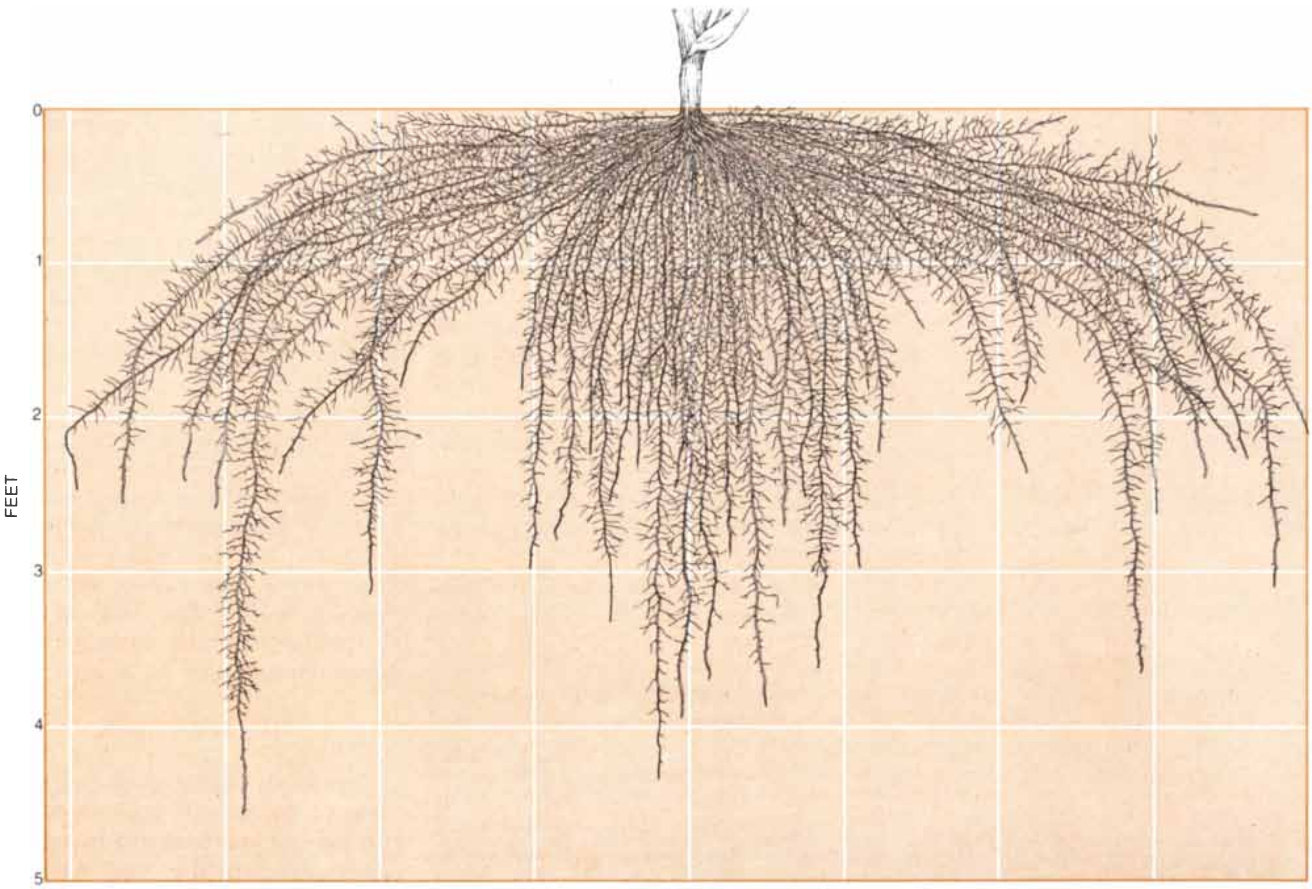
merse roots in a solution of radioactively labeled potassium chloride; the labeling might be done with potassium 42 or chlorine 36 or both. After a period of immersion under conditions of temperature and aeration that are favorable to active cell metabolism we remove the roots from the solution and rinse them well. This washes away not only the labeled solution that has adhered to the root surface but also the solution that has diffused into the porous cell-wall spaces of the epidermis and cortex.

Now, if the plasma membranes of the epidermal and cortical cells have been effective permeability barriers and have barred the way to the further diffusion of the mineral ions, the well-rinsed roots would be expected to show no sign of radioactivity. Moreover, if one postulates instead that for some reason or other the plasma membranes actually do not bar the diffusion of ions, then the rinsing should have served to leach out any ions that had entered the cell cytoplasm and again the roots would now show no sign of radioactivity. In actuality the roots turn out to be demonstrably radioactive. Some of the ions have indeed penetrated the plasma membrane. Furthermore, they have made a one-way trip; they cannot now recross the membrane in the opposite direction.

Quantitative experiments of this kind reveal many fascinating features of the transport of ions across the plasma membrane. For example, one can expose barley roots that contain virtually no chloride to a solution containing radioactive chlorine at a concentration of only two-tenths of a part per million. Within an hour the concentration of chloride within the roots rises to 40 parts per million, a level 200 times higher than the level in the solution. When the roots are now transferred to a solution that contains no chloride, the labeled chloride absorbed by the roots does not diffuse out of them, even though the activity gradient favors such a diffusion. With certain plant materials, exposed to solutions for several days, the ratio of the internal concentration of the mineral to the concentration of the external solution can become as high as 10,000 to one. The process of concentration obviously requires energy; the source of the energy is the metabolism of the root cells.

That the absorption of ions by roots is a metabolically active process is demonstrated by experiments of another kind. If nitrogen is bubbled through the experimental solution instead of air, thereby depriving the roots of oxygen, the absorption of ions ceases. The same is true when for experimental purposes any one of a wide range of metabolic poisons is added to the solution or when the solution is chilled so that the root's metabolic processes are slowed. Dennis R. Hoagland of the University of California at Berkeley, who was the pioneer investigator of ion absorption in the 1930's and 1940's, emphasized this metabolically active aspect of ion absorption by terming the process "active transport."

But how does this active transport work? From the 1930's on plant physiologists have used as a working hypothesis the concept that chemical entities called "carriers" assist the ions in crossing the plasma-membrane barrier. The hypothesis as it stands today can be outlined as follows: When the carriers are positioned at the outer surface of the membrane and are bathed by the external solution, they are able to form "complexes" with the ions in the solution. It is in the form of a carrier-ion complex, and not as a free ion, that the mineral nutrient moves across the membrane to its inner surface. There, as a



DENSE NETWORK of corn roots is twice as wide as it is deep. The total area of contact between the soil and the roots and root hairs of plants can be huge; a single rye plant, grown in a small box, developed roots that had nearly 7,000 square feet of surface area.

result of a change in the molecular conformation of the carrier, the ion is released into the cytoplasm of the cell. The unfreighted carrier then cycles back to the outer surface of the membrane, reassumes the conformation that will enable it to form a complex with another ion, and repeats the process. The ions carried through the membrane in this way are now trapped behind its inner surface. They cannot diffuse outward through the membrane, because of its impermeability to charged particles. Nor can they ride back to the outer surface by forming complexes with other carriers, because the conformation of the carriers at the inner surface of the membrane is such as to release ions rather than to form complexes with them.

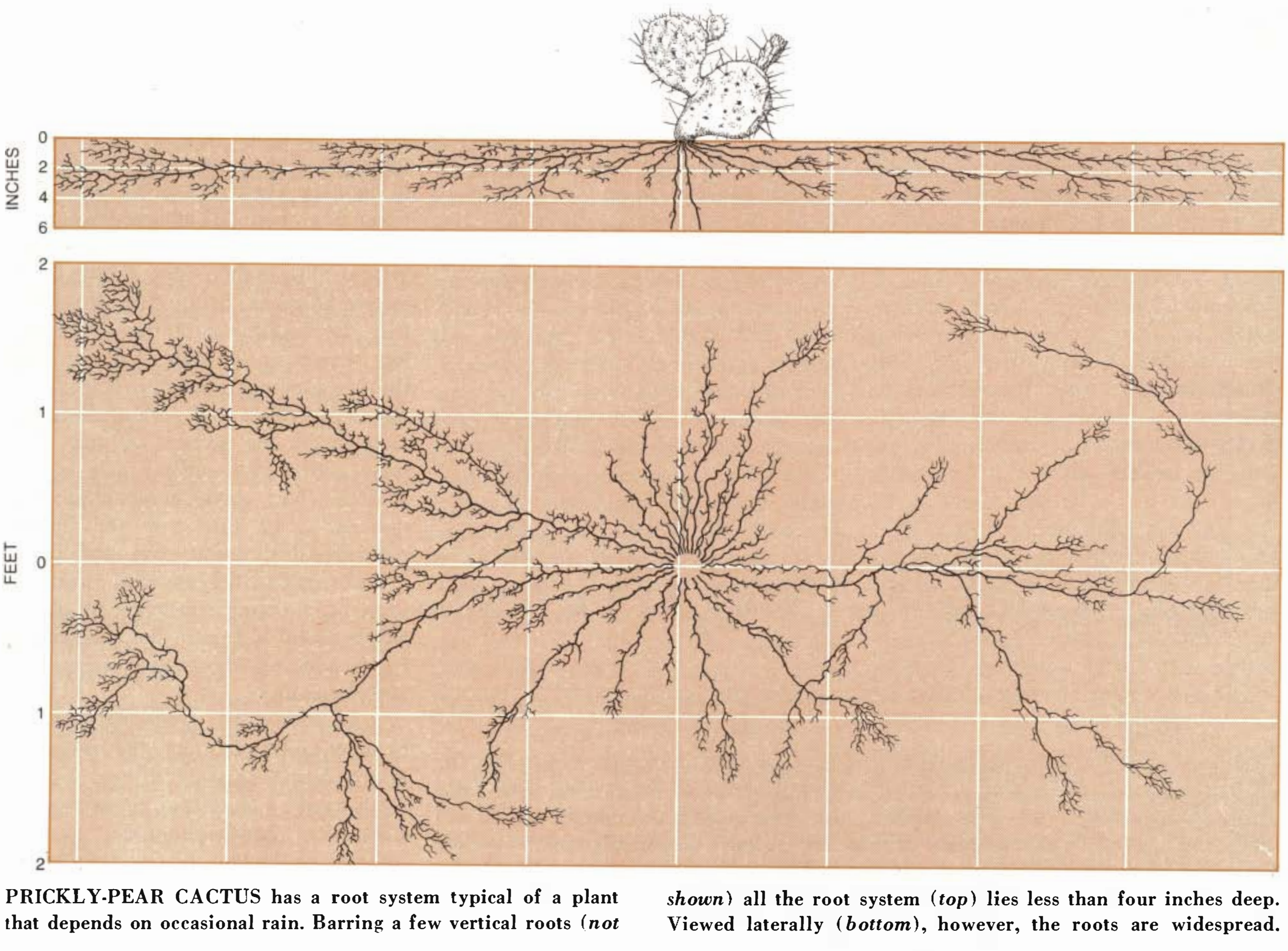
Transport Anomalies

The rate of absorption of many ions depends in a distinctive way on the concentration of the ions in the surrounding solution. At ion concentrations ranging from very low values to about two hundredths of a millimole per liter the rate of absorption rises steeply in relation to the increase in concentration. As the concentrations approach a value of one-tenth of a millimole per liter the rate of absorption levels off; it remains almost level until the concentration reaches one millimole per liter. This essentially steady rate of absorption over a considerable range of concentrations indicates that the carrier mechanism has been saturated.

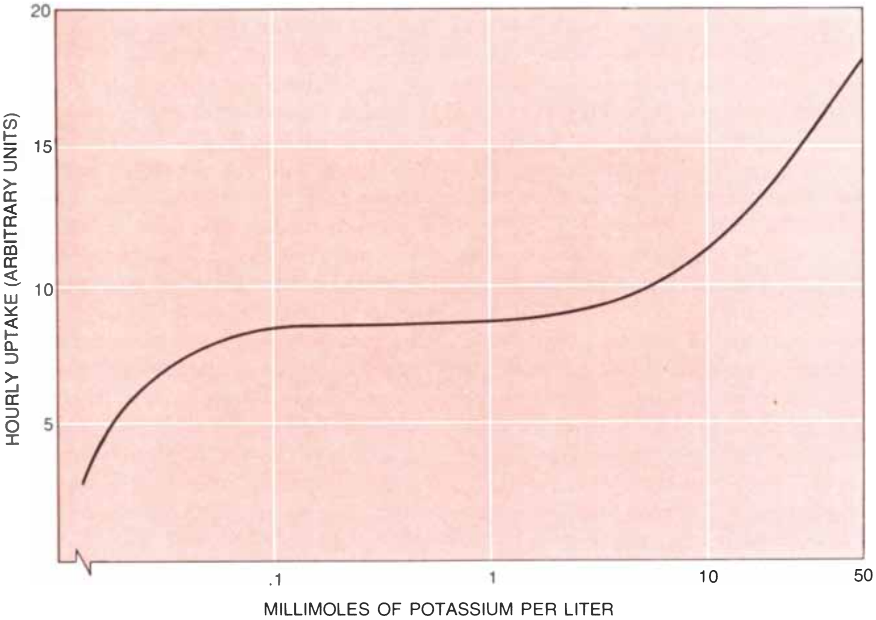
At still higher concentrations, however, the rate of absorption increases to values well above that level. This suggests that at ion concentrations in the solution greater than one millimole per liter a second transport mechanism goes into operation. The low side of the range of concentrations corresponds to the levels of nutrient concentration most commonly found in the soil solution. Higher concentrations, however, are by no means unusual.

Observations of what has been termed the "selectivity" or "specificity" of transport lend force to the hypothesis that there are two different transport mechanisms, one working at low ion concentrations and the other at high concentrations. For example, measurements have been made of the absorption of potassium by barley roots from solutions containing various concentrations of potassium and sodium. When the concentration of potassium in the solution is low, the rate of potassium uptake by the barley roots is about the same regardless of whether the concentration of sodium in the solution is high, low or zero. Sodium and potassium are close chemical relatives; they are neighbors in the first group in the periodic table. That being the case, the transport mechanism that operates at low levels of potassium-ion concentration is evidently highly specific for potassium and is not inhibited by the presence of large quantities of the element's close relative, sodium. This is not true of the transport mechanism that operates at high potassium concentrations. When the potassium in the solution rises above one millimole per liter, the presence of sodium in the solution strongly depresses the absorption of potassium.

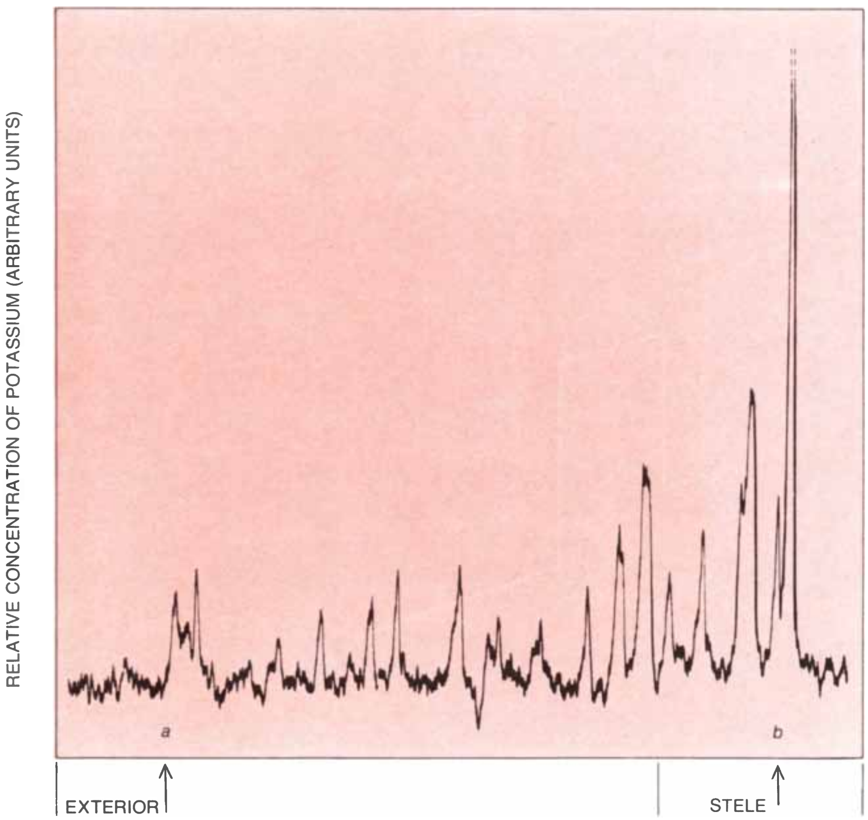
How might such reactions prove to be advantageous to a plant under natural circumstances? With regard to the mechanism that operates at low potassium levels, the ability to absorb potassium selectively could be vitally important to plants in many situations. Of all the minerals needed by plants potassium is the one required in the greatest quantity. In many soils, however, the concen-



PRICKLY-PEAR CACTUS has a root system typical of a plant that depends on occasional rain. Barring a few vertical roots (*not shown*) all the root system (*top*) lies less than four inches deep. Viewed laterally (*bottom*), however, the roots are widespread.



RATE OF ABSORPTION of potassium by plant roots rises in proportion to the external concentration at low levels, becomes steady at intermediate levels and rises again at higher levels. This suggests the existence of two separate carrier mechanisms for the transport of mineral ions across the plasma membranes of plant cells. Evidently one carrier mechanism becomes saturated when the concentration exceeds one-tenth of a millimole per liter. At concentrations above one millimole per liter, however, the rate of absorption rises to values in excess of the maximum rate obtained by the mechanism that acts at lower concentrations, suggesting that a second mechanism begins to operate at these higher ion concentrations.



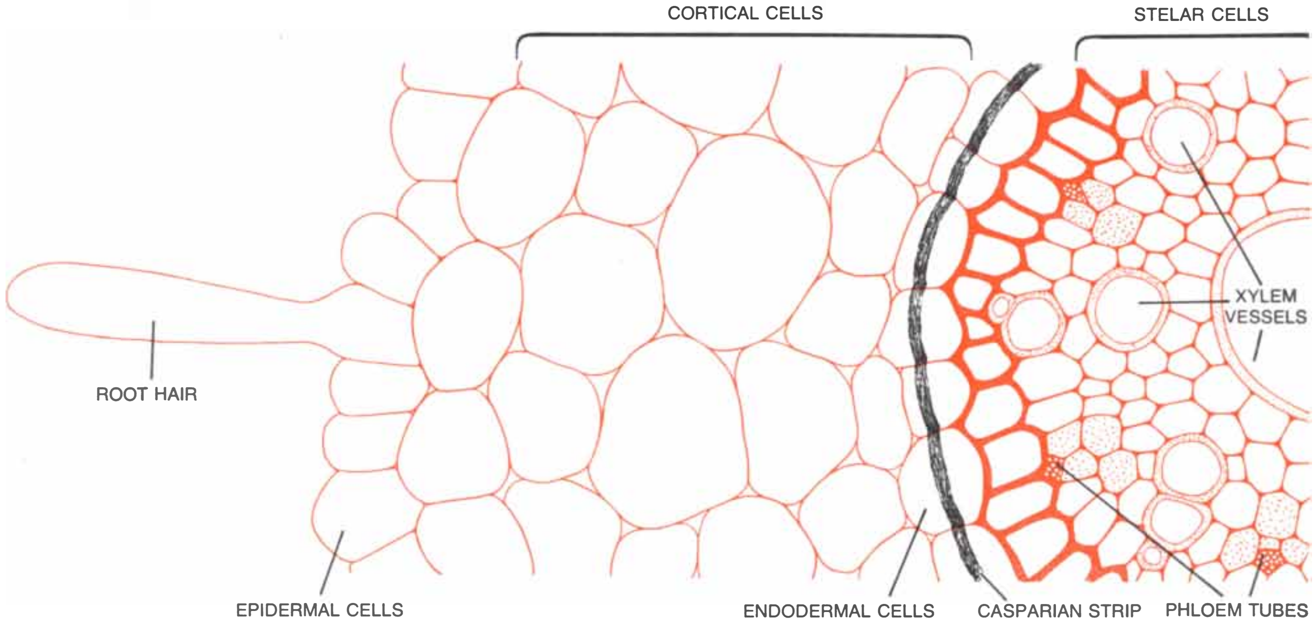
CONCENTRATION OF POTASSIUM in different parts of a root sample is far higher in the stelar cells adjacent to xylem vessels (*b, right*) than elsewhere in either the stele or the cortex (*center and left; "a" marks outside of root*). This suggests that mineral ions do not diffuse through the plasma membrane of stelar cells to enter the transpiration stream but are instead accumulated in stelar cells adjacent to the vessels and then "pumped" across the membrane by special carrier mechanisms (*see illustrations on page 58*). Measurements are by A. Lauchli, A. R. Spurr and R. W. Wittkopp of the University of California at Davis.

tration of potassium in the soil solution is quite low: one millimole per liter or less. In the arid and semiarid regions of the world the concentration of sodium in the soil solution often exceeds this level. Thus a potassium-selective mechanism of active transport serves to keep plants in such environments supplied with the potassium they need while preventing them from being swamped with sodium, which most plants need only in small quantities or not at all.

The hypothesis that carriers are the agents responsible for the transport of mineral ions across the plasma membrane of plant cells has been developed on the basis of experiments in spite of the absence of any direct evidence of the carriers' chemical nature. This is not a unique situation in the biological sciences. Genes were postulated as the basic factors in inheritance long before they were chemically identified as deoxyribonucleic acid. By the same token, enzymes were recognized as the catalysts of biological transformations well in advance of their identification as proteins. The work of Thomas K. Hodges and his collaborators at Purdue University now makes it likely that the carriers responsible for ion transport in plants are certain enzymes of the adenosine triphosphatase (ATPase) group: proteins that utilize the energy of adenosine triphosphate (ATP) for the energy-demanding transport of mineral ions across the cell membrane.

ATPases have been implicated in the transport of ions across cell membranes in animals. Their separation from plant tissue has been difficult, however, and the ATPases of plants evidently differ a good deal from their counterparts in the animal kingdom. In any event it is likely that the hypothesis of transmembrane ion carriers in the roots of plants is on the verge of being confirmed by direct chemical evidence. If such confirmation is provided, the carriers will join company with genes and enzymes as substances whose mode of operation was understood or suspected well before their chemical identity was established. (The reverse is the case with auxin, the growth-promoting substance in plants, whose chemical identity has been known for decades but whose mode of operation still eludes us.)

Even though the epidermal and cortical cells of the root play the key role in accumulating mineral nutrients, the root retains only a fractional amount of the minerals it absorbs from the soil. The majority of the ions migrate from cortex to stele and thence into the vessels of the xylem, where they move upward into



CASPARIAN STRIP bars mineral ions from moving by diffusion through the porous wall spaces of cortical cells (*left*) into the wall spaces of the stelar cells (*right*) and thence into the xylem vessels of the stele. The strip, connecting the walls of adjacent endodermal cells (*black*), is impregnated with waxlike substances and therefore is impermeable both to water and to ions in solution.

the plant shoot as passengers in the transpiration stream. Plant physiologists are fairly well agreed that the first step in the journey is the one-way passage through the plasma membrane described above, followed by movement through successive cells by way of the strands of cytoplasm that connect one cell to the next. There is less agreement, however, about what happens after that.

Obviously if they are to reach a xylem vessel the traveling ions must once more travel across a plasma membrane; in this instance it will be the membrane of a stelar cell and the passage will be outward, from the cytoplasm into the porous cell-wall space beyond. But how do the ions cross the membrane? Obviously there must be a difference between stelar-cell membranes and cortical-cell membranes. Cortical-cell membranes transport ions inward, from the wall space to the cytoplasm, whereas stelar-cell membranes allow the ions to travel in the opposite direction.

Second Crossing

Until recently the most favored hypothesis to explain the exit of ions from the cytoplasm of stelar cells was that, unlike the membranes of cortical cells, the membranes of stelar cells were permeable. The movement from stelar-cell cytoplasm to xylem vessel would therefore be a matter of simple diffusion. Now, however, four groups of investigators (one in England, one in Germany, one in Australia and our own group at Davis) have doubts about this hypothesis. We think it is more likely that the membranes of stelar cells possess "reverse" carrier mechanisms that operate basically in the same way that the carriers of cortical-cell membranes do but in the opposite direction, transporting ions from the cytoplasm into the cell-wall space.

The most direct evidence in support of this belief comes from experiments with the electron-probe analyzer. In this instrument a narrow beam of electrons is made to travel across a section of the plant tissue under investigation; when the electrons impinge on atoms of the mineral elements present in the tissue, X rays of wavelengths characteristic of the elements are emitted. We have scanned sections of root tissue in this way to determine their potassium content, beginning at the epidermis and moving across to the stele. The concentration of potassium proves to be much higher in stelar cells adjacent to xylem vessels than it is anywhere else [*see bottom illustration on opposite page*].

This finding makes it very unlikely that the living cells of the stele are passive and leaky. They seem to play an active role instead, accumulating high concentrations of nutrients and "pumping" them across the plasma membrane into the porous cell-wall space beyond. Once there the ions are free to move into the xylem vessels by simple diffusion. With this second passage across a plasma membrane a crucial series of operations has been concluded; the root has carried mineral nutrients to the vessels that lead to the part of the plant that is out of touch with the soil, making the nutrients available to the shoot and, indirectly, available to any animal that consumes the plant.

Accumulating Minerals

What amounts of minerals do the roots of plants extract from the soil each year? The quantity of minerals present in plants is variable; it depends not only on the kind of plant but also on the soil the plant grows in and other factors. It is a conservative estimate, however, that 5 percent of the weight of dry plant material consists of the eight elements potassium, calcium, magnesium, nitrogen, phosphorus, sulfur, chlorine and silicon. Several other "micronutrient" elements are also present but in such limited quantities as to constitute only a minute fraction of the mineral accumulation in most plants.

Ecologists estimate that the total annual production of matter by land plants is a dry weight of 100 billion metric tons. Using the estimate of 5 percent for the plants' mineral content, this means that plants "mine" some five billion metric tons of minerals each year. Since 5 percent is a conservative estimate of the mineral contents of plants, the actual tonnage mined may be considerably higher. How does this compare with man's mining endeavors? The Department of the Interior *Minerals Yearbook* for 1970 states that the worldwide pro-

SOME THINGS ARE CHANGING FOR THE BETTER:

A lot has happened to interferometry since Professor Michelson.

In 1892 Professor A. A. Michelson used his recently invented interferometer to determine the number of wavelengths of cadmium light in the standard meter. His error was 5 x 10^{-8} meter. It took another 70 years until the invention of the laser permitted the application of interferometry with this accuracy to everyday industrial measurements.

Today, HP's laser interferometer has found a place in metalworking and integrated circuit manufacturing as the tool to guarantee that a given dimension is precisely achieved. This modern version of Michelson's invention is enhancing accuracies of all types of new and existing machines and, as a consequence, lowering inspection costs and reducing rejection rates.

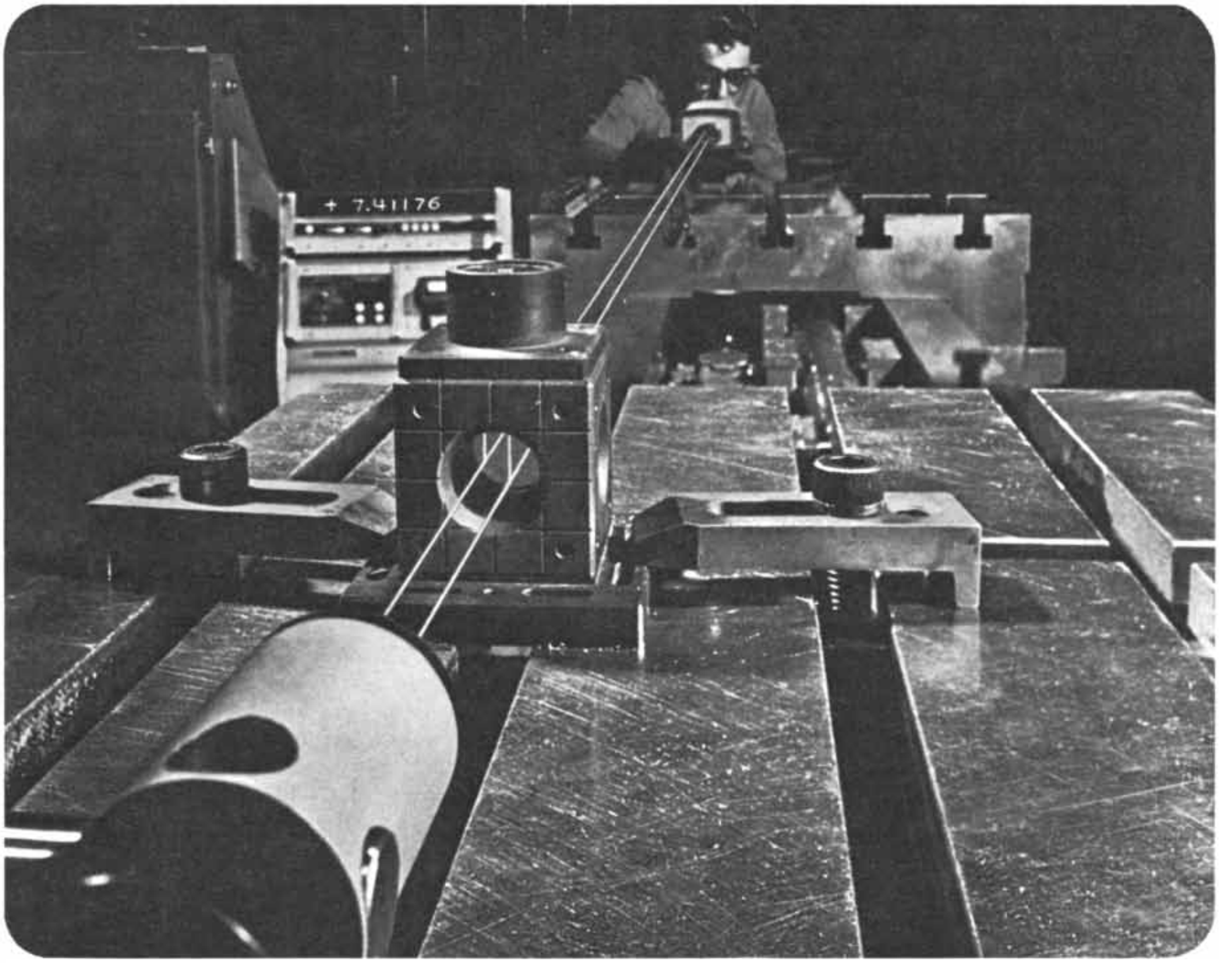
A little over a year ago, a group of HP engineers seeking to improve the quality and flexibility of this measurement technique developed the "HP Magic Cube." In essence, it is an interferometer that can have a laser beam aimed at it from some distance away.

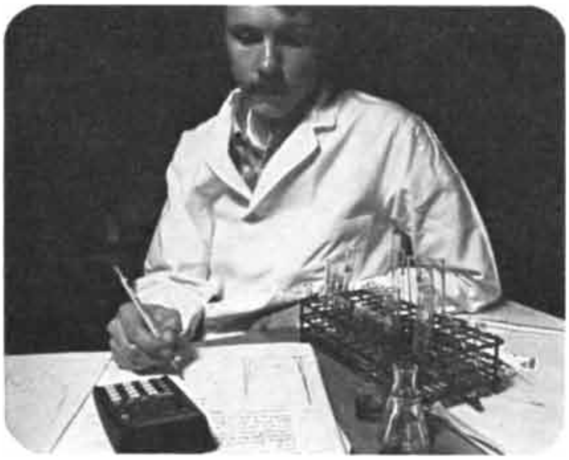
With it many kinds of measurements can now be taken without moving the laser head – pitch, yaw, flatness, straightness, squareness, and non-contact measurements – all that's needed to define a machine tool or measuring machine.

It can even make measurements around a corner, in tight places, or simultaneously measure x and y axes.

Moreover, it's portable. You can set it up, plug it in and start making error-free measurements in minutes wherever it's required in your metrology and quality control laboratories.

If you would like to know more about laser interferometry, we'll be glad to mail you our comprehensive brochure on the HP 5526A Laser Measurement System.





High-performance liquid chromatography: a new HP analytical tool for life scientists.

Gas chromatography (GC) has revolutionized analytical chemistry in practically every discipline. However, it has been of somewhat limited help to the life scientist because he works mostly with large molecules that cannot be vaporized for GC without breaking them down.

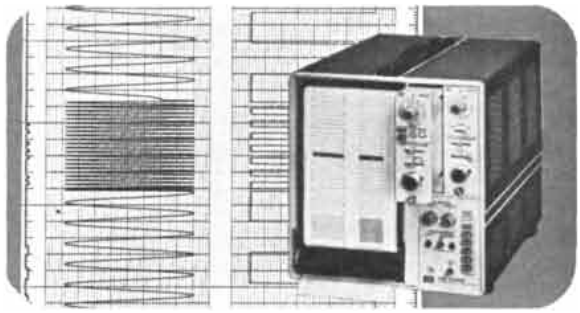
As a result, liquid chromatography (LC) has always been more attractive to the life scientist because it separates a sample in the liquid state at relatively low temperatures. But only recently has LC instrument technology advanced enough to give him the analytical speed and sensitivity that he also needs.

At the forefront of this new technology is HP's new Model 1010 Liquid Chromatograph. Already well established with European chromatographers, the 1010 is capable of performing all forms of LC, including ion-exchange and gel permeation. It delivers constant flow through the columns at pressures up to 3700 psig. A septum-less injector permits reproducible sample injection against full-column pressure; column change and solvent flush are exceptionally fast and convenient. Completely modular in design, the 1010 can be equipped for solvent programming, gradient elution and fully automatic preparative operation. Ask for full information.

Just a better trace for your money. That's all.

There's a great temptation to talk about the features on our new HP 7402A two-channel oscillographic recorder as though they were the last word in recorder design. But as we both know, the field is in a great state of flux, with new technology appearing in every model.

Common sense prevails. We simply want you to understand that the HP 7402A delivers one of the clearest, most readable, traces you will find on any currently available machine irrespective of price. Yet the price of the HP 7402A can be as little as $1740 for a complete recorder.



For your money, you get a big, wide, easy-to-read 50 mm trace on each channel. A high-pressure inking system puts down an instant drying, fine image with no overshoot, no skipping and no smear at any speed (in tests, the carbide tips showed no wear after 2,000 miles of travel over 80,000 feet of chart paper; you can expect them to last the life of the machine).

You also get the flexibility and economics of plug-in amplifier modules and a wide range of options to accommodate your exacting recording needs.

We'll be happy to send you more information on this sensible recorder.

00304

Mail to: Hewlett-Packard, 1502 Page Mill Road, Palo Alto, CA 94304
Please send me information on the following:
( ) HP 5526A Laser Measurement System
( ) HP Model 1010 Liquid Chromatograph
( ) HP 7402A Oscillographic Recorder

Name__________

Title__________

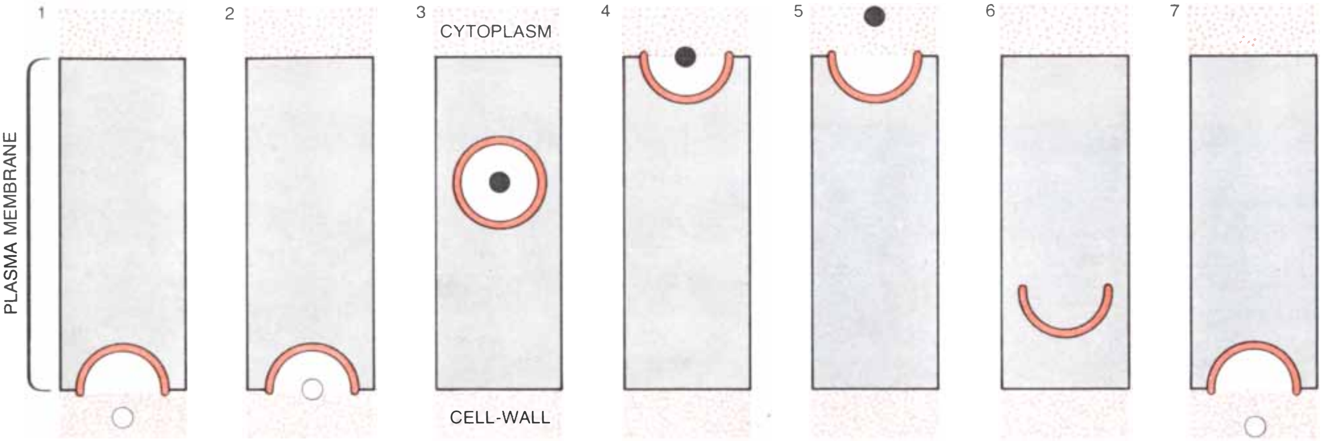
Company__________

Address__________

City__________ State______ Zip______

HEWLETT hp PACKARD

Sales, service and support in 172 centers in 65 countries.
Palo Alto, California 94304. Offices in principal cities throughout the U.S.

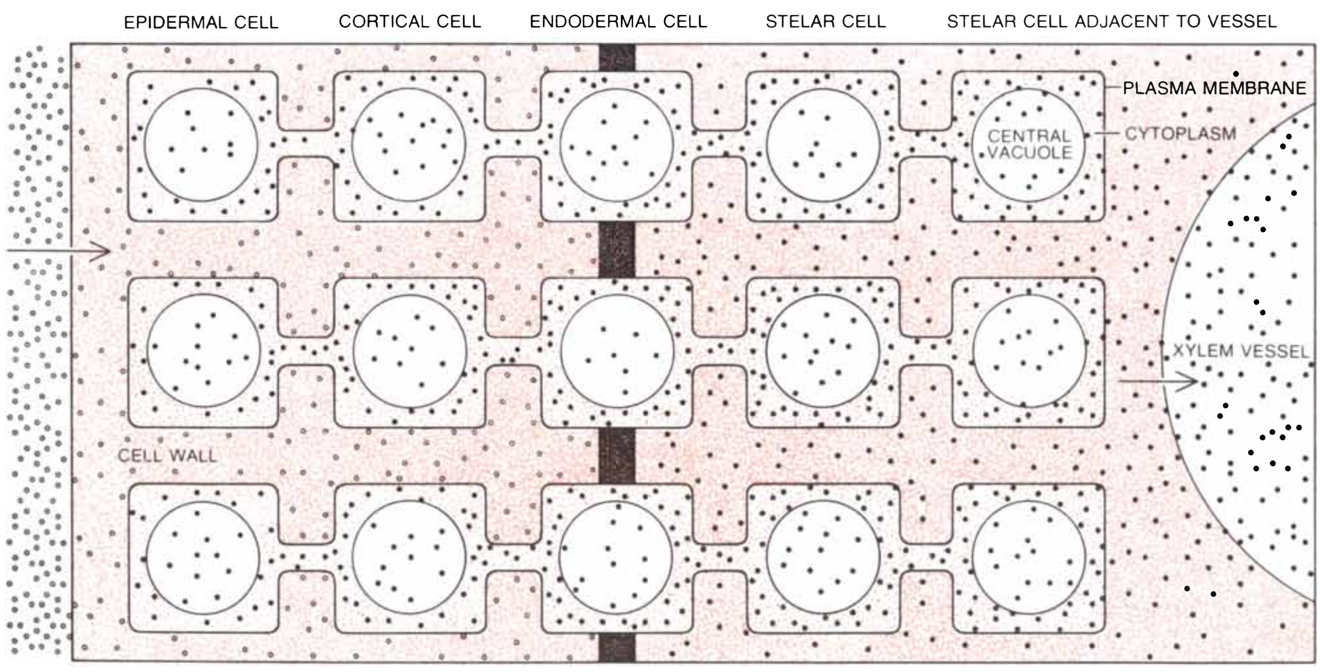


CARRIER MECHANISMS that transport mineral ions across the plasma membrane and into the cytoplasm of plant cells are thought to operate as shown in this schematic diagram. When the carrier is located at the outer surface of the membrane (*1*) its conformation is such that it can form a complex with a nearby ion (*2*). Carrier and ion then migrate across the membrane (*3*). At the inner surface the carrier changes conformation (*4*) and releases the ion (*5*) into the cytoplasm. The carrier then repeats the cycle (*6, 7*).

duction of iron ore in that year was about three-quarters of a billion metric tons. Even when the 1970 totals for the three major nonferrous metals (copper, lead and zinc) are added to the iron-ore tonnage, the worldwide ore output still comes to less than a billion metric tons. In these terms, then, the plant kingdom is outproducing man by better than five to one.

The comparison does not tell the whole story because it does not take into account the difference in the quality of the ores mined by plants and men respectively. Iron, the metallic element mined by man in the greatest quantity, is not economical to extract unless the iron content of the ore is about 30 percent. Potassium is the metallic element that plants extract in the greatest quantity. More often than not it is present in the soil at a concentration of less than 1 percent.

The capacity of plants to capture the energy of sunlight to produce organic matter is often discoursed on in an almost covetous fashion, and the hope is held out that, if only we could duplicate the process, an ample supply of food would be assured for all mankind. But more than energy is needed to sustain life: potassium is needed, phosphorus, zinc, molybdenum and many more mineral nutrients. In the sole ubiquitous source of these elements, the soil, they are only available at very low concentrations. Hence as long as we can now foresee, terrestrial life will depend on those living lacy threads, the roots of plants, and on the integrity of that exceedingly complex system of soil and water wherein roots live.



DIFFUSING IONS in the porous cell walls of the cortex (*open dots*) reach the xylem vessels of the stele in spite of the barrier presented by the Casparian strip. First (*left*), carrier mechanisms transport the ions through the plasma membrane into the cytoplasm of epidermal and cortical cells. The transported ions (*solid dots*) then move from cortex to stele along the many strands of cytoplasm that run from cell to cell, bypassing the Casparian strip. When the ions reach a cell near a xylem vessel (*right*), "reverse" mechanisms are thought to transport ions across the membrane once more. Free to diffuse at random, many ions enter xylem vessel.

Opel: Number one seller in Germany.
Opel Manta Luxus: $2,816.50 in America.

The Opel outsold every other car in Germany in 1972. The plush Manta Luxus is one of the reasons why. For the above price you get all the standard features listed on the price sticker below. Give up a little plush and Manta is $2,622.50. Or go sporty with Manta Rallye—$2,805.50. All prices are Manufacturer's Suggested Retail Prices, including dealer new vehicle preparation charges. Optional equipment, State and local taxes and transportation charges are additional.

Manta. The personal-size sport coupe proven in more than 2 billion miles of actual driving. Manta. With rack and pinion steering. A 1.9 liter engine. Disc brakes up front. A grand-touring ride. Manta. Luxurious. Economical. Comfortable. Really ready for America.

BUICK MOTOR DIVISION
GENERAL MOTORS CORPORATION
OPEL
MANUFACTURER'S SUGGESTED RETAIL PRICE

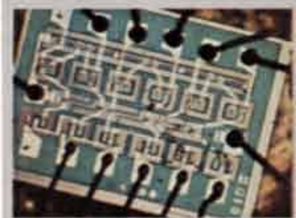
HOMETOWN BUICK, INC.
HOMETOWN, U.S.A.
ARLINGTON, VIRGINIA
OPEL
BALTIMORE, MARYLAND
57L IDR MANTA LUXUS CPE

THE FOLLOWING ITEMS ARE STANDARD EQUIPMENT ON THIS 1973 MODEL OPEL: 1.9 LITER ENGINE DESIGNED FOR NO-LEAD, LOW-LEAD FUEL, WITH HYDRAULIC VALVE LIFTERS AND AUTOMATIC CHOKE • POWER BRAKES WITH FRONT DISCS, SELF ADJUSTING • CARPETED TRUNK COMPARTMENT, PACKAGE SHELF AND FLOOR • CUSTOM CLOTH TRIM WITH VINYL EDGING, AND COLOR COORDINATED INTERIOR • HINGED REAR QUARTER WINDOWS • RUBBER BUMPER GUARDS AND STRIPS • ELECTRIC CLOCK • RALLYE TYPE WHEELS AND STEERING WHEEL • ADJUSTABLE FRONT SEAT BACKS.

TOTAL AMOUNT

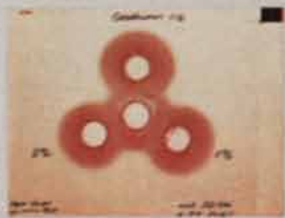
German precision imported by General Motors.
Sold and serviced by 2200 Buick dealers.

·ou find out where you're going.



Integrated circuit X10

Architectural model



Ouchterlony immunodiffusion gel

Creative graphic design

Valuable stamp

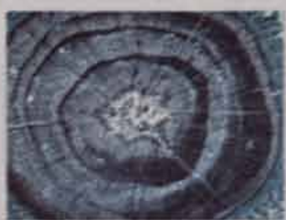
Dental records

Jewelry for catalog

Latent fingerprint

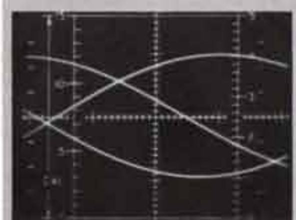
Diatomes X825

Fucus filiformis X150

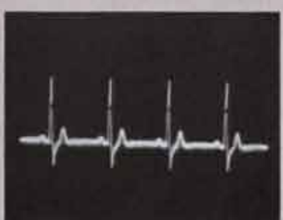


Pine stem X312

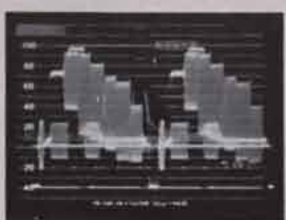
Ant (dark field illumination) X25



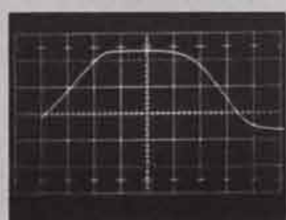
Audio waveform



Electrocardiogram



TV signal waveform



Signal generator waveform

"Polaroid" and "Polacolor"®

THE EVOLUTION OF THE INDIAN OCEAN

A 75-million-year magnetic record in the floor of the Indian Ocean reveals how India was rafted northward for 5,000 kilometers before it collided with Asia. The encounter gave rise to the Himalayas

by D. P. McKenzie and J. G. Sclater

The creation of the Himalayas, the highest and most dramatic mountain range in the world, long presented a puzzle to geologists. Like the Alps and the heavily eroded Appalachians, it was clear that the Himalayas consisted of sedimentary rocks, laid down over many millions of years in shallow seas, then uplifted and heavily deformed by mighty tectonic forces. But in what sea were the Himalayan rocks deposited and how did they get sandwiched between the subcontinent of India and the great Asian landmass to the north? The geology textbooks of a decade ago could provide no satisfactory answer.

With the rapid development and acceptance of the concept of sea-floor spreading, continental drift and plate tectonics the origin of the Himalayas is no longer a mystery. The vast Himalayan range was created when a plate of the earth's crust carrying the landmass of India collided with the plate carrying Asia some 45 million years ago, having traveled 5,000 kilometers nearly due north across the expanse now occupied by the Indian Ocean. Thus the Indian Ocean and the Himalayas have a common origin.

The principal evidence supporting this theory has come from oceanography rather than from classical geology. A geologist's life is too short for him to study a large part of a major mountain system in any detail. Furthermore, the rocks that form the Himalayas have been so strongly deformed and eroded that it is not possible to discover the extent of the relative motion of India and Asia from a study of the rocks themselves.

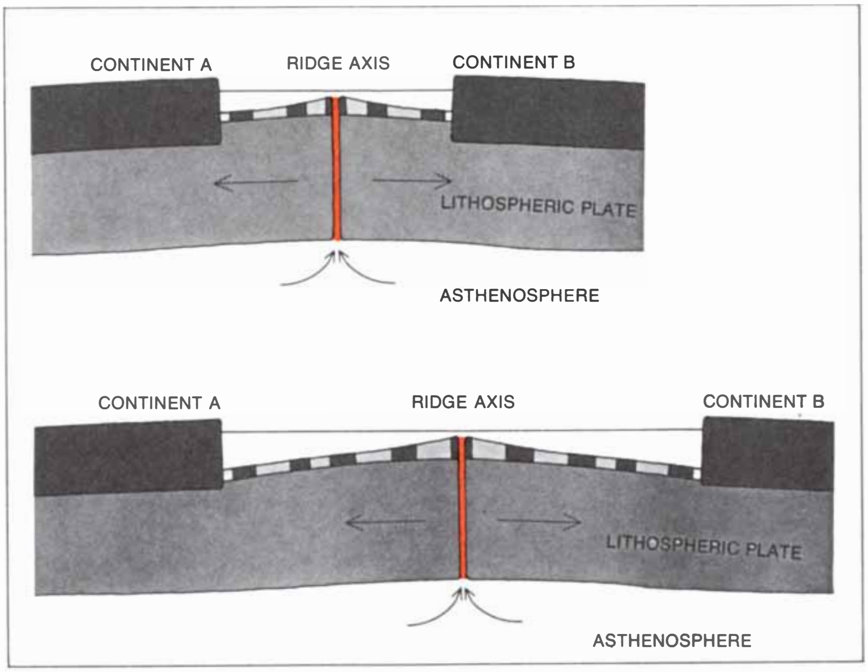
That left the ocean as the only other place to look for evidence of a collision. It seemed unlikely that a large continent could cross an entire ocean basin and leave no trace of its passage. Yet as recently as a decade ago oceanographers who had studied the Indian Ocean saw no evidence indicating the passage of a continent and therefore were mostly opposed to the idea of continental drift. The solution to this difficulty and the beginning of our present understanding was provided about 1960 by the late Harry H. Hess of Princeton University. Hess suggested that the continents do not plow their way through the ocean floor but move with it much like coal on a conveyor belt. He proposed that the sea floor is being continuously generated along linear fissures running for thousands of kilometers through the major ocean basins; molten rock wells up into the fissures from the mantle below and quickly hardens [*see top illustration on next page*]. Elsewhere on the earth the ocean floor plunges into a trench and then sinks to great depths in the mantle.

The difference between the concepts of sea-floor spreading and plate tectonics is mostly one of emphasis. Sea-floor spreading is primarily concerned with the processes by which the ocean floor is created and destroyed, and much less with the continents and the passive parts of the ocean floor. The inactive areas, on the other hand, are the principal subject matter of plate tectonics, which starts from the idea that all areas of the earth's surface free from earthquakes are not at present being deformed. The concept is useful because most earthquakes occur within narrow belts that delineate the regions where large plates of the earth's crust are being subjected to irresistible tectonic forces. It turns out that most of the earth's surface can be divided up into only a few large, rigid plates.

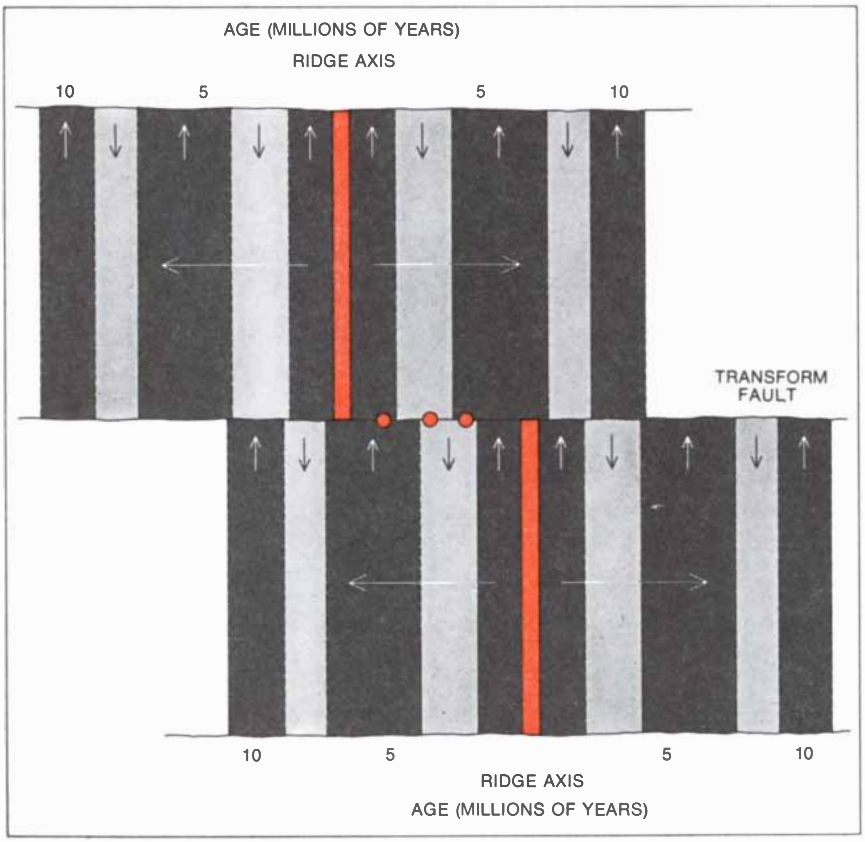
Since the plates are rigid, one can work out the relative motions everywhere of any two plates simply by knowing how they move at any two points along a common boundary. One must emphasize "relative motion" because there is as yet no method for defining absolute motion. The simplest type of motion on a plane is when two plates slide past each other with no relative rotation. A more general type of motion also involves rotation, and it must then be everywhere parallel to circles centered on the center of rotation. Wherever plates are moving apart, the separating edges must be growing by the upwelling of material from below. Such accretions almost always add to each retreating plate at the same rate, and commonly they form boundaries known as ridges. Because fresh material usually adds to the two separating plates at the same rate, ridges normally spread symmetrically. The reason for this symmetric behavior is not fully understood. Furthermore, there is no geometric requirement that the ridge should form at right angles to the direction of motion of the plates, but that is usually the case.

When two plates simply slide past

HIMALAYAN RANGE looking to the west was photographed from *Apollo 7* in October, 1968. Nepal is at the left of the range; China is at the right. Mount Everest is near the S-shaped lake near the middle of the picture. The river at the left is the Ganges. The Himalayas were created 45 million years ago when India collided with the underside of Asia.



BASIC MECHANISM OF CONTINENTAL DRIFT involves a rift in the ocean floor that allows molten rock to well up from the asthenosphere and form a spreading ridge. The continents are rafted apart at rates up to 20 centimeters per year. As the molten rock solidifies at the ridge axis it becomes magnetized in stripes of alternating polarity (*light and dark bands*) as a consequence of intermittent reversals in the earth's magnetic field.

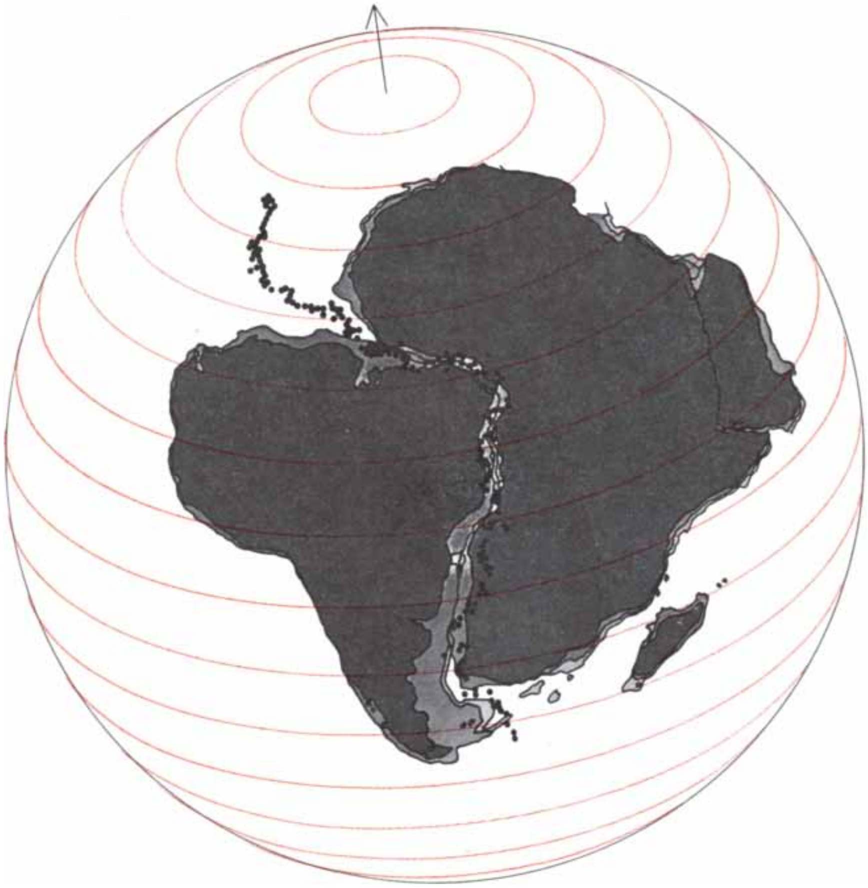


REVERSALS OF EARTH'S MAGNETIC FIELD produce a zebra-striped pattern of magnetization in the crust of the ocean floor. The reversals (*small arrows*) have been dated back about 80 million years by combining radioactivity measurements of rocks with identification of organisms in deep-sea cores. Transform faults often offset the pattern of magnetic anomalies. Friction between the sliding blocks causes shallow earthquakes (*colored dots*).

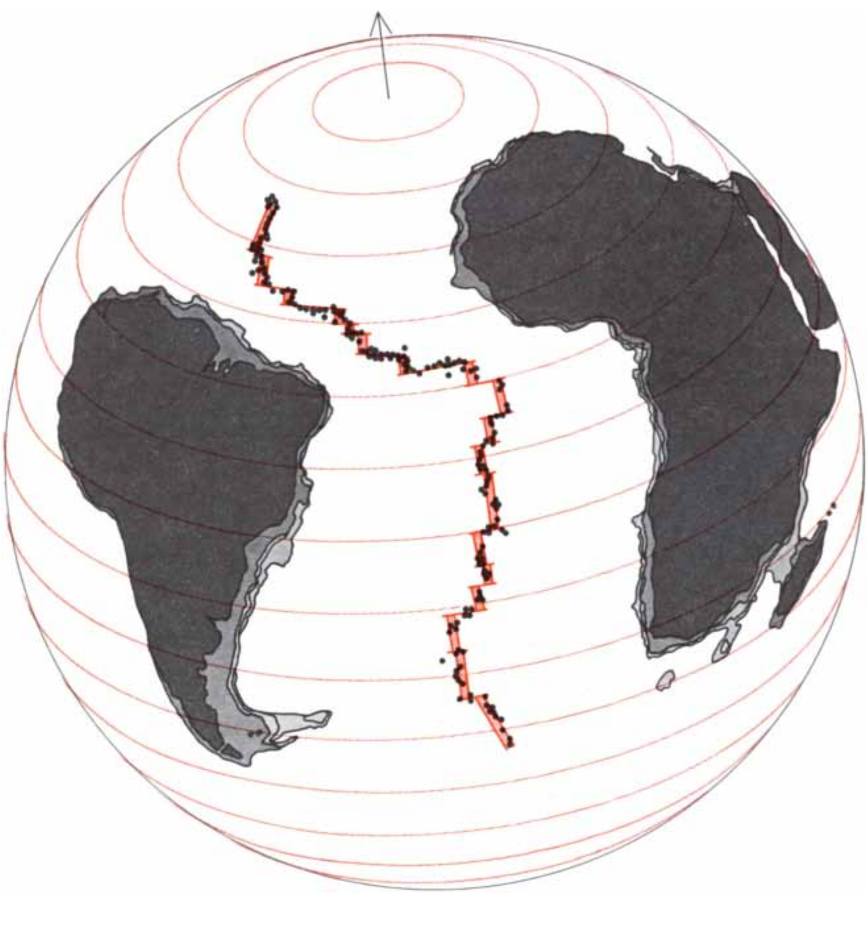
each other, nothing is added to or subtracted from either plate. Plate boundaries of this type, known as transform faults, are always parallel to the direction of relative motion between the two plates. Such faults must therefore be segments of circles around the center of rotation. When two plates are moving toward each other, one eventually underrides the other and is consumed by plunging into a trench. Simultaneously the edge of the overriding plate is typically uplifted, creating a belt of crumpled and folded rocks parallel to the trench. There is, however, no simple relation between the downward direction of a trench and the relative direction of motion between the plates.

The motion of plates on the earth is slightly more complicated because the earth is spherical. Their relative motion, however, can still be described by a rotation about an axis. Since the plates must remain on the earth's surface, this axis must pass through the earth's center. The motion must again be parallel to circles about the rotation axis. The axis for the opening of the South Atlantic is shown in the illustrations on the opposite page, together with circles drawn with the axis as the center. The earthquake epicenters mark the plate boundary, which is in places a ridge and in places a transform fault parallel to the direction of relative motion marked by the small circles. The original relative position of Africa and South America can be obtained by rotating the two continents toward each other until they meet. The earthquakes have not been moved, but each continent has been rotated through an equal angle toward the ridge axis.

The reason the earthquakes then mark the original break is that the ridge in the South Atlantic has added to the African and South American plates at the same rate. South America and Africa can thus be fitted together rather accurately by a simple rigid rotation. Neither has been deformed during the formation of the South Atlantic, and since both continents consist largely of very ancient rocks, neither has been appreciably deformed in the past 300 million years. Together they form a spherical cap whose diameter is about equal to the radius of the earth. If the earth had expanded appreciably in the past 600 million years, both continents would have been broken into many fragments, since the spherical cap they originally formed would no longer fit on an expanded earth. Plate tectonics therefore shows



AFRICA AND SOUTH AMERICA were locked together as part of the primitive continent of Gondwanaland until about 180 million years ago. The black dots mark the positions of earthquakes that have occurred on the Atlantic mid-ocean ridge in the past 10 years. Their location coincides with the location of the primitive cleavage line between South America and Africa when the two continents are rotated toward each other by an equal amount along the "latitude" lines centered on the axis of rotation (*arrow*). India, Antarctica and Australia were also part of Gondwanaland, but there is still no agreement on how several pieces fitted together.

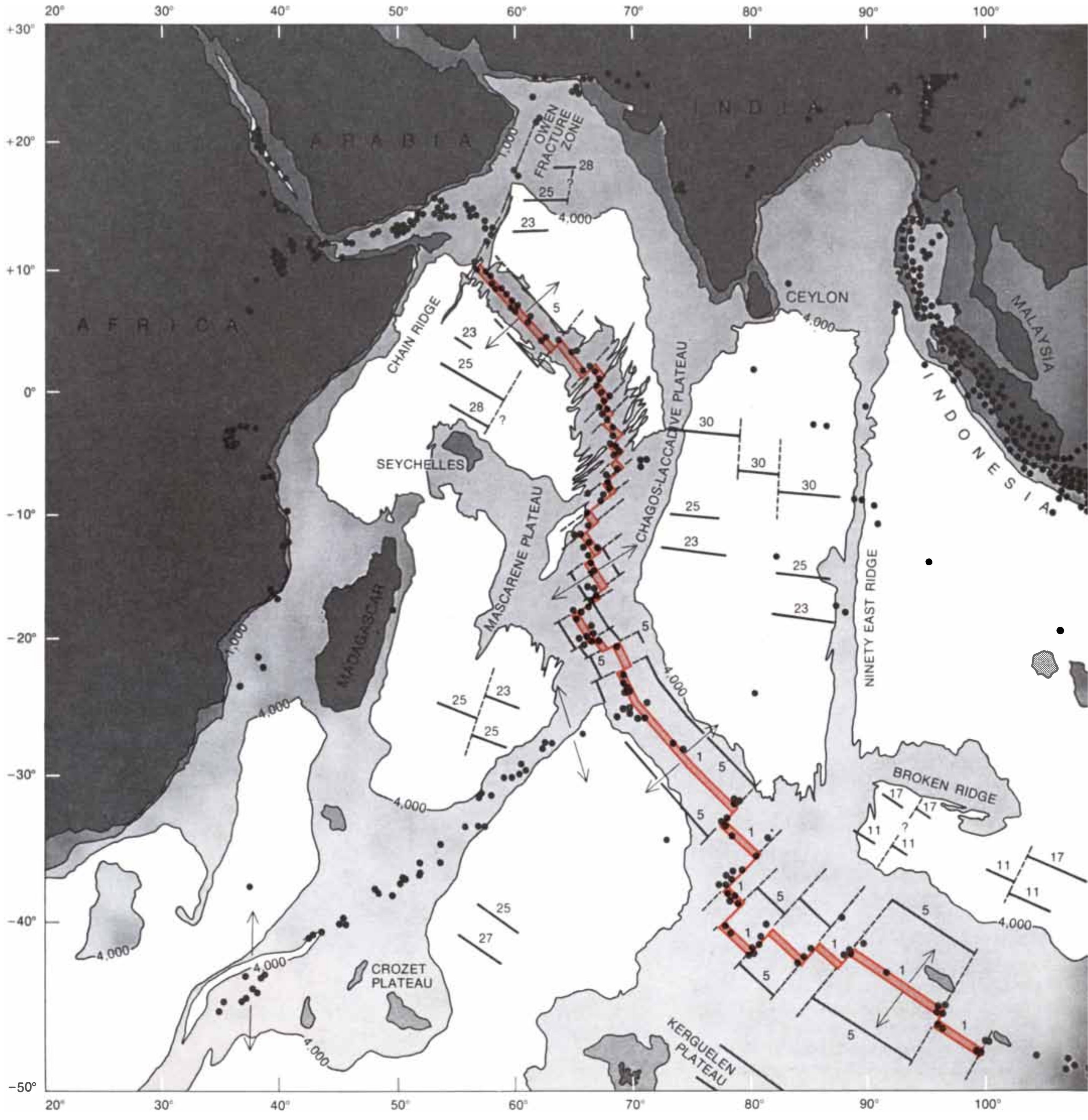


PRESENT LOCATION of Africa and South America shows the two continents to be equidistant from the still-active spreading ridge (*color*) in the mid-Atlantic. Evidently molten rock flowing into the ridge has increased the size of each plate by an equal amount for the past 180 million years. The two continents are still being pushed apart at the rate of about three centimeters a year. The transform faults developed because the initial break was not everywhere at right angles to the circles of rotation. As required by the theory of plate tectonics, the transform faults lie parallel to the circles.

that geologically important expansion has not occurred, although it does not rule out changes in radius of a few tens of kilometers.

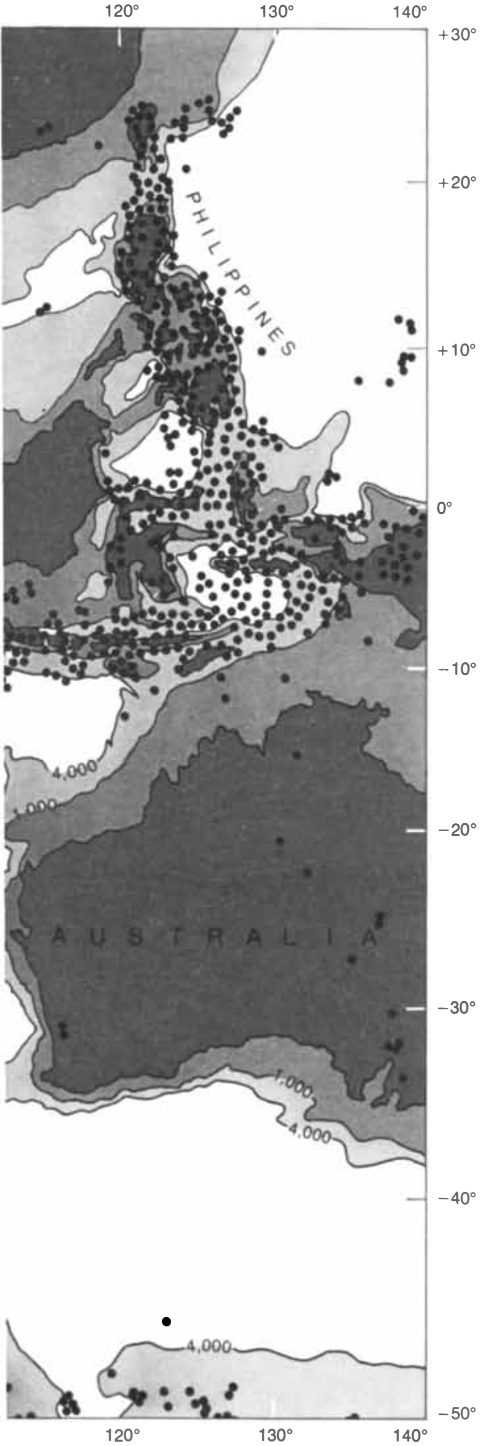
The South Atlantic also shows how transform faults can originate. The original break between South America and Africa was a jagged line that followed old lines of weakness in the Precambrian shield and did not form a plate boundary at right angles to the direction of motion between the plates. After separation had started, the plate boundary changed from a jagged ridge to a series of straight ridge sections joined by transform faults. This change maintained the general shape of the ridge but allowed the ridges to be at right angles to the direction of relative motion.

To a continental geologist it is quite surprising that the simple ideas of plate



MAGNETIC ANOMALIES IN INDIAN OCEAN are represented by straight lines. The numbers on the lines start with No. 1 on the present ridge axis (*two closely spaced lines*) and extend to No. 33, which identifies molten rock that solidified about 80 million years ago. Earthquakes recorded over a recent seven-year period are indicated by black dots. They can be seen to cluster along the present ridge axis and on lines where the axis is offset by transform faults. The greatest number of earthquakes, however, take place around the island arc between Southeast Asia and Australia where the Indian Ocean plate and the Pacific Ocean plate plunge into deep trenches and are consumed. The major features on the floor of the Indian Ocean are the three large ridges that run north and

tectonics work so well. The rocks he studies were generally severely deformed by plastic flow when they were near a plate boundary. Thus there was opposition to the new ideas from investigators who had worked only on land. That opposition has now largely ceased.



south: Ninety East Ridge, the Chagos-Laccadive Plateau and the Mascarene Plateau. Two parallel ridges, Broken Ridge and the Kerguelen Plateau, run approximately east-west near Australia. Such features aided in reconstructing evolution of Indian Ocean.

The early tests of plate tectonics all involved the present motions of plates measured at plate boundaries as defined by the location of earthquakes. This approach is obviously powerless to determine how plates may have moved in the geologic past. Providentially the rocks of the ocean floor carry a permanent record of their own history. It was noted about 1960 that the oceanic crust is not uniformly magnetized. Instead magnetometer surveys revealed a remarkably regular striped pattern of magnetic variation: stripes of normal polarization (with north pointing to the magnetic north pole) alternate with stripes of reverse polarization.

In 1963 F. J. Vine and D. H. Matthews of the University of Cambridge proposed that when lava wells up in the oceanic ridges, it is magnetized as it cools by the magnetic field of the earth. The alternations of polarity indicate that the magnetic field of the earth reverses at infrequent and irregular intervals. As a result the entire ocean floor is covered with long linear blocks of alternating magnetic polarity [*see bottom illustration on page 64*]. The shape of the blocks is exactly the shape of the ridge that formed them. We are doubly fortunate: the period between reversals of the magnetic field is generally sufficient to produce stripes at least 10 kilometers wide and, equally important, the reversals come at irregular intervals, so that the stripes vary in width. Hence the magnetic anomalies can be matched where they have been shifted by transform faults.

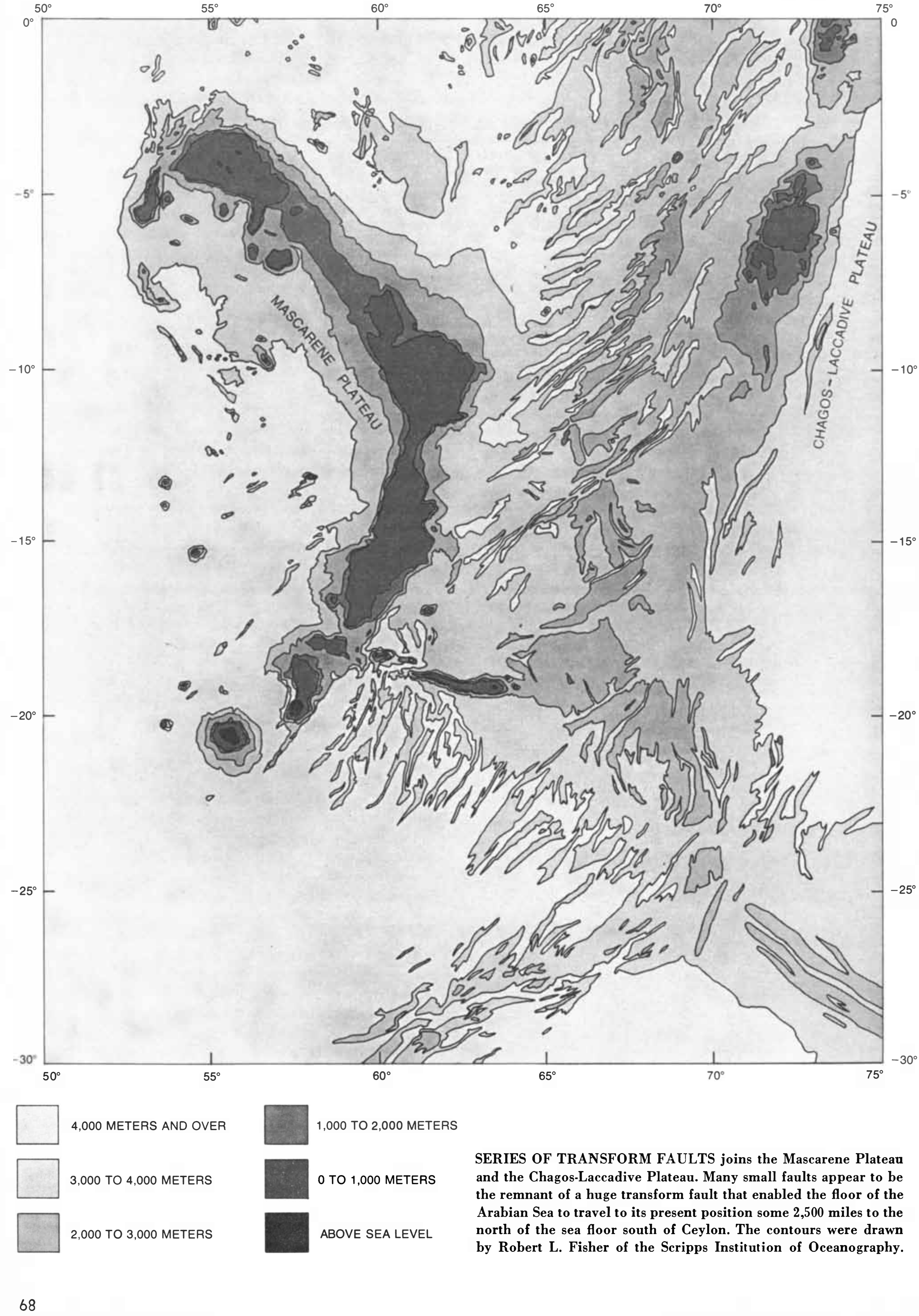
If the present ridge axis is designated No. 1, the magnetic stripes that flank it on each side can be numbered sequentially up to No. 33, formed about 80 million years ago. Older anomalies are known, but their age and relation to the most recent 33 are not yet established. Magnetic surveys can be made with a magnetometer towed behind either a ship or an airplane. The mapping and identification of magnetic anomalies now provides a fast and accurate method of studying the evolution of an ocean.

In regions where there are no magnetic data or where the anomalies are not clear the depth of the sea can often provide a rough estimate of the age of the sea floor. The youngest portion of a plate forms the peak of a ridge system that rises about 3,000 meters above the flatter and older portions of the sea floor. The reason is that the recently intruded rock is hotter and therefore less dense than the older material that has cooled and contracted as it moved away from the plate boundary. Sea-floor profiles calculated on the basis of rock-cooling models agree well with profiles established by accurate depth surveys.

The separation rate recorded by the anomalies indicates the relative motion between two plates. It is also possible to measure the motion between one plate and the earth's magnetic pole. If the magnetic pole is assumed to coincide approximately with the rotational pole, the measurement of plate motion can be used to find the latitude and orientation of the plate on which the observations were made. If the rocks are magnetized near the Equator, the magnetization is horizontal; the closer the rocks are to the poles when they are magnetized, the more vertical is the pitch, or dip, of the magnetization. With many paleomagnetic measurements made over the past two decades one can plot the apparent migration of the pole going back tens of millions of years. One cannot decide, of course, whether the pole itself has moved or whether the pole has remained fixed while the plates have moved. For convenience we take the poles as being fixed and move the plates to a latitude corresponding to the pitch of their magnetization. Naturally any reconstruction of the positions of the plates must be consistent with respect both to latitude and to north-south direction of polarization.

When we started our work on the Indian Ocean, the theory of plate tectonics had been proposed and tested but its success was not yet generally recognized. Like many scientific projects, ours started by chance. When we began in 1968, the Indian Ocean was the only major ocean in which no one was trying very hard to use marine geological and geophysical data to construct a tectonic history. It happened, however, that Robert L. Fisher of the Scripps Institution of Oceanography was doing detailed work on the Central Indian Ridge and had organized an expedition that would spend six months collecting marine-geological and geophysical information in the Indian Ocean. One of us (Sclater) was chief scientist on that expedition for two months and proposed that the other (McKenzie) join the ship for a month, more to show a theoretical geophysicist how cruises are run than with the intention of putting him to work.

The magnetic anomalies observed on that cruise left little doubt in our minds that the plate carrying India had crossed the Indian Ocean. To see if we could strengthen the hypothesis we collected all the available magnetic profiles from ship and airplane traverses of the ocean.



SERIES OF TRANSFORM FAULTS joins the Mascarene Plateau and the Chagos-Laccadive Plateau. Many small faults appear to be the remnant of a huge transform fault that enabled the floor of the Arabian Sea to travel to its present position some 2,500 miles to the north of the sea floor south of Ceylon. The contours were drawn by Robert L. Fisher of the Scripps Institution of Oceanography.

Fortunately the theory of plate tectonics changed very little in the three years that the two of us and three co-workers spent interpreting the observations. When we were done, the evolution of the Indian Ocean emerged clearly.

The shapes of the magnetic anomalies show that the evolution of the Indian Ocean was much more complicated than the evolution of the South Atlantic. The topography of the floor of the Indian Ocean is correspondingly more complex. The floor shows four large north-south ridges, two of which are still actively spreading [*see illustration on pages 66 and 67*]. It is hard to see any relation between the two huge inactive ridges (Ninety East Ridge and the Chagos-Laccadive Plateau) and the ridges that have remained active. It is also difficult to understand how the various ancient continental fragments, such as the island of Madagascar, the Seychelles islands and perhaps also Broken Ridge west of Australia, became isolated from the surrounding landmasses. Islands and ridges of this type are not found in the Atlantic and the Pacific. Their existence in the Indian Ocean is baffling.

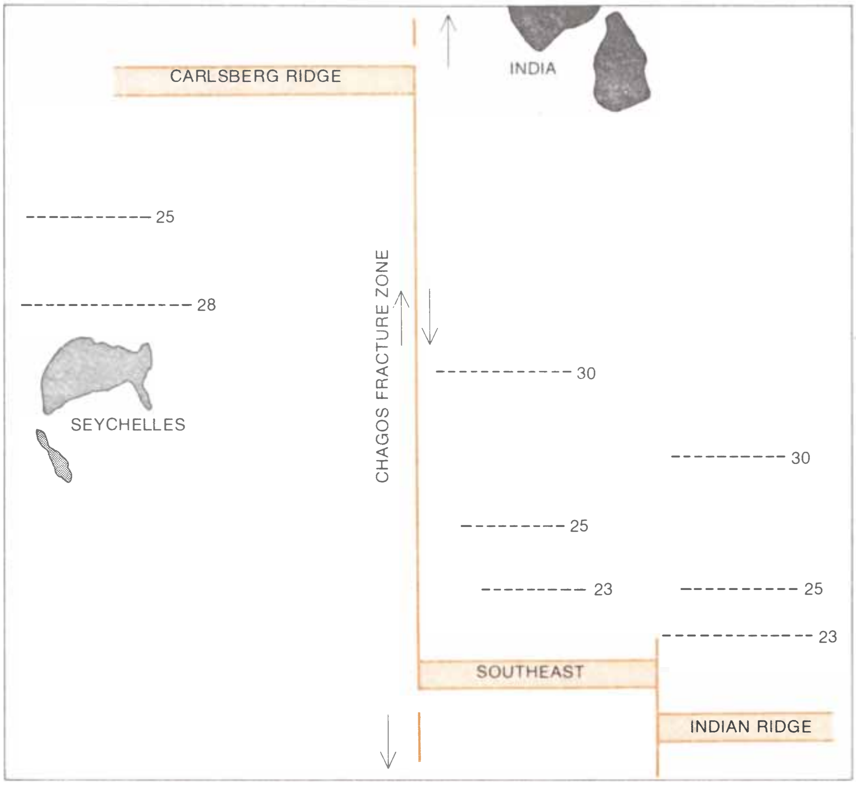
A feature of particular interest is the plate boundary that runs north and south between the Chagos-Laccadive Plateau and the Mascarene Plateau in the northwestern part of the Indian Ocean. In this area Fisher's work has shown that the short spreading ridge segments are offset by many transform faults, none of which crosses the two inactive ridges on each side [*see illustration on opposite page*]. These faults therefore cannot have been produced by the shape of the original break, as the faults in the South Atlantic were. Their origin becomes clear only when the anomalies between anomaly No. 23 and anomaly No. 29 are used to map the shape of the ridge axis that produced them.

The anomalies in the Arabian Sea south of Iran run east and west but are about 2,500 kilometers north of the similarly numbered anomalies south of Ceylon. Hence there must at one time have been a huge transform fault joining the ridge in the Arabian Sea to the one south of Ceylon [*see illustration on this page*]. Some 55 million years ago relative motion south of Ceylon slowed from 16 centimeters per year to about six centimeters and probably stopped altogether in the Arabian Sea. Then about 35 million years ago movement resumed in the Arabian Sea, but its direction was no longer parallel to the ancient transform fault. New sea floor was again generated, and the boundary changed from a ridge oblique to the spreading direction to a series of ridge segments that are perpendicular to the direction of motion and are joined by transform faults [*see illustrations on next page*].

As in the Atlantic, the shape of the present plate boundary preserves the shape of the ancient boundary and the north-south section is the image of the old transform fault preserved by the new structures. This region and the central Atlantic thus show how the complex geometry of plate boundaries often results from the simple behavior of ridges and transform faults. To be sure, not all ridges behave so simply. In some cases a plate breaks and forms a ridge in a new place. Such "jumping" ridges are fortunately rare. Since they produce very muddled patterns of anomalies, they are difficult to identify.

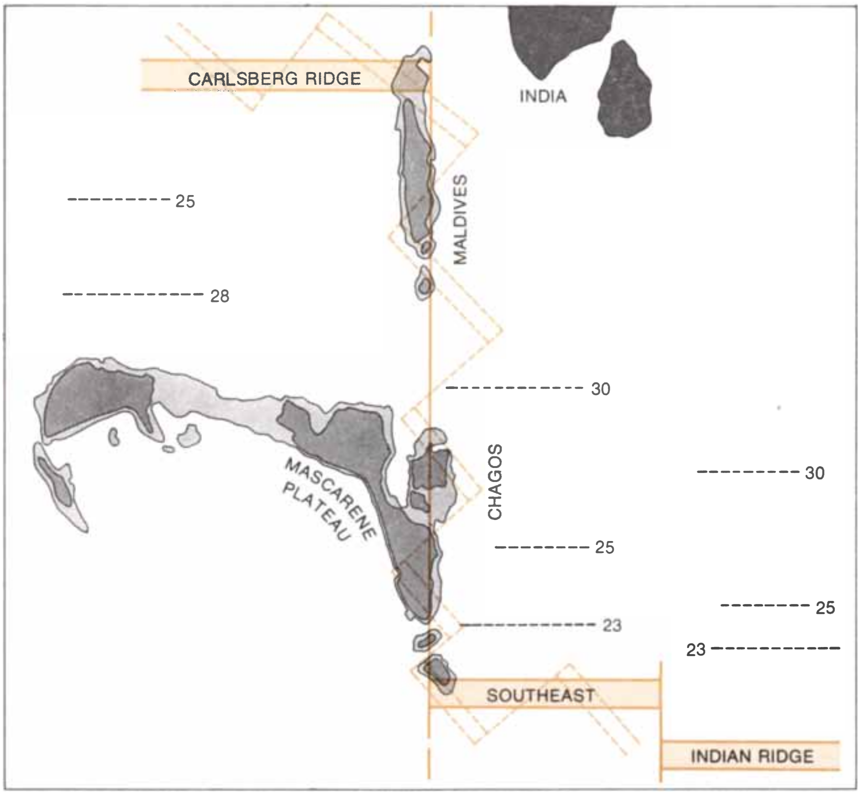
A second area of much interest is the region south of Ceylon. Here we found anomalies from anomaly No. 32 to anomaly No. 21 that were formed between 75 million and 55 million years ago. These anomalies are now just south of the Equator, but the matching anomalies on the other side of the ridge are now in the far southwest Indian Ocean, northeast of the Crozet Islands. The two segments of anomaly No. 32 are now separated by about 5,000 kilometers. India must therefore have traveled this distance at the rate of 7.5 centimeters per year after breaking away from Antarctica 75 million years ago.

The separation of the anomalies between anomaly No. 30 and anomaly No. 22 shows that the separation rate during this period was more than twice as rapid: 16 centimeters per year. At that rate a plate would be carried completely around the world in only 250 million years, which is a very short span of time by geological standards. Even faster rates, however, have been measured on the present spreading ridge in the southeast Pacific, where the plates are moving apart at nearly 20 centimeters per year.

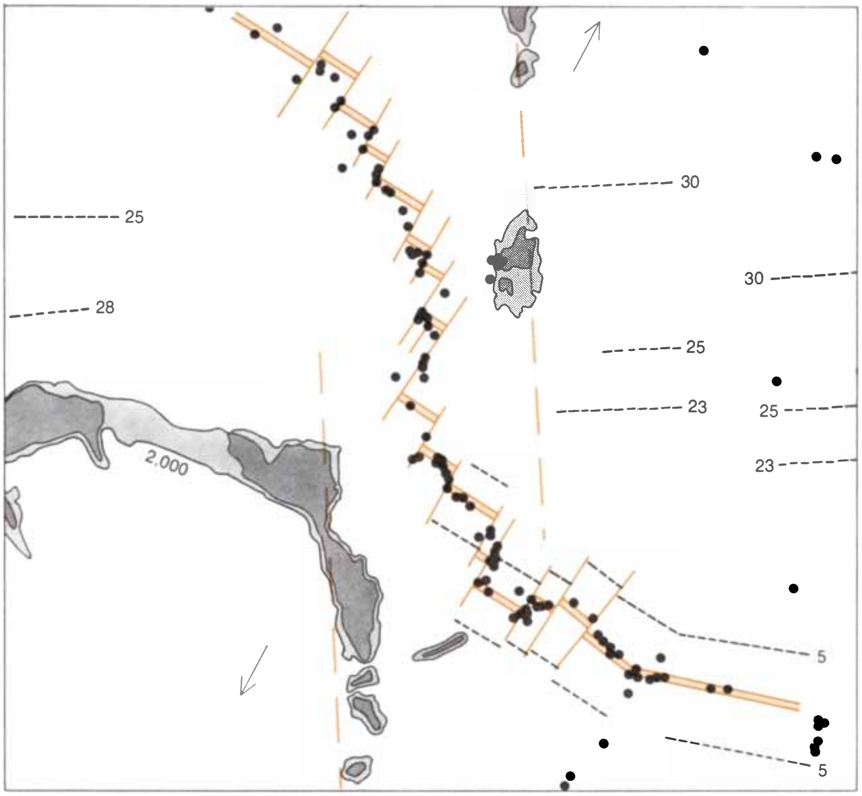


FIFTY-FIVE MILLION YEARS AGO a huge transform fault, the Chagos fracture zone, divided the floor of the Indian Ocean into two large segments. Between 80 million and 55 million years ago the segment carrying India traveled about 16 centimeters a year.

We can now use the magnetic lineations to reconstruct the relative positions of the different plates underlying the Indian Ocean. We can also employ the paleomagnetic observations from the neighboring continent of Australia to place all the plates in their correct position with respect to the magnetic poles, which we believe coincide approximately with the earth's rotational poles. The motion of each plate can be described by a rotation, so that the complete reconstruction process consists of a series



THIRTY-FIVE MILLION YEARS AGO, after a 20-million-year period during which India's northward drift slowed significantly, the single north-south fault of the Chagos fracture zone evidently broke up into a number of ridge segments connected by short faults.



WHEN NORTHWARD MOTION RESUMED about 35 million years ago, new sea floor was extruded along ridges that cut obliquely across the line of the ancient Chagos fault. Thus the direction of sea-floor spreading was rotated about 45 degrees from the original direction. This map repeats a central portion of the ridge-and-fault structure illustrated on page 66.

of rotations. Unlike the single rotation that suffices to close the South Atlantic, however, the closing of the Indian Ocean requires the rotation of several plates that have moved at varying rates and in different directions since India started its rapid movement northward.

The earliest reconstruction we can obtain is governed by the oldest anomaly we can identify: anomaly No. 32, formed about 75 million years ago. At that time the ridge south of Ceylon was spreading in such a way as to propel India generally a little east of north, and Ninety East Ridge was an active transform plate, facilitating the northward slippage of the Indian plate [*see illustration on opposite page*]. Since Asia was evidently more or less stationary, or at least not moving north nearly as fast as India was approaching, the northern edge of the Indian plate must have been consumed in a trench lying somewhere to the south of Asia, but little is known about the position or history of that trench. Australia was then still joined to Antarctica, forming a single continent; Africa and South America had already separated.

During the next 20 million years India moved rapidly northward between two great transform faults: Ninety East on the east and a fault now called the Owen fracture zone on the west. Some 55 million years ago the relative movement between plates slowed down or stopped. Only the ridge between India and Antarctica continued to produce new plate, and the rate of separation dropped from 16 centimeters per year to six centimeters or less. Forty-five million years ago there was a rupture between Australia and Antarctica, setting Australia free for its northward migration. Soon afterward all the plate boundaries in the Indian Ocean became active again, but they were now moving in new directions [*see illustration on page 72*].

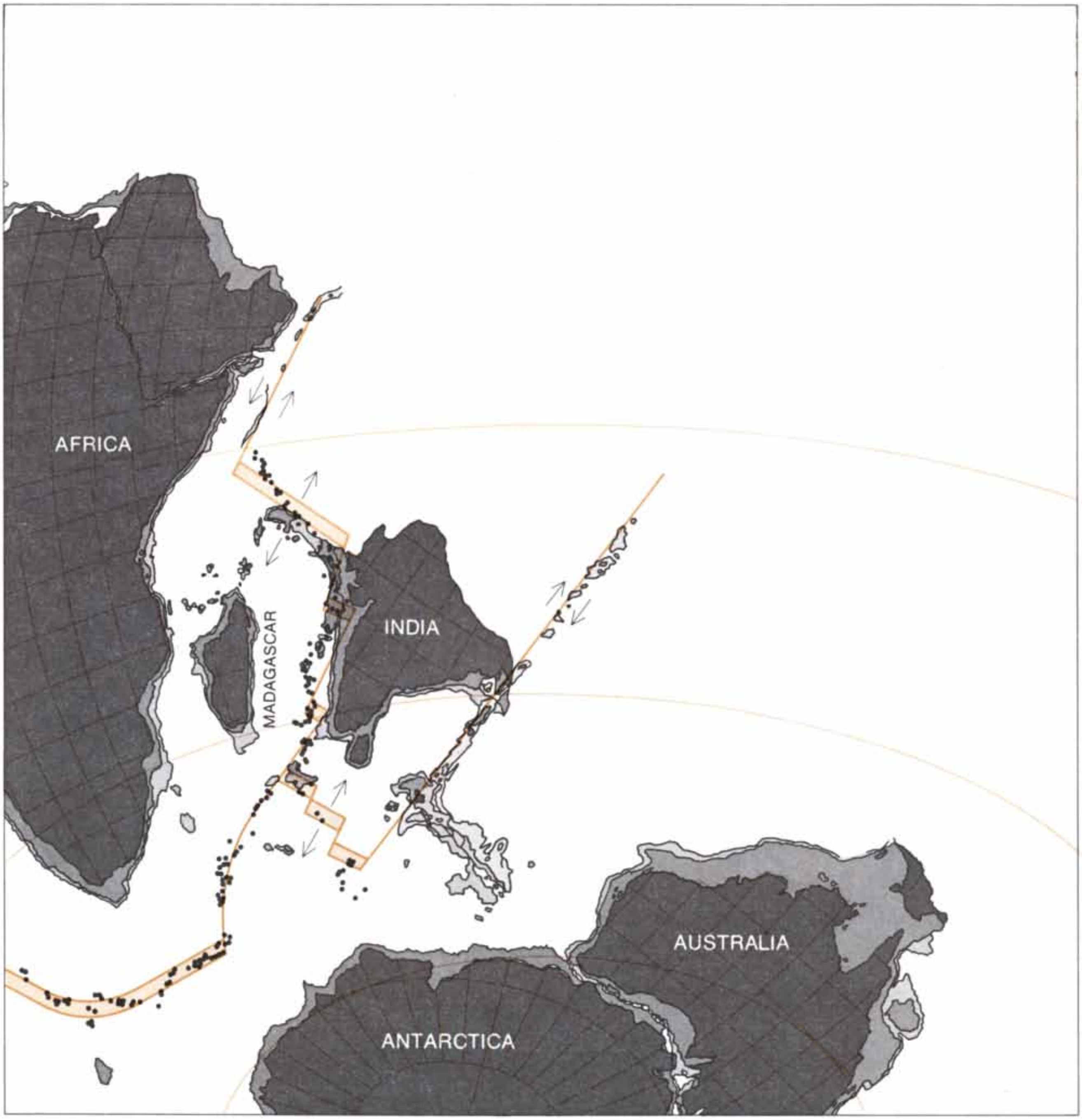
The result was the complicated pattern of ridge segments and transform faults that is still active in the central part of the western Indian Ocean. Motion on the Ninety East plate boundary ceased about 45 million years ago. The most recent events in the history of the Indian Ocean have been the formation of the Red Sea and the Gulf of Aden when Arabia separated from Africa, probably about 20 million years ago.

Beginning about 45 million years ago the northern edge of the Indian plate must have begun crumpling up the many layers of shallow-water sediments, known as a geosyncline, laid down over millions of years on the continental shelf that bordered the southern edge of Asia.

The result was the upthrusting of the Himalayas. The application of plate-tectonic concepts to both the evolution of the sea floor and the building of mountains is encouraging. It should eventually be possible to tie the two phenomena together by relating the composition and historical movement of rocks in the mountain ranges to records left in the sea floor.

Many attempts have been made to guess precisely how South America, Africa, India, Antarctica and Australia were once joined to form the primitive continent known as Gondwanaland. There is as yet no general agreement as to how this should be done. The fit between South America and Africa, as is well known, is excellent. The fit between Australia and Antarctica is good. The arrangement of all five major units, however, is controversial, and the original position of Madagascar is unknown.

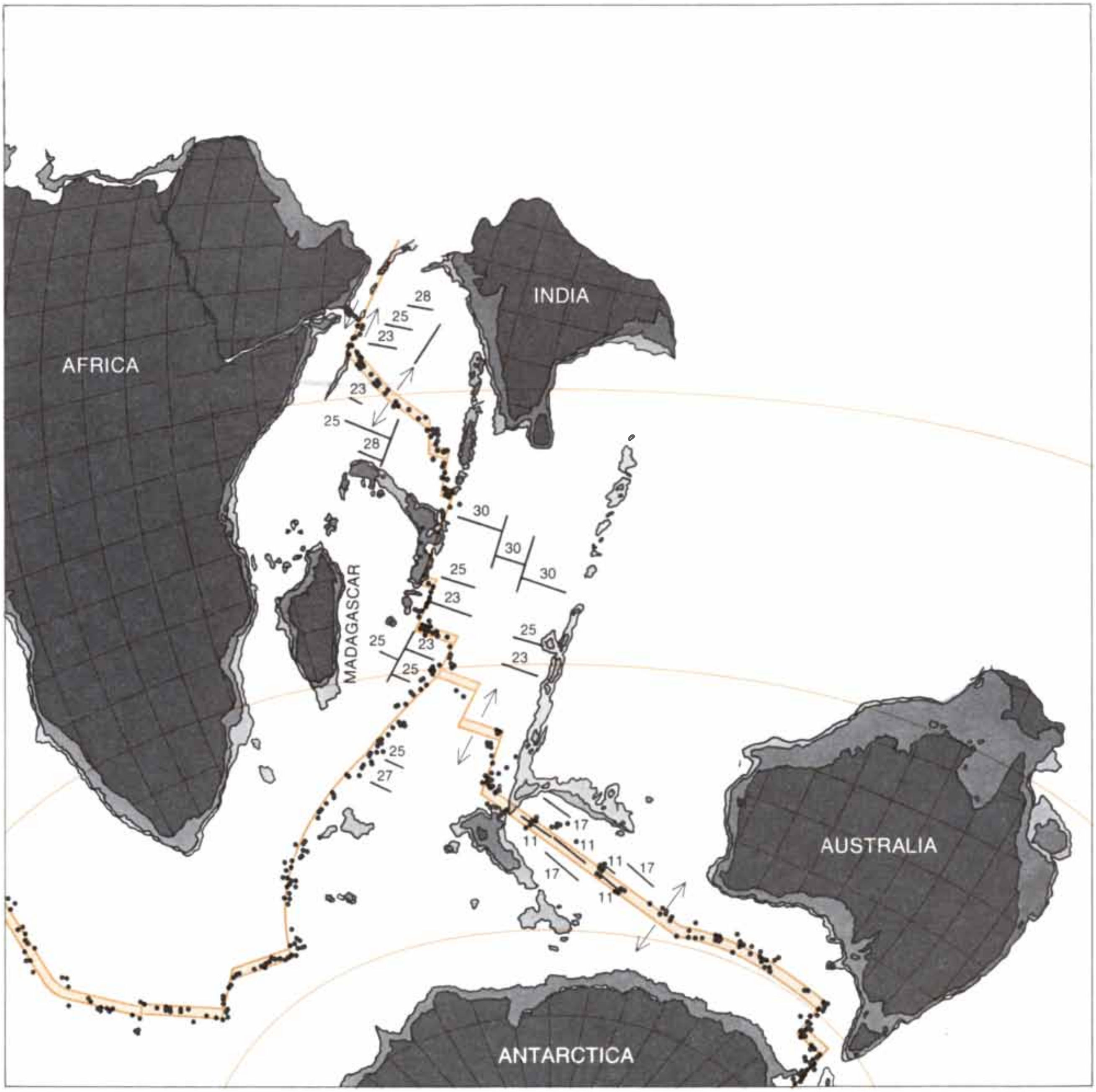
The principal difficulty is that no magnetic lineations have yet been discovered on the older parts of the floor of the Indian Ocean between the continents. We therefore cannot continue to reassemble the continents by the same methods we have used to trace the movement of



SEVENTY-FIVE MILLION YEARS AGO, after breaking away from the ancient primitive continent named Gondwanaland, the plate carrying India started to move rapidly toward the northeast. The motion was accompanied by the generation of new plate along the ridges indicated by pairs of parallel lines. Large transform faults east and west of the connected series of ridges enabled the Indian plate to move unimpeded. The original fit of India, Africa and Madagascar with Antarctica-Australia is not yet established.

India during the past 75 million years. There are also no other structures like Ninety East Ridge, which was recognized as a transform fault even before the magnetic lineations were mapped. Fortunately the area of sea floor in which the record presumably lies hidden is not great. Last year a series of deep holes were drilled in the floor of the Indian Ocean by the drilling vessel *Glomar Challenger*. The data from these holes have confirmed and amplified our reconstruction of the history of the ocean. They have also added to the evidence needed to reconstruct Gondwanaland.

Meanwhile one can speculate about the original juxtapositions of India, Antarctica and Australia. One guess is that existing reconstructions are wrong because they have attempted to remove practically every piece of sea floor between the continents. That approach has been favored because all the continents believed to have formed Gondwanaland show evidence of having been covered by a huge ice cap some 270 million years ago. We know from the recent glaciation in the Northern Hemisphere, however, that continental ice caps can simultaneously cover landmasses that are separated by oceans. It may be that a small ocean basin, comparable perhaps to the Arctic Ocean, was nestled somewhere among the southern landmasses 270 million years ago. It may be our ignorance of its existence and shape that is preventing the successful reconstruction of Gondwanaland.



THIRTY-FIVE MILLION YEARS AGO India had been carried some 4,000 kilometers to the north at rates that evidently varied from 16 centimeters per year to seven centimeters or less. It was about this time that the direction of relative motion among several of the major plates changed quite drastically to produce the complicated sea-floor geometry we see today. Australia has been set adrift from Antarctica and has started on its way to its present position. At this point Arabia has still not split off from Africa.

Premise:
Pollution control begins with effective sensing and measurement.

Honeywell offers a practical way to detect harmful chemical vapors.

The job of cleaning up air pollution involves far more than control of smokestacks and auto engines. Many industrial processes are the source of other potentially toxic gases and vapors.

For example, the use of paint products results in concentrations of vapors such as toluene or benzene. Degreasing operations produce vapors of other hydrocarbons. Dry cleaning and refrigeration plants must control perchloroethylene and ammonia.

Charcoal filters, after-burners or scrubbers neutralize such pollutants. However, an effective control system requires dependable sensors to detect harmful vapor levels and provide for alarm control functions. Thus far, most common methods of sensing and measuring have been less than satisfactory or too costly.

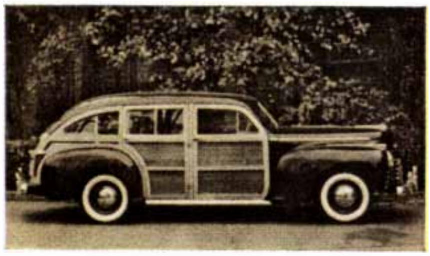
Honeywell technology has developed simple, reliable detectors for many noxious gases and vapors. One such instrument, using the principle of ultra-violet absorption, can be calibrated to sense such pollutants as ammonia, benzene, hydrogen sulfide, perchloroethylene, trichloroethylene and sulfur dioxide. For a wide range of applications, Honeywell sensors can provide an economical, practical means of monitoring industrial atmosphere.

Honeywell fully recognizes that pollution control is a difficult and complex task. We'd like to help by sharing our specialized knowledge in the field of sensing and measurement. For details on the subject above, or other information, please write Mr. Richard Boyle, Director of Environmental Services, Honeywell, 2701 Fourth Ave. So., Minneapolis, Minn. 55408.

Honeywell

The Chrysler Town & Country Wagon was created to be the most luxurious wagon in the world.

That idea has never changed.



The original Chrysler Town & Country wagon was introduced in 1941. It was one of the finest wagons of its day, offering such luxuries as deep-cut pile carpeting in the rear seat, an automatic transmission, even an electric clock and directional signals, all as standard equipment.

Beyond the shadow of a doubt, it was the most luxurious wagon money could buy.

Today's 1973 Town & Country follows this same tradition of luxury. It's big. And comfortable. With such standard features as a powerful 440 cubic inch, four-barrel V-8, TorqueFlite automatic transmission, power steering, power front disc brakes, and a three-way tailgate.

In front, there is a split-bench seat with a center armrest and passenger recliner. All of the seats are upholstered with rich vinyl, and cushioned with thick foam padding.

The floor is covered wall-to-wall with deep-loop pile nylon carpeting. Including the entire cargo area. And the Chrysler Town & Country has more room for cargo than most other station wagons in the world.

If this sounds like a wagon that comes up to your standards, see your Chrysler-Plymouth dealer for a test drive.

You'll be surprised at how easy it is to own the worlds most luxurious wagon for a little more than the price of an ordinary wagon.

CHRYSLER
Plymouth

Chrysler Town & Country

Extra care in engineering...it makes a difference.

CONFIRMATION

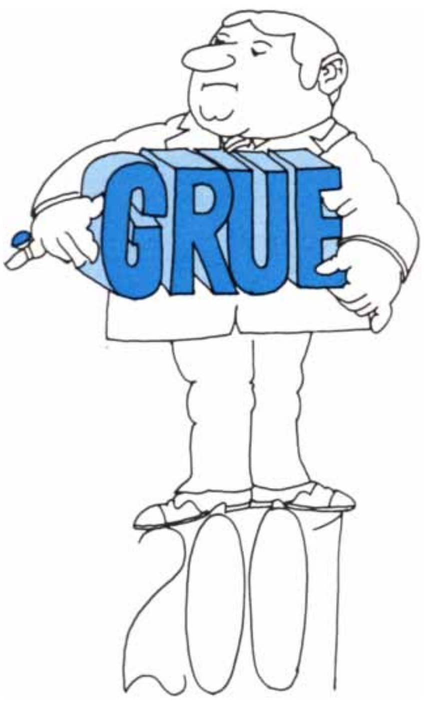
For centuries investigators have been testing, confirming and disconfirming hypotheses by observation and experimentation. Yet the logic of the process is still far from being understood

by Wesley C. Salmon

A scientific hypothesis is confirmed, one generally assumes, by finding that it has true consequences. Newton's gravitational theory, for example, implied that any two material bodies experience a mutual attraction. The theory was confirmed by observations of the behavior of falling objects, the motions of the planets and the ebb and flow of the tides; it received direct experimental confirmation when Henry Cavendish devised a torsion balance with which the attraction between two objects could be demonstrated in the laboratory. Such findings do not verify a scientific generalization conclusively. In spite of the many different experimental confirmations of Newton's theory, it is no longer considered totally acceptable. Disconfirming evidence turned up, and now Newtonian physics has been superseded by Einsteinian relativity. No matter how securely a theory seems to be established, one cannot be sure it will not suffer a similar fate. The most one can say is that experimental results tend to confirm a theory and that in some cases accumulated confirming evidence elevates a general hypothesis to the status of at least provisional acceptability. For example, the law of the conservation of energy is currently taken, on the basis of much confirming evidence and no compelling disconfirmations, to be a sound scientific generalization.

The foregoing characterization of scientific testing of hypotheses may seem straightforward and free of problems, but it is an oversimplification. In the

PARADOX OF THE RAVENS suggests the subtleties of confirmation theory. If all ravens are black, surely non-black things must be non-ravens. The generalizations are logically equivalent, so that any evidence that confirms one must tend to confirm the other. Hence the observation of a green vase seems to confirm the hypothesis that all ravens are black. Even a black raven finds it strange.



GRUE-BLEEN PARADOX posits that an object is "grue" if it is green until a certain time (say the end of the 20th century) and blue thereafter. Observation (in 1973) of a green emerald confirms that emeralds are green, but it confirms equally that emeralds are grue. If the man has an emerald in 1973, will it be green or grue (blue) in 2001?

missing details there lurk a host of fundamental difficulties. A good way to expose some of the underlying problems is to consider a series of simple examples, each of which has some puzzling or counterintuitive feature. Bertrand Russell once remarked: "A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science." Although Russell's statement was intended to apply primarily to deductive logic, in particular as it enters into pure mathematics, I think it is equally apt when applied to the logic of confirmation in the empirical sciences.

Logical puzzles are not new to confirmation theory. Ever since the 1940's philosophers have been exercising their heads on two very famous ones. The first is Carl G. Hempel's "paradox of the ravens," which can be described as follows:

Observations of black ravens (in the absence of any observations of ravens of other colors) would normally be taken as confirmation of the generalization "All ravens are black." This statement is logically equivalent to a second generalization, "All non-black things are non-ravens." The observation of non-black things that are non-ravens (in the absence of any observations of non-black ravens) would seem to confirm this second generalization. Since the two generalizations are logically equivalent to each other, whatever counts as evidence for one must also count as evidence for the other. Hence the observation of a green vase (a non-black non-raven) appears to be evidence for the hypothesis "All ravens are black." There seems to be something wrong.

The second puzzle is Nelson Goodman's "grue-bleen paradox." Two peculiar color terms, "grue" and "bleen," are defined. Consider an arbitrary future point in time t_0. For an object existing at any time t, we shall say that the object is grue at t if the object is green and the time t is earlier than t_0, but if the time t is later than t_0, the object must be blue in order to qualify as grue. If we take midnight on December 31, A.D. 2000, as t_0, an object that exists during a period that extends into both the 20th and the 21st century is grue during the entire period if it is green during the 20th century but changes to blue at the beginning of the 21st century and remains blue thereafter. "Bleen" is defined analogously: An object is bleen during the entire time span if it is blue prior to the end of the 20th century and green thereafter. Grue and bleen are strange terms, but they are perfectly well defined.

Now, we would normally say that observations of green emeralds (in the absence of observations of emeralds of other colors) tend to confirm the generalization "All emeralds are green." Since the present date of writing is earlier than t_0, however, each emerald observed to be green is also observed to be grue—at least now—so that these same observations confirm the hypothesis "All emeralds are grue." What, then, should we predict regarding 21st-century emeralds? Will they be green? Or will they be grue—which is to say blue? (As Henry E. Kyburg, Jr., has remarked, this is one of the most pressing problems in the entire logic of confirmation, since we have only 27 years in which to solve it.)

The usual immediate reaction to this grue-bleen terminology is that it is "positional": it involves an arbitrary reference to a particular point in time. That certainly is true if we start with ordinary English words. As Goodman has observed, however, if we start with the grue-bleen terminology, then our ordinary color words turn out to be positional: "green" means "grue prior to t_0 but bleen thereafter," whereas "blue" means "bleen prior to t_0 but grue thereafter." Is there any logical reason for preferring the English terms, or is our preference for them as opposed to Goodman's grue-bleen terminology merely the result of historical accident?

These two puzzles have been a standard part of confirmation theory for the past 30 years. An enormous literature has grown up around them. By now every theorist believes he has the definitive answer to each of them, but there is conspicuous lack of agreement on what the correct answers are. Instead of adding to this discussion I shall deal with some other puzzles that are less familiar but to my mind no less fundamental.

In setting out to test a hypothesis an investigator uses the hypothesis to predict some phenomenon whose occurrence or nonoccurrence can be ascertained by observation. A general hypothesis by itself, however, entails no observable facts: the hypothesis must be applied to some particular situation whose description constitutes a set of "initial conditions." In order to predict an eclipse, for instance, the astronomer needs to know not only the laws of motion that govern the earth and its natural satellite but also the relative positions of the earth, the moon and the sun at some particular time; from the laws of motion and the initial conditions together he can deduce the time and place of a total solar eclipse. Often it is necessary to manipulate circumstances in order to achieve a set of initial conditions suitable for the testing of a given hypothesis; that is what is involved in performing an experiment.

A puzzle proposed by Russell is directly pertinent to the kind of experimental scheme I have just outlined. Consider the hypothesis "Pigs have wings." In conjunction with the observed fact (initial condition) that pork is good to eat, we deduce the consequence—we predict—that some winged creatures are good to eat. When we see that people enjoy eating ducks and turkeys, we observe that the consequence—or prediction—is true; we appear to have a confirmation of the original hypothesis.

By no stretch of the imagination can the verification of such a consequence be taken to lend any weight to the hypothesis in question. If we call such an outcome a "positive instance," meaning that it agrees with the deduced prediction, we must conclude that positive instances do not necessarily lend the slightest credibility to the hypothesis. This apparently silly example points a profound and significant moral: There is more to scientific confirmation than merely finding true consequences. (This is a point that should be kept firmly in mind when evaluating such work as that of Immanuel Velikovsky on the basis of allegedly true predictions.)

If we reason deductively from the premises "All mammals have hair" and "Whales are mammals" to the conclusion "Whales have hair," we can be sure that the conclusion will be true if the premises are true (as in fact they are); this is essentially the defining characteristic of valid deductive inference. In deduction, however, it is an elementary logical error (known as "the fallacy of affirming the consequent") to argue backward from the truth of the conclusion to the truth of the premises. When it comes to scientific induction, on the other hand, it sounds quite respectable to suppose that the observation of hair in the embryonic whale (which is known to be a mammal) is evidence for the generalization that all mammals have hair. This double standard led Morris R. Cohen to the facetious characterization of logic texts as books that are divided into two parts: in the first part [on deduction] the fallacies are explained and in the second part [on induction or scientific method] they are committed (with apparent impunity). Russell's pigs-have-wings example shows that instances of the fallacy of affirming the consequent do not, however, automatically qualify as sound scientific confirmations.

The relation of logical entailment has the obvious and important property of transitivity: if A entails B and B entails C, then A entails C. If deducibility of true consequences were the whole story regarding confirmation, then it too would be a transitive relation. If C were to confirm B because it follows from B, and if B were to confirm A because it

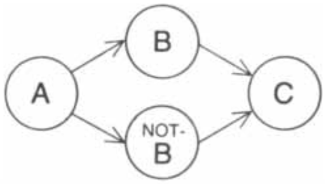
"PIGS HAVE WINGS" is the hypothesis. The investigator knows that pork is good to eat. On the basis of the hypothesis he predicts that some winged creatures must be good to eat. Testing the hypothesis, he tries duck, and finding it good to eat, considers the hypothesis confirmed. (The pig, punning in French, knows this true consequence is irrelevant.)

LOGICAL

IF *A* ENTAILS *B* AND *B* ENTAILS *C*, THEN *A* ENTAILS *C*.

PROBABILISTIC ANALOGUE

PROBABILITY (*C*, GIVEN *A*) = PROBABILITY (*B*, GIVEN *A*) × PROBABILITY (*C*, GIVEN *A* AND *B*) + PROBABILITY (NOT-*B*, GIVEN *A*) × PROBABILITY (*C*, GIVEN *A* AND NOT-*B*).



TRANSITIVITY is a property of logical entailment: If being a whale entails being a mammal and being a mammal entails having hair, then being a whale entails having hair. The probabilistic analogue is more complex: The probability that a human (*A*) will get lung cancer (*C*) equals the probability that he (*A*) is a smoker (*B*) times the probability that a human smoker (*A* and *B*) gets lung cancer plus the probability that he (*A*) is a nonsmoker (not-*B*) times probability that a human nonsmoker (*A* and not-*B*) gets lung cancer.

follows from *A*, then *C* would have to confirm *A*, since *C* would follow from *A* by transitivity of deductive entailment. Indeed, it might seem intuitively that confirmation is a transitive relation. The following argument was actually cited as an example of sound inductive reasoning in a book published a dozen years ago: "Since, if there was smoke here there was very probably fire here, and if there is soot here there was very probably smoke here, and there is soot here, there was probably fire here." That argument, however, has essentially the same logical structure as the following one: Since it is very probable that any scientist who ever lived is alive today (it has been estimated that 90 percent of all scientists are still alive), and since it is very probable that any organism alive today is a microorganism, then, given that Smith is a scientist, it is likely that he is a microorganism.

This example shows unmistakably that *A* (being a scientist) can lend confirmation to *B* (being alive at present), and *B* can in turn lend weight to *C* (being a microorganism), whereas *A* not only fails miserably to confirm *C* but also is actually incompatible with it. The example illustrates a serious risk involved in thinking about confirmation: the danger of drawing unfounded analogies with deduction. It is easy to assume intuitively that properties of deductive relations apply more or less exactly to the logic of confirmation. That is not the case. Some of the most fundamental properties of deductive relations fail absolutely and completely when one departs in the least from deductive logic itself. The relation of transitivity is an excellent case in point.

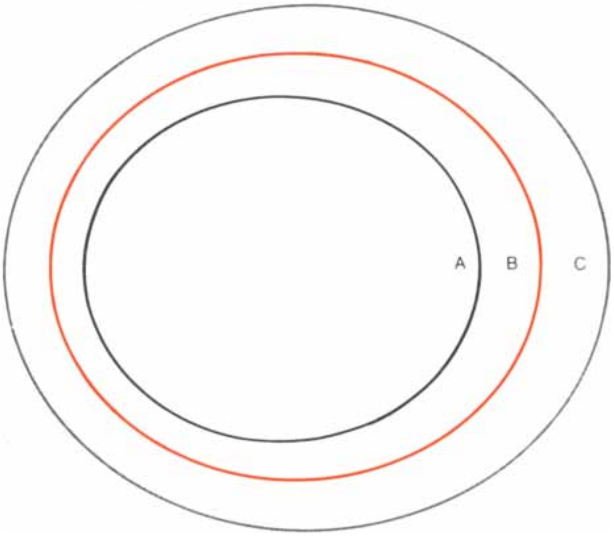
The best antidote for errors that arise from intuitions about confirmation is, in my opinion, to pay close attention to the mathematical theory of probability [*see bottom illustration on opposite page*]. If confirmation can be regarded as a type of probability, it is easy to show that confirmation relations are not transitive. It can also be shown that the truth of a consequence of a hypothesis does not necessarily enhance the probability of that hypothesis. Even if we bring the probability calculus into full play, however, there is still danger of serious confusion based on a simple ambiguity.

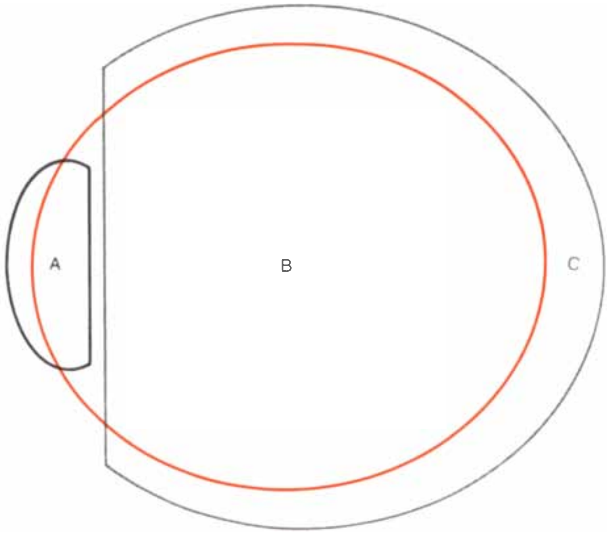
To say, for example, that the special theory of relativity has been confirmed might mean either of two things. Either some item of evidence, such as the recent test of the relativistic retardation of moving clocks, has increased the theory's credibility to some degree, or the theory has been supported by such a large and varied body of evidence that it can be considered, for the time being at least, a scientifically acceptable law. A general hypothesis that initially has a relatively low credibility can have it raised to some extent by new evidence without thereby attaining a high degree of confirmation. In such cases a hypothesis can obviously be confirmed in one sense without being confirmed in the other sense. Let us refer to raising the probability of the hypothesis as "incremental" confirmation and to the attainment of a high probability as "absolute" confirmation. In order to emphasize the distinction let us confine the unqualified term "confirmed" to the incremental sense and always use a phrase such as "highly confirmed" for the absolute sense. We would say, then, that the special theory of relativity was confirmed by the clock-retardation experiment and that it is highly confirmed by the total body of evidence supporting it.

Although the distinction between these two senses of confirmation is obvious and has long been acknowledged, its implications have not always been clearly recognized. As a matter of fact, the incremental sense has some strange properties that are easily overlooked because they are not shared by the absolute concept. Consider the following example, which, although it is totally fictitious, nevertheless exhibits the logical possibilities.

Jones consults his physician about a respiratory ailment. After a preliminary examination the physician says that he thinks Jones has pneumonia but that he is not sure whether it is bacterial, viral or (a rare possibility) both types together. Further testing is required. Jones is given the prescribed test. The physician tells him the test has confirmed the hypothesis that he has bacterial pneumonia and has also confirmed the hypothesis that he has viral pneumonia but has disconfirmed the hypothesis that he has pneumonia! Most people would find it understandable if at this point Jones sought the advice of a different physician.

Yet the physician may be on solid ground. Suppose that on the basis of the superficial examination he concludes there is a 96 percent chance that Jones has pneumonia, but he has no indication of whether it is bacterial or viral (assuming those are the only two types). Moreover, he decides there is a 2 percent chance that Jones has both types. Consequently the probability that Jones has bacterial pneumonia is 49 percent and the chance that he has viral pneumonia is also 49 percent. Suppose further that there is a test that quite reliably picks out the rare cases where both types are present together. When the test is administered to Jones, the result is positive, making it 89 percent certain that he has both types. Assume, moreover, that this test seldom comes out wrong for someone who has only one type of pneumonia; that is, if the result is positive and the individual does not have both types, he is very likely to have neither type. In





"ALL" AND "ALMOST ALL" can be very different. Given that all *A* are *B* and all *B* are *C*, it must be true that all *A* are *C* (*left*). However, given that almost all *A* are *B* and almost all *B* are *C*, it may happen that no *A* are *C* (*right*). In the example in the text, although most scientists (*A*) are living things (*B*) and most living things (*B*) are microorganisms (*C*), no scientists at all are microorganisms.

particular the positive test result means there is a 90 percent chance that Jones has bacterial pneumonia, a 90 percent chance that he has viral pneumonia and consequently a 91 percent chance that he has one or the other or both [*see illustration on page 81*].

This fanciful example illustrates the possibility that evidence can confirm each of two hypotheses and yet disconfirm their disjunction (the one-or-the-other-or-both combination). In this case the result of the special test increased the probability that Jones had bacterial pneumonia and also increased the probability that he had viral pneumonia, and at the same time the result decreased the probability that he had one or the other—in other words, since there are only the two types, that he had pneumonia!

To see how this is possible we must look at the addition rule for probabilities [*see illustration below*]. In order to compute the probability for a disjunction of nonexclusive alternatives we add together the probabilities of each of the two alternatives taken separately and then subtract the probability of their joint occurrence. For example, the probability of getting on a draw from a standard bridge deck either an honor card (an ace, a king, a queen, a jack or a 10) or a trump (a spade, say) is equal to the probability of an honor card (20/52) plus the probability of a spade (13/52) minus the probability of a spade honor (5/52). The subtraction is necessary because the spade honors were, so to speak, counted twice: once as honors, once as spades. The answer is 28/52 [*see illustration on page 82*]. In the pneumonia example the probability of the disjunction decreases whereas the probability of each alternative increases by virtue of the high (89 percent) probability of the joint (or twice-counted) occurrence after the test in contrast with its low (2 percent) probability before the test.

Results such as these are characteristic of incremental confirmation, and they are disconcerting partly because there is a natural tendency to confuse the incremental concept with the absolute. Confirmation in the absolute sense means that a hypothesis has a high probability of being correct. If viral pneumonia is highly confirmed in the absolute sense, then pneumonia is at least as highly confirmed. That is because having pneumonia (of some kind) is a logical consequence of having viral pneumonia, and a basic rule of mathematical probability states that the probability conferred on a logical consequence of any proposition

ADDITION	PROBABILITY (*A* OR *B*) = PROBABILITY (*A*) + PROBABILITY (*B*) − PROBABILITY (*A* AND *B*).
LOGICAL CONSEQUENCE	IF *A* ENTAILS *B*, THEN PROBABILITY (*B*) IS EQUAL TO OR GREATER THAN PROBABILITY (*A*).
MULTIPLICATION	PROBABILITY (*A* AND *B*) = PROBABILITY (*A*) × PROBABILITY (*B*, GIVEN *A*).
BAYES'S THEOREM	PROBABILITY (*A*, GIVEN *B*) = $\frac{\text{PROBABILITY }(A) \times \text{PROBABILITY }(B\text{, GIVEN }A)}{\text{PROBABILITY }(B)}$.

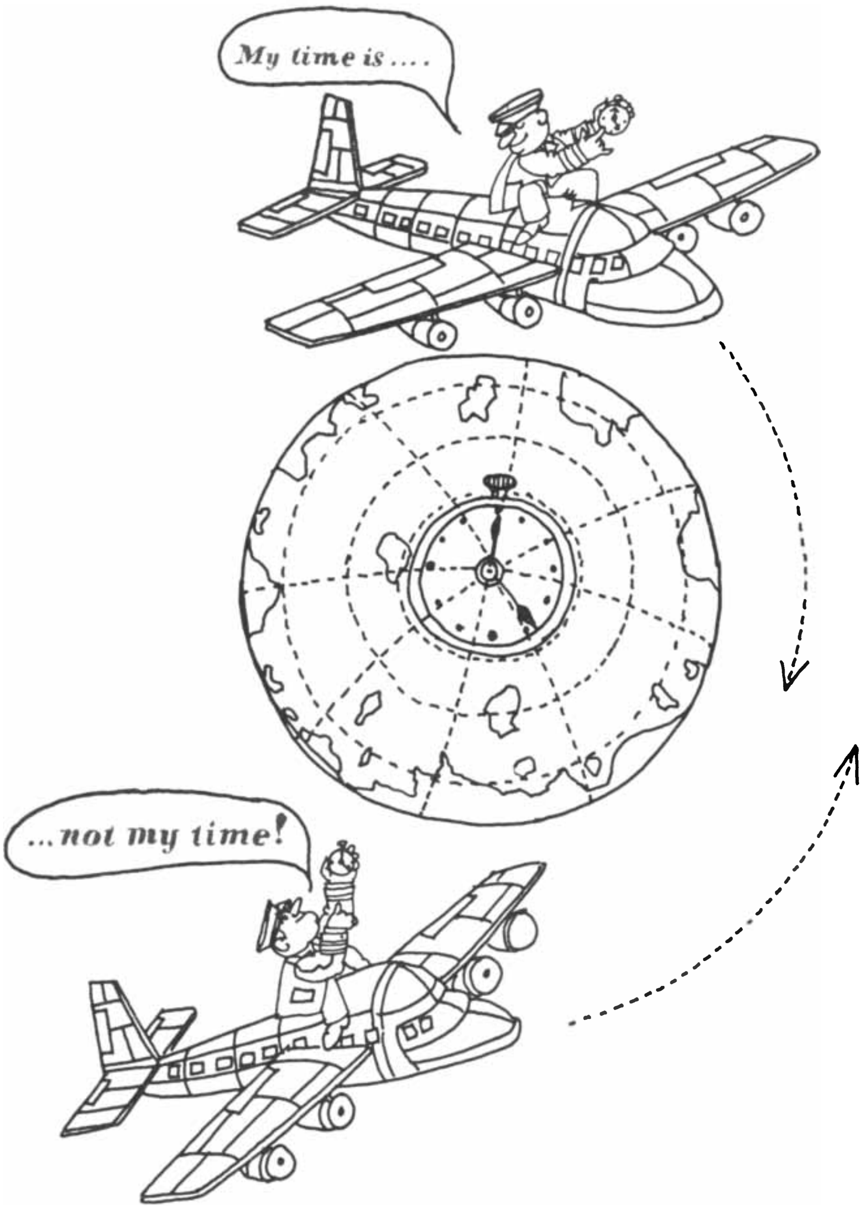
CONCEPT OF PROBABILITY does for empirical confirmation of hypotheses what logical deduction does for mathematical proof. Some of the probability rules referred to in the text are listed. Another, for transitivity, is given in illustration on the opposite page.

is as great as or greater than the probability of the proposition itself. I have not violated that condition in describing Jones's case: the probability of pneumonia is greater than that of viral pneumonia on the basis of the preliminary diagnosis and also on the basis of the special test. Nevertheless, when it comes to incremental confirmation (a change in confirmation), it turns out that new evidence can increase the probability of a proposition and yet decrease the probability of one of its consequences. This is one striking logical contrast between incremental and absolute confirmation.

Even stranger things can happen. Suppose two investigators set out to test a hypothesis. Each goes to his own laboratory to perform an experiment and each achieves a positive finding: a confirmation of the hypothesis. Could it ever happen that although each finding confirms a hypothesis, the conjunction of them disconfirms the same hypothesis? It is logically possible.

Let A and B be two atoms of an imaginary isotope that can decay radioactively in any one of three ways. Given that disintegration has occurred, there is a probability of .7 that an alpha particle was emitted, a probability of .2 that a negative electron was emitted and a probability of .1 that a positive electron (a positron) was emitted. Suppose both of these atoms have just disintegrated and the ejected particles are approaching each other. Consider the hypothesis that they will annihilate on meeting, an event that will occur if and only if one particle is a negative electron and the other is a positron. Given no further information, the probability of annihilation is .2 times .1 plus .1 times .2, or .04. Suppose one physicist finds that atom A has ejected an electron; on the basis of this evidence the probability that annihilation will occur is .1, since that is the probability that B emits a positron. Suppose another physicist finds that atom B has ejected an electron; on the basis of this evidence (but without the evidence obtained by the first physicist) there is likewise a probability of .1 that annihilation will occur. Yet on the basis of both pieces of evidence in conjunction, it is clear that annihilation is impossible.

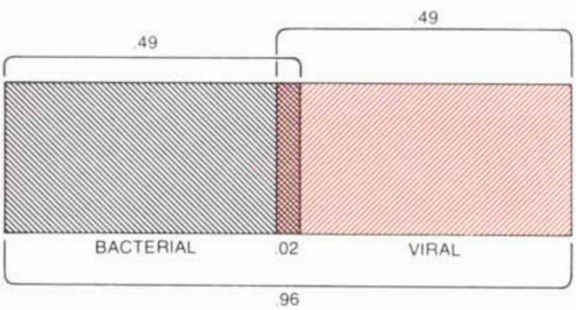
Each of two pieces of evidence separately confirms the annihilation hypothesis (each raises its probability from .04 to .1). In conjunction, however, they not only disconfirm it, they actually refute it. Could such a thing ever happen in the course of an investigation? In an experiment on the Compton scattering of photons by electrons one physicist might measure the frequency of photons scattered at a particular angle while another measures the energy of the recoiling electrons. Even if each set of measurements confirms the hypothesis of Compton scattering, how can one be sure that taken together they do so? In the case of Compton scattering it happens that the conjunction of the findings does confirm the hypothesis, but this does not follow automatically from the mere fact that separately the findings are confirmations; it depends on more circumstances, including the fact that the conjunction itself is one of the predictions of the theory. The annihilation example shows, nevertheless, that there are broad and basic questions about the legitimacy of supposing that the accumulation of many confirming test results must inevitably enhance the credibility of scientific hypotheses.



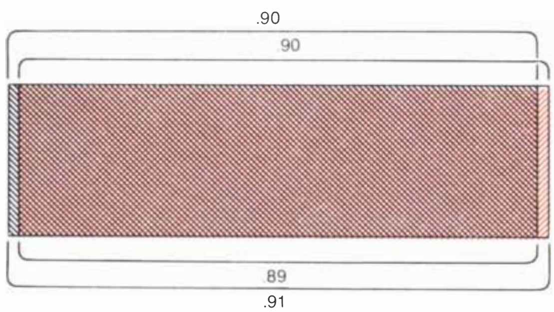
RELATIVITY THEORY predicts that the faster a clock is transported, the slower it runs. Last year Joseph C. Hafele of Washington University and others reported on an experiment in which cesium clocks were carried around the earth in opposite directions on commercial jet airplanes. In relation to fixed stars one was moving some 600 miles faster than the earth and the other 600 miles slower; one should lose time and one should gain time in relation to reference clocks on the ground. The time discrepancies were close to the predicted values, providing confirmation of the theory by the classic prediction-and-experiment method.

So far we have spoken only of confirmation, or of the case in which the outcome of the test is positive. It remains to discuss the negative test result: the case in which the prediction of the hy-

PROBABILITY (PNEUMONIA) = PROBABILITY (BACTERIAL) + PROBABILITY (VIRAL) − PROBABILITY (BOTH), OR .49 + .49 − .02 = .96



PROBABILITY (PNEUMONIA) = PROBABILITY (BACTERIAL) + PROBABILITY (VIRAL) − PROBABILITY (BOTH), OR .90 + .90 − .89 = .91



WHAT CONFIRMS each of two hypotheses may at the same time disconfirm their disjunction. After a superficial examination the physician decides there is a 96 percent chance that Jones has pneumonia; he does not know if it is bacterial or viral, but guesses there is only a 2 percent chance that both types are present. The various probabilities are diagrammed by the overlapping bars (*left*). A test is ordered. The result greatly increases the probability that Jones has bacterial pneumonia and also that he has viral pneumonia—but it reduces to 91 percent the probability that he has pneumonia (*right*). The reason lies in the addition of probabilities.

pothesis turns out to be false. It has often been argued that there is a strong asymmetry between the positive and the negative cases because of a simple fact in logic: While the inference from a true conclusion to the truth of the premises is a case of the fallacy of affirming the consequent, the inference from the falsehood of a conclusion to the falsity of at least one premise is perfectly valid; it is known as *modus tollens,* or denying the consequent. Since a valid deduction is defined as one whose conclusion must be true if it has true premises, we can indeed conclude that a valid deductive argument with a false conclusion cannot have premises that are all true. And so it seems that a negative outcome not only disconfirms a hypothesis but also actually refutes it conclusively.

The situation is not quite that simple. As Pierre Duhem has pointed out, in most, if not all, cases auxiliary hypotheses are involved when one sets out to test a scientific hypothesis. In the testing of an astronomical theory telescopes and other instruments are likely to be used, so that the laws of optics and other laws are involved in the effort to establish a connection between a spot on a photographic plate and a celestial body. In a sophisticated science the initial conditions required to test a hypothesis are hardly ever ascertainable by direct observation; auxiliary hypotheses are needed to relate what is actually observed to suitable initial conditions. Moreover, predicted outcomes may be somewhat remote from direct observation, and again auxiliary hypotheses come into play.

The net result of these complications is that the negative outcome of an experimental test of a hypothesis cannot be taken automatically as a refutation of that hypothesis. The negative test result shows only that something is wrong somewhere. It may be that the hypothesis being tested is false or it may be that some auxiliary hypothesis is false. Strictly speaking, the negative result only disproves the conjunction of the auxiliary hypotheses and the hypothesis being tested; it does not refute the tested hypothesis by itself. (As an interesting historical example, consider the false predictions about the motions of Uranus that emerged from Newtonian mechanics in the 19th century. Instead of refuting Newtonian physics they led to the discovery of Neptune; the negative outcome was attributed to the auxiliary hypothesis or perhaps to the initial conditions themselves. Later, however, irregularities in the orbit of Mercury led not to the discovery of a new planet as predicted on the basis of Newtonian mechanics but rather to the eventual downfall of Newtonian mechanics: the precession of the perihelion of Mercury was a crucial early item of evidence for Einstein's general theory of relativity.)

In view of Duhem's fundamental insight what moral about confirmation and disconfirmation should be drawn regarding a negative test outcome? Surely either the hypothesis under test is disconfirmed or the auxiliaries are disconfirmed to some extent; surely neither the auxiliaries nor the main hypothesis can be confirmed by the negative outcome. Both assumptions sound plausible but they are wrong. It is logically possible for an experimental outcome that conclusively refutes a conjunction of two hypotheses to confirm each of them individually. This possibility is closely related to the radioactive-decay example.

Let us assume the same two-atom setup as in that example: atoms *A* and *B* each have three possible decay modes, with the probabilities .7 for the alpha particle, .2 for the negative electron and .1 for the positron. This time it is the annihilation of the two particles that is the observed evidence. We consider the hypothesis that atom *A* emitted a negative electron. Since we know by virtue of the annihilation that one of the atoms emitted a negative electron, but we do not know which, the hypothesis that it was emitted by *A* has probability .5. The same goes for the hypothesis that *B* emitted a negative electron. The fact that annihilation occurred makes it impossible, however, that both of the atoms ejected negative electrons; it therefore refutes the conjunction of the two hypotheses. Nevertheless, this very evidence confirms each hypothesis separately, because in each case it raised the probability from .2 to .5.

Imagine what this kind of thing might mean to scientific methodology. Scientist Smith comes home late at night after a hard day at the laboratory. "How did your work go today, dear?" asks his wife.

"You know the Smith hypothesis on which I've staked my entire reputation? Well, today I ran an experimental test, and the outcome was negative."

"Oh, dear, what a shame! Does that kill your favorite hypothesis and send your reputation down the drain?"

"Not at all. In order to carry out the test I made use of some auxiliary hypotheses."

"Oh, what a relief—saved by Duhem! Your hypothesis wasn't refuted after all." Mrs. Smith breathes a deep sigh.

"Better than that," Smith continues. "I actually confirmed the hypothesis."

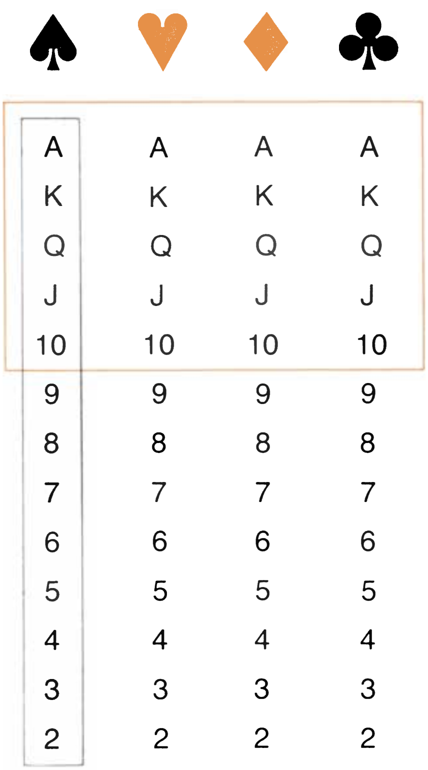
"Why, that's wonderful, dear," replies Mrs. Smith. "You must have found that by rejecting the auxiliary hypothesis you could show the test actually supported your hypothesis. How ingenious!"

"No," Smith continues, "it's even better. I found I had confirmed the auxiliary as well!"

Such outlandish possibilities may seem to make a shambles of scientific methodology. The fact that actual scientists in actual practice do not get involved in such difficulties may be taken by many to show that confirmation is a matter of scientific intuition and resists all efforts at formalization. I do not believe such a conclusion is justified, and I shall argue my case by offering a parallel in the history of the concept of proof in mathematics.

The idea of mathematical demonstration had emerged by 600 B.C. Thales of Miletus is credited with bringing geometry from Egypt to Greece and in the process transforming it into a mathematical science. Although the Egyptians had applied geometry in surveying, there is no evidence that they actually proved any geometrical theorem; Thales is believed to have proved, among other theorems, that the base angles of an isosceles triangle are equal. By about 300 B.C. Euclid had recast geometry as an axiomatic system in which all theorems are to be deduced from a small number of axioms or postulates. Some elementary portions of deductive logic were developed in antiquity by Aristotle and the Stoic philosophers. Yet it was not until 1879 that Gottlob Frege developed a deductive logic that could begin to be adequate to characterize deduction in mathematics. At the very minimum, then, it took 2,500 years from the time mathematical proof was first employed for logicians to come to any clear understanding of its nature.

Mathematical logic has now flourished for 100 years, and many deep results have been established. The process has not been without its vicissitudes. For example, Russell found a famous contradiction in the very logic on which Frege had tried to base all mathematics. It arose from considering puzzles similar to the famous barber paradox: In a certain town there is a barber who shaves every man who does not shave himself. Who shaves the barber? [see "Paradox," by W. V. Quine; SCIENTIFIC AMERICAN, April, 1962]. The fact that working mathematicians were not constantly embroiled in contradictions did not prevent Russell's paradox from constituting a crisis in the foundations of mathematics. Other astonishing developments, such as Kurt Gödel's proof of the essential incompleteness of arithmetic [see "Gödel's Proof," by Ernest Nagel and James R. Newman; SCIENTIFIC AMERICAN, June, 1956], were disquieting to say the least, although they did not exhibit outright inconsistency.



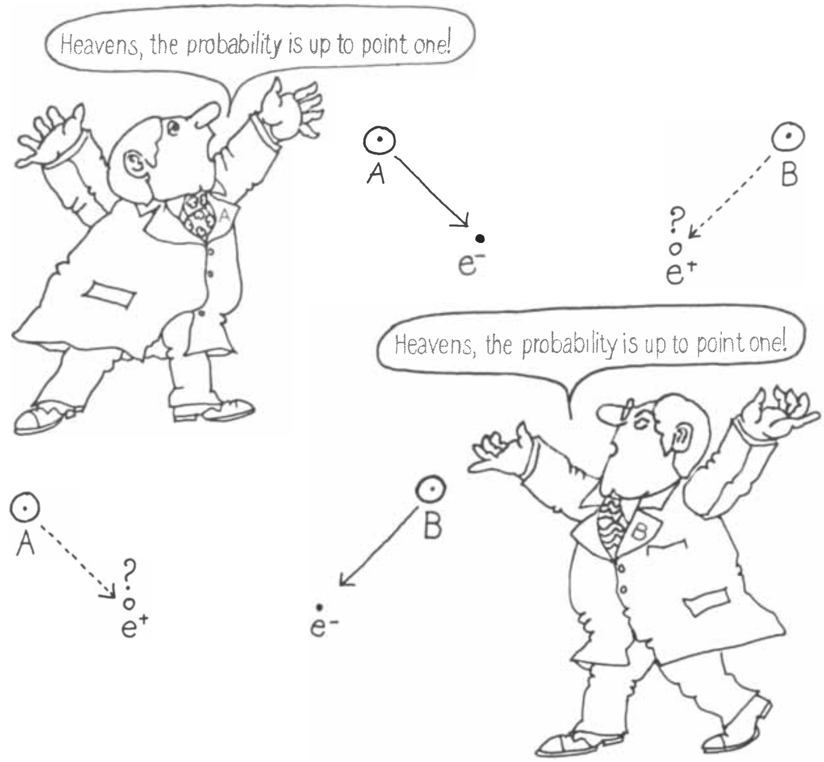
PROBABILITY (HONOR OR SPADE) = PROBABILITY (HONOR) + PROBABILITY (SPADE) − PROBABILITY (SPADE HONOR), OR 20/52 + 13/52 − 5/52 = 28/52.

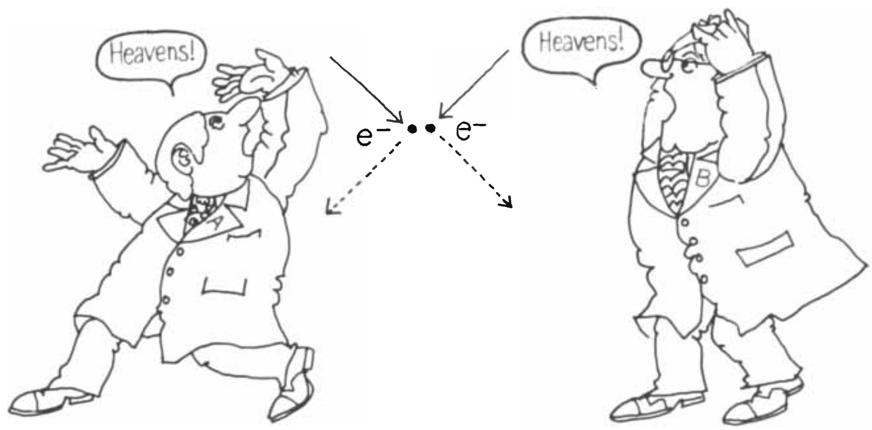
PROBABILITY of drawing an honor or a spade is determined by addition of probabilities as shown. The subtraction of the five spade honors is required because, as the rectangles here indicate, they are otherwise counted twice. The 89 percent probability of joint occurrence of bacterial and viral pneumonia in the preceding example is analogous to the spade-honor occurrence.

Empirical scientists have been making observations and performing experiments in order to test sophisticated hypotheses ever since the rise of modern science in the 16th and 17th centuries. When it comes to drawing conclusions from the results of these observations and experiments, we are far from having a clear understanding of the kind of reasoning involved. We are now in a situation analogous to that of mathematics during the millenniums in which mathematical proof was used often and with good results while the logic behind it remained basically mysterious. Current work in confirmation theory and inductive logic is attempting to remedy the situation.

There is much difference of opinion as to the best course to follow in trying to deal with the puzzles of confirmation. Two resources seem to me to offer considerable promise of help. The first of these is Bayes's theorem, a simple theorem in the mathematical calculus of probability [*see bottom illustration on page 79*]. Bayes's theorem is often called the "rule of inverse probability." Given the probability that certain evidence would obtain if a particular hypothesis is true (and given some other probabilities as well), Bayes's theorem enables one to compute the probability that the hypothesis is true given that the aforementioned evidence is found. In at least certain cases it can be used to ascertain the probability that some particular cause was operative, given that a certain effect has occurred.

Bayes's theorem has been widely exploited in recent years by statisticians who called themselves Bayesians, notably the late L. J. Savage. Bayes's theorem contributed to confirmation theory a scheme that seems far more adequate to inference in science than the fallacy of affirming the consequent can ever hope to be. (It holds a key to puzzles arising from the pigs-have-wings and scientist-as-microorganism examples, which were based on an oversimplified notion of scientific confirmation.)

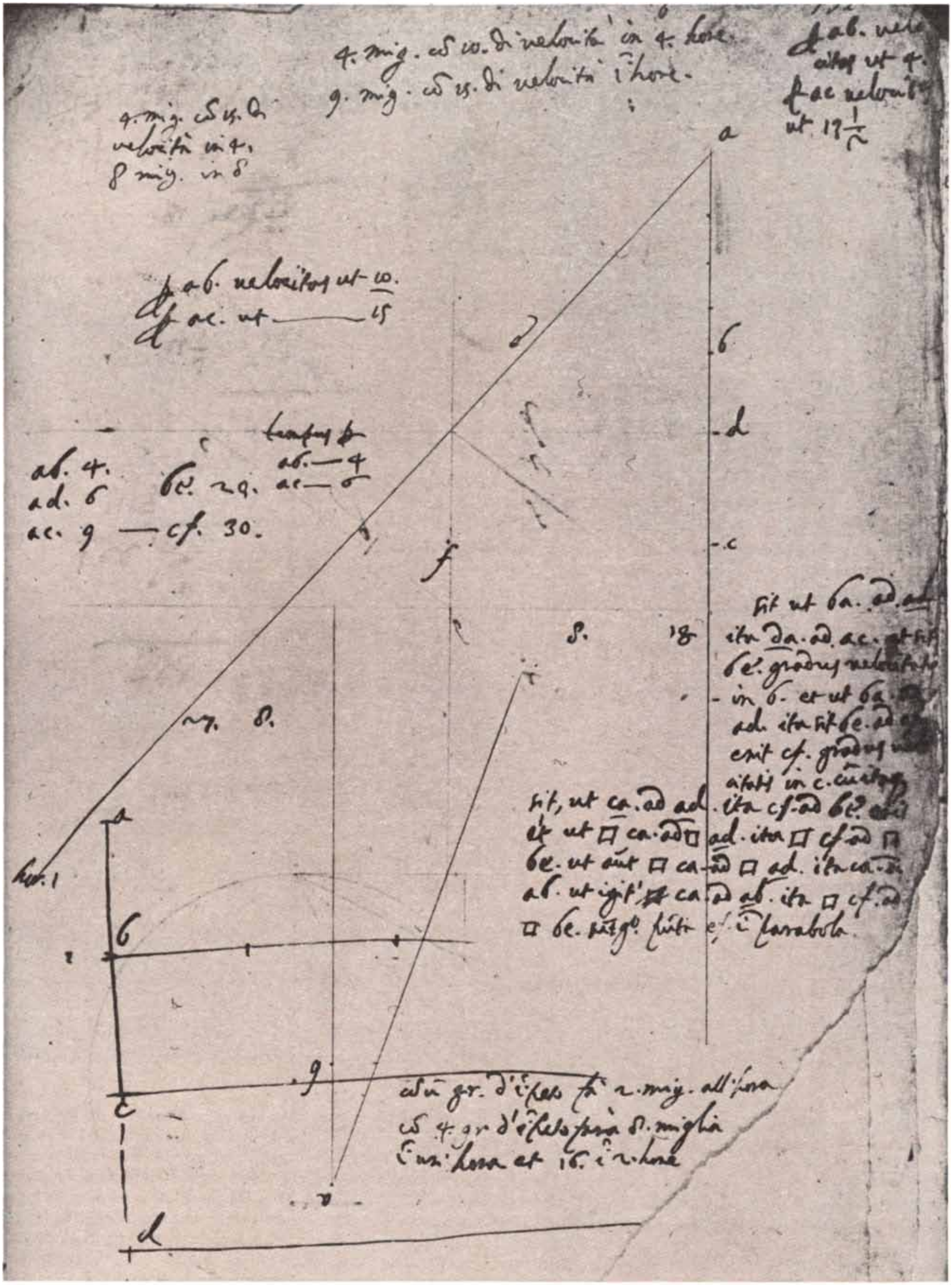




CONJUNCTION of two confirmations may refute a hypothesis. If atom *A* or *B* decays, the probability is .2 that it emits an electron and .1 that it emits a positron. The probability of annihilation, which will occur if one atom emits an electron and the other a positron, is therefore $.2 \times .1 + .1 \times .2$, or .04. If scientist *A* observes only that atom *A* has emitted an electron, he considers that the probability of annihilation is raised to .1 (the probability that atom *B* emits a positron). Scientist *B*, observing only that atom *B* has emitted an electron, reports the same increase in probability. The conjunction of both observations (the emission of two electrons), however, means that there can be no annihilation.

The second resource arises from a clear recognition of the incremental concept of confirmation as opposed to the absolute concept. Incremental confirmation involves change of probability, which is basically a concept of probabilistic relevance. The way is thus open to defining a measure of relevance based on the mathematical characteristics of probability, by means of which incremental confirmation can be studied in a precise and systematic fashion. Such a measure was defined and elaborated by Rudolf Carnap in 1950, but insufficient attention seems to have been paid to it. The pneumonia puzzle and the two annihilation puzzles were devised by relying on Carnap's treatment of relevance, and further careful attention to a formal concept of incremental confirmation will, I believe, strip such examples of their paradoxical air.

At the present stage of development studies in confirmation theory and inductive logic have produced more paradoxes and puzzles than convincing or widely accepted solutions of fundamental problems. Further work on such puzzles should, however, yield rich insights into the logic of the empirical sciences, much as studies in the foundations of mathematics have paid rich rewards in the understanding of that discipline's logic.



UNPUBLISHED MANUSCRIPT records Galileo's first discovery of the law of free fall, before October, 1604. It is reproduced by courtesy of the National Central Library in Florence, where it is preserved as *f. 152r* in Volume 72 of the Galilean manuscripts. The crucial diagram is the large one at the top; the lower diagram and the bottom paragraph refer to calculations for horizontal motion.

Galileo's Discovery of the Law of Free Fall

It has been thought that he erroneously assumed that the velocities of a falling body were proportional to distances. A new manuscript shows that he treated them correctly as being proportional to time

by Stillman Drake

The modern era in physics opened with the publication of Galileo's *Discourses on Two New Sciences* in 1638. It reported basic discoveries he had made 30 years earlier. In describing his book to a friend in January, 1639, the old and blind author dictated these words (here, of course, translated from the Italian): "I assume nothing but the definition of that motion of which I wish to treat and whose properties I then demonstrate.... I declare that I wish to examine the essentials of the motion of a body that leaves from rest and goes with speed always increasing...unifomly with the growth of time.... I prove the spaces passed by such a body to be in the squared ratio of the times.... I argue from supposition about motion defined in that manner, and hence even though the consequences might not correspond to the events of natural motion of falling heavy bodies, it would little matter to me.... But in this, I may say, I have been lucky; for the motion of heavy bodies, and the properties thereof, correspond point by point to the properties demonstrated by me."

Not even Galileo's severest critics attribute his discovery of the law of free fall to sheer luck; hence it may seem odd that I, one of his most fervent admirers, should do so now. I believe Galileo meant his remark quite literally. Evidence for this belief exists in an early manuscript of his that has never before been published in full. That document unfolds a fascinating story of scientific discovery through a combination of error, good luck, persistence and mathematical ingenuity.

Historians of science have searched the writings of earlier men for the possible origins of Galileo's analysis of accelerated motion because no documentary evidence in his own hand had been found. Nothing surviving from classical antiquity offers a plausible source. In the 14th century, however, there were some very interesting developments in the application of mathematics to physical questions. In particular William Heytesbury and Richard Swineshead of Merton College, Oxford, and Nicole Oresme of Paris analyzed accelerated motion. The roots of medieval investigations lay in a theological problem—the increase of charity in a man—and in its philosophical implications, which spread into the general problem of rate of change. The results that were achieved are impressive, and it seems strange that their authors never thought to apply them to the problem of the changing speed of a freely falling object.

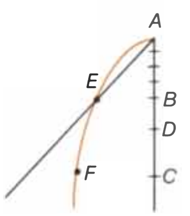
Medieval English writers adopted an arithmetical approach, from which they developed the mean-degree or mean-speed theorem, often called the Merton rule. Under this rule the speed at the middle instant was taken as being representative. Uniform motion at this mean speed over a fixed time was declared to be equivalent to uniformly accelerated motion from rest over that same time. It followed that in any uniformly accelerated motion from rest, one-fourth of the total distance was traversed in the first half of the time. This fact yielded the ratio 3 : 1 for the distances traversed in the later and earlier halves. Oresme proved the rule geometrically, and in another manuscript he extended the ratio to the progression 1, 3, 5, 7 and so on for equal times. Not even Oresme, however, connected uniform acceleration with free fall, nor did any medieval writer announce that the distances covered are proportional to the squares of the times, a fact that is deducible from the above progression. The prevailing view among historians of science was recently summarized by Edward Grant in his book *Physical Science in the Middle Ages:*

"Oresme's geometric proof and numerous arithmetic proofs of the mean-speed theorem were widely diffused in Europe during the fourteenth and fifteenth centuries, and were [then] especially popular in Italy. Through printed editions of the late fifteenth and early sixteenth centuries, it is quite likely that Galileo became reasonably familiar with them. He made the mean-speed theorem the first proposition of the Third Day of his *Discourses on Two New Sciences,* where it served as the foundation of the new science of motion."

Puzzles nevertheless remained. I again cite Grant's words: "The Mertonians arrived at a precise definition of uniform acceleration as a motion in which an equal increment of velocity is acquired in any whatever equal intervals of time, however large or small." Yet "Galileo, as late as 1604, mistakenly assumed that velocity is directly proportional to distance rather than to time, as he later came to realize." If we were to assume that the medieval writings were Galileo's source, it would be hard to explain why he accepted and extended the earlier results while rejecting the only definition on which they were based. Similarly, if he came on the medieval writings later in life, why did he still make no use of the Merton rule in his proof of the proposition mentioned, either in his notes or in his book?

I once suggested two possible alternative sources for Galileo's times-squared law. The first was that he might have discovered that spaces traversed in equal times follow the odd-number rule by roughly measuring the distance traveled by an object rolling down a gently inclined plane, using the first distance as his unit. I thought he might have made

<table>
<tr><th>GALILEO'S CALCULATIONS</th><th>EXPLANATION</th></tr>
<tr><td>(1) 4 miles with 10 degrees of speed in 4 hours</td><td>Galileo first assumed that one more degree of speed is consumed in each mile of distance traversed: the speed is of one degree in the first mile, two in the second, three in the third and four in the fourth. That gave him 4 miles with 10 degrees (1 + 2 + 3 + 4) of speed consumed. The time elapsed was arbitrarily put at one hour for each different degree of speed.</td></tr>
<tr><td>(2) 9 miles with 15 degrees of speed in $5^?$ hours</td><td>$5^?$ means that the time was to be examined. Although Statement 2 was intended to follow from Statement 1, and the next increment of speed to be added was indeed five degrees, the number of miles to be added was one and not five. Had his intention been to add a greater increment of speed for each hour, he would have written "$9^?$ miles" instead of "$5^?$ hours," and Statement 1 would still have contradicted Statement 2. (An unsuccessful attempt to get a ratio of overall speeds appears in the top right corner of f. 152r. In it Galileo tried multiplying the ratios of times and distances together.)</td></tr>
<tr><td>(1a) 4 miles with 10 degrees of speed in 4^6 hours
(3) 4 miles with 15 degrees of speed in 4 hours
(3a) 8 miles in 8 hours at 15 degrees of speed</td><td>To get a ratio of distances and times Galileo substituted 6 hours for the 4 hours in Statement 1. With this change, in order to cover 4 miles in 4 hours it is necessary to travel at 15 degrees of speed. Galileo noted the implication that if "15 degrees" is an overall rate, then in order to cover twice the distance (8 miles) it would take twice as long (8 hours), which would contradict Statement 2.</td></tr>
<tr><td>(4)

</td><td>In an effort to resolve the contradiction Galileo drew a vertical line and first lettered it A, B and C. He divided the distance AB into four units; AC is nine units. D he added so that the distance AD would be the mean proportional between them at six units ($\sqrt{4} \times \sqrt{9} = 2 \times 3 = 6$). He let this represent the time taken in traversing AC and let AB represent the time taken in falling from A to B.</td></tr>
<tr><td>(5) through AB speed 10
through AC as . . . 15
<table>
<tr><th>distance</th><th>speed acquired</th><th>time through</th></tr>
<tr><td>AB 4</td><td></td><td>AB – 4</td></tr>
<tr><td></td><td>BE 20</td><td></td></tr>
<tr><td>AD 6</td><td></td><td></td></tr>
<tr><td>AC 9</td><td>CF 30</td><td>AC – 6</td></tr>
</table></td><td>Using the mean proportional AD, Galileo tabulated new times for the two original distances. A new working hypothesis had emerged in which the original time through AB at 10 degrees of speed was again 4 hours, but the time through AC at 15 degrees of speed had become 6 hours, the mean proportional of the distances from the rest. It happens that if two objects fall from rest through distances that have the ratio 4 : 9, their respective average speeds do have the ratio 10 : 15. In other words, like the times (4 hours and 6 hours), the average speeds have the ratio 2 : 3. The entries for BE and CF, made later, represent the acquired speeds at the end of the falls AB and AC and are exact doubles of the speeds previously assigned through those distances.</td></tr>
<tr><td>(6)


As BA to AD, let DA be to AC, and let BE be the degree of speed at B; and as BA to AD, let BE be to CF; CF will be the degree of speed at C. And since as CA is to AD, so CF is to BE, then as the square of AC to the square of AD, so will be the square of CF to the square of BE; and further, since as the square of CA to the square of AD, so CA is to AB, the square of CF will be to the square of BE as CA is to AB. Therefore the points E and F are in a parabola passing through A.</td><td>Galileo next drew the line AE and placed point E so that BE would represent the speed acquired at B. He wrote out his conclusion, assuming that the ratio of the acquired speeds would be the same as the ratio of the overall speeds through falls from rest. He expected that other termini of horizontal lines representing acquired speeds would fall on line AE, just as the terminus of BE did. When he calculated the placement of point F by the ratio he had developed in his conclusion (BA is to AD as BE is to CF), he found that F lay not on line AE but on a parabola through points A and E. The values BE 20 and CF 30 later added to his tabulation in Statement 5 would increase the horizontal scale of his diagram by about five to one as compared with the vertical scale.</td></tr>
</table>

GALILEO'S CALCULATIONS (*left column*) for the law of free fall are explained in detail (*right column*). Portions of Galileo's calculations in color do not appear on *f. 152r* but are inserted for clarity. The term "degrees of speed" is an arbitrary one and is used in much the same way that doctors refer to degrees of burn injury. Galileo's conclusion is analyzed in the chart on page 88.

this discovery incidentally in testing an earlier (and mistaken) belief of his that acceleration is only a temporary event at the very beginning of motion.

Alternatively, it seemed to me that Galileo might have arrived at the odd-number rule by pure reasoning, as Christiaan Huygens did many years later. For example, suppose that acceleration adds an equal increment of distance in each equal time. Then in the sequence of numbers representing spaces the ratio of the first number to the second number must be the same as the ratio of the first two numbers to the second two numbers, which in turn must be the same as the ratio of the first three numbers to the second three numbers, and so on. Why must these equalities be preserved? Because we have arbitrarily chosen to use a certain unit of time, and we might have used its double or its triple instead.

Since the number representing each distance must be uniformly larger than the preceding one, the numbers will have to be in arithmetic progression. Does such a progression exist? It is certainly not the progression of consecutive integers. The ratio of the first two integers (1 and 2) is $1 : 2$. The ratio of the first two integers added together $(1 + 2)$ to the second two added together $(3 + 4)$ is $3 : 7$, which is clearly not the same as $1 : 2$.

The progression of the odd integers alone, however, does follow the rule we have set up. First, each number is uniformly two greater than the preceding one $(1 + 2 = 3,\ 3 + 2 = 5,\ 5 + 2 = 7$, and so on). Second, the sum of the first two numbers $(1 + 3)$ has the ratio of $1 : 3$ to the sum of the next two numbers $(5 + 7)$, and $1 + 3 + 5$ has the same ratio $(1 : 3)$ to $7 + 9 + 11$, and so on. Moreover, the progression of odd integers is the only arithmetic progression that meets these conditions, as Galileo pointed out in 1615 to the same friend to whom he was writing in 1639. (There are plenty of other number sequences, such as 1, 7, 19, 37, 61, 91 . . . , that meet the ratio test above, but in no such sequence is each number uniformly greater than the preceding one.)

Alas! My two earlier suggestions must now be rejected along with the notion that Galileo got his idea for the law of free fall from medieval writings. The document I shall present shows no more trace of experimental evidence or of arithmetical reasoning than it does of the Merton mean-speed rule. Moreover, that document cannot be dated after October, 1604, when Galileo wrote to his friend Paolo Sarpi in Venice clearly stating the times-squared law and also worked out a curious proof of it. In the letter he remarked that he had known the law for some time but had lacked any indubitable principle from which to prove it. Now, he believed, he had found a very plausible principle, which was borne out by his observations of pile drivers, namely that the speed of an object is proportional to the distance it has fallen. In the ordinary meaning of the words this statement is simply false.

The proof Galileo gave Sarpi has been a headache to historians of science for a long time. In his later *Discourses on Two New Sciences* Galileo correctly stated that the speed of a falling object is proportional to the time of fall. We could forgive him for having started with the wrong assumption and having found the right one before he had published his account. That does not, however, seem to be what happened. Apparently he had reached the right assumption earlier and had then written out for Sarpi a demonstration based on the wrong one.

Even that is not the worst of the puzzle. Galileo candidly admitted in his last book that he had long believed it made no difference which assumption was chosen. Now, if Galileo ever actually believed such seeming nonsense, there should be among his notes at least one instance where he made some obvious mistake traceable to the wrong choice. Yet no such mistake is found, even in notes that can be positively dated to the long period during which he adhered to the proof composed for Sarpi. (Even 10 years later he had a version of that proof copied for use in a book on motion that he intended to publish.) Galileo was either pulling the reader's leg in 1638 by confessing to a pretended error, or we are missing some essential point in what he said in 1638 and have been misinterpreting his earlier demonstration concerning the pile drivers, although both seem so clear as to be incapable of being misunderstood.

The answer to all these puzzles now shows that Galileo was no less candid in 1638 than he had been ingenious in 1604. What we have been missing is at last revealed by a document designated *f. 152r* in Volume 72 of the Galilean manuscripts preserved at the National Central Library in Florence [*see illustration on page 84*]. I do not think anyone could have guessed the real answer. Furthermore, if anyone had guessed it, his conjecture would have been laughed at in the absence of a document in its support.

How could such a document have escaped notice for so long? All Galileo's notes were edited and published around the turn of this century. The distinguished editor, Antonio Favaro, omitted only those sheets (and portions of sheets) that contain nothing but diagrams and calculations whose meaning was uncertain. There are many such sheets; Galileo was a paper saver, and until the time of his death he kept many calculations he had made 40 years earlier. So when Favaro published *f. 152r,* he retained only two coherent sentences at the right center and bottom center of the sheet, together with a much modified version of the diagrams. These extracts meant little unless they could be dated. Favaro despaired of putting the 160 sheets of Volume 72 in the order of their composition after they had been chaotically bound together long ago.

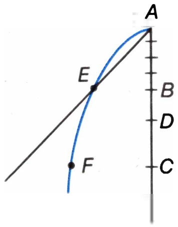
Through the generosity of the John Simon Guggenheim Memorial Foundation and the University of Toronto, I was enabled to spend the first three months of 1972 studying the manuscripts in Florence. My purpose was to attempt a chronological arrangement in order to annotate a new English translation of the *Discourses on Two New Sciences.* Like Galileo, I was lucky. It turned out that the watermarks in the paper he had used provided an essential clue. He had lived at Padua until the middle of 1610 and then had moved back to Florence. It was therefore reasonable to expect that there would have been a change in his source of paper. An inspection of his dated correspondence revealed that there was no duplication of watermarks between the letters written from the two cities. Watermarks on the undated sheets thus separate the earlier ones from the later ones.

It also happens that 40 of the 160 sheets in Volume 72 were copies in the handwriting of two of Galileo's pupil-assistants in Florence. Every one of the 40 sheets shows the same watermark, one that also appears on Galileo's letters between 1615 and 1618. It is evident that the copies were made at his house under his direction in the writing of a book on motion. Most of the originals survive, and their watermarks confirm the Paduan origin of the theorems copied, whereas the copies supply some early theorems of which the originals are lost.

A probable order of Galileo's notes thus began to emerge. The watermark evidence sorted out Galileo's early work. Additional clues were provided by his handwriting. Finding such clues presented a harder problem to the skilled editor of 50 years ago than to the relative amateur of today. He could have compared samples of handwriting only by leafing back and forth in a bound volume, whereas I could work with a Xerox copy made from microfilm. I could therefore not only place sheets side by side but also cut them up for the closer comparison of individual entries.

At that stage Galileo's pages of calculations began to fit in order, enabling me to recognize the origin of the undistinguished-looking *f. 152r.* It is pretty certain that the document is Galileo's first attempt at the mathematics of acceleration. His train of thought has now been traced out, and it is summarized in the chart on page 86.

The first two lines on the sheet, written neatly across the top and centered, certainly have nothing to do with any actual experiment. The units–miles, hours and "degrees of speed"–are quite arbitrary. Acceleration in free fall was

GALILEO'S CONCLUSION IN EUCLIDEAN RATIOS	GALILEO'S CONCLUSION REPHRASED IN MODERN TERMINOLOGY
	New definitions: s_1 denotes distance *AB* s_2 denotes distance *AC* $\sqrt{s_1s_2}$ denotes *AD*, the mean proportional of *AB, AC* t_1 denotes the time *AB* for an object to fall through *AB* t_2 denotes the time *AD* for an object to fall through *AC* v_1 denotes the velocity (speed) *BE* acquired by an object at the end of the fall s_1 v_2 denotes the velocity (speed) *CF* acquired by an object at the end of the fall s_2
By construction of the diagram, *AB* : *AD* : : *AD* : *AC*.	$s_1/\sqrt{s_1s_2} = \sqrt{s_1s_2}/s_2$ by the definition of a mean proportional.
Let *BE* be the degree of speed at *B*. Then *CF* will be the degree of speed at *C* if we let (7*a*) *AB* : *AD* : : *BE* : *CF*. When each of two ratios is equal to a third, they are equal to each other, so that *AD* : *AC* : : *BE* : *CF*.	Let v_1 be the velocity of an object at the end of s_1. Then v_2 is the velocity at the end of s_2 if we let (7*b*) $s_1/\sqrt{s_1s_2} = v_1/v_2$. From the above statements it can be said that $\sqrt{s_1s_2}/s_2 = v_1/v_2$.
It follows that if both sides are squared, $(AD)^2 : (AC)^2 : : (BE)^2 : (CF)^2$.	It follows that if both sides are squared, $(\sqrt{s_1s_2})^2/s_2^2 = v_1^2/v_2^2$.
Furthermore, since *AD* is the mean proportional of *AB* and *AC* and $(AD)^2 : (AC)^2 : : AB : AC$, and since ratios equal to the same ratio are equal, then (8*a*) $(BE)^2 : (CF)^2 : : AB : AC$.	Furthermore, since $(\sqrt{s_1s_2})^2/s_2^2 = s_1s_2/s_2^2 = s_1/s_2$, then this statement with the previous one gives (8*b*) $v_1^2/v_2^2 = s_1/s_2$.
Therefore the points *E* and *F* are on a parabola passing through *A*.	Although Galileo did not expressly mention time in his conclusion, he had tabulated the times t_1 and t_2 (*AB* and *AD*) through the falls s_1 and s_2 (*AB* and *AC*) in such a way that they exhibited the ratio $t_1/t_2 = s_1/\sqrt{s_1s_2}$, and that is precisely how he habitually spoke of times and calculated with them. Thus it is not an exaggeration to say that a direct implication of Statement 7*b* is $t_1/t_2 = v_1/v_2$. This, together with Statement 8*b*, at once implies (9) $s_1/s_2 = t_1^2/t_2^2$. In other words, the ratio of the distances traversed by two falling objects is equal to the ratio of the squares of the times from rest. Therefore if the velocities, which are proportional to the times, are plotted against the distances on Galileo's original diagram, the result would be a parabola, as he said. (A parabola results whenever one variable is proportional to the square of the other.)

GALILEO'S CONCLUSION is expressed in his own terms of Euclidean ratios (*left column*) and is rephrased in modern notation (*right column*). Euclidean ratios, such as "*AB* : *AD* : : *AD* : *AC*," are read in the manner "*AB* is to *AD* as *AD* is to *AC*." The order of the quantities in some of his ratios has been changed for consistency when this does not interfere with his train of thought.

what basically interested Galileo, but his first step was to seek a general rule of proportionality for uniform growth of distances, times and speeds. In his first working hypothesis he assumed for the sake of argument that one degree of speed was gained for each unit of distance fallen. Thus he assumed that four miles were traversed with one degree of speed consumed in the first unit of distance, two in the second, three in the third and four in the fourth. This gave him four miles with 10 degrees ($1 + 2 + 3 + 4$) of speed consumed. He arbitrarily put one hour as the time elapsed for each different degree of speed. Consistent times for each speed (or distance) could not be, and did not need to be, determined in the initial working hypothesis.

Galileo next wrote: "9 miles with 15 of speed in 5? hours." The question mark is Galileo's, not mine; it indicates that this time was to be examined, as we would today write "*x* hours." I mention the point because in the original manuscript the number looks more like a 1 than a 5, as 5's written in those days often do. This second statement was doubtless meant to be a part of the same working hypothesis as the first, Galileo's purpose being to have two different examples in order to compare ratios. In actuality the two statements are not even consistent. If the increment of speed to be added next was five degrees, bringing the total to 15, then the number of miles to be added was one, not five. I believe the ambiguity of the rule "One more degree of speed for each additional mile" tricked Galileo into adding five miles for the new five degrees of speed. The slip was careless, but it was not fatal; far from it. As James Joyce remarked, a man of genius makes no mistakes; his errors are portals of discovery.

Now that Galileo had two distances, two speeds and two times to work with, he proceeded to apply the Euclidean theory of proportion, which was the only device he trusted for the application of mathematics to physics. His first step was to reduce both hypothetical motions to the same speed in order to compare ratios of distances and times. Accordingly he wrote 6 above the 4 in the phrase "4 hours" of his first statement. In this way he redefined "10 [degrees] of speed" so that four miles at that speed would take six hours rather than four. In order to cover the four miles in four hours, then, a body would have to travel at 15 degrees of speed. Galileo noted this fact to the left of his original statements and continued with its implication: If the meaning of "15 degrees of speed" is "going 4 miles in 4 hours," then in order to cover eight miles at 15 degrees of speed the same body would take eight hours. (This would be true of any motion, however irregular, if the meaning of "speed" is fixed by total time and total distance.) That, however, immediately contradicted his earlier hypothesis that nine miles were traversed at 15 degrees of speed in only five hours.

Here it will be useful to recall that all Galileo was trying to do at this stage was to find a mathematically consistent rule for the use of the phrase "degrees of speed." He was not yet concerned with attaching any physical meaning to the phrase, a task that would seem pretty pointless to him until its use in ratios was first made possible. In order for the phrase to be useful in ratios, any given overall speed must carry the moving body through proportional distances in proportional times.

Faced with an apparent contradiction from a working hypothesis that he (mistakenly) thought he had expressed consistently in his first two statements, Galileo did not go back to see whether or not he had made an error. If he had discovered his error and corrected it, he would have ended up with the consistent but useless formula "4 miles at 10 degrees of speed in 4 hours; 5 miles at 15 degrees of speed in 5 hours." Such a formula would have equalized the ratios of distances and of times under acceleration, contradicting good sense and the basic idea of acceleration itself, and it would have told him nothing at all about ratios of speeds. In any case he did not check back. Instead he looked for the source of the trouble by drawing a vertical line and lettering it *A*, *B* and *C* to represent distances from rest. He indicated four units from *A* and *B*. The distance from *B* to *C* is supposed to be five units, making *AC* equal to nine units, as is shown in his tabulation.

Quite by accident, in my opinion, the two distances in Galileo's working hypothesis were both square numbers: 4 and 9. If Galileo had had any previous inkling of a rule involving squares and square roots when he wrote his first two statements, he would surely have put in the numbers 2 and 3, either for the speeds or for the times. I have already noted three ways in which he might have suspected that a square-root law existed, and there may be still other ways. His choice of numbers here, however, seems to me to exclude all of them. I believe that he first discovered the squaring relation in precisely this search for consistent ratios, and that he confirmed it afterward by experimental test.

It is the fact that 4 and 9 both happen to be square numbers that accounts for Galileo's addition of point *D* between *B* and *C* in his diagram. The distance *AD* is equal to six units. So far Galileo's problem had been one of conflicting ratios, and this was a difficulty that could be forever eliminated by introducing continued proportion. To the mind of any mathematician of the time, continued proportion was immediately suggested by the squares of two integers. Between any two such numbers there is always an integral mean proportional that is the product of the two square roots. The ratio of the lesser square to the mean proportional is then equal to the ratio of the mean proportional to the greater square.

In Galileo's case the mean proportional is equal to the product of the square root of 4 and the square root of 9, that is, the product of 2 and 3, which is 6. The ratio $4 : 6$ is equal to the ratio $6 : 9$, that is, $2 : 3$. Galileo entered point *D* on his vertical line six units distant from *A* just because it created a continued proportion. It then occurred to him to use the mean proportional to solve the puzzle of the time ratios. If the distance *AB* represented the time in his first statement (4 hours), then *AD* (6) represented the time in the second statement (originally 5? hours.) From the mean proportional a new working hypothesis had emerged: The original time through the shorter distance *AB* at 10 degrees of speed was again four hours, but now the time through the longer distance *AC* at 15 degrees of speed had become six hours, the mean proportional of the distances from rest.

From that day forward Galileo had in his hands the principal analytical device that he applied in all his reasoning about free fall. There is no logic to the foregoing procedure except for the logic of discovery. Galileo perceived that this arrangement of numbers would preserve a consistency of ratios, and that was the necessary first move. Whether or not the rule would also agree with observable facts he left until later; in this, as he later remarked, he had been lucky.

I have said that he was just lucky in having two square numbers to start with. If that is so, can anything else be said of the ratio $10 : 15$ chosen for the speeds assigned to those distances at the outset of the inquiry? When two objects fall from rest through distances that have the ratio $4 : 9$, their respective average speeds do have the ratio $10 : 15$. In other words, the speeds, like the times, are in the ratio 2 : 3. If, however, Galileo had started with any other two numbers, even two squares other than 4 and 9, then neither the ratio 10 : 15 nor the ratio of numbers obtained by adding up "degrees of speed consumed" would have agreed with the ratio of the smaller number to the mean proportional between the two numbers selected.

Two such lucky coincidences may strain the reader's credulity. All I can say is that I have not succeeded in finding any better—or indeed any other—reconstruction that will account reasonably well for all the entries on *f. 152r*. Moreover, coincidences among very small numbers are not as improbable as they may seem.

Another line from point *A* stretches off to the left at an angle from the original vertical line. Galileo probably intended it to represent the speeds acquired in acceleration. These speeds at the end of the falls *AB* and *AC* happen to be exact doubles of the speeds properly assigned *through* those distances. Galileo seems to have not yet been aware of that relation when he placed point *E* on his slanted line so that the distance *BE* would represent the speed acquired at *B* and wrote out his long conclusion. He did assume, however, that the ratio of acquired speeds would be the same as the ratio of his "overall" speeds through falls from rest. He apparently drew the slanted line in the expectation that other termini of horizontal lines representing acquired speeds would fall on it just as the terminus of *BE* did. Here a surprise was in store for him. When he calculated the length of *CF* (the speed acquired through *AC*) according to the ratio he had developed in his conclusion, he found that point *F* lay not on the slanted line *AE* but on a parabola through points *A* and *E*.

Galileo's conclusion on *f. 152r* is in a way the starting point of the modern era in physics. In it he correctly related the acquired speeds to the mean proportional of distances from rest, and he obtained explicitly or implicitly all the essential rules governing acceleration in free fall. Galileo's conclusion is analyzed in his terms of Euclidean ratios of proportion and in 20th-century terminology in the chart on page 88.

Galileo's conclusion on *f. 152r* did not mention time as such. Here I wish to point out once more that the only acceleration he ever applied was the acceleration of free fall. Thus whereas we generally think of acceleration as a function of two variables, distance and time, Galileo's acceleration was completely determined by either one. A rule for speeds in terms of distances, assuming one universal constant acceleration, left no freedom of choice of times. The relations Galileo expressed were thus complete and correct; *f. 152r* implied all the significant relations of distance, time and speed that exist in free fall.

Although Galileo did not expressly mention time in his conclusion, he had explicitly tabulated the times through the two distances from rest using a mean-proportional relation between those distances, and that is precisely how Galileo

COMPARISON OF SARPI PROOF WITH PROOF IN *f. 152r*

Galileo arranged matters so that the points representing speeds on his diagram would fall on a straight line rather than on a parabola. All he had to do was replace his original ratio $v_1^2/v_2^2 = s_1/s_2$ with the ratio $V_1/V_2 = s_1/s_2$, where $V_1 = v_1^2$ and $V_2 = v_2^2$. In his original proof of the law of free fall for Paolo Sarpi, Galileo often applied the phrase "degrees of speed" to speed acquired at a point but not to speed through a distance. He considered speed through a distance in the Sarpi proof as being proportional to V^2 (instead of proportional to v, as he had done in *f. 152r*). Thus he gave the phrase two senses: In *f. 152r* "speed through a distance" was an overall speed, a kind of average treated as being proportional to terminal speed. It is denoted w below. In the Sarpi proof "speed through a distance," which is denoted W, represents Galileo's provisional idea of a "total speed," for which we have no word or concept.

Using these convenient denotations (as well as the previous ones s, t and v), Galileo's two treatments can be displayed along with their modern counterpart:

f. 152r	*Modern usage*	*Proof for Sarpi*
$v_1/v_2 = w_1/w_2$	$v_1/v_2 = t_1/t_2$	$V_1/V_2 = s_1/s_2$
$w_1/w_2 = t_1/t_2$	$s_1/s_2 = v_1t_1/v_2t_2$	$W_1/W_2 = V_1^2/V_2^2$
(8*b*) $s_1/s_2 = v_1^2/v_2^2$	$s_1/s_2 = v_1^2/v_2^2$	W "in contrary proportionality to t," that is $\sqrt{W_1}/\sqrt{W_2} = t_1^2/t_2^2$
(9) $s_1/s_2 = t_1^2/t_2^2$	$s_1/s_2 = t_1^2/t_2^2$	$s_1/s_2 = t_1^2/t_2^2$

Galileo's choice of $V_1/V_2 = s_1/s_2$ in the Sarpi proof does not mean that he had to reject $v_1^2/v_2^2 = s_1/s_2$ (Statement 8*b*). It did mean, however, that *velocità*, defined for Sarpi as being proportional to distances traversed, could not be the same physical entity as the "speed" that had been implicitly defined as being proportional to time (the times-squared law) in *f. 152r*. The basic relations that can be derived from Galileo's two treatments are:

For *f. 152r*:	$v_1/v_2 = t_1/t_2$	$v_1^2/v_2^2 = s_1/s_2$	$s_1/s_2 = t_1^2/t_2^2$
For Sarpi:	$V_1/V_2 = t_1^2/t_2^2$	$V_1/V_2 = s_1/s_2$	$s_1/s_2 = t_1^2/t_2^2$

Galileo sought a proof for only the third and last relation, which remained unchanged by his new definition of "velocity." He believed that he could derive this law as well from his new definition $V_1/V_2 = s_1/s_2$ as from the old one $v_1/v_2 = t_1/t_2$, and he did. Later he found an experimental reason for adopting only the latter (older) definition of "velocities" in free fall.

"Contrary proportionality" in the proof for Sarpi relates the root of one variable to the square of the other. Since Galileo concluded from $W_1/W_2 = s_1^2/s_2^2$ that $s_1/s_2 = t_1^2/t_2^2$ by invoking "contrary proportionality," then his phrase meant that $\sqrt{W_1}/\sqrt{W_2} = t_1^2/t_2^2$. This relation is not derivable from *f. 152r* alone, since there $w_1/w_2 = v_1/v_2 = t_1/t_2 = \sqrt{s_1}/\sqrt{s_2}$. Yet the relation does arise from *f. 152r* and the proof for Sarpi *taken together*, that is, from the simultaneous assumption that $W_1/W_2 = V_1^2/V_2^2$ and $w_1/w_2 = v_1/v_2$. Hence it was *only* if Galileo had already perceived the fact that speeds in free fall are proportional to times ($v_1/v_2 = t_1/t_2$), and retained that knowledge when he was writing the proof for Sarpi, that he could validly invoke "contrary proportionality." Only on this same basis can we understand the fact that in all his notes and books Galileo never made use of the assumption that our ordinary velocity is proportional to distance ($v_1/v_2 = s_1/s_2$), as he has often been accused of doing.

PROOF OF THE LAW OF FREE FALL that Galileo wrote in a letter to his friend Paolo Sarpi can be analyzed in modern terms. The quantities s_1, s_2, v_1, v_2, t_1 and t_2 retain the same meaning that they have in the chart on page 88. To Galileo the relation $v_1/v_2 = \sqrt{s_1}/\sqrt{s_2}$ seemed anything but certain. He was dissatisfied that the velocities of falling bodies, when plotted against the distances they had fallen, followed a parabola instead of a straight line. The Sarpi proof redefined "velocity" by taking its square in order to make the velocities fall along a straight line and to justify Galileo's times-squared law of free fall ($s_1/s_2 = t_1^2/t_2^2$). The proof does not, as has been supposed up until now, begin by assuming that the velocities of falling bodies are proportional to the distances through which they have fallen.

habitually spoke of times and calculated with them in the rest of his notes and published books. Hence I do not think it is exaggerating to say that a direct implication of *f. 152r* is that the ratio of the speeds at two points in fall from rest is also the ratio of the elapsed times of fall. In modern terminology we write this ratio as $v_1/v_2 = t_1/t_2$, where v_1 and v_2 are velocities at two points and t_1 and t_2 are the respective times that have elapsed from rest. This ratio at once implies that the ratio of the distances is equal to the ratio of the times squared.

That Galileo himself saw the implication is indicated by the fact that the times-squared law was the very proposition he offered to prove for Sarpi, saying he had known it for some time but had lacked any indubitable principle from which to demonstrate it. The obvious conclusion to be drawn is that in 1604 Galileo did not consider a simple statement equating ratios of times and speeds under acceleration to be of such a nature as to be acceptable as an "indubitable principle."

On the contrary, in writing for Sarpi he took as his principle the seemingly contradictory assumption that speeds acquired are directly proportional to distances fallen from rest. Now, if we take the word "speed" here in the modern sense, which it had had in *f. 152r* and was to have later in *Discourses on Two New Sciences,* then the principle assumed for Sarpi was not only false but also inconsistent with Galileo's own conclusion in *f. 152r,* where it is not the speeds but the squares of the speeds that are proportional to the distances. Thus it must seem at this point that we have no choice but to say that Galileo first knew the right answer and then turned his back on it in favor of the wrong one, and that only years later did he return to the position that had been implied in his very first try at the mathematics of acceleration. That is even worse than what all historians of science have been saying up to the present.

I said "It must seem at this point" because things are not going to turn out that way at all. We have already seen what a very chancy business Galileo's investigations on *f. 152r* had been. His conclusion entirely lacked objective evidence, nor is there any reason that we should think Galileo remained unaware of the flimsy character of his first stab at a perplexing problem. Even if we now know, in a kind of absolute way, that his first result was the right one, there was nothing sacred about it to Galileo, who had arrived at it merely as an exercise in proportionality. He would have had every right to turn his back on it for something better, if indeed he had done so, which he did not. All that the result on *f. 152r* represented to Galileo was an internally consistent meaning for "speed," and for all he knew it might just be one among many.

The result that the ratio of the distances was proportional to the ratio of the velocities squared through those distances ($s_1/s_2 = v_1^2/v_2^2$) was anything but certain. There was as yet no way to measure a speed directly, and the square of a speed was hard to even make physical sense of. Even if the result was rewritten so that the ratio of the speeds was proportional to the ratio of the square roots of the distances ($v_1/v_2 = \sqrt{s_1}/\sqrt{s_2}$), it did not even look probable. How would a speed go about adjusting itself to a quantity represented by a distance through which it had already passed?

Galileo's other relation, that the distances fallen were proportional to the squares of the times through those distances ($s_1/s_2 = t_1^2/t_2^2$), may also have been puzzling, but it differed from the others in one respect: it could be physically verified. Galileo proceeded to verify it, probably in the way I formerly thought he might have discovered it, and in his letter to Sarpi he mentioned the odd-number rule that holds for successive spaces in equal times. Hence when he wrote to Sarpi, Galileo was far more certain of the truth of the times-squared law than he was of the validity of the derivation that had led him to it.

The most surprising implication of the documents, however, is this: Galileo never did adopt distance-proportionality for speeds in acceleration except as a temporary working hypothesis, and it led him at once to time-proportionality. We have been wrong in supposing that he rejected time-proportionality in favor of distance-proportionality in his proof for Sarpi late in 1604, in spite of its wording. Moreover, what he told the public in 1638 was the literal historical truth: For a long time he had thought it was a matter of indifference whether speeds in free fall were defined as proportional to the elapsed times or were related to the distances traversed. He merely neglected to add the statement, perhaps supposing that it was as obvious to others as it was to him: "Provided that the rest of your mathematics remains unaltered by your choice of assumption." It now remains for us to see how Galileo himself managed this, and to consider what it tells us about what he thought was the physical meaning of the word "speed."

We can start by noting what seems at first to be the most sophisticated thing in *f. 152r:* Galileo's reference to the fact that under the normal definition, points representing speeds would fall on a parabola when they are plotted against distances of fall. If we had no idea of the date of this entry and only knew about Galileo's later achievements respecting the parabolic trajectory, we might imagine that this remark greatly pleased Galileo. Although his work on the trajectory was still four or five years in the future when he wrote his conclusion in *f. 152r,* he already knew a great deal about parabolas. His first paper in 1587 dealt with the centers of gravity of paraboloids, and he applied for a chair of mathematics on the strength of it. The parabola as such did not dismay him. In Galileo's lifelong view, however, nature always acts in the simplest way. Since the simplest rule would put the point *F* on the straight line *AE,* I fancy that Galileo wrote the last words of *f. 152r* not with the joy of discovery but with something like disgust, and that he regarded them as casting serious doubt on the reliability of this chancy speculation about acceleration in free fall.

It was not hard, however, to rearrange matters so that the points representing speeds acquired by a falling body would fall on a straight line. All Galileo had to do was replace his mean-proportional treatment of speeds in his conclusion with a linear treatment, requiring only a change in his definition of "speed." The substitution would simply make $s_1/s_2 = V_1/V_2$, where the new "velocity" V_1 represents v_1^2 and V_2 represents v_2^2. The relation between velocity and distance would then become linear instead of parabolic. Careful examination shows that Galileo did exactly this when he composed the demonstration for Sarpi. He also found a physical justification for the new definition of velocity.

One thing that historians of science have all overlooked is that no definition of "speeds" in acceleration had ever been clearly made in terms of ratios of distances and of times. Archimedes had done it only for the case of uniform motion. Galileo took it from there, seeking a ratio definition for the ever changing speeds under acceleration. The Merton rule that represented overall velocities by their mean speeds threw no light on the problem of these ratios but ingeniously circumvented it. When Galileo tackled it, no physical measure for each and every speed in acceleration had ever been assigned. He therefore felt free to define the measure of speed in any way he pleased as long as experience bore him out, or at least did not contradict him. I might add that it was a long time, probably more than 20 years, before Galileo realized what the medieval writers had always assumed, namely that there does exist a uniform motion equivalent to any uniformly accelerated motion from rest. No trace of this realization is to be found in his manuscript notes. It first appeared as Theorem I in Book II of the Third Day in his *Discourses on Two New Sciences,* a theorem that, with all due respect to Professor Grant, does not employ any mean speed to represent accelerated motion in free fall.

In his letter to Sarpi, Galileo observed that pile drivers strike twice as hard when the weight falls twice as far. That remark shows that what Galileo meant by his word *velocità* in the Sarpi proof could not be the same as what he had meant by *gradus velocitatis* in *f. 152r.* In *f. 152r* the speeds would be in the ratio of $\sqrt{2} : 1$ at the end of falls in the ratio of $2 : 1$. In the conclusion of *f. 152r* Galileo essentially stated that the velocities were proportional to the square roots of the distances from rest ($v_1/v_2 = \sqrt{s_1}/\sqrt{s_2}$), so that speeds in the ratio of $\sqrt{2} : 1$ would be found at the end of falls whose lengths were in the ratio of $2 : 1$. In the letter to Sarpi, however, Galileo appealed to observations of pile drivers. The effect of a pile driver, involving kinetic energy, is governed not by speed but by its square. Hence the effect does not support a distance-proportionality assumption in the sense that we have always imputed to Galileo's words. What it does support is the mean-proportional relation between distances and speeds derived in *f. 152r* and altered to linear form by Galileo in the Sarpi proof by a simple redefinition of "velocity." In modern terms the Sarpi proof is equivalent to the proof of *f. 152r* with the substitution of V_1 for v_1^2 and V_2 for v_2^2. Galileo did this, I believe, in order to place the acquired *velocità* along a straight line rather than on a parabola. With the transformation Galileo's earlier ratio $s_1/s_2 = v_1^2/v_2^2$ becomes $s_1/s_2 = V_1/V_2$ [*see illustration on page 90*].

Galileo's choice of the ratio $s_1/s_2 = V_1/V_2$ when he wrote to Sarpi did not mean, as historians have all previously supposed, that he had to reject his original relation $s_1/s_2 = v_1^2/v_2^2$. It did mean, however, that the *velocità* defined for Sarpi as being proportional to distance could no longer be the same physical entity as the "speed" implicitly defined as proportional to time in *f. 152r.* The times-squared law, however, remained unchanged by Galileo's new definition of *velocità.* He believed that he could derive the times-squared law as well from $s_1/s_2 = V_1/V_2$ as he could from $v_1/v_2 = t_1/t_2$. In fact he could, and he did. Later he found an experimental reason for adopting only the second (older) relation and defining "velocity" as we do.

It is easy enough to say at this point: "But there must have been some criterion of choice at every stage, since in fact the speeds in free fall do not increase according to distances fallen but do increase according to elapsed times." This statement, however, assumes a physical definition of "speed." What we know as "speeds" do indeed increase in that way. In Galileo's proof for Sarpi, however, he chose to use something else, based on his observations of pile drivers, which is the square of our notion of speed. Let us call Galileo's choice *velocità.* We could not very well argue that *velocità* fails to increase according to the distance fallen, basing our argument on the fact that *velocità* in Italian means, after all, the same thing as "speed" in English. It would be necessary to change the usual methods of measuring time and distance to make *velocità* behave like our speed. Galileo made no alteration in his relations of time and distance, however, so that his *velocità* behaved differently. The worst we can say of his definition is that we prefer another one: the same one that he himself later adopted. Galileo reasoned that the effect of a pile driver could change only when the falling weight acquired a different speed. In effect he decided, for a time, to think of "*velocità*" as "whatever it is that changes the striking power of a body falling from different heights." This quantity he could measure, and it does behave like our v^2. Later he found a way of directly observing speed in our sense of the word, but he had no such way in 1604.

In his proof for Sarpi, Galileo invoked a "contrary proportionality" between speeds and times, a curious relation that equates the ratio of the roots of one variable to the ratio of the squares of the other. With this tool he concluded that the square root of what he calls the "total speed" was proportional to the time squared [*see illustration on page 90*]. Did Galileo have any ground for asserting such a relation? He certainly presented none in the proof for Sarpi, since time was not mentioned there until he made his appeal to contrary proportionality. Nor can such a relation be derived from *f. 152r* alone.

Yet the relation does arise from the two documents together, and it could only arise because Galileo assumed that what we call "speed" remained proportional to time even when his new *velocità* was made proportional to distance. It was only because he assumed this when he wrote the proof for Sarpi that he could validly invoke "contrary proportionality" as he did. And that he did assume it is borne out by the solid fact that throughout the 160 sheets of notes on motion, written over a period of 30 years, there is not one single instance in which Galileo ever made use of the assumption that speeds in the ordinary sense are proportional to distances in free fall.

This, then, is the new picture: Galileo obtained the result that speeds in free fall are proportional to times elapsed from rest in the course of his very first try at the mathematics of accelerated motion, probably in the middle of 1604. He never abandoned that conception, although for a time he altered his definition of *velocità* in the interests of elegance and the simplicity of natural phenomena, supported by reasoning about an observed phenomenon of kinetic energy. Ultimately a classical experiment, still unpublished (*f. 116*), induced him to reject the alternative definition.

We do not have, and we do not need, any special name for the physical quantity *velocità* of the Sarpi proof. If we did give it a name, that word (say "punch") would occur frequently in our discussions of energy, and it would seem to us that this newly named physical entity should enter such discussions directly, instead of as the square of something else. We might think of a falling body's velocity as "how fast it will go horizontally if deflected," and its "punch" as "how hard it will hit vertically." Whether there really are two physical entities (velocity and "punch") or whether there is only one entity (velocity) that enters into some calculations as itself (v) and into others as its square (v^2 or V) would be hard to decide.

This question was bound to crop up in some form after Galileo, convinced that nature had forced his hand, suppressed his alternative definition. In one form the question arose late in the 17th century, when Leibniz christened the neglect of v^2 ("punch") as "the memorable error of Descartes." After decades of heated argument the entire problem was recognized to be a merely semantic controversy.

The Infrared Receptors of Snakes

The snakes of two large families have sensitive organs that can detect the heat radiation emitted by their prey. The performance of these detectors is investigated with the aid of an infrared laser

by R. Igor Gamow and John F. Harris

A boa constrictor will respond in 35 milliseconds to diffuse infrared radiation from a carbon dioxide laser. A sensitive man-made instrument requires nearly a minute to make what is essentially the same measurement. Reflecting on this comparison, one wonders what feat of bioengineering nature has performed to make the snake's sensor so efficient. One also wonders if a better understanding of the animal's heat-sensing apparatus would provide a basis for improving the man-made ones. It was the pursuit of these questions that gave rise to the somewhat unusual situation in which a group of workers in an aerospace engineering laboratory (our laboratory at the University of Colorado) was investigating snakes.

Snakes belong to one of the four large orders that comprise the living members of the class Reptilia. The order Testudinata contains such members as the turtles, the tortoises and the terrapins. The order Crocodilia contains the crocodiles and the alligators. The third order, the Rhynchocephalia, has only one member: the tuatara of New Zealand. In the fourth order, the Squamata, are the lizards and the snakes.

On the basis of outward appearance one might suppose that the lizards are more closely related to the crocodiles and the alligators than to the snake. Evolutionary evidence, however, clearly indicates that the snakes arose from the lizard line. Although the lizard is therefore the snake's closest relative, the two animals have developed pronounced differences during the course of evolution. Most lizards have limbs and no snakes have limbs, although vestigial ones are found in certain snakes. Most lizards have two functional lungs, whereas most snakes have only one. Again a few snakes have a small second lung, which is another indication of the direction of evolution from the lizard to the snake.

Today most herpetologists would agree that the first step in the evolution of the snake occurred when the animal's ancestral form became a blind subterranean burrower. In evolving from their lizard-like form the ancestral snakes lost their limbs, their eyesight and their hearing as well as their ability to change coloration. Later, when the animals reappeared on the surface, they reevolved an entire new visual system but never regained their limbs or their sense of hearing.

Today the snakes constitute one of the most successful of living groups, being found in almost every conceivable habitat except polar regions and certain islands. They live in deep forests and in watery swamps. Some are nocturnal, others diurnal. Some occupy freshwater habitats, others marine habitats. Certain snakes are arboreal and survive by snatching bats from the air, others live in the inhospitable environment of the desert. Their success is indicated by the fact that their species, distributed among 14 families, number more than 2,700.

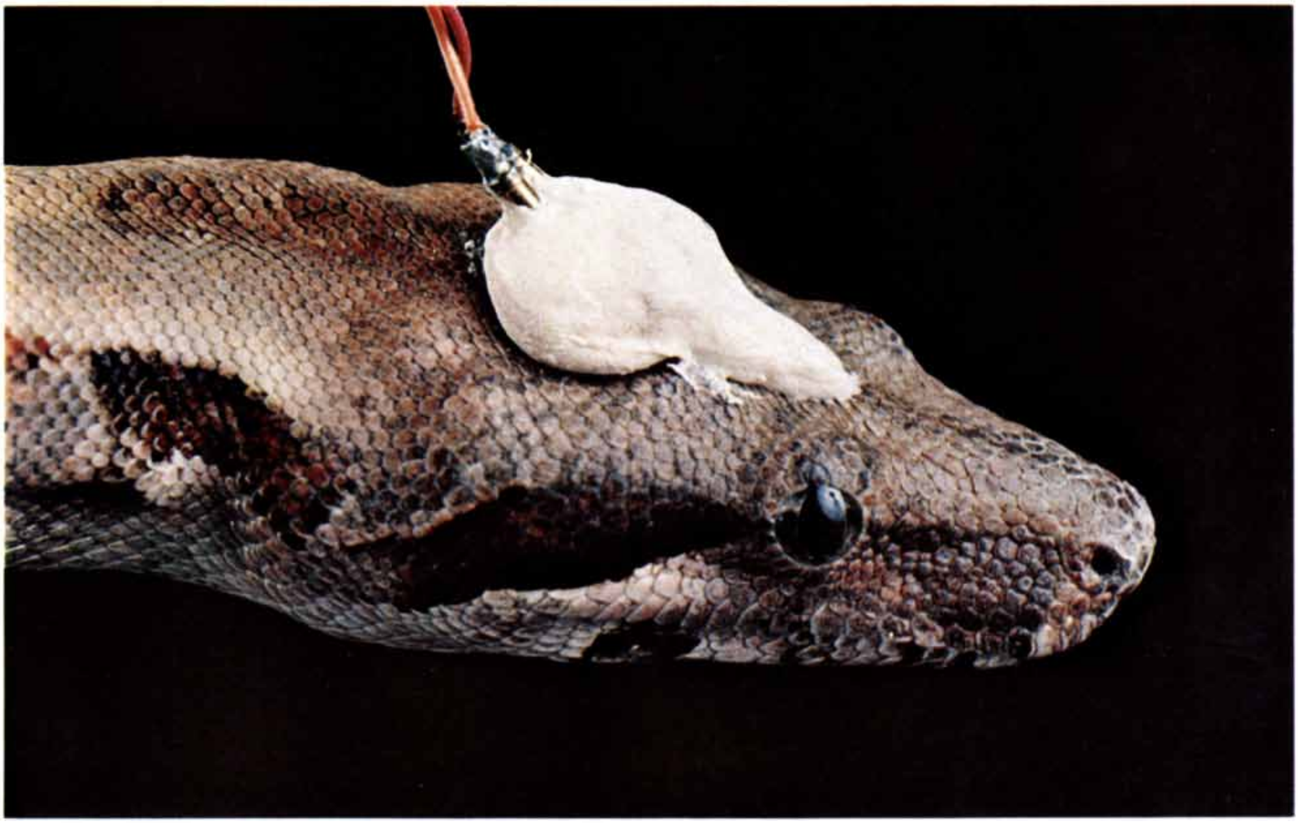
Two of the 14 families are distinguished by the fact that all their members have heat sensors that respond to minute changes of temperature in the snake's environment. The snake employs these sensors mainly to seek out and capture warm-bodied prey in the dark. It seems probable that the snake also uses the sensors to find places where it can maintain itself comfortably. Although snakes, like all reptiles, are cold-blooded, they are adept at regulating their body temperature by moving from place to place. Indeed, a snake functions well only within a rather narrow range of temperatures and must actively seek environments of the proper temperature. A case in point is the common sidewinder, which maintains its body temperature in the range between 31 and 32 degrees Celsius (87.8 and 89.6 degrees Fahrenheit). One advantage of a heat sensor is that it enables the snake to scan the temperature of the terrain around it to find the proper environment.

One of the families with heat receptors is the Crotalidae: the pit vipers, including such well-known snakes as the rattlesnake, the water moccasin and the copperhead. The other family is the Boidae, which includes such snakes as the boa constrictor, the python and the anaconda. Although all members of both families have these heat receptors, the anatomy of the receptors differs so much between the families as to make it seem likely that the two types evolved independently.

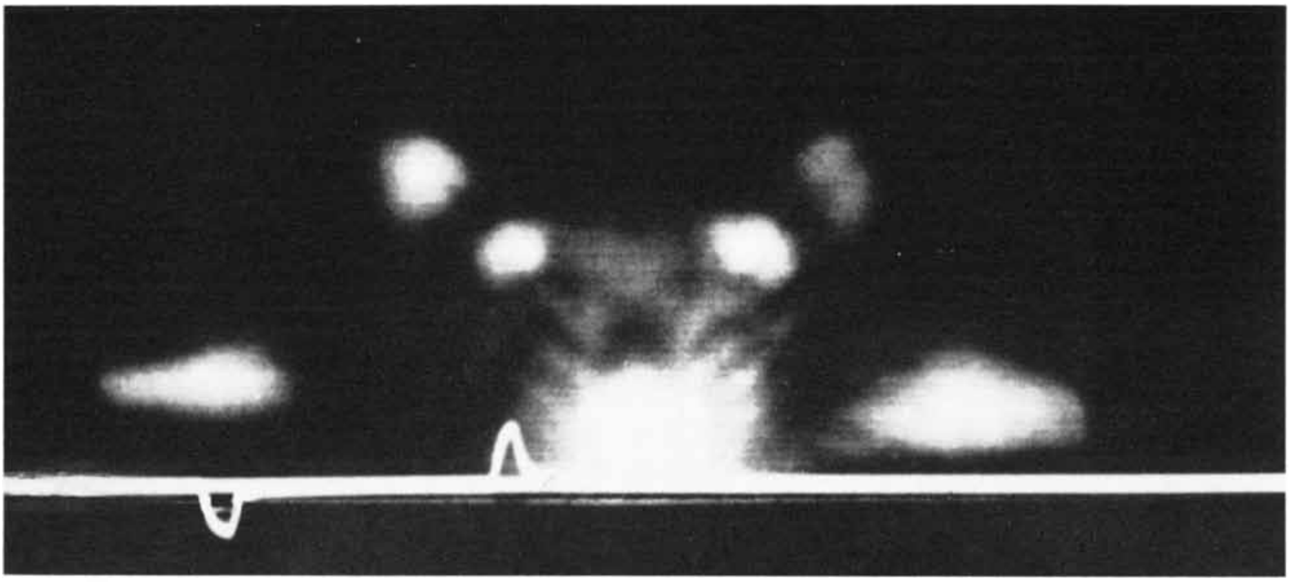
In the pit vipers the sensor is housed in the pit organ, for which these snakes are named. There are two pits; they are located between the eye and the nostril and are always facing forward. In a grown snake the pit is about five millimeters deep and several millimeters in diameter. The inner cavity of the pit is larger than the external opening.

The inner cavity itself is divided into an inner chamber and an outer one, separated by a thin membrane. A duct between the inner chamber and the skin of the snake may prevent differential changes in pressure from arising between the two chambers. Within the membrane separating the chambers two large branches of the trigeminal nerve (one of the cranial nerves) terminate. In both snake families this nerve is primarily responsible for the input from the heat sensor to the brain. Near the terminus the nerve fibers lose their sheath of myelin and fan out into a broad, flat,

GREEN PYTHON of New Guinea (*Chondropython viridis*) is a member of the family Boidae that has visible pits housing its infrared detectors. The pits extend along the jaws. Photograph was made by Richard G. Zweifel of the American Museum of Natural History.



BOA CONSTRICTOR is a boid snake with infrared detectors that are not visible externally, although they are in the same location as the green python's. This boa wears an apparatus with which the authors recorded responses of the brain to infrared stimuli.



INFRARED VIEW OF RAT suggests what a snake "sees" through its infrared detectors when it is stalking prey. Snakes with such detectors prey on birds and small mammals. This view was obtained with a Barnes thermograph, which detects infrared radiation. In a thermogram the coolest areas have the darkest appearance and the warmest areas, such as the nose of the rat, appear as white spots.

BOA IN LASER BEAM was tested in the authors' laboratory at the University of Colorado for responses to infrared radiation. The carbon dioxide laser, emitting in the infrared at 10.6 microns, appears as a glowing area in the upper right background. Its beam is spread by a lens so that the snake, even when moving about in the cage, is doused in infrared radiation delivered in occasional pulses lasting eight milliseconds each. A brain signal recorded by the electrode assembly on the snake's head goes to a preamplifier and then to an oscilloscope and to a signal-averager. Electroencephalograms recorded in this way appear in the illustration on page 100.

palmate structure. In this structure the nerve endings are packed full of the small intracellular bodies known as mitochondria. Evidence obtained recently by Richard M. Meszler of the University of Maryland with the electron microscope strongly suggests that the mitochondria change morphologically just after receiving a heat stimulus. This finding has led to the suggestion that the mitochondria themselves may constitute the primary heat receptor.

In the family Boidae there are no pit organs of this type, although somewhat different pits are often found along the snakes' upper and lower lips. Indeed, it was once thought that only the boid snakes with labial pits had heat sensors. An extensive study by Theodore H. Bullock and Robert Barrett at the University of California at San Diego has shown, however, that boid snakes without labial pits nonetheless have sensitive heat receptors. One such snake is the boa constrictor.

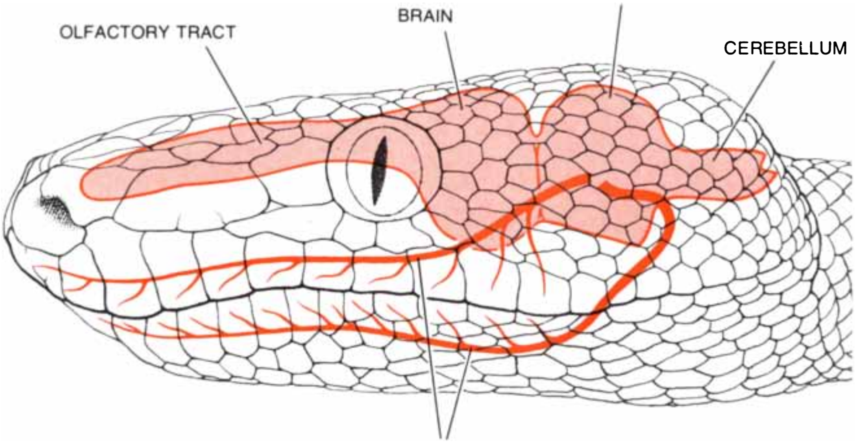
For experimental purposes the boid snakes are preferable to the pit vipers because the viper is certain to bite sooner or later, and the bite can be deadly. The boids, in contrast, can be described as friendly, and they get along well in a laboratory. When our laboratory became interested several years ago in the possibility of using an infrared laser as a tool to help unravel the secrets of the mode of operation of the snake's heat sensor, we chose to work with boid snakes.

Bullock and his collaborators have done most of the pioneering work on the heat receptors of snakes. In their original experiments, using the rattlesnake, they first anesthetized the animal and then dissected out the bundle of large nerves that constitute the main branches of the trigeminal nerve. It is these branches that receive the sensory information from the receptor.

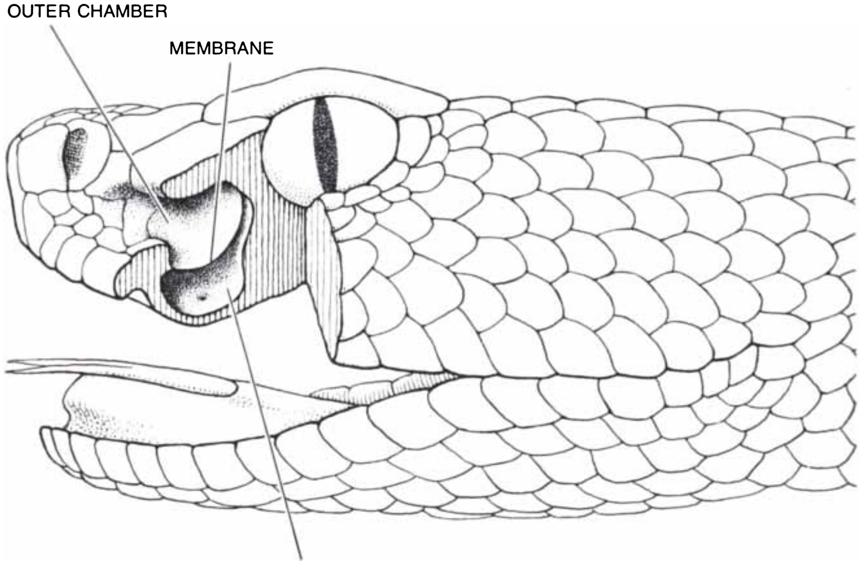
Bullock and his colleagues found by means of electrical recording that the frequency of nerve impulses increased as the receptor was warmed up and decreased as it was cooled. The changes were independent of the snake's body temperature; they were related only to changes of temperature in the environment. The Bullock group also determined that the operation of the sensor is phasic, meaning that the receptor gives a maximum response when the stimulus is initiated and that the response quickly subsides even if the stimulus is continued. (Many human receptors, such as the ones that sense pressure on the skin, are phasic; if they were not, one would be constantly conscious of such things as a wristwatch or a shirt.)

Our work was built on the foundations laid by Bullock and his associates. In addition we had in mind certain considerations about electromagnetic receptors in general. Biological systems utilize electromagnetic radiation both as a source of information and as a primary source of energy. Vision is an example of electromagnetic radiation as a source of information, and photosynthesis is a process that relies on electromagnetic radiation for energy.

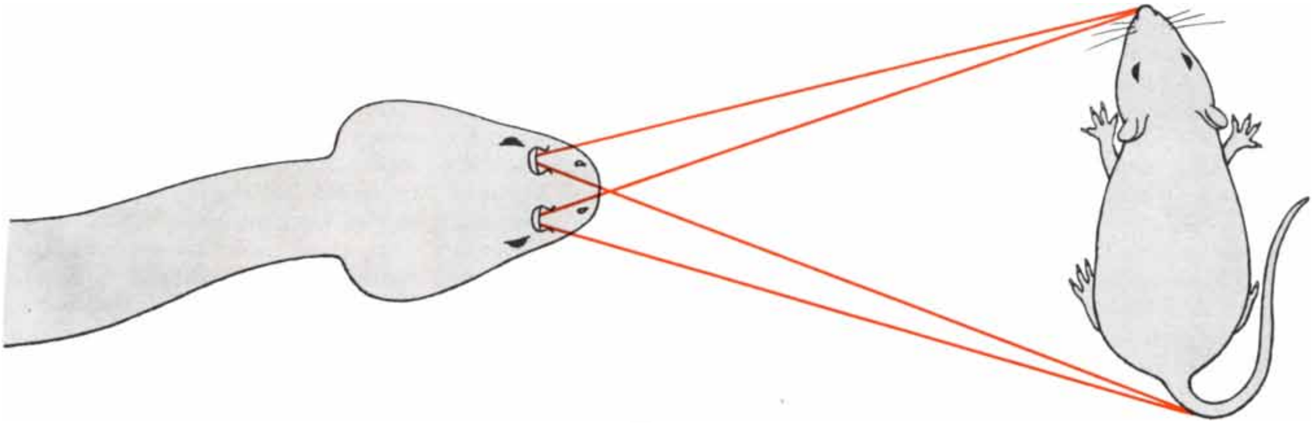
All green plants utilize light as the source of the energy with which they build molecules of carbohydrate from carbon dioxide and water. To collect this energy the plants have a series of pigments (the various species of chlorophyll molecules) that absorb certain frequencies of electromagnetic radiation. Indeed, green plants are green because they absorb the red part of the spectrum and reflect the green part. Because the chlorophyll molecule absorbs only a rather narrow spectral frequency, it can



ANATOMY OF RECEPTOR in a boa is indicated. The scales along the upper and lower jaws have behind them an elaborate network of nerves, which lead into the two branches of the trigeminal nerve shown here. When the system detects an infrared stimulus, the trigeminal nerve carries a signal to the brain. A response can be recorded from the brain within 35 milliseconds after a boid snake receives a brief pulse of infrared radiation.



STRUCTURE OF PIT in a pit viper, the rattlesnake *Crotalus viridis*, differs from the anatomy of the infrared receptor in boid snakes. A pit viper has two pits, located between the eye and the nostril and facing forward. Each pit is about five millimeters deep, with the opening narrower than the interior. The elaborate branching of the trigeminal nerve is in the thin membrane that separates the inner and outer chambers of the pit organ.



DIRECTIONALITY OF SENSOR in a pit viper is indicated by the location of the two pits on the snake's head and by the geometry of the pits. It appears certain that in stalking its prey, which include birds and small mammals, such a snake can establish the direction in which the prey lies by shifting its head as it does in using eyesight. Rattlesnake and copperhead are among the pit vipers.

be called a frequency (or wavelength) detector.

The eye is also a frequency detector, but it does not use radiation as an energy source. The incoming radiation triggers the release of energy that has been stored in the nerves previously, having been produced by normal metabolism. The eye, like other frequency detectors, operates within a narrow band of the electromagnetic spectrum, namely at wavelengths from about 300 to about 1,000 nanometers (billionths of a meter). One can see how narrow the band is by recalling that man-made instruments can detect electromagnetic wavelengths from 10^{-20} meter to 10^{5} meters, a full 25 orders of magnitude, whereas the range of the human eye is from $10^{-6.4}$ to about $10^{-6.1}$ meter. Within this range the eye can resolve thousands of different combinations of wavelengths, which are the number of shades of color one can recognize. Although the eye is a good frequency detector, it is a poor energy detector: a dim bulb appears as bright as a bright one to the dark-adapted eye, which is to say that the eye adjusts its sensitivity according to the conditions to which it has become adapted.

Why has nature chosen this frequency range for its photobiology? From an evolutionary point of view the answers seem clear. One reason is that 83 percent of the sunlight that reaches the surface of the earth is in that frequency range. Moreover, it is difficult to imagine a biological sensor that would detect X rays or hard ultraviolet radiation, because the energy of the photons would be higher than the bonding energy of the receiving molecules. The photons would destroy or at least badly disrupt the structure of the sensor. Low-frequency radiation presents just the opposite difficulty. The energy of long-wavelength infrared radiation and of microwaves is so low that the photons cannot bring about specific changes in a molecule of pigment. Hence the sensor must operate in a frequency range that provides enough energy to reliably change biological pigment molecules from one state to another (from a "ground" state to a transitional state) but not so much energy as to destroy the sensor.

Early workers on the heat detectors of snakes had determined that the receptor responded to energy sources in the near-infrared region of the spectrum. The work left unanswered the question of whether the sensor contained a pigment molecule that trapped this long-wave radiation, thus acting as a kind of eye, or whether the sensor merely trapped energy in proportion to the ability of the tissue to absorb a given frequency and was thus acting as an energy detector. We therefore directed our experiments toward trying to resolve this issue.

To make sure that the response we obtained was maximal, we wanted to work with snakes that were functioning as close to their normal physiological level as possible. First we studied the normal feeding behavior of boa constrictors that were healthy and appeared to be well adjusted. The work entailed seeing how the snake sensed, stalked and captured prey animals such as mice and birds. Since the snake can capture prey in complete darkness as well as in light, it is clear that the heat receptors play a crucial role.

Barrett, while he was a graduate student working with Bullock, went further with this type of behavioral study. He found that the snake would strike at a warm sandbag but not at a cold, dead mouse. On the other hand, the snake would swallow the cold mouse (after a great deal of tongue-flicking and examination) if the mouse was put near the snake's mouth, but it never tried to swallow a sandbag. Barrett concluded that the snake has a strike reflex that is triggered by the firing of the heat receptors, whereas another set of sensory inputs determines whether or not the snake will swallow the object.

In searching for a reliable index that would tell us whether or not the heat receptor was responding to an infrared stimulus, we first tried measuring with an electrocardiograph the change in heartbeat after the snake received a stimulus. This venture ran afoul of the difficulty of finding the heart in such a long animal. (It is about a third of the way along the body from the head.) A more serious difficulty was our discovery that a number of outside influences would change the rate of the heartbeat, so that it was hard to establish a definite stimulus-response relation.

We next turned to a method that proved to be much more successful. It entailed monitoring the electrical activity of the snake's brain with an electroencephalograph. A consistent change in the pattern of an electroencephalogram after a stimulus has been received by the peripheral nervous system is called the evoked potential. When a neural signal from a sensory receptor arrives at the cortex of the brain, there is a small perturbation in the brain's electrical activity. When the signal is small, as is usually the case, it must be extracted from the electrical background noise. The process is best accomplished by averaging a substantial number of evoked potentials. This procedure results in a highly sensitive measure of a physiological response.

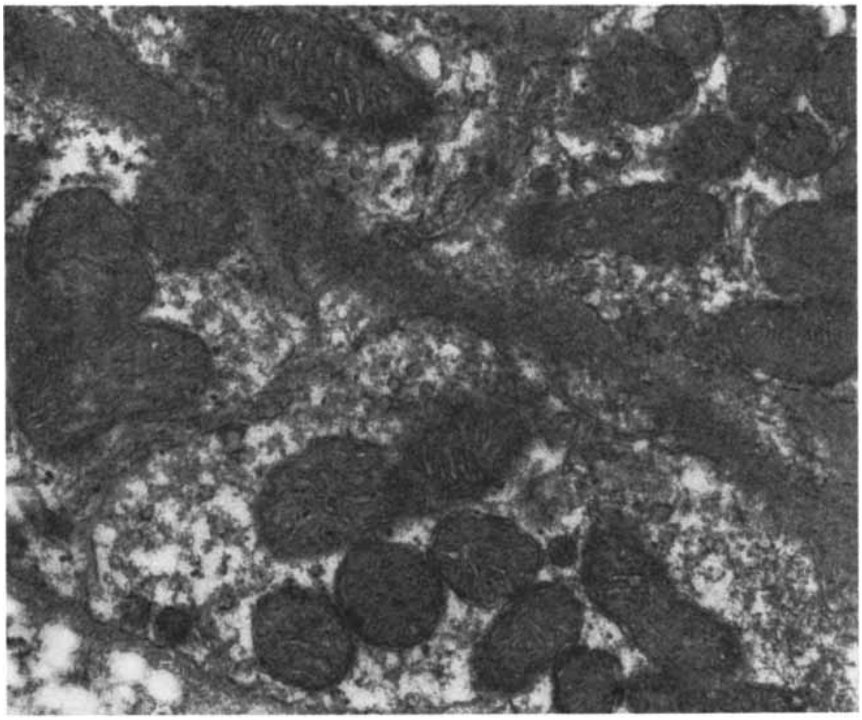
The boa constrictors used in our study ranged from 75 to 145 centimeters in length and from 320 to 1,200 grams in weight. For several weeks before we involved them in experiments they lived under normal conditions in our laboratory. To prepare a snake for the experiments we anesthetized it with pentobarbital and then installed an electrode assembly on its head. After a postoperative recovery period the animal appeared to behave in the same way as snakes that had not been operated on.

A brain signal recorded by this apparatus went to a preamplifier and then to an oscilloscope and to a signal-averager. The signal-averager, which is in essence a small computer, is the workhorse of our system. By averaging the electroencephalogram just before and just after a stimulus it extracts the evoked potential, which would otherwise be buried in the background noise of the brain. In general we average the evoked potentials from about 20 consecutive stimuli.

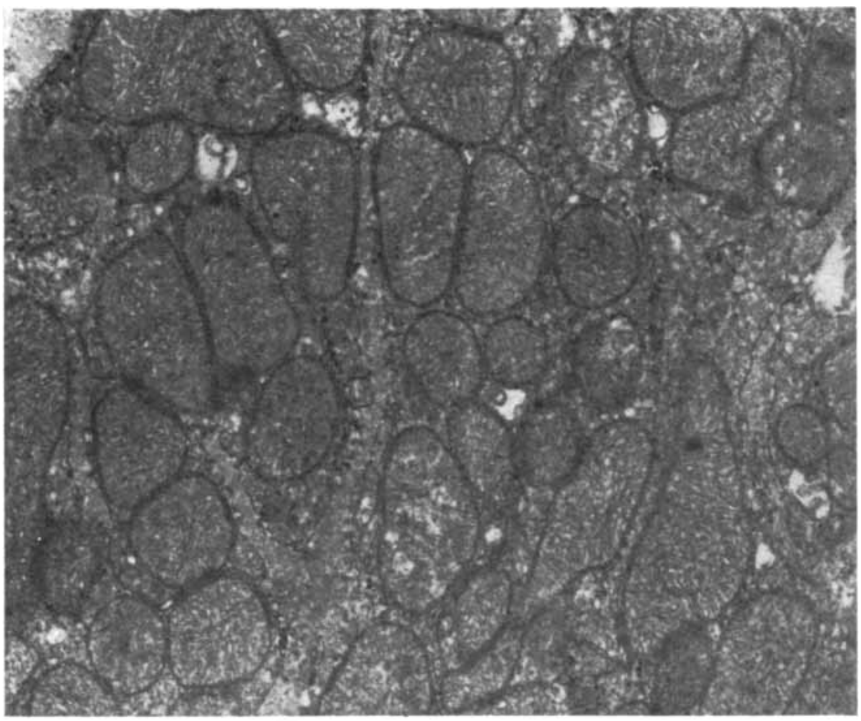
The birds and mammals that the boa constrictor hunts emit infrared radiation most strongly at wavelengths around 10 microns. A carbon dioxide laser is ideal for our experiments because it produces a monochromatic output at a wavelength of 10.6 microns. We pulse the laser by means of a calibrated camera shutter so that it will deliver a stimulus lasting for eight milliseconds. The opening of the shutter also triggers the signal-averager, thus establishing precisely the time when the stimulus is delivered.

After the beam passes through the shutter it is spread by a special infrared-transmitting lens, so that the entire snake is doused in the radiation. The intensity of the radiation is measured by a sensitive colorimeter placed near the snake's head. This is the instrument we mentioned at the outset that takes nearly a minute to measure the power, whereas the snake gives a maximal response within 35 milliseconds after a single eight-millisecond pulse. Another indication of the sensitivity of the snake's receptor can be obtained by putting one's hand in the diffused laser beam; one feels no heat, even over a considerable period of time.

In order to verify that the responses of the snake resulted directly from stimulation of the heat receptor, we repeated the entire procedure with a common garter snake, which has no heat sensor. Even at laser powers far exceeding the



APPEARANCE OF MITOCHONDRIA in the nerve endings of the infrared receptor of a cottonmouth moccasin (*Agkistrodon piscivorus piscivorus*) after exposure of the receptor to an infrared stimulus is shown in this electron micrograph made by Richard M. Meszler of the University of Maryland. The enlargement is 34,000 diameters. In contrast to the mitochondria in the micrograph below, which was made when the receptor was exposed to a cold body, these mitochondria are condensed, as shown by the dense matrix and the organization of the inner membrane. Change in morphology of the mitochondria after a heat stimulus has led to the suggestion that they constitute the primary receptors in the detector.



CONTRASTING APPEARANCE of the mitochondria in the infrared detector of a cottonmouth moccasin when the receptor was exposed to a cold body is evident in this electron micrograph made by Meszler. The enlargement is 27,000 diameters. A cold body, in contrast to a warm one, is known to reduce the firing of discharges by the heat-sensitive receptor.

stimulus given to the boas, we found no response in recordings from the garter snake. On the other hand, both species showed clear responses to visible light.

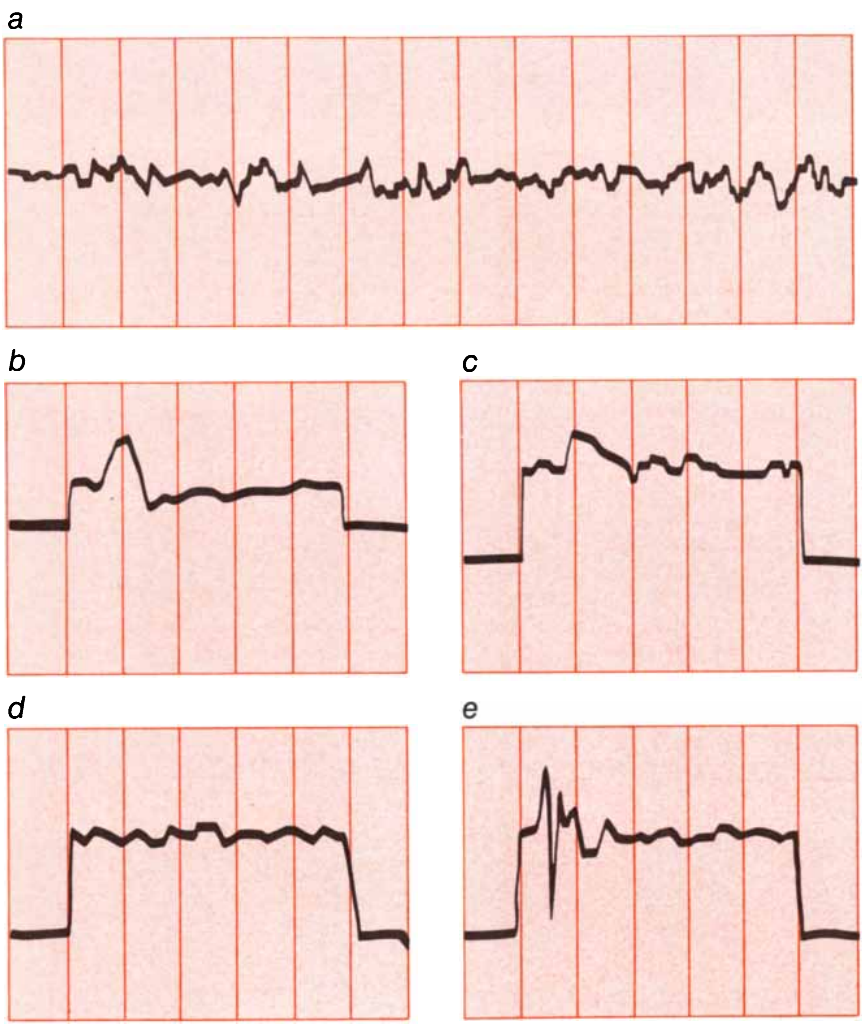
Our data strongly suggested an answer to the question of whether the receptor is a photochemical frequency detector like an eye or is an energy detector. The answer is that the receptor is an energy detector. One argument supporting this conclusion is that the stimulus is so far out in the low-energy infrared region of the spectrum (10.6 microns) that it would not provide enough power to activate an eyelike frequency detector, and yet the snake shows a full response. Another argument has to do with the 35-millisecond interval between the stimulus and the response. Photochemical reactions are quite fast, occurring in periods of less than one millisecond. Although the time a nerve impulse from the eye takes to reach the cortex is about the same (35 milliseconds) as the time the nerve impulse from the heat sensor takes, the neural geometry of the two systems is quite different. The visual pathway incorporates a large number of synapses (connections between neurons), which account for most of the delay. In the trigeminal pathway no synapses are encountered until the signal reaches the brain. We therefore believe the delay found in the heat-receptor response is largely a result of the time required to heat the sensor to its threshold.

We also tested the snake's receptor in the microwave region of the spectrum, where the signals have longer wavelength, lower frequency and lower energy than in the infrared. The reason was that in view of the many problems that have arisen in contemporary society about exposure to radiation we wanted to see whether an organism experienced physiological or psychological effects after being exposed for various periods of time to low-energy, long-wavelength radiation. There is no question that high-intensity microwave radiation can be detected not only by snakes but also probably by all animals; after all, a microwave oven can cook a hamburger in a matter of seconds. Our concern was with the kind of exposure arising from leaky microwave appliances such as ovens and from the increasing use of radar.

Testing the snakes with microwave radiation as we did with infrared, we obtained a clear-cut response [*see illustration on this page*]. Our result provides what we believe is the first unambiguous physiological demonstration that a biological system can indeed be influenced by such low-energy microwave radiation. Our conviction that the snake's heat receptor functions entirely as an energy detector is therefore reinforced.

The question of how much energy is required to activate the detector can be answered with certain reservations: it is approximately .00002 (2×10^{-5}) calorie per square centimeter. The reason for the reservations is that it is difficult to obtain an absolute threshold of sensitivity for any biological phenomenon. For one thing, a biological system shows considerable variability at or near its threshold of response. Moreover, there is always a certain amount of variation in the amount of energy put out by our sources of energy. With these reservations we have determined that the snake can easily and reliably detect power densities from the carbon dioxide laser ranging from .0019 to .0034 calorie per square centimeter per second. Since this density is administered in a short time period (eight milliseconds), the total energy that the snake is responding to is about .00002 calorie per square centimeter. The density of microwave power that is needed for a reliable response from the snake is about the same as the amount of laser power.

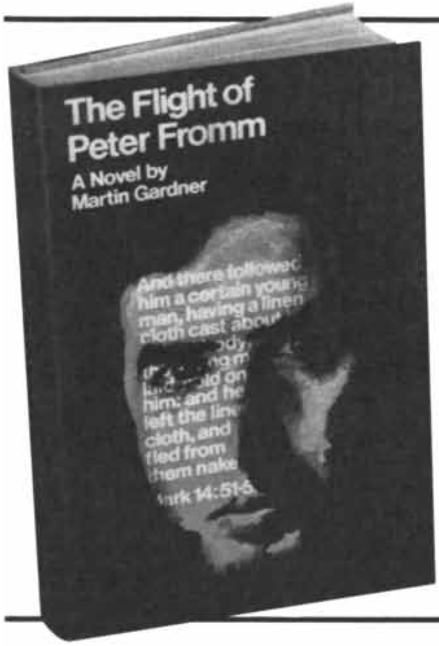
Our studies have shown that the heat-sensing snakes have evolved an extremely sensitive energy-detecting device giving responses that are proportional to the absorbed energy. It will be interesting to see whether the growing understanding of the snake's heat sensor will point the way toward an improvement in man-made sensors.



ELECTROENCEPHALOGRAMS of a boa constrictor were recorded under various conditions. The normal activity of a boa's brain (*a*) was traced directly on a strip recorder; the interval of time between each pair of colored vertical lines was 100 milliseconds. The remaining electroencephalograms were recorded through a signal-averager that reflected both the stimulus and the response. In each case the first rise shows the time of the stimulus and the next rise, if any, shows the response. The time interval is 100 milliseconds in every tracing but *b*, where it is 50 milliseconds. Traces show the averaged evoked response after an infrared stimulus (*b*) and a microwave stimulus (*c*), in the control situation in which the snake was shielded from the stimulus (*d*) and after a series of visible-light flashes (*e*).

OUTSTANDING NEW BOOKS FROM

WILLIAM KAUFMANN, INC. • ONE FIRST STREET / LOS ALTOS / CALIFORNIA 94022



The Flight of Peter Fromm

A Novel by Martin Gardner

A searching story of the spiritual pilgrimage of a young fundamentalist during his years at the University of Chicago and in the US Navy in World War II. A probing, profound book about religion in our time by the versatile author of the monthly *Mathematical Games* feature for SCIENTIFIC AMERICAN and of many books on science, mathematics and literature.

An ordained clergyman who read the manuscript wrote: ". . . so descriptive of my own seminary experience and my ministry that reading it was almost painful . . . Protestantism has been walking the razor's edge of poetic scriptural interpretation long enough. 'Lord, I believe; help Thou my unbelief' is an apology whose time of crisis has come—in the publication of this book."

(Is this *really* a novel?) *340 pages* *May 1973* *cloth, $8.95*

STALKING THE WILD TABOO

"*. . . a refreshing, stimulating book*," by Garrett Hardin, Professor of Human Ecology, University of California, Santa Barbara.

"*. . . one of the intellectual leaders of our time*" —SCIENCE.

Discusses the nature of taboo in sophisticated societies. The author shows how to uncover and attack the taboos that interfere with the rational evaluation of such important subjects as abortion, religion, technology and competition.

240 pages

Published February 1973

Cloth, $8.95

Edited by Alfred Heller

Published July 1972

120 pages

Cloth, $7.95

Paper, $2.50

THE CALIFORNIA TOMORROW PLAN

A comprehensive plan for the future conservation and development of California.

". . . It may be the most important document ever published in California."—*Sierra Club Bulletin*. ". . . an ideal teaching tool for many classrooms where today's environmental issues are considered, whether from a behavioral, biological, business, economic, geographic, political, or sociological point of view. One of its chief values both as a classroom text and a citizens' handbook is that it will raise questions, stimulate new thinking and debate, and thus help generate new and better planning. . . ."—*American Scientist*.

A (soft) systems approach to problem-solving, creativity and design

THE UNIVERSAL TRAVELER

by Don Koberg and Jim Bagnall, California Polytechnic State University, San Luis Obispo

". . . a fascinating product . . . a helpful guide to those who wish to think more creatively and productively."

Just published in a paperbound experimental edition, 8½ x 11, 120 pages, $2.95

TO: WILLIAM KAUFMANN, INC. ONE FIRST STREET / LOS ALTOS / CALIFORNIA 94022

An Order for

The Flight of Peter Fromm ______ copies @$8.95 each $______
Stalking the Wild Taboo ______ copies @$8.95 each ______
The California Tomorrow Plan ______ copies @$7.95 (cloth) each ______
______ copies @ $2.50 (paper) each ______
The Universal Traveler ______ copies @ $2.95 each ______

Total Purchases ______
Less applicable Discount (see below) ______
Net amount due for books ______
California Residents add 5% tax $______
Enclosed is full payment of $______

Save $. If you order two or three books *and include full payment with your order*, you may deduct 10% from the price of each book.

If you order four or more books *and include full payment with your order*, you may deduct 20% from the price of each book.

Postage and handling will be paid by the Publisher.

Send Books to:

NAME

ADDRESS

CITY

STATE

Bill me ☐ ZIP

MATHEMATICAL GAMES

A new miscellany of problems, and encores for Race Track, Sim, Chomp and elevators

by Martin Gardner

From time to time I present a collection of miscellaneous problems that should be new to most readers and that do not require a knowledge of advanced mathematics. All will be answered next month.

1. Sketch a large 4-by-4 checkerboard on a sheet of paper, obtain 16 paper matches and you are set to work on this new solitaire puzzle. The matches represent arrows that point in the direction of the match head. Put a single spot on both sides of one match, two spots on both sides of eight matches and three spots on both sides of seven matches. When a match is placed on a square of the board pointing north, south, east or west, the single spot means it points to the immediately adjacent cell, two spots mean it points to the second cell and three spots mean it points to the third cell.

Seven matches can be placed to map a closed tour [*see illustration below*]. Start at any match of the seven and place your finger on the cell to which it points. The arrow on that cell gives the next "move." Follow the arrows until you return (in seven moves) to where you started. The problem is to place all 16 matches, one to a cell, so that they map a closed tour that visits every cell. There are just two solutions, not counting rotations and reflections.

The tour will have a length of $1 + (2 \times 8) + (3 \times 7) = 38$. It is not hard to prove that this is the longest closed tour that can be made on the board by using any combination of the three types of arrows. Brian R. Barwell, a British engineer who introduced the problem in the *Journal of Recreational Mathematics* (October, 1969), found that only one other maximum-length tour is possible. It requires six 3-arrows, 10 2-arrows and no 1-arrow. Readers are invited to search for all three patterns.

The arrows are of course merely a convenient way to map a maximum-length, closed tour by a chess rook, which lands on each cell exactly once. (Queen tours of this type are less interesting because there are so many of them, bishop tours cannot close and cannot visit all cells and knight tours cannot vary in length.) The 2-by-2 field is trivial and the 3-by-3 is easily analyzed. (Its maximum tour has a length of 14.) As far as I know, the 5-by-5 and all higher squares have yet to be investigated.

2. My wife and I recently attended a party at which there were four other married couples. Various handshakes took place. No one shook hands with himself (or herself) or with his (or her) spouse and no one shook hands with the same person more than once.

After all the handshakes were over I asked each person, including my wife, how many hands he (or she) had shaken. To my surprise each gave a different answer. How many hands did my wife shake? (From Lars Bertil Owe of Lund in Sweden.)

3. An unlimited number of cardboard squares and equilateral triangles, each with unit sides, are assumed to be available. With these pieces it is easy to form convex polygons with from three to 10 sides [*see top illustration on opposite page*]. Can you make an 11-sided convex polygon with the pieces? And what is the largest number of sides a convex polygon formed by the pieces can have? (Both questions are based on Michael Goldberg's solution in *The American Mathematical Monthly* for May, 1969, to a problem posed the year before by Joseph Malkewitch.)

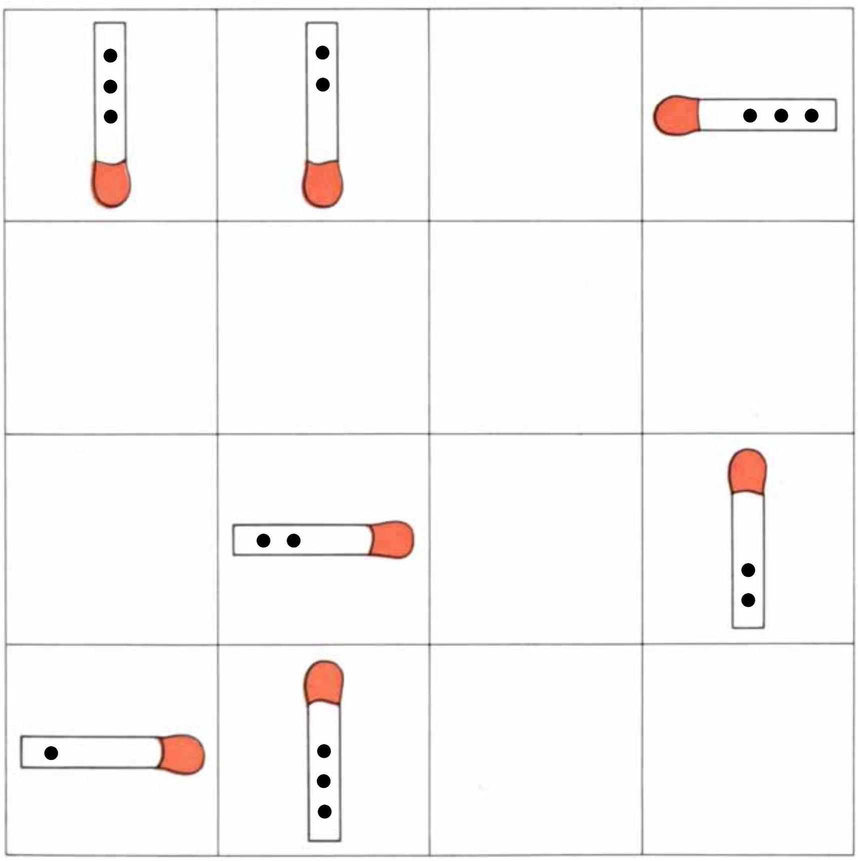
4. Evaluate each of the 10 statements as to its truth or falsity. (From David L. Silverman of West Los Angeles.)

(1) Exactly one statement on this list is false.

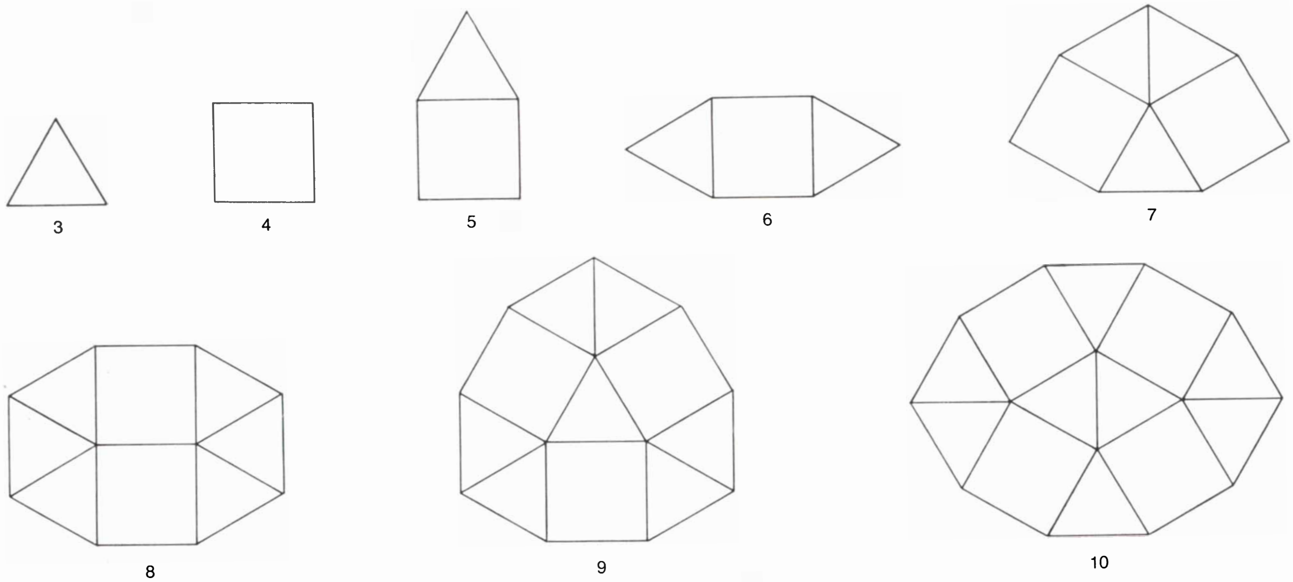
(2) Exactly two statements on this list are false.

(3) Exactly three statements on this list are false.

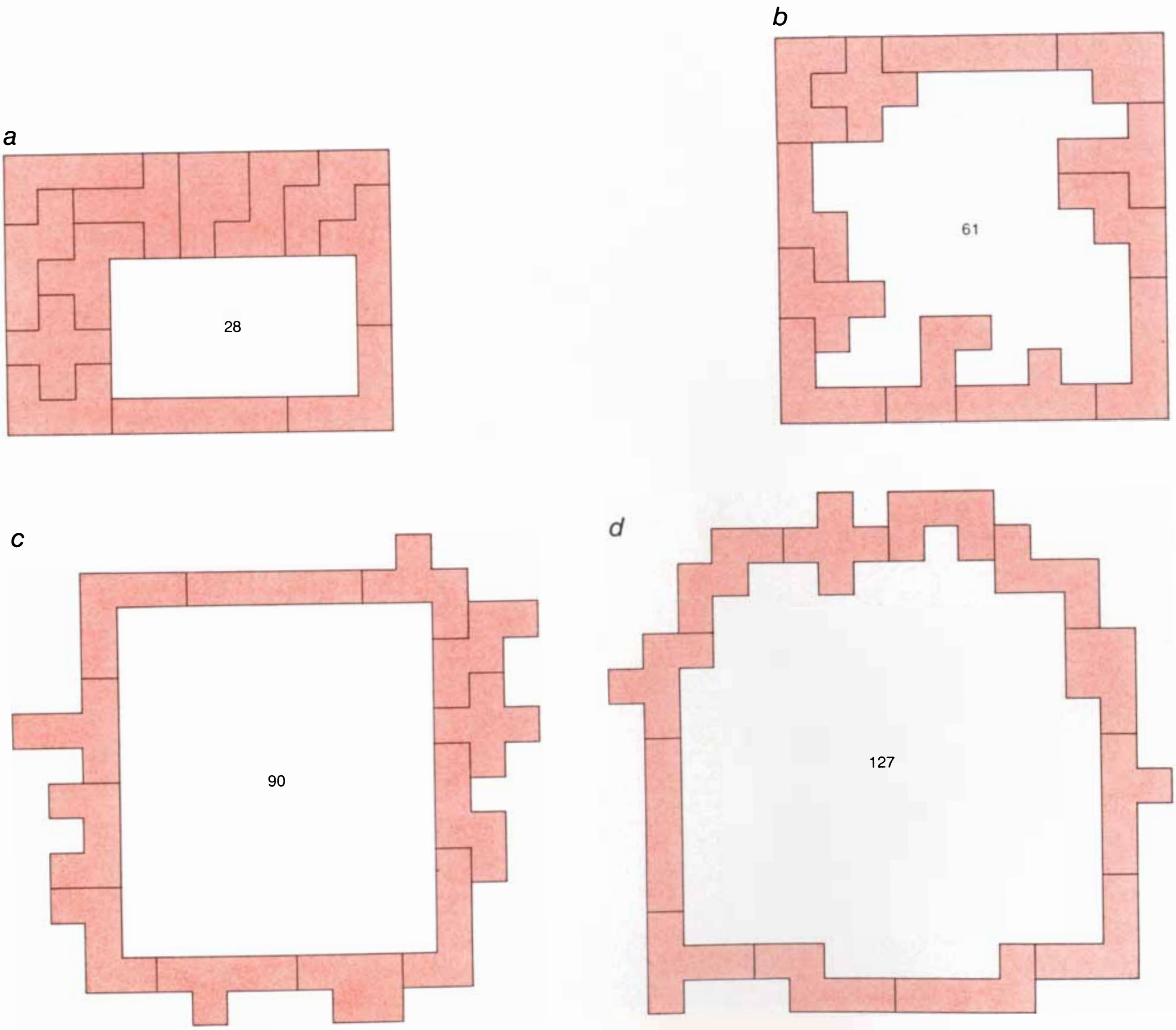
(4) Exactly four statements on this list are false.



A closed arrow tour



Convex polygons with from three to 10 sides



Pentomino fence problems

(5) Exactly five statements on this list are false.

(6) Exactly six statements on this list are false.

(7) Exactly seven statements on this list are false.

(8) Exactly eight statements on this list are false.

(9) Exactly nine statements on this list are false.

(10) Exactly 10 statements on this list are false.

5. Victor G. Feser of Saint Louis University has proposed four maximum-area problems, each using the full set of 12 pentominoes. Three have been solved and the fourth is probably solved.

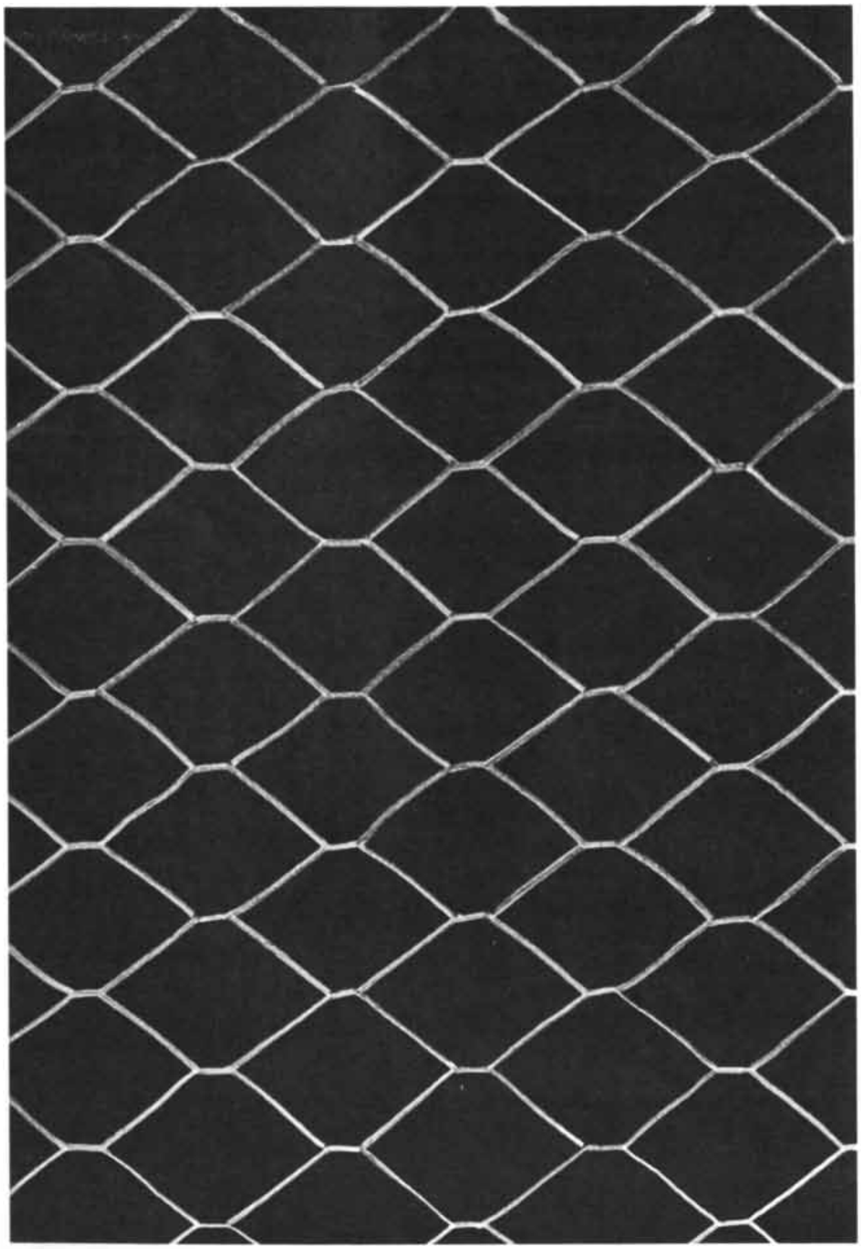
(1) Form a rectangular "fence" around the largest rectangular field. The 4-by-7 has been proved maximum [*see "a" in bottom illustration on preceding page*].

(2) Form a rectangular fence around the largest field of any shape. The maximum is 61 unit squares [*see "b" in illustration*].

(3) Form a fence of any shape around the largest rectangular field. The 9-by-10 is maximum [*see "c"*].

(4) Form a fence of any shape around the largest field of any shape. (As in the preceding problems, the fence must be at least one unit thick at all points.) This is the most difficult of the four. Chris Lindstedt of Göteborg in Sweden, Wade E. Philpott of Lima, Ohio, and I found solutions of 127 squares, such as the one shown at *d* in the illustration. This was believed to be the maximum until Donald E. Knuth, the Stanford computer scientist, recently raised it to 128. Knuth has an informal proof that 128 cannot be exceeded.

Readers will find it a pleasant and difficult task to find a 128 solution. If anyone can exceed it or prove it to be the maximum by computer, I would welcome the news. For information on the first and third problems, see Feser's article "Pentomino Farms" in the *Journal of Recreational Mathematics* for January, 1968, pages 55–61, and discussions in October, 1968, page 234, and July, 1939, page 187.



Chicken-wire folds

6. A kitchen has an uneven floor. There are no "steps," but the continuous random waviness of the linoleum is such that when one tries to place on it a small square table with four legs, one leg is usually off the floor, causing the table to wobble. If one does not mind the table top's being on a slant, is it always possible to find a place where all four legs are firmly on the floor? Or can a floor wave in such a way that no such spot is available? The problem can be answered by a simple, elegant proof. (From Miodrag Novakovic of Belgrade and Ken Austin of Chesham in England.)

7. This strange parlor trick comes from Tan Hock Chuan, a Chinese professional magician who lives in Singapore. He described it in a letter to Johnnie Murray, an amateur conjuror of Portland, Maine, who passed it on to me.

A blank sheet of paper about eight by five inches (half a sheet of typewriter paper works nicely) is initialed by an onlooker so that later it can be identified. The magician holds it behind his back (or under a table) for about 30 seconds. When he brings it back into view, it is covered with creases that form a regular hexagonal tessellation [*see illustration at left*]. How is it done? The performer is usually accused of pressing it against a piece of chicken wire but the creasing actually is done without using anything except the hands.

8. Some of the most beautiful of modern chess problems are those that require retrograde analysis: deduction of the moves that must have preceded the given position.

The problem shown in the illustration on the opposite page is an excellent introduction. It is charmingly simple and at first glance seemingly without a solution. The problem was sent to me two years ago by Robert Rosenwald of New Hope, Pa., who did not know its origin. The illustration shows a position in a legal chess game just after the white king has been accidentally knocked off the board. Where was the king and what was White's last move?

9. By convention, the value of a ladder of exponents such as

$$2^{2^{2^{2^2}}}$$

is computed by starting at the top and working down. The highest pair equals 4, then $2^4 = 16$, and $2^{16} = 65{,}536$. How large is $2^{65{,}536}$? A few years ago Geoffrey W. Hoffmann of West Germany sent me a computer printout of this number. It starts 20035...and has 19,729 digits. Adding another 2 to the ladder gives a number that will never be calculated because the answer, as Hoffmann put it, would require the age of the universe in computer time and the space of the universe to hold the printout. Even a ladder as short as three 9's is $9^{387{,}420{,}489}$, a number of more than 360 million digits. In 1933 E. Skewes published a paper in which he showed that if $\pi(x)$ is the number of primes less than x, and $\mathrm{li}(x)$ is the logarithmic integral function, then $\pi(x) - \mathrm{li}(x)$ is positive for some x less than

$$10^{10^{10^{34}}}$$

an integer said to be the largest known to play a role in a nontrivial theorem.

In 1971 Aristid V. Grosse, one of the pioneer atomic chemists at Columbia University in 1940 (he is now president of Germantown Laboratories, Inc., affiliated with the Franklin Institute), began an investigation of exponential ladders of identical numbers that are calculated in the opposite direction (up) and their relations to down ladders. He coined the term "polypowers" for ladders of both types. Ladders of two x's are called "dipowers," of three x's "tripowers," and so on according to the Greek prefixes. The value of x can be rational or irrational, transcendental, complex or entirely imaginary. In most cases the polypowers are single-valued, continuous and differentiable. Since 1 to any polypower of 1 is 1, all these functions and their derivatives, when graphed against x, cross one another at $x = 1$, and their values at 0 are the limits as x approaches 0. Grosse's notes, which already fill many volumes, lead into a lush jungle of unusual theorems as well as new classes of numbers.

Up and down dipowers obviously are identical, but for all higher polypowers the two directions give different numbers. The triplet of 9's, for example, when calculated upward is a number of only 77 digits. The number of ways of "powering" a ladder (by different placements of parentheses) is 2^{n-2}, where n is the number of x's. In all cases going "up all the way" gives the minimum number and "down all the way" gives the maximum. In what follows, the arrows indicate these maximum and minimum numbers.

What happens when up and down ladders of different lengths are equated? If an up triplet of x's equals a down triplet of x's, $x = 2$. (We exclude $x = 1$ as being trivial.) Each additional x on the up ladder increases the value of x by 1. If three down x's equal four up x's, $x = 3$; if three down equals five up, $x = 4$, and so on.

As an introduction to polypowers, readers are asked to solve the three equations below, which begin a series with down tetrapowers on the left:

$$\overleftarrow{x^{x^{x^x}}} = \underrightarrow{x^{x^x}}$$

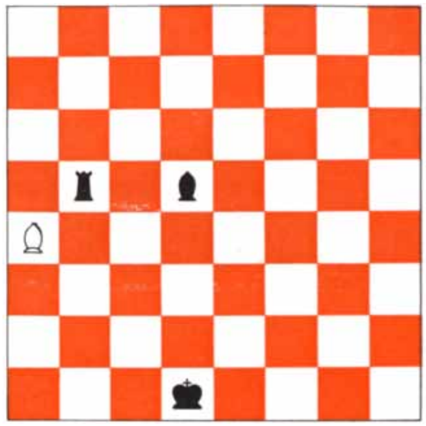
$$\overleftarrow{x^{x^{x^x}}} = \underrightarrow{x^{x^{x^x}}}$$

$$\overleftarrow{x^{x^{x^x}}} = \underrightarrow{x^{x^{x^{x^x}}}}$$

Readers may enjoy investigating ladders of fractional x's, reciprocals of x and more exotic forms. Grosse has also developed the concept of a perfect polypower, that is, x to the xth power (up or down) an x number of times. (Example: π to the π'th power, π times up, is 588,916.326+.) The reverse operation to polypowers he calls "polyroots." Have these fields been investigated before? In spite of considerable effort, neither he nor I have uncovered references. We should both be grateful to learn of any.

The main problem last month was to devise algorithms for converting a signed binary number to its negabinary equivalent and vice versa, using Napier's counting board.

To change a signed binary number to negabinary: (1) Express the number in binary on row 2. (2) If the number is positive, move all counters that have negative values (in negabinary) down like rooks to the first row. (On a standard chessboard this means moving down all counters on white squares.) If the number is negative, move down all counters of positive value (those on black squares). (3) Regard both rows as negabinary numbers. Subtract the first row from the second, using the procedure explained last month for negabinary subtraction. (4) Clear the bottom row by negabinary rules.



Where was the white king?

To convert a negabinary number to a signed binary: (1) Express the number in negabinary on row 2. (2) If the number is positive (an odd number of digits), move down all the negative counters (white squares). If the number is negative (even number of digits), move down all positive counters (black squares). (3) Regard both rows as binary numbers. Subtract the first row from the second, using a binary procedure. (4) Clear the answer by binary rules and prefix the proper sign (plus if the original number was positive, minus if it was negative).

There are other algorithms for conversion but these seem to be the simplest to use on Napier's board. The answer to the question about negabinary palindromes is that the smallest composite number that is palindromic in negabinary is 21. Its plus form is 10101, its minus form is 111111.

January's three games prompted a variety of interesting letters. Many readers felt that Race Track rules should not allow one car to win if another car on the same move could also cross the finish line. They suggested giving the win to the car farthest from the finish line at the end of the move. Joe Crowther was the first of many readers who proposed drawing one or two patches on the roadway to represent oil slicks. Cars are required to move at a constant speed and direction when passing wholly or partly through each patch. J. P. Schell, in addition to oil slicks, proposed adding upgrades and downgrades on which cars are forced to speed up or slow down, as well as pretty girls located along the track to distract drivers.

David Popoff suggested a fast-acceleration move. Any time that a car slows to a full stop it can on the next move go any desired distance in either or both of the two directions. Tom Gordon, who welcomed the game as a teaching device for his high school physics students, added a power-braking option that allows a car to reduce both coordinates by two units provided the move continues the preceding move in a straight line.

C. R. S. Singleton described two novel variants of the game: (1) Instead of a track, numbered gates are marked on the graph. Cars must pass through the gates in numerical order. (2) A series of numbered checkpoints are substituted for the track. Cars must visit each checkpoint by ending a move on that point.

Michael D. Greenberg and his friends at the Westinghouse Aerospace Division in Baltimore adopted two rules to offset the advantage of a first move: (1) Slant the starting line (as in actual racing) and allow the second player to choose between the two starting points. (2) Allow cars to occupy the same point at the same time. They also preferred to draw the track along grid lines to avoid arguments over whether a point was on the track or inside it. Two British readers, Giles Vaughan-Williams and John Kinory, devised rules allowing cars to brake and skid when rounding a sharp curve at high speed.

The game of Sim on a complete graph for five points is now known to be a draw if played rationally. (All draws are topologically equivalent to a pentagram of one color inscribed in a pentagon of the other color. Think of the points as balls connected by elastic strings. If one pattern can be changed to another, they are considered identical.) A complete game tree for five-point Sim was hand-constructed by Eugene A. Herman of Grinnell College and Leslie E. Shader of the University of Wyoming. Jesse W. Croach, Jr., of West Grove, Pa., was able to draw the tree by extracting information from his computer printout for six-point Sim. The first computer program written specifically for five-point Sim was by Ashok K. Chandra of the Artificial Intelligence Laboratory at Stanford University. It produced a complete tree in a few seconds. The results were confirmed by Michael Beeler's program at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology.

	16	8	4	2	1
1	16	8	4	2	1
3	48	24	12	6	3
9	144	72	36	18	9
27	432	216	108	54	27

Chomp as a divisor game

Both Chandra and Herman noticed that a good strategy in five-point Sim is to form a closed circuit of four edges of your color, with a fifth edge attached to any of the four dots. This guarantees your win. Herman noticed that as soon as a dot has three edges of the same color attached to it a draw becomes impossible. Variations and generalizations of Sim came from several readers.

The most surprising letter (to put it mildly) was from G. J. Westerink of Veenendaal in the Netherlands disclosing that the game of Chomp is isomorphic with a number game invented by the late Fred Schuh, a mathematician at Delft Technical College. It is one of the prettiest isomorphisms I have ever encountered in recreational mathematics. The game does not appear in Schuh's *Master Book of Mathematical Puzzles and Recreations* (Dover, 1968), but he explained it in an article, "The Game of Divisors," in *Nieuw Tijdschrift voor Wiskunde*, Volume 39, 1952, pages 299–304. Two players agree on any positive integer, N. A list is made of all the divisors (including N and 1), then players take turns crossing out a divisor and all *its* divisors. The person forced to take N loses. Planar Chomp corresponds to this game when N has exactly two prime divisors, solid Chomp to the game when N has three prime divisors, four-dimensional Chomp when N has four prime divisors, and so on.

This is best made clear with an example. Consider $N = 432$, a number that prime-factors to $2^4 \times 3^3$. Draw a rectangular Chomp field with sides of 5 and 4 (the exponents raised by 1), label the four rows with powers of 3 and the five columns with powers of 2. Counters have values that are products of their row and column [*see illustration at left*]. The equivalence of Chomp to the divisor game is now readily apparent. Moreover, any integer whose prime factors have the formula $m^4 \times n^3$ will correspond to the same Chomp field. Incredibly, most of the theorems discovered by David Gale for his game of Chomp (including the beautiful proof of first-player win) had been discovered by Schuh in arithmetical form!

Schuh offered to play readers by correspondence, using $N = 720$. Because the factors of 720 are $2^4 \times 3^2 \times 5^1$, it corresponds to Chomp on a 5-by-3-by-2 field. This proved to have two winning first moves (counters 36 and 48 when numbered according to the system explained). Like Gale, Schuh was unable to find a strategy for first-player win, a way to determine a winning first move short of constructing the game tree, or a two-prime (planar Chomp) game that had more than one winning first move.

The first counterexample to the conjecture that all planar games have unique winning first moves was found by Ken Thompson of Bell Laboratories. His computer program produced many examples of fields with two first-move wins, the smallest being 8 by 10. The winning moves leave either five columns of 8 and five of 4, or eight columns of 8 and two of 3. This has been confirmed by Beeler.

Cubical Chomp is an interesting challenge. It is easily seen that the winning first move on the order-2 cube is to take the order-1 cube from the corner. Westerink's analysis of the order-3 cube reveals that the winning first move is to take an order-2 cube from the corner. In January I gave Gale's simple proof that the winning move for any square of order n is to take a square of order $n - 1$. Do winning first moves on all cubes of order n consist in taking a cube of order $n - 1$? If so, does this generalize to n-space cubes?

Robert W. Floyd's elevator problem (February) was solved in 11 steps by such a large number of readers that listing names is impossible. Most solvers proved 11 to be minimal. I should not have included nine as a possible minimum because Floyd's lower-bound argument concerns a much more general form of the problem. In this special case it is trivially obvious that the elevator must go to the top and back down (10 steps).

A sample 11-move solution (it is not unique) is: 23456 to Floor 2, 33445 to 3, 444556 to 4, 255566 to 5, 122666 to 2, 566666 to 6, 123455 to 5, 123344 to 4, 112333 to 3, 11122 to 2, 11111 to 1.

Solomon W. Golomb was the first to inform me (by telephone) of an 11-step solution. Later he sent a 14-page typescript on the general problem when floors, number of people per floor and elevator capacity all equal k. As he (and others) proved, $2k - 2$ steps (minimal even with unlimited elevator capacity) can be achieved only if k is less than 5. For k greater than 4, $2k - 1$ is the lower bound. Floyd's lower bound shows that $2k - 1$ is impossible if k is 14 or greater. It is not yet known for what values of k, from 7 through 13, the minimum of $2k - 1$ can be achieved.

THE AMATEUR SCIENTIST

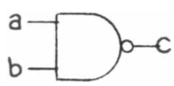
Microelectronic flip-flops and NAND gates add up to a high-speed counter

Conducted by C. L. Stong

Essential elements of the pinpoint landings on the moon and of other feats during recent years are a pair of electronic gadgets known as the flip-flop and the NAND gate. Flip-flops, which can count events and remember the count, and NAND gates, which can make decisions and control the action of flip-flops, are basic building blocks of digital computers. Recently a number of amateurs have undertaken to convert these devices into such things as intrusion alarms, precision clocks and detectors for making audible the supersonic cries of bats.

The flip-flop is analogous to a light switch of the pull-chain type. Successive pulls on the chain alternately turn the lamp on and off. Similarly, when pulses of voltage are applied sequentially to the single input terminal of a flip-flop, they cause pulses of voltage to appear alternately at each of two output terminals.

NAND gates have two or more input terminals but only one output terminal. When the voltage falls to zero at any or all of the input terminals of the NAND gate, the voltage rises to a maximum at its output terminal. Conversely, the application of a voltage to all input terminals causes the voltage to fall to zero at the output terminal [*see illustration on this page*].



V_{in}		V_{out}
a	b	c
0	0	1
0	1	1
1	0	1
1	1	0

Inputs and output of NAND gate

Physically both devices consist of microscopic deposits of metallic films and chemical compounds on minute "chips" of silicon that function as electrical networks of transistors and resistors. Usually several chips are enclosed by a terminal-studded wafer of plastic or ceramic. Circuit designers tend to ignore the internal details of the structures. They combine flip-flops and NAND gates with accessories to form computer systems by following a set of rules that amateurs can easily learn. An introductory project that involves the design of a high-speed counter for use in experiments of many kinds is described by Barry Shackleford, an electronics engineer with the Hughes Aircraft Company. Shackleford writes:

"All electronic digital computers process information in the form of binary digits, a system of numbers that is familiar to all grade-school children who have completed two years of the 'new math.' The binary system is similar to the decimal system but differs in two details. First, in the decimal system the positions for numerals, from right to left, represent increasing powers of 10. In the binary system the corresponding positions represent increasing powers of 2. Second, the decimal system requires 10 symbols: 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9. The binary system requires only two: 0 and 1. In the decimal system the quantity twenty-five is written 25, which reading from right to left represents the sum of 5×10^0 plus 2×10^1. The same quantity is written in the binary system as 11001. Again, reading from right to left, the binary number represents the sum of 1×2^0 plus 0×2^1 plus 0×2^2 plus 1×2^3 plus 1×2^4, or $1 + 0 + 0 + 8 + 16 = 25$.

"Numbers can be symbolized by electrical switches. For example, a rotary switch that includes a wiper arm for making contact with 10 different terminals can be used to express each numeral of the decimal system. Large numbers can be expressed by providing a similar rotary switch for each power of 10. Such switches are costly and difficult to maintain. Binary numbers can be expressed by a simple switch that merely closes and opens a circuit. The closed position can symbolize a 1 and the open position a 0. Electronic switches of this kind, which can be made cheaply, have no moving parts.

"A switch of particular interest that is found in all modern computers is known as the *J-K* flip-flop. It is available commercially in the form of a rectangular wafer of plastic about three-quarters of an inch long, a quarter of an inch wide and an eighth of an inch thick [*see illustrations on opposite page*]. One end of the wafer is marked by a small notch.

"In one popular form of the device, which is known as a 'dual-in-line package,' terminals that extend from the long edges of the wafer are bent at right angles toward one side (the bottom) to form evenly spaced rows. The terminals are identified by a standard numbering scheme that applies to all similarly packaged devices. Terminal 1 is located adjacent to the notch at the left (as the unit is viewed from the top). The remaining terminals are numbered sequentially down the left side and up the right side.

"Power in the form of direct current at a potential of five volts is applied to one terminal of the package, which is designated *Vcc*. The device is connected to ground by another terminal, designated *Gnd*. In addition each flip-flop has a minimum of six operating terminals. One of the six, which is conventionally identified as the clock terminal, serves as the input. (The term 'clock' derives from the frequent use of flip-flops to divide sequences of precisely timed voltage pulses.) Two of the five remaining terminals connect to the *J* and *K* circuits of the flip-flop. These are control terminals. They can be used to inhibit the operation of the device.

"Since a flip-flop switches between two possible states, an output voltage appears at one or the other of two output terminals but never at both simultaneously when the device is supplied with power. Thereafter the output voltage shifts alternately from one output terminal to the other as pulses of voltage are applied sequentially to the input terminal. Usually an output voltage is taken only from the terminal designated Q. The second output terminal, which is designated $\bar{Q}$ (meaning 'not-Q'), can be used for control functions. The remaining terminal is designated 'reset.'

"The device operates normally when potential is maintained on the reset terminal. When the voltage on the reset terminal falls to zero, potential on the Q output either remains at zero or automatically returns to zero. Symbolically the presence of a potential in excess of two volts on the Q output terminal can be interpreted as the binary digit 1 or alternatively as the logical 1: 'yes' or 'true.' Conversely, a potential of less than two volts symbolizes the binary 0 or the logical 0: 'no' or 'false.' In apparatus of conventional design potentials that are referred to as 1 and 0 actually approach five volts and zero volts respectively.

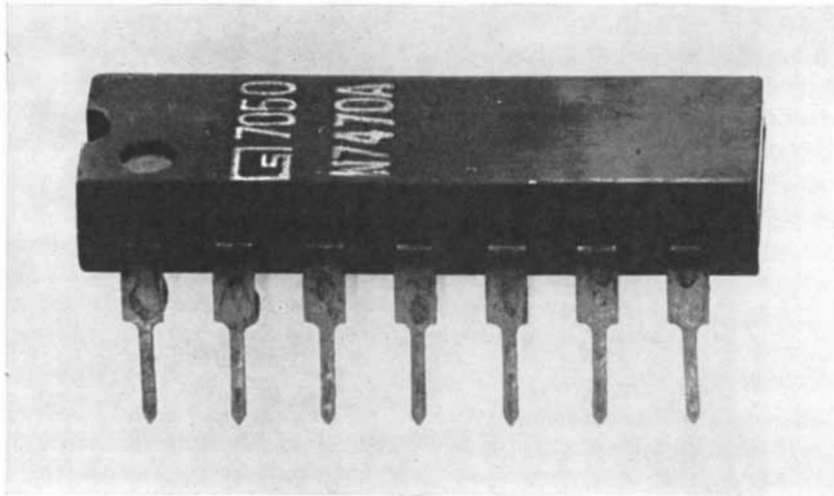
"An apparatus for counting up to 16 events can be assembled by interconnecting four flip-flops [*see top illustration on next page*]. (The diagram illustrating the connections has been simplified by omitting the *Vcc*, *J*, *K* and 'reset' terminals.) It is assumed that the Q outputs of all flip-flops are initially at zero potential.

"The push button is operated to demonstrate the counting action of the circuit. As the contacts close, a potential of five volts is applied almost instantly to the input of flip-flop *A* of the series. The rise in potential from zero to five volts causes no apparent change in the state of the device.

"When the push button is released, the potential at the input terminal falls abruptly. As the potential approaches zero the flip-flop changes state. The potential at output terminal Q_a rises to five volts, which symbolizes a binary 1, and the potential at $\bar{Q}_a$ falls to zero. The potential at Q_a is applied to the input of flip-flop *B*.

"Note that the input voltage rose sharply, remained constant for a time and then returned quickly to zero. A graph made by plotting potential against time would have the form of a rectangle. An electrical signal of this form is known as a square wave. Note also that flip-flop *A* changed state in response to the trailing edge of the pulse.

"Flip-flops are designed to ignore



A dual J-K *flip-flop*

pulses that are not almost perfectly rectangular. To trigger a change in state the signal voltage must drop from maximum to near zero in less than a millionth of a second. This characteristic enables the devices to distinguish between a true signal and the voltage pulses of random shape known as noise. Flip-flops will not in fact respond reliably to pulses that are generated by an ordinary push button. An appropriate switch for generating square electrical pulses will be described below.

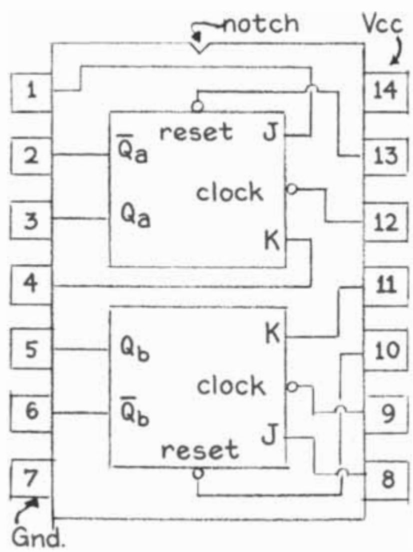
"At the end of the first pulse the voltage states of the four Q outputs can be represented by a sequence of four corresponding binary digits: 0001 (reading left to right). Another pulse is now applied to the input. Its trailing edge causes flip-flop *A* to change state. The potential falls to zero at Q_a and also at the input of flip-flop *B*. Thus the sequence of two square pulses applied to the input of flip-flop *A* generates a single pulse at the input of flip-flop *B*. Flip-flop *B* changes state. The state of the four Q outputs is now represented by the binary number 0010. Translated into decimal notation the binary number is equal to $0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0$, or 2.

"The trailing edge of the next pulse, the third in the series, causes flip-flop *A* to change state, resulting in the expression at the Q outputs of the binary number 0011, which is equivalent to decimal 3. The fourth pulse induces the output state 0100, equivalent to decimal 4, and so on. After the 15th pulse the voltage appears on all four Q outputs: 1111. The trailing edge of the 16th pulse returns all outputs to zero.

"The state of all Q outputs after each pulse in a sequence of 16 pulses is depicted by the accompanying set of graphs [*middle illustration on next page*], which were prepared by plotting voltage against time. The maximum quantity that can be expressed by a series of flip-flops is equal to $2^n - 1$, in which n is the number of flip-flops. Thus a string of four flip-flops can register $2^4 - 1 = 15$, or 1111 in binary.

"A series of flip-flops can also be used for dividing numbers by any selected power of 2. For example, a string of three flip-flops can divide a number by 8. Each sequence of eight pulses applied to the input of the string generates one pulse at the output of Q_c.

"Division by numbers other than the powers of two can be accomplished by resetting all flip-flops to zero at the end



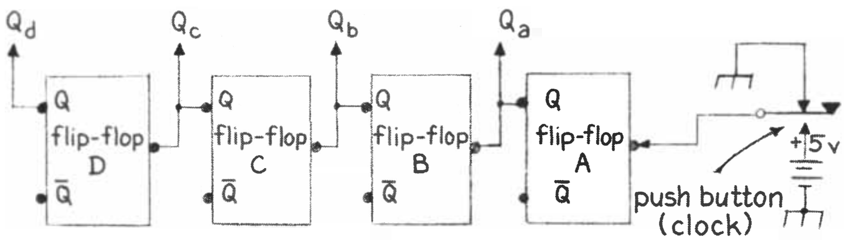
Terminals of J-K *flip-flop*

of any desired count. The scheme is made possible automatically by the 'reset' feature of the *J-K* flip-flop and the use of a NAND gate. Since flip-flops reset automatically to the $Q = 0$ state if the potential on the reset terminal is reduced below two volts, and the voltage falls to zero at the output terminal of a NAND gate when a voltage is applied to all its input terminals, division by any number can be performed by de-energizing the 'reset' terminals of the flip-flop through an appropriately connected NAND gate.

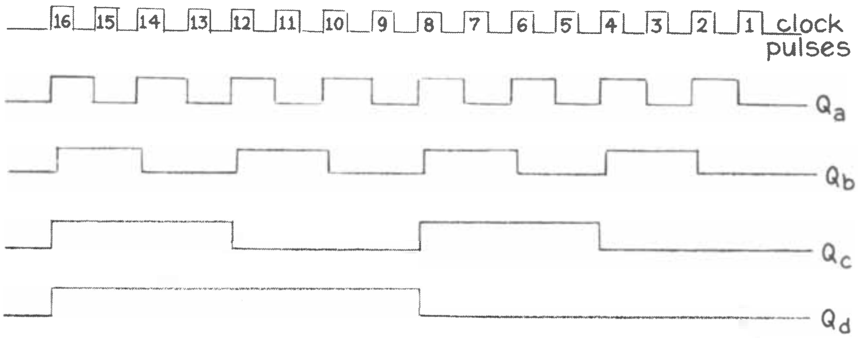
"The scheme is illustrated by the accompanying circuit [*bottom illustration below*], which has been designed to divide by 9. The output terminal of the NAND gate sends a voltage to the 'reset' terminals of all the flip-flops until a potential appears simultaneously at output terminals Q_a and Q_d, which is the state described by the binary number 1001. Hence at the count of nine the output potential of the NAND gate and the 'reset' terminal falls to zero, thus resetting all Q outputs to logical 0 in preparation for the next series of nine incoming pulses.

"Circuits for division by any number that is not a power of 2 can be devised by following a few simple rules. Express the divisor as a decimal number. Find the next higher power of 2. Diminish the exponent by 1 to find the required number of flip-flops. For example, if the divisor is 25, the next higher power of 2 is 32. Since 32 is 2^5, the required number of flip-flops is four.

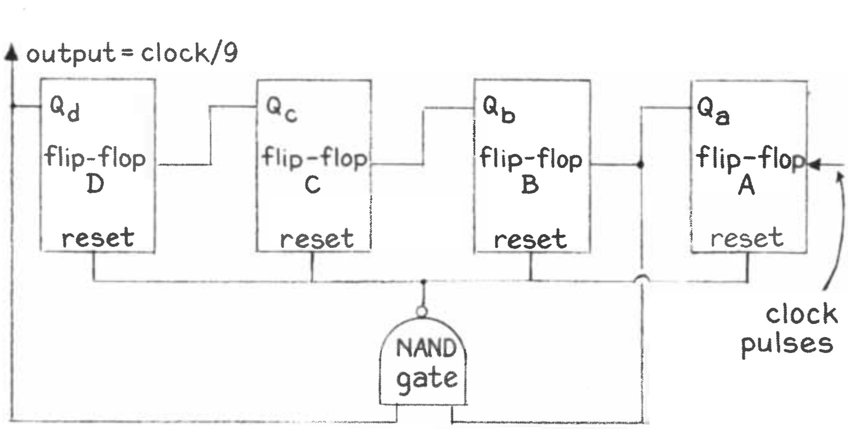
"Connect the Q output of each flip-flop to the input of the next flip-flop in



Flip-flop sequence for counting 16 events



Voltage states at outputs of four-stage flip-flop counters



Divide-by-9 counter

the sequence. Connect the 'reset' terminals of all flip-flops to the output of a NAND gate. Express the divisor as a binary number. Connect the input terminals of the NAND gate to the Q outputs of all flip-flops that are at the state of logical 1 in the binary number that expresses the divisor. For example, the three input terminals of a NAND gate connected to Q_e, Q_d and Q_a of a series of five flip-flops would cause the apparatus to recycle on the count of 25 (11001).

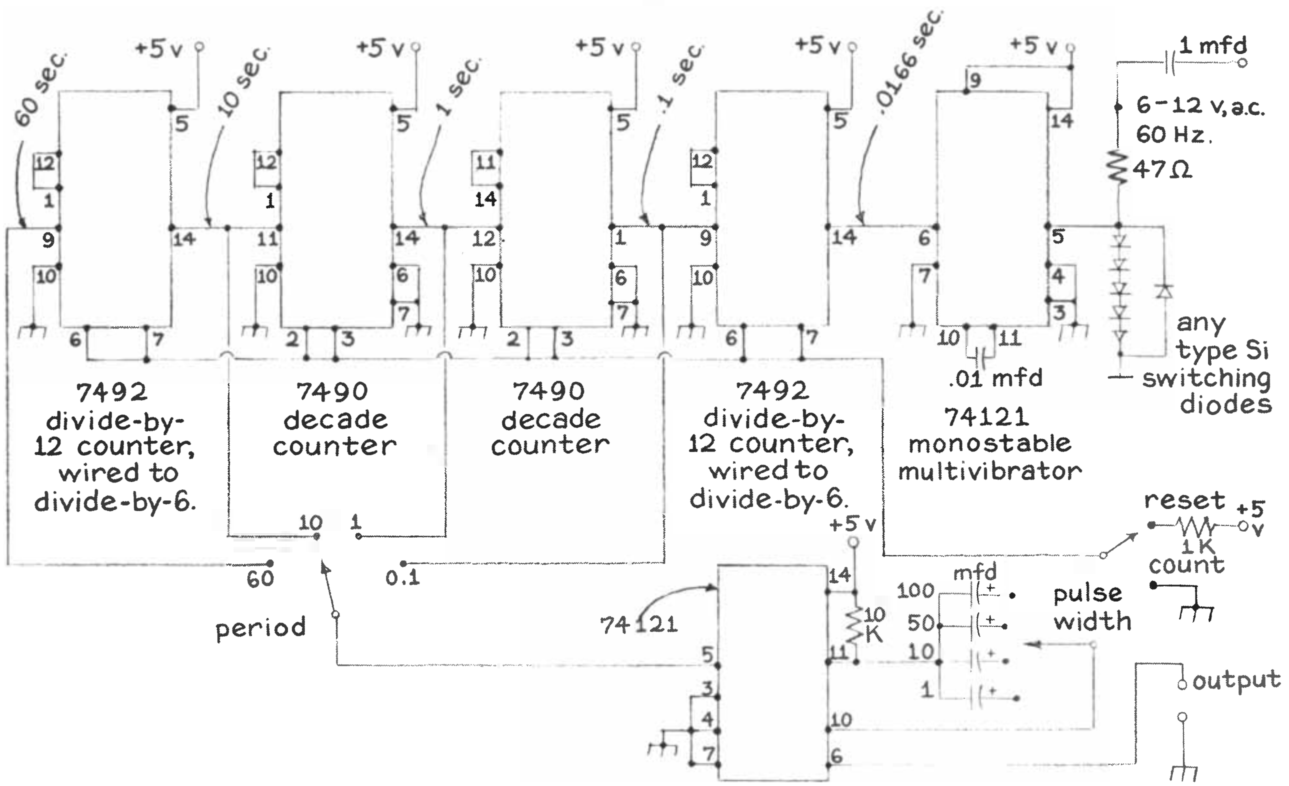
"Experimental counters of this general kind can be assembled with flip-flops of the 7473 *J-K* master-slave type. These units, like all integrated-circuit devices in the 7400 series, are commercially available in the form of dual-in-line packages. The 7473 package contains two separate flip-flops and costs $1.68 in lots of from one to nine units. The Type 7400 package contains four separate NAND gates, each with two input terminals, and costs 96 cents in lots of from one to nine. A gate with an odd number of input terminals can be made by connecting the unused input terminals to a potential of five volts through a resistor of 1,000 ohms.

"Combinations of flip-flops and NAND gates interconnected as described are known as ripple counters. The reason is that input pulses 'ripple' through the string of flip-flops sequentially, each device triggering the next one in the series. Because of variables in manufacturing techniques the devices operate at slightly different speeds. For this reason the maximum rate at which a ripple counter can operate is determined by the performance of the slowest device in the circuit.

"Ripple counters assembled from devices in the 7400 series can divide reliably up to about a million pulses per second. To divide at higher rates the devices are interconnected to operate as a synchronous counter. In this scheme incoming pulses are applied simultaneously to the input terminals of all the flip-flops. Depending on the state of the count, certain of the flip-flops are disabled by applying a voltage from selected $\bar{Q}$ terminals to either or both of the *J-K* terminals.

"Synchronous counters can reliably divide more than 20 million pulses per second. The circuit is somewhat tedious to design. Synchronous counters are available commercially as self-contained devices in the 7400 series. Here are some examples of their use.

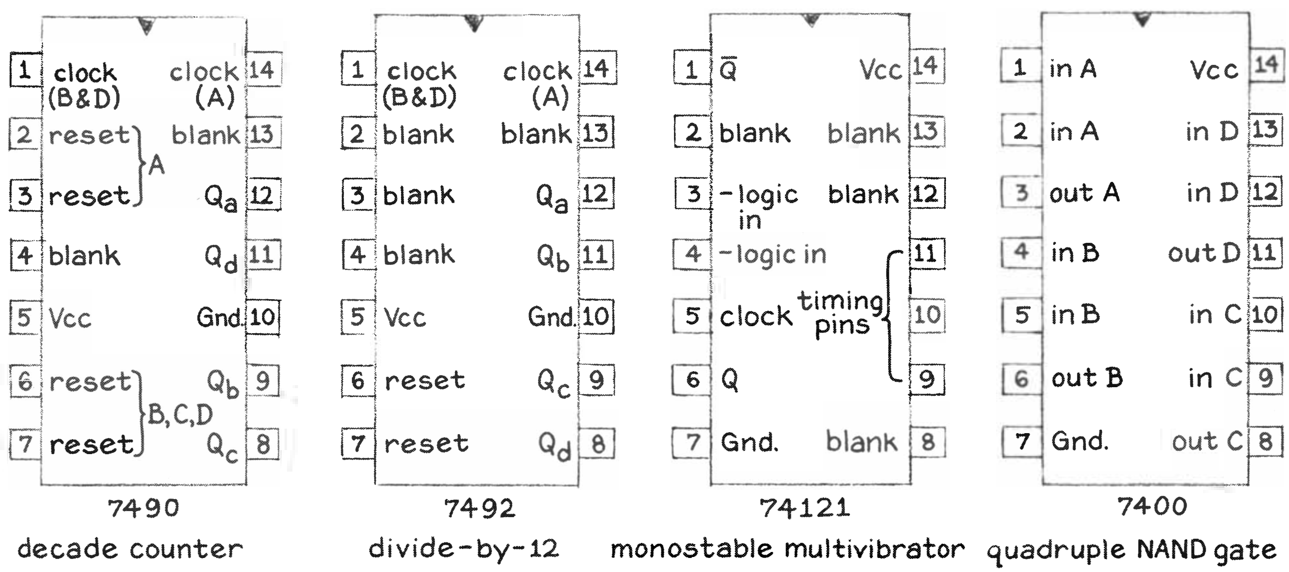
"A convenient source of pulses for testing counters can be generated with an inexpensive apparatus that includes a



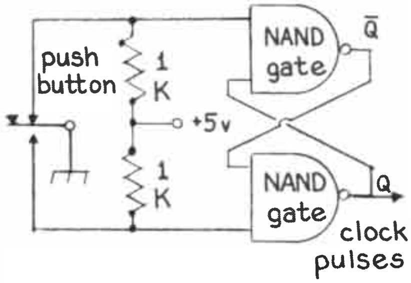
Circuitry of voltage clipper

pair of resistors, a pair of NAND gates and a single-pole, double-throw mechanical switch in the form of a push button [*see top illustration on next page*]. Mechanical switches do not generate single pulses reliably because their contact springs tend to vibrate and thus to create unwanted pulses. The suggested circuit is known as a latch; indeed, it is analogous to the latch on a screen door. When the door is slammed shut, the latch closes and prevents the door from bouncing open. Similarly, when the push button of the suggested circuit is pushed, the voltage normally present on $\bar{Q}$ is transferred to the output, Q, and it remains constant on Q even if the contacts of the push button thereafter open and close repeatedly. When the button is released, the output voltage falls to zero in less than 20 billionths of a second and remains at zero, thereby forming a single pulse with an abrupt trailing edge to which *J-K* flip-flops are sensitive by design.

"A counter for generating time signals that appear at a desired rate and per-



Terminal arrangement, top view



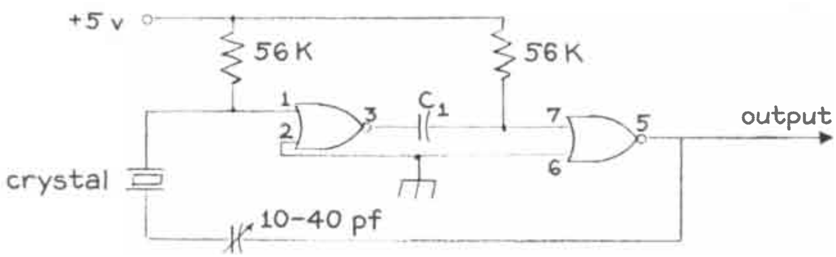
Latch circuit for clock pulses

sist through a desired interval has many experimental applications. For example, voltage pulses that mark equal intervals of time can fix the speed of a synchronous motor for driving such a mechanism as a clock or a telescope and actuating a pen motor that inscribes time pips on a seismogram, can establish a calibrated time base for an oscilloscope and can generate an audio signal for timing a lunar occultation, processes in the photographic darkroom and so on. A pair of counters that have been particularly useful in my own experimentation illustrate the functions of several integrated-circuit devices and some of the techniques that have been developed for assembling the devices into useful instruments.

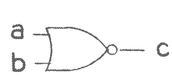
"Particularly useful is a counter that divides the 60-hertz frequency of a power line into periods of .1 second, one second, 10 seconds and one minute. To operate the counter the undulating pulses of voltage from the power line are reshaped into rectangular pulses by two devices. One device, which is known as a voltage clipper, includes a set of five silicon diodes connected in series. A sixth diode is connected across the set in opposite polarity [*see top illustration on preceding page*].

"Voltage is applied to the diodes through a resistor by a step-down transformer that reduces the power-line potential to 12 volts. When an alternating current is applied to the clipper, the potential across the string of five diodes rises until, at three volts, the diodes begin to conduct. Thereafter the voltage remains constant until the applied potential falls below three volts. Similarly, when the current from the power source reverses, the potential across the single diode increases to about .6 volt and thereafter remains constant until the applied potential falls below .6 volt. The result is a pulse of 3.6 volts with a flat top and steeply sloping sides. This pulse is used to operate the second pulse-shaping device: an electronic switch called a monostable multivibrator. When the switch is triggered by the incoming pulse, it closes, then remains closed for an adjustable interval and opens automatically. Typically the interval during which the voltage rises or falls is measured in billionths of a second.

"Pulses thus formed are applied for division to a series of four synchronous counters, each of which contains four flip-flops together with NAND gates and a network of interconnections. The first



NOR-gate crystal oscillator



V_{in}		V_{out}
a	b	c
0	0	1
0	1	0
1	0	0
1	1	0

NOR-gate inputs and output

frequency		value of C_1
1	MHz	30 pf
.5	MHz	430 pf
.1	MHz	1,000 pf

Capacitor values

device in the series, a Type 7492 divide-by-12 counter, contains four flip-flops. Three of them are interconnected to form a synchronous divide-by-6 counter. The fourth flip-flop, which is not used in this application, is available as a separate *J-K* flip-flop that divides by 2.

"The divide-by-6 portion of the device reduces the output frequency of the monostable multivibrator to 10 hertz, meaning individual periods of .1 second each. At this frequency the output is available externally for use as a timing signal. It is also applied for division by 10 to the input of a Type 7490 synchronous decade-counter.

"In addition to reducing the frequency to one cycle per second, this device was wired to equalize the interval during which the pulse persists and the interval between pulses. Therefore voltage at the output terminal of the counter is on for half a second and off for half a second. The scheme is known as symmetrical frequency division. Another Type 7490 decade-counter similarly divides the one-hertz frequency to one pulse per 10 seconds. The final division is made by a 7492 divide-by-12 counter that is wired to divide by 6, thus reducing the frequency to one pulse per minute.

"The output of each counter is connected by a four-position switch to the input of a Type 74121 monostable multivibrator. An identical four-position switch connects any one of four capacitors to the multivibrator. The capacitors control the duration of pulses generated by the multivibrator. In this instrument pulses can be selected that persist for intervals of .007 second, .07 second, .35 second and .7 second.

"Some experiments require a time base of greater accuracy than is provided by the frequency of power lines, particularly in communities served by generating stations that are not connected to interstate power grids. A time base of sufficient accuracy for most experiments can be generated by a simple oscillator consisting of a quartz crystal ground to vibrate at, say, a million cycles per second, two capacitors, two resistors and a pair of Type 7402 NOR gates. This device, which contains four independent NOR gates, is priced at $1.25 in lots of one to nine.

"The circuit generates pulses of approximately rectangular shape that will operate a multivibrator [*see illustrations at bottom of opposite page*]. The output of the multivibrator can be divided by any desired combination of counters in a circuit similar to the one used for dividing the frequency of a power line. Make at least the first division by a synchronous counter to ensure reliable operation.

"Devices that display the output of digital circuits are as diverse as the applications for which the circuits are designed. Perhaps the simplest readout device consists of a light-emitting diode connected between the output and the ground through a resistor of about 25 ohms.

"A few special materials and procedures have been developed that simplify the construction of integrated-circuit devices. Plastic sheets about a sixteenth of an inch thick perforated with a rectangular grid of small holes spaced .1 inch apart are commonly used for mounting dual-in-line packages. The perforated sheets are available commercially as the Vector Integrated Circuit Board, Pattern *P*.

"I mount 7400-series devices side by side, spaced half an inch apart, by pushing the terminals through the holes. Two lengths of straight, bare No. 14 copper wire are installed parallel to the ends of the devices on the terminal side of the board. One length of wire interconnects the power supply and the five-volt terminals of all devices. The second wire serves as the common ground connection. Circuit connections are made by pushing bare No. 24 copper wire into the hole that contains the terminal and fastening the joint with a dab of solder. The soldering is done with a pencil-size iron that draws less than 35 watts. I use 60/40 tin-lead solder wire .05 inch in diameter. The wiring is insulated as required with PVC plastic tubing, size No. 24.

"I test circuits repeatedly during construction, beginning with the power supply. If the power supply does not work, nothing else will. It is easier to track down an error in a small section of a circuit than to locate trouble when the entire system fails. According to Murphy's law, the probability of making an error increases with the square of the number of connections that must be made, and so does the difficulty of locating errors in a circuit after the job is finished.

"I buy new components from Allied Electronics Corporation (2400 West Washington Boulevard, Chicago, Ill. 60612). A selection of new surplus devices is available at a fraction of the list price from dealers such as Babylon Electronics (P.O. Box 4, Carmichael, Calif. 95608) and Poly Paks (P.O. Box 942R, Lynnfield, Mass. 01940)."

BOOKS

The failed mission of Apollo 13, African social predators and the forms of clouds

by Philip Morrison

THIRTEEN: THE FLIGHT THAT FAILED, by Henry S. F. Cooper, Jr. The Dial Press ($5.95). In spite of doubt and joke, astronauts found on the moon not Russians but plenty of glassy rock; their half-dozen voyages were arduous, beautiful and precise. Those who value exploration and new knowledge, those who share in some way in the technical mastery displayed will cherish the memories of the costly missions: the floating journeys, the lunar lope, the hammer and feather dropped wittily in the vacuum, the seismometers, the magnificent suite of moon samples. (Perhaps the falling maintenance budget can keep them sorted and dry.) Yet public attention lagged steadily; one network preferred playbacks of old football games to the astronauts live on the moon last December. Drama requires uncertainty.

This brief, adroit, suspenseful, opinionated narrative is the most exciting book for the general reader to arise out of Apollo, in part because its topic is the one flight that failed. The emergencies, the improvisations, the alternatives examined reveal the meaning and tasks of the whole as sharply as pathology illuminates health. Cooper, who writes with verve and clarity, neither knows much about nor cares a great deal for technical matters. There is one rather naïve diagram, no more. Here and there his book shows the limitations of the reporter's skill at describing what people do without quite being at home in their world; his mention of gravityless space, stainless-steel insulation, the width of radio waves, moving east or west on a given longitude show a man—or his editors—anxious to avoid jargon but less anxious to understand it. Yet his insight and his dramatic chronicle of how people acted and why rise above such slips.

The chunky Mission Control Center near Houston has no windows "as there are none on the lower decks of a ship—and, in a sense, [it] really *was* the lower decks of a ship, the upper deck being a couple of hundred thousand miles away in space. Astronauts are more like officers aboard a large ship than like solitary heroes." The "thin sheet of vapor, like a cirrus cloud," that astronaut Lovell first saw was "as if a ship's captain had reported seawater rushing in through the hull." Worse.

It all began in plain view of Houston. The three-part spacecraft—service module, command module, lunar module—linked, a day away from moon orbit, two days out and 200,000 miles from earth. A party of engineers in a rooftop observatory 200 yards from the Control Center had been following the sunlit spacecraft on a television display fed from a modest telescope. The ship-star had faded, and the brighter, tumbling booster, like a variable star, had dwindled in its turn. Now they saw a sudden unexpected bright spot appear in mid-screen, grow to the size of a dime and slowly fade. The television had been working poorly, and so they blamed this flare-up on their circuits and went home.

At that moment the astronauts felt a shudder, perhaps even a bang. The pulse rate nearly doubled in all three men. It took about 13 minutes before the flight controllers, far removed from that shudder, confident in the redundancy of the design, aware of the ambiguities of interpreting the telemetry of a couple of hundred instruments to learn the unexpected about so complex a distant ship, knew that both fuel cells that fed power to the command module were dead. That flare in space had been scattered sunlight on the oxygen escaping from one tank of that nutrient liquid, where wires, whose insulation had been unknowingly burned off in a ground test, had arced when the tank was routinely stirred. The rupture was an explosive one. The astronauts finally saw a wide gash the entire length of the module when that silvery cylinder swam away once it was cut adrift a few hours before the return to the earth.

It is wrong to detail a suspense plot, even though we all recall from real time how *Apollo 13* limped back safely. There were three great problems of the space-wreck. First, the "lifeboat," the lunar module, was short on power and on instrument-cooling water; never mind the cold hands and the thirst of the astronauts. No lifeboat on the open sea has so inhospitable a black vacuum around it. Supplies were urgent minute by minute; even navigation depended on gyroscope spin and power. Ampere-hours were rationed, not mere biscuits. Second, the stars (which guide mariners in space too) were diffused and obscured by sun glare on the specks of ice and debris leaking uncontrollably from the wounded ship. Third, there was a subtle crisis of momentum. That leaking fluid steadily passed recoil momentum to the ship as the days went by; the controllers never could anticipate in detail the motion of the craft—an eerie mystery to men whose Newtonian confidence had been built up with intact ships whose mass and momentum changed only on command.

Still, they mastered every problem, the young men at the computer consoles, the back-room groups writing unheard-of checklists and trying them out on the simulators around the clock. (They improvised an indispensable airflow connection with a space-suit hose, some plastic checklist cards and a couple of storage bags—"like putting together a model airplane, but... it would look more like an R.F.D. mailbox." The astronauts got the idea precisely, and with the box they could keep the air breathable.) The controllers here are people, with names and styles, such as Charles Deiterich, David Reed, Seymour Liebergot, men with training as engineers or fliers, hasty veterans of a five-or-six-year career with the intricate Apollo systems. Surely all the canons of the romance of adventure, in fact or fiction, are fully satisfied by this tale of a failure, overcome without tragedy, filled with the intimate trappings and jargon of a limitless strange sea and of the men who sail it, whether in body or indirectly, through design-anticipation or at the end of an S-band down link.

It came time to "punch off the LM."

Before ordering any irreversible action, the flight director had to check in turn over the telephone link with all the controllers in their rows. Eugene Kranz called it "going around the horn."

"He ticked off the names rapidly—RETRO, FIDO, GUIDO, CONTROL, TELMU, GNC, EECOM, Surgeon, and INCO—and the answers came back just as fast: 'Go,' 'Go,' 'Go.' "

THE SPOTTED HYENA: A STUDY OF PREDATION AND SOCIAL BEHAVIOR, by Hans Kruuk. The University of Chicago Press ($15). SERENGETI: A KINGDOM OF PREDATORS, by George B. Schaller. Alfred A. Knopf ($12.95). The caldera of Ngorongoro is a grassy plain 10 miles across, encircled by steep, forested walls 1,600 feet high. The walls enclose "amazing numbers of animals"—nearly 100 large wild mammals per square kilometer, plus two Masai villages, their cattle and a few tireless ethologists. Moving up Olduvai Gorge from the crater one comes at last to Serengeti National Park, a vast grassy land with some treeless steppe, host to an animal population that is less dense but much larger than that of the tiny enclave of the crater. These two books, the one a technical but fascinating report on a strange species and the other a popular, beautifully illustrated picture book with 100 color photographs, together amount to a personal and intimate survey of the life of the hoofed beasts and those who hunt them across these remarkable African countrysides.

Kruuk watched the wild hyenas of the region for three years and more. He kept a free-roving pet hyena, who treated the house as its den; he marked hyenas and mapped their movements; he followed one steadily, by the signal of a radio collar, for 12 days, night and day, the task shared among two men and a Land-Rover. He himself scavenged on the hyenas' kill for meat. He used all the tools of his trade: the Piper Super-Cub, hypodermic darts, infrared spotting 'scopes, plenty of chi-squared tests and above all a keen eye and a sharp mind to find out the secrets of this ugly, despised, misunderstood, night-prowling beast. His report is remarkable. Whether, as he thinks, he is close to the track of early man (who was, after all, another social carnivore), even though the hunters he watches are not primates, only time can tell. (The chimpanzees have turned out to be carnivores of opportunity themselves.) The four-footed social hunters are fascinating on their own terms.

The spotted hyena is misjudged by most men who know it. It is not mainly a scavenger; those packs are powerful hunters by night, running with maniacal whoops and wild laughter. Hyenas share game with the lion; where the hyenas are many, the lion is as likely to scavenge from the kill of the hyena as the other way around. Indeed, the niches these two social carnivores occupy are subtly differentiated; we do not quite know how—perhaps by the different selection of prey that their sudden rushes at the grazing herds of gazelle or wildebeeste induce.

The hyena is wonderfully well suited to scavenging. It uses well its versatile talents for the crushing and full digestion (to a white powder) even of big bones, whether at its own kills, at those of others by day or at the refuse and spoil of man. But the wild hyena packs live mainly by skillful hunting; doglike, they eat at their living prey to bring it down, sometimes during a run at 40 miles an hour, rather than suffocating or strangling it in one mouthing as the big cats do. All social carnivores, it appears, hunt in social groups to enlarge the size range of the prey that can make a meal; the loners—leopards, cheetahs, even the striped hyena—live chiefly on animals considerably smaller than they are.

Aristotle thought the hyena was a hermaphrodite. There is, of course, no literal truth in the idea. The hyena is a mammal closely related to the cats, although its behavior converges a good deal toward the dogs. Hyenas never climb trees; they do like water and swimming, and their "whole bodies are shaped for running rather than stalking." In mating and parental behavior they are more like cats; in hunting, feeding and the use of claws they are more like dogs.

But the sexual dimorphy of the hyena is unique. The females seem to run about 10 percent larger than the males and are clearly dominant in the pack. More than that, the external genitalia of male and female are strikingly similar: the clitoris is as large and erectile as the penis, which it exactly resembles, and the female has two sacs filled with fibrous tissue that look much like a scrotum. No wonder Aristotle was confused. These organs are almost as large in the cubs as they are in the adult. Kruuk thinks he has understood the function of this remarkable pattern, which is not at all sexual. In meeting, hyenas mutually and regularly display erect penis or clitoris in a ceremonial fashion. He concludes that the elaborate scheme has the function of keeping the fierce animals close for a time while allowing initial tensions to subside, a useful thing for animals that are "very social but often also solitary." The selective "symbolic" advantage of a familiar, complex and conspicuous structure may have been what has caused it to evolve in cubs and females, in whom it has no direct sexual function.

Kruuk supplies many graphs, maps, tables and black-and-white photographs. ("The behavior of hyenas is literally shrouded in darkness.")

The Schaller book, a companion piece to Schaller's own monograph on the Serengeti lion no less than to Kruuk's report, shows us by bright day the animals and the landscape of the region: the green grass, the striped zebra, the red tooth and claw, the eager dog pack and the immense grazing herds stretching to the flat horizon. In a few text pages about each predator species Schaller summarizes what is known about lion, leopard, hyena, cheetah and hunting dog. For the method, the detailed argument, the unfolding of the science, the Kruuk report is indispensable; for strong impressions and the main results, the general reader will enjoy the colorful pages of the Schaller volume. Both men write well, without pretense; the splendid pictures of the Schaller book are bordered by text set in a handsome but eccentric typographical style, which may or may not entice the reader's eye.

CLOUDS OF THE WORLD: A COMPLETE COLOR ENCYCLOPEDIA, by Richard Scorer. Stackpole Books, Harrisburg, Pa. ($29.95). A devoted Victorian amateur, the Hon. Ralph Abercromby, traveled twice "round the world to assure himself that the same basic cloud forms were to be found everywhere." He extended to 10 the famous Latinate classification of the pharmacist Luke Howard in 1803: cirrus, cumulus, stratus and their many compounds. That system became the basis of the elaborate international atlases of cloud forms, beautiful if wholly denotative, that served well the one function of enabling untrained observers to describe the clouds for a distant analyst. The last international atlas, with its 10 main forms and subdivisions, came out in 1956.

Richard Scorer is no amateur but a distinguished and energetic theorist of the Imperial College of Science and Technology. He has lovingly collected and provided with personal and vivid text this large atlas, with some 410 photographs of clouds (mainly snapshot-size color prints and altogether beautiful). This is a cloud atlas with a difference, however. Nomenclature is reduced to a minor tool, the rationale for the juxtapositions is more physical than visual, and for every form there is a penetrating account of its formation and meaning. The work is almost entirely nonmathematical, but it is not simplistic. The ideas of cyclonic rotation, mixing and shear, the distinction between wave and material motion, the latent heat of phase change and similar fundamentals of atmospheric fluid mechanics permeate the explanations. The tougher problems are often elucidated with a sequence of line drawings. The result is a book that can engage a high school student for a long time and yet will very likely keep discussion going among groups of professional meteorologists. Professor Scorer disdains mere pedantry: "Nimbostratus is not illustrated: it is a grey amorphous raining sky, and who wants to get a good camera wet to photograph that!"

Scorer's camera has seen fine days and wet; a great many of these pictures are his own. Others come from people all over the world. ("I wish I had been there," he writes.) Aircraft and mountaintop play leading roles in providing width of view; there are quite a few satellite photographs and some radar-screen displays. The collection is worldwide: one sees the sky over Kanchenjunga and Ullswater, the Antarctic and Osaka, Hawaii and the oases of the Sahara. Clouds differ little the world over, but there are favored habitats for each type. Stereo pairs are presented for a number of examples, printed both for cross-eyed viewing (directions and trial photographs given) and with duplicates for conventional stereo viewers. He sees cloud processes too in smokestack plume and dust devil, in high contrail and in dew on the grass; his final chapter is the hurricane's, an organization of clouds on an awesome scale.

The explanations are the more credible because they are deeper than the usual presentations of the introductory texts. The dust devil is of course a case of air converging horizontally, conservation of angular momentum causing the rotation (in either direction, cyclonic or anticyclonic) to speed up as the air moves inward. The central pressure falls because of the centrifugal forces; only frictional losses then allow further convergence. The dust devil starts on hot sunny surfaces of low thermal conductivity, such as low-density sand. No part of such a surface can supply thermally rising air for more than a short time, but a moving vortex can slide along such ground, gaining hot air close to the surface where friction makes strong convergence possible, and the air is fed up the central column, keeping the pressure low and the dervish whirling.

The speedy flow raises the dust; on the side where rotation augments the wind speed the losses are greater and there is more convergence. And so "the whirl moves across the wind towards that side, along a curved track." There it goes, past a hut in Uganda, and in the next pages waterspouts, tornadoes and finally hurricanes make a Federal case out of the same phenomena plus the interplay of buoyancy and the production of latent heat. There are 14 such connected visual narratives, including one on optical phenomena (what sun dogs, halos and glowing crosses!) and a very detailed story of contrails and their negatives.

The first few pages on thermals and the cumulus clouds they produce make a strong case for Scorer's view that "beautiful though these pictures are, it is through simplified understanding that we come to like them." The forms made by thermal currents from the ground come clear in origin, and from them in turn can grow cloud-capped turrets: cumulus growing not from thermals rising past the level of droplet condensation but from the buoyancy of condensation heat alone. The features of the cloudscape are more readable to Professor Scorer than any landscape, since the flow of air is more apparent than the growth of land forms. "The Earth is our home, and it is a delight to stand back in wonder at its beauty.... Life has made the atmosphere what it is." This book is a scholar's solid gift to the general reader, but one gathers that the author would have delighted in making it even if no one were to scan its colorful and meaningful pages.

COLOR: ITS PRINCIPLES AND THEIR APPLICATIONS, by Frederick W. Clulow. Fountain Press, London. Available from International Publications Service, New York ($17.50). "Brown, russet, citrine, olive green and other dull-colored surfaces are really dark yellow or dark orange or dark green materials and so on." Yet the skin of an orange looks orange under bright or under dull illumination. This nonmathematical text (there are plenty of graphs and some simple algebraic relations) is one of the few honest and understandable introductions to the complex topic of color vision and color reproduction. It says little about postretinal processes or about the scanning motions of the eye; apart from the cones-and-rods distinction, it remains at an external level. Yet it makes plain the rough theory of color vision at a resolution adequate for the general scientific reader and for most of the users of color in the arts and crafts; its arguments are coherent, based on the facts of experience, and it does not blink the recognition that the simple theory is an approximation.

The sensible identification of the color brown is only one test of such a text. (Many popular accounts fail there.) But Clulow, a lifelong teacher to the English printing and photographic community, passes many more. He fills the book with useful and interesting illustrations (the obligatory spectrum and various fatigue and contrast effects are shown in color) and many arguments illustrated by graphs. A sequence of actual spectrophotometric curves, with the graphed samples pictured beside them, is most helpful. Black needle paper and white typewriter bond differ by about a factor of 15 in reflectance all across the spectrum, while a mid-gray is the geometric mean between the two. A scarlet pigment, mercuric sulfide, is in fact a low-frequency reflector, reflecting about equally everything redward of a half-reflecting point in the orange. A bright cadmium yellow does the same, except that its half-reflectance point is in the green. A tomato and a lemon behave the same way, whereas flesh tone has a blue secondary peak, perhaps typical of an English ambience.

The book begins with a few chapters of the general kind, which show the quality of the text by their breadth of interest. There is a long chapter on color measurement, with the standard charts and with considerable attention to the modern instruments available for numerical characterization and matching of color. The most detailed chapters are those on additive and subtractive methods of reproduction; these excel in penetrating explanations, not just in elaboration of technique. The need for subtractive methods in printing is argued cleanly from the light loss on reflection from paper and the compensating gain by the use of subtractive primary pigments, which (speaking crudely) remove a third of the light rather than two-thirds, as additive primaries do.

Then we see plotted the spectrophotometric graphs of the three idealized subtractive primaries: low-frequency reflective, ideal yellow—everything from green to the deepest red sent back; the high-frequency component, ideal cyan, which sends back all the light, and only the light, blueward of the orange; the central-band pigment, ideal magenta, which fully absorbs green, yellow and orange but supplies an unbiased return both at the red and at the blue end. The real pigments are compared with these, and we see why all three printed together do not produce a black. And so the fourth impression, in black, of most color-printing processes gains some of its rationale. Lacking black, shadows become too colorful! More than that, the black printer controls crispness, which depends more on lightness differences than it does on color differences. Here there is an analogy with color television, in which the luminance—the three-color sum—is transmitted separately, along with two color-difference signals, rather than three color signals directly. (The scheme has other practical advantages in television as well.)

So the volume goes. Of modest length, it carries a reader far enough into the topics, whether on gloss and luster, on day-glo or on the several national-standard color-television systems, to meet most needs short of the fully specialized and quantitative literature.

The Photographs of Thomas Eakins, by Gordon Hendricks. Grossman Publishers ($30). Motion pictures, on the big screen or the little one, are an obviously integral part of modern sensibility. It is just 100 years since the ingenious photographer Eadweard Muybridge began to settle Leland Stanford's bet as to whether a horse in full gallop ever lost all contact with the ground. (E. J. Marey, a French physiologist, already had chart recordings—although not yet photographs—that decided the point.) These pioneers of the movie, it turns out, had very early won the serious attention of a gifted painter, the Philadelphian Thomas Eakins.

While Muybridge was at work for Stanford, Eakins had sought Muybridge's sequential photographs to document an unusual representation of the feet of trotting horses in the lively painting *A May Morning in the Park (The Fairman Rogers Four-in-hand).* Rogers, influential and well-to-do, himself invited Muybridge to come to the Franklin Institute to display his pioneer movies in 1883. The *Philadelphia Record* reported "a large and interested audience" for the motion pictures of "horses, cows, sheep, dogs, deer and men." By the end of the summer the University of Pennsylvania had invited Muybridge to come there to work, and it was in Philadelphia that he completed his many-volume work of sequential photographs, *Animal Locomotion,* using an arkful of models from the Fairmount Park zoo. Eakins was no mere spectator to this new image-making. He himself (after Marey) designed and built a "drop shutter" and a "revolving disc with the hole in it which can be run by hand or by electric motor." With this kind of apparatus he took sequence photographs of men in motion, many of which can be seen in this large, scholarly, beautifully reproduced volume.

"In pursuance of my business and professional studies I use the naked model," Eakins wrote in a famous statement. So he did for pictures with his Marey wheel: men pole vaulting, running, walking, jumping. "The History of a Jump," nine overlapping exposures at 1/8-second intervals of a naked broad jumper, with technical notations by Muybridge on one print, is genuinely beautiful, clearly reminiscent of one of Eakins' paintings, which so well convey the pleasure and the inward order of sport. It becomes evident that the artist was devoted to photography, for itself and for the aid it gave his painting. Here we see the very swimming hole of one of his paintings, the youths themselves at play; the wrestlers flat on the mat; the Arcadian scenes of faun piper and playing children; even the graybeard Walt Whitman, all of them more or less like the final paintings. Hendricks is at pains to sort out how much of the relation between canvas and dry plate is literal, but the reader finds himself throughout in Eakins' world. That world was caught with a four-by-five view camera from 1880 onward; 300 photographs are here, including, in a forgivable extension of the title, 50-odd photographs of the artist, made over almost 60 years of his life. (Even the camera is shown, now in a collection.) The shy, grave, sharp-eyed young woman who became Eakins' wife is here sitting in the shade of some bushes. Walt Whitman in his rocking chair in Camden looks into the lens; he "delighted in" Eakins although he was bored with photographers.

This is a work of art, an evocation of a time and a man, and a flash of insight into the early growth of that now intimate relation between the images wrought by photochemistry and those made by the hand and eye of the artist.

Oceanographic Ships Fore and Aft, by Stewart B. Nelson. Office of the Oceanographer of the Navy. U.S. Government Printing Office ($4.50). Patently the loving work of a devotee, this good-looking volume is an illustrated narrative of the oceanographic ships of the U.S., both Navy and otherwise—"an unbelievable potpourri of ships." It begins with Benjamin Franklin. Franklin heard about taking advantage of the set of the Gulf Stream from his cousin, a Nantucket sea captain, and published a chart that is reproduced here. The book ends with models (catamarans) and unfunded proposals, plus a flip-ship or two. It has the fascination of the specific that attaches to all catalogues of ships and their masters across two centuries, here nearly free of tragic overtones. Matthew Fontaine Maury, the first of naval oceanographers, put it so: "Navies are not all for war. Peace has its conquests, science its glories, and no Navy can boast of brighter chaplets than those gathered in the field of geophysical exploration and physical research."

Observe, then, the *Albatross,* the first ship built for marine research by any government. Commissioned in 1882, she was an iron-hull steamer, the first Government vessel with electric lights throughout. (An Edison dynamo of 51 volts worked 120 lamps.) For 40 years she collected marine organisms. Observe the *Tuscarora:* she first carried Lord Kelvin's new piano-wire sounding machine. Kelvin had designed it for H.M.S. *Challenger,* but it was not installed on that ship; as Kelvin said, "innovation is very distasteful to sailors." The Tuscarora Deep, a record in 1875, was where Jules Verne's moon voyagers plunged to the earth. Here are today's National Science Foundation ships, such as the converted cargo vessel R/V *Eltanin,* steadfast support of U.S. Antarctic research and one of the ships whose magnetic records were pieced together to build the arguments that have established seafloor spreading and the rest of our modern picture of the earth.

The *Alpha Helix,* built in 1966 as a biological research vessel, represents the newer trend. The newest Navy research vessels, such as the sister ships R/V *Melville* and R/V *Knorr* (for the Pacific and the Atlantic respectively), have four times the displacement of the *Alpha Helix,* with a special design for station-keeping and control even at zero speed. These diesel ships have a vertical shaft at each end that drives a disk holding half a dozen feathering propeller blades for fingertip maneuverability. New and more economical craft are on the horizon: smaller utility boats, modified for special tasks to augment the few big ships we now have. Aircraft, already much used, hydrofoils and captured-air-bubble vehicles are also on the minds of planners.

The unique *Glomar Challenger,* stitching her path along the sea bottom with oil driller's gear, is not mentioned. No ship in the book is more beautiful than the U.S. schooner *Experiment,* here drawn as she lay to for her ship's boat, working in the Coast Survey in the 1830's.

INDEX OF ADVERTISERS

MAY 1973

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

CRYPTOGRAPHY AND COMPUTER PRIVACY

COMMUNICATION THEORY OF SECRECY SYSTEMS. C. E. Shannon in *Bell System Technical Journal,* Vol. 28, pages 656–715; October, 1949.

MODERN CRYPTOLOGY. David Kahn in *Scientific American,* Vol. 215, No. 1, pages 38–46; July, 1966.

THE CODEBREAKERS. David Kahn. The Macmillan Company, 1967.

ELEMENTARY CRYPTANALYSIS. Abraham Sinkov. Random House, Inc., 1970.

THE ROLE OF THE HEARTBEAT IN THE RELATIONS BETWEEN MOTHER AND INFANT

THE IMPORTANCE OF THE HEARTBEAT RHYTHM TO HUMAN NATURE: THEORETICAL, CLINICAL, AND EXPERIMENTAL OBSERVATIONS. Lee Salk in *Proceedings of the Third World Congress of Psychiatry.* University of Toronto Press, 1961.

MOTHERS' HEARTBEAT AS AN IMPRINTING STIMULUS. Lee Salk in *Transactions of the New York Academy of Sciences,* Vol. 24, No. 7, Series II, pages 753–763; May, 1962.

HEARTBEAT RHYTHM AND MATERNAL BEHAVIOR. I. Hyman Weiland in *Journal of the American Academy of Child Psychiatry,* Vol. 3, No. 1, pages 161–164; January, 1964.

BIOLOGICAL FREUDIANISM—LASTING EFFECTS OF EARLY ENVIRONMENTAL INFLUENCES. René Dubos, Dwayne Savage and Russell Schaedler in *Pediatrics,* Vol. 38, No. 5, pages 789–800; November, 1966.

PATTERNS OF MOTHER-INFANT CONTACT: THE SIGNIFICANCE OF LATERAL PREFERENCE. I. Hyman Weiland and Zanwil Sperber in *The Journal of Genetic Psychology,* Vol. 117, Second Half, pages 157–165; December, 1970.

TWO-DIMENSIONAL MATTER

THE SOLID-GAS INTERFACE. Edited by E. Allison Flood. Marcel Dekker, Inc., 1967.

EXISTENCE OF LONG-RANGE ORDER IN ONE AND TWO DIMENSIONS. P. C. Hohenberg in *The Physical Review,* Vol. 158, No. 2, pages 383–386; June 10, 1967.

CRYSTALLINE ORDER IN TWO DIMENSIONS. N. D. Mermin in *The Physical Review,* Vol. 176, No. 1, pages 250–254; December 5, 1968.

MOTIONAL NARROWING IN MONOLAYER-He^3-FILM NMR. R. J. Rollefson in *Physical Review Letters,* Vol. 29, No. 7, pages 410–412; August 14, 1972.

SHORT-RANGE ORDER AND MELTING ANOMALIES IN THIN FILMS. J. G. Dash and M. Bretz in *Journal of Low Temperature Physics,* Vol. 9, Issues 3/4, pages 291–306; November, 1972.

ROOTS

MINERAL NUTRITION OF PLANTS: PRINCIPLES AND PERSPECTIVES. Emanuel Epstein. John Wiley & Sons, Inc., 1972.

TRANSLOCATION OF INORGANIC SOLUTES. Andre Lauchli in *Annual Review of Plant Physiology,* Vol. 23, pages 197–218; 1972.

MECHANISMS OF ION TRANSPORT THROUGH PLANT CELL MEMBRANES. Emanuel Epstein in *International Review of Cytology,* in press.

THE EVOLUTION OF THE INDIAN OCEAN

THE EARTH AS A DYNAMO. Walter M. Elsasser in *Scientific American,* Vol. 198, No. 5, pages 44–48; May, 1958.

THE EAST PACIFIC RISE. Henry W. Menard in *Scientific American,* Vol. 205, No. 6, pages 52–61; December, 1961.

THE PRINCIPLES OF PHYSICAL GEOLOGY. Arthur Holmes. The Ronald Press Company, 1965.

REVERSALS OF THE EARTH'S MAGNETIC FIELD. Allan Cox, G. Brent Dalrymple and Richard R. Doell in *Scientific American,* Vol. 216, No. 2, pages 44–54; February, 1967.

SEA-FLOOR SPREADING. J. R. Heirtzler in *Scientific American,* Vol. 219, No. 6, pages 60–70; December, 1968.

THE ORIGIN OF THE OCEANIC RIDGES. Egon Orowan in *Scientific American,* Vol. 221, No. 5, pages 102–119; November, 1969.

CONFIRMATION

ENQUIRY CONCERNING HUMAN UNDERSTANDING. David Hume. Open Court Publishing Company, 1900.

LOGICAL FOUNDATIONS OF PROBABILITY. Rudolf Carnap. The University of Chicago Press, 1962.

CHOICE AND CHANCE: AN INTRODUCTION TO INDUCTIVE LOGIC. Brian Skyrms. Dickenson Publishing Co., Inc., 1966.

THE FOUNDATIONS OF SCIENTIFIC INFERENCE. Wesley C. Salmon. University of Pittsburgh Press, 1967.

PROBABILITIES, PROBLEMS AND PARADOXES: READINGS IN INDUCTIVE LOGIC. Edited by Sidney A. Luckenbach. Dickenson Publishing Co., Inc., 1971.

CONFIRMATION AND RELEVANCE. Wesley C. Salmon in *Minnesota Studies in the Philosophy of Science,* in press.

GALILEO'S DISCOVERY OF THE LAW OF FREE FALL

DIALOGUES CONCERNING TWO NEW SCIENCES. Galileo Galilei, translated by Henry Crew and Alfonso De Salvio. Dover Publications, Inc., 1966.

GALILEO GALILEI. Ludovico Geymonat, translated by S. Drake. McGraw-Hill Book Company, 1965.

GALILEO STUDIES: PERSONALITY, TRADITION, AND REVOLUTION. Stillman Drake. University of Michigan Press, 1970.

PHYSICAL SCIENCE IN THE MIDDLE AGES. Edward Grant. John Wiley & Sons, Inc., 1971.

THE INFRARED RECEPTORS OF SNAKES

PROPERTIES OF AN INFRA-RED RECEPTOR. T. H. Bullock and F. P. J. Diecke in *The Journal of Physiology,* Vol. 134, No. 1, pages 47–87; October 29, 1956.

MEN AND SNAKES. Ramona and Desmond Morris. Hutchinson of London, 1965.

RADIANT HEAT RECEPTION IN SNAKES. T. H. Bullock and R. Barrett in *Communication in Behavioral Biology,* Part A, Vol. 1, pages 19–29; January, 1968.

THE PIT ORGANS OF SNAKES. Robert Barrett in *Biology of the Reptilia—Morphology B: Vol. II.* Academic Press, 1970.

SNAKE INFRARED RECEPTORS: THERMAL OR PHOTOCHEMICAL MECHANISM? John F. Harris and R. Igor Gamow in *Science,* Vol. 172, No. 3989, pages 1252–1253; June 18, 1971.

THE AMATEUR SCIENTIST

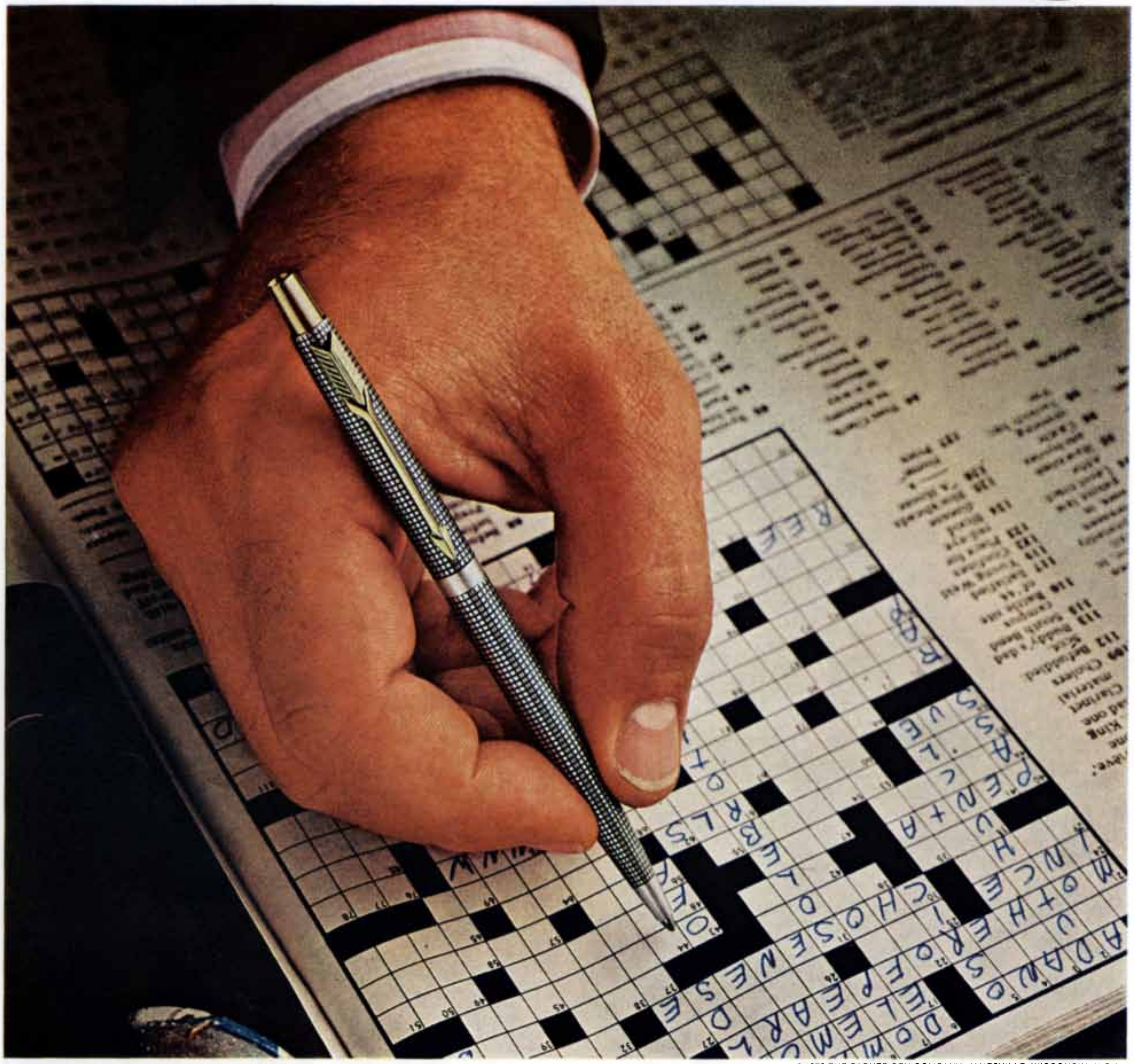
DESIGNING WITH TTL INTEGRATED CIRCUITS. Staff, Texas Instruments Incorporated, 1971.

FAIRCHILD SEMICONDUCTOR TTL DATA BOOK. Fairchild Semiconductor, 1973.

THE LOGIC DESIGN OF TRANSISTOR DIGITAL COMPUTERS. Gerald A. Maley and J. Earle. Prentice-Hall, Inc., 1963.


The Silver Martini.
For people who want a silver lining
without the cloud.
SMIRNOFF
SILVER
Ste Pierre Smirnoff Fls
VODKA
90.4 PROOF
DISTILLED FROM GRAIN
DISTILLED FROM GRAIN 90.4 PROOF VODKA
DISTILLED BY STE PIERRE SMIRNOFF FLS
HARTFORD, CONN
Smirnoff Silver
Ninety point four proof. Smirnoff leaves you breathless.

PARKER 75



A pen can't tell you everything about a man. But it can tell a lot.

The Parker 75 Classic Ball Pen.

Its clean lines tell you he likes functional design. And, with him, you know the sterling silver case isn't a pretense. He knows it adds heft and balance, and he simply sees no reason why a durable thing has to be ugly.

This pen also tells you he respects precision. He may not realize that the tip has been microscopically textured to grip any writing surface. But he does know this pen is made so precisely it writes a clear, crisp track every time.

Is he concerned about the pen's guarantee? Probably not. Nevertheless: If this pen fails to perform due to defects during its owner's lifetime, Parker will repair or replace it—free. Finally, the fact he does crossword puzzles *in ink* tells you he has confidence far beyond most men. The Parker 75 Classic Ball Pen: for the man who expects perfection, it makes a near-perfect gift.

PARKER
World's most wanted pens

The Parker 75 Classic Ball Pen in solid sterling silver is $12. Also in vermeil, $25; 22K gold electroplate, $8.50, brushed stainless steel, $5 Matching pencils. You'll find the distinctive arrow clip trademark on every Parker, from the famous $1.98 Jotter Ball Pen to the $150 Parker 75 Presidential Pen