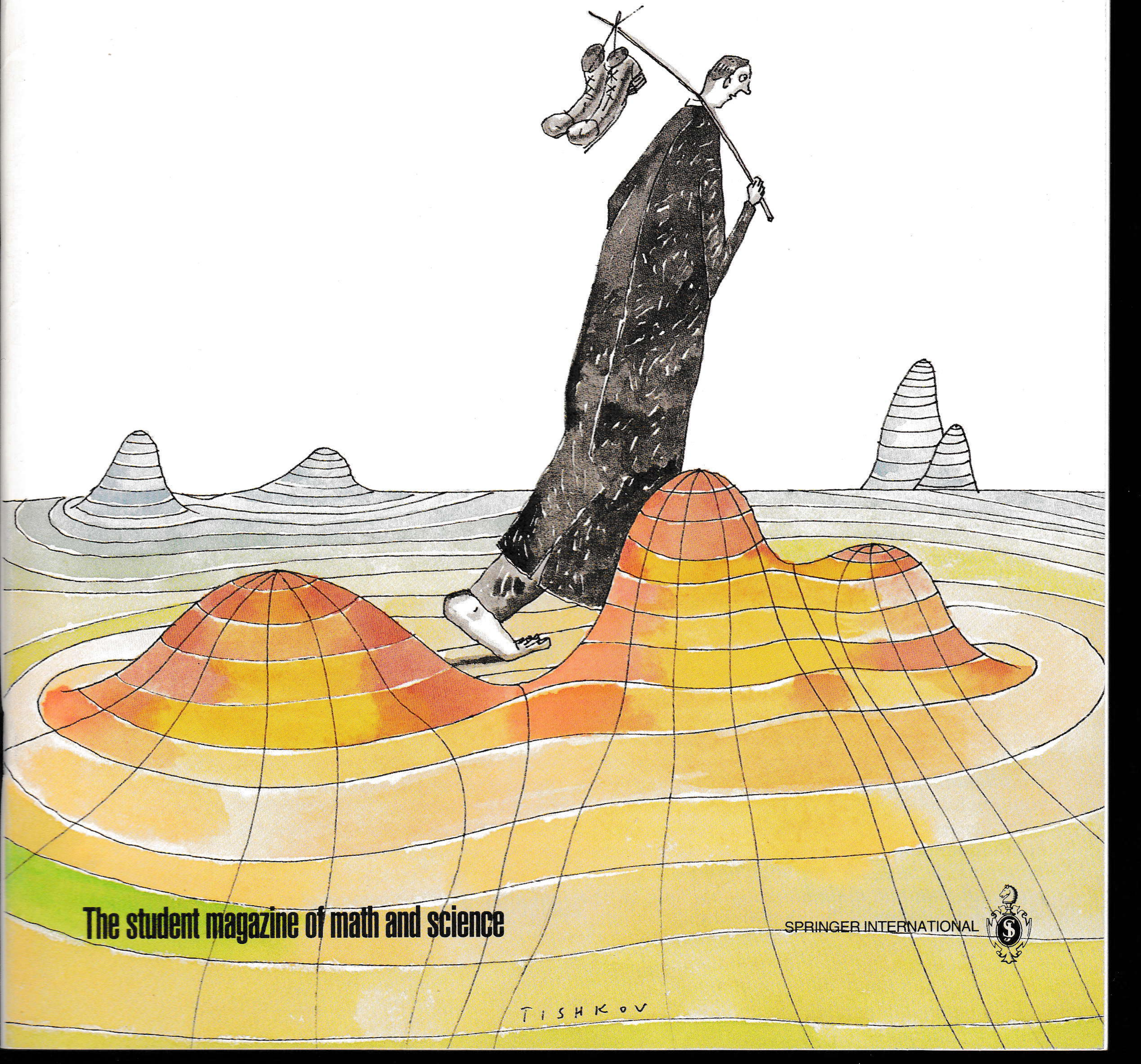
QUANTUM

SEPTEMBER/OCTOBER 1992

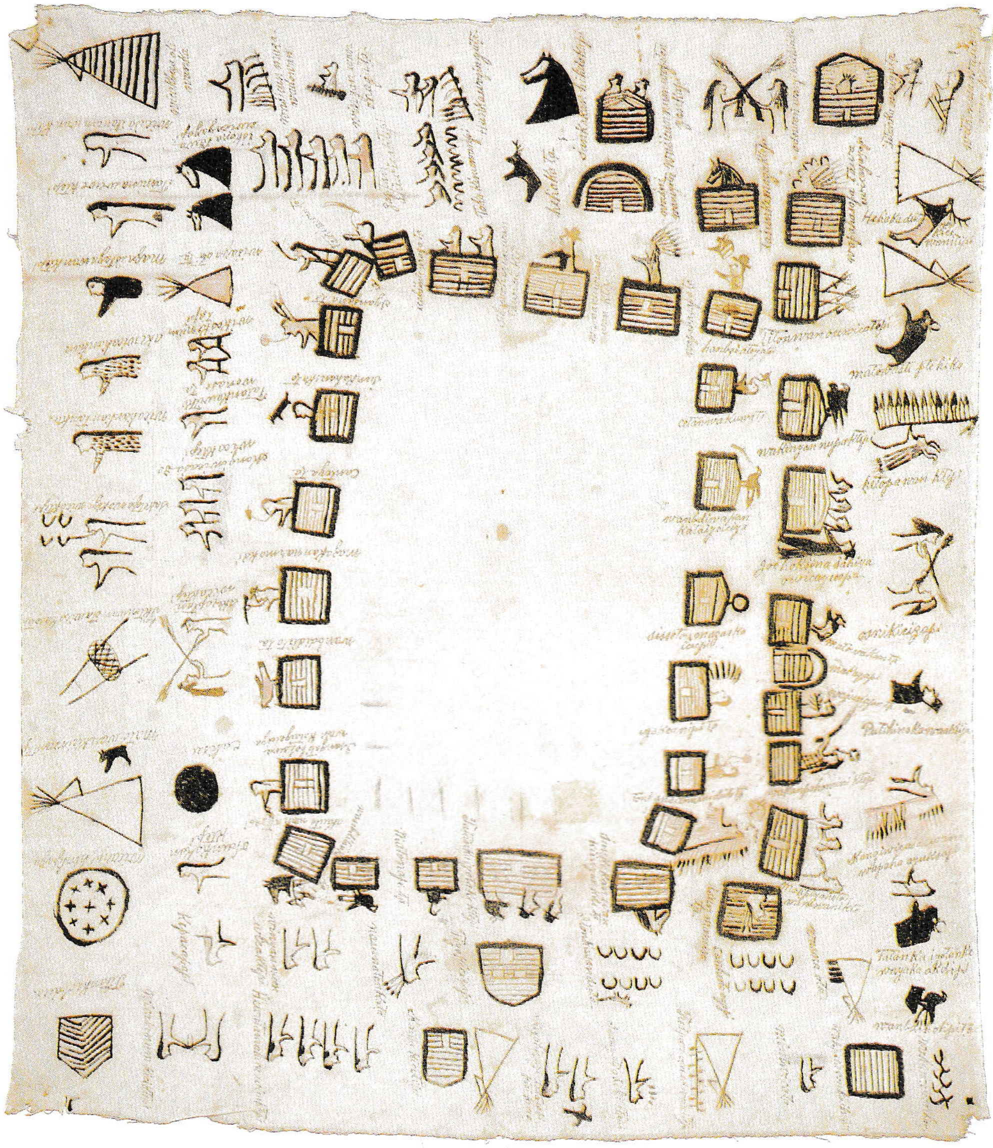
$5.00



The student magazine of math and science

SPRINGER INTERNATIONAL

GALLERY



© The Detroit Institute of Arts (Gift of Mr. and Mrs. Richard A. Pohrt)

Yanktonai Winter Count (ca. 1900–10)

AMONG THE INDIANS OF the American frontier, a winter count was a history in calendar form. Each winter the count keeper would consult with the members of his tribe and select the most important event of the past year. He would then depict that event on an animal hide or, later, on muslin cloth. The winter count thus served as a pictorial aid to tribal memory.

The Upper Yanktonai (Nakota) who created this winter count were part of the Sioux confederation of tribes inhabiting the northern Great Plains and western prairies. It is curious that the count was not kept line by line, but rather in spiral fashion, beginning in the lower-right corner. The count keeper was destined to run out of room in the center of the history.

The winter count spans at least two generations of Yanktonai life. The first entry corresponds to the year 1823 and is labeled "they left the corn standing." In that year the Yanktonai attacked an Arikara village in conjunction with a punitive federal campaign led by Colonel Leavenworth. The Yanktonai took all but the bad corn from the fields nearby. In 1833, according to the pictograph and the caption, "the stars fell." The famous Leonid meteor shower was visible that year throughout North America. The entries for 1837 and 1838 translate as "great scabs on them" and "scabs on them." A smallpox epidemic had claimed many victims in the Yanktonai community during those years.

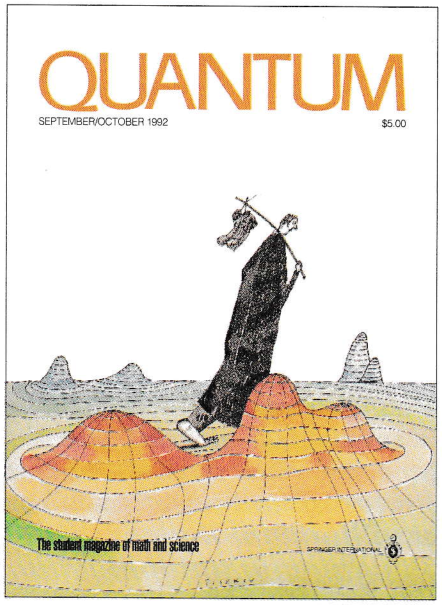
It is evident that a drastic change occurred in 1877 (the pictograph in the second row from the top, just right of center). A thick layer of snow covers a cabin, representing an unusually harsh winter—the Yanktonai's first on a reservation. A reservation-style cabin is depicted for each year thereafter, and most of the entries mark the death of members of the tribe. The pictographs were supplemented by captions some time after this, when transliteration of Nakota words was standardized.

Like all histories, written or depicted, the winter count is meant to take us back in time. The unbroken, linear form of the count is much like a strict chronological narration, but unlike the sentences in most languages, the viewer moves backward as easily as forward in this pictorial account.

Mental time travel is commonplace, though perhaps no less mysterious because of it. What about physical time travel? See page 50.

QUANTUM

SEPTEMBER/OCTOBER 1992 VOLUME 3, NUMBER 1



Cover art by Leonid Tishkov

A mathematician trekking across a landscape? Or a map maker wandering in an abstraction? Whether he's a topologist or a topographer, he's driven by a sense of place (*topos* in Greek).

The word "topic" stems from the same root. In the branch of classical philosophy devoted to rhetoric, topics were "places" where various types of arguments were "stored." So-called special topics were restricted in application (to the law courts, for instance), but common topics, or commonplaces, were valid in any situation. One could use a commonplace like *cause and effect* to explore and speak about anything.

A more colloquial meaning of commonplace is "an obvious statement." It may harbor ambiguities, but its virtue lies in being widely accepted or acceptable at first blush. A commonplace can serve as a nice starting point for serious inquiry.

A commonplace in science says that theories are based on data, not vice versa. Stephen Jay Gould makes a case for vice versa (page 10). And everyone knows that topologists and topographers toil in two completely different fields. Mikhail Shubin takes their shared root as his point of departure in "Topology and the Lay of the Land."

FEATURES

DEPARTMENTS

Science and fanaticism

Reflections on policy in a political season

SCIENCE HAS ALWAYS served humankind. From the times when natural philosophy and technology were inextricably linked, to the present, when science and engineering both have numerous fields and subfields, their primary purpose remains to serve human welfare. In a real sense, such service of a human activity to human needs is but a reflection of our place among the thousands of life forms seeking evolutionary survival. We are driven to such modes of behavior because it sustains our very existence.

The education of one generation of humans by the previous generation is the means by which our cultures of science and technology advance. Each new generation steps in and advances our basic knowledge and the technology that sustains and supports our lives. But each new generation also has the capacity to bring forth ignorance and superstition of massive proportions, leading us to another thousand years of intellectual backwardness, like the Dark Ages of our not-so-distant past.

Several events in recent times may be signs that ignorance, superstition, and a complete failure to understand science may move our society toward such a dark and dangerous period. When that ignorance and superstition is manifested as fanatical or deeply prejudiced political action on the part of powerful groups or individuals, we all have much to fear.

I see two areas of concern in the United States at this time. First are the "animal rights" proponents; then there are those in the fanatical religious right, especially those who have been able to influence political decisions of the Bush Administration.

Animals in research

The animal rights groups assert that human welfare is no more important than that of other animals on Earth and that, therefore, animals should not be used for research that is in any way harmful to them, even when such research offers great promise in saving human lives or leads to important discoveries that improve the lives of suffering human beings. There are literally thousands of examples of how research on animals has directly benefitted humans, from better designs in engineering to life-saving antibiotics and antiviral vaccines. It is sadly ludicrous to assign some special restrictions on human utilization of other life forms for food or research, given the web of interdependence, often quite violent, among the thousands of species that depend on other species for their survival.

None of this requires gratuitous or clearly unnecessary imposition of pain or death on other life forms. But human survival as a species is just as important to the survival of other life forms as others might be in the chain of ecological interdependence. When we behave so as to sustain this species, we are fulfilling our obligation to life itself.

Fetal tissue research

The other, perhaps more serious, area of concern is the threat from the fanatical religious right, especially as it relates to research on human fetuses.

In the United States, as well as in most other industrialized nations of the world, abortion is a legal right. Millions of abortions are performed in the US each year. Some of these abortions are performed because of serious defects or problems with the fetus. In other cases, women choose to have an abortion for personal reasons, and such reasons are entirely private. I believe that, under our constitution and laws, no one has a right to question the privacy of these decisions.

In June 1992, President George Bush, acceding to the demands of the fanatical religious right, along with those of the so-called right-to-life proponents, vetoed a law that would have allowed fetal tissue from abortions to be used in medical research. He declared that, in effect, any such tissue should come from defective fetuses only. One obviously flawed argument put forward by the Bush Administration was that somehow, if fetal tissue were used for medical research, women would get pregnant more often and have abortions more often.

In making this veto, President Bush acted against the desires and will of the vast majority of Americans, and in opposition to the majority of both the House of Representatives and Senate. It is sad that the bill could not receive enough votes to be overridden in the House. And so the President and a small minority of House members have successfully doomed thousands of living human beings to suffering and early death. They have helped delay the chance of preventing and controlling some of the horrible diseases for which fetal tissue research offers such great promise. Parkinson's disease and diabetes are the two areas where fetal tissue research appears to be most promising. There is evidence that both diseases may be controlled by fetal tissue implants, allowing people with these diseases to avoid the deterioration that leads to early disability and death.

I am particularly outraged at the unthinking and cruel act of the President of the United States because my own grandson has diabetes. This gentle little boy must face daily bloodletting from his fragile fingers, shots twice a day, and the regular crises of control, not to mention future loss of limbs, blindness, and a premature death, when we are on the verge of fetal research that could give him a normal life. Let Mr. Bush justify his unfeeling and thoughtless political act to this little boy!

As future scientists, you must be ever on guard against political pressure from fanatics, or others who do not understand science, who seek to interfere with your legitimate research. As the brightest of our next generation, you will be the ones who save, if not my grandson, then perhaps his children, from the dreaded diseases that plague us.

—Bill G. Aldridge

QUANTUM

THE STUDENT MAGAZINE OF MATH AND SCIENCE

A publication of the National Science Teachers Association (NSTA)
& Quantum Bureau of the Russian Academy of Sciences
in conjunction with
the American Association of Physics Teachers (AAPT)
& the National Council of Teachers of Mathematics (NCTM)

Quantum (ISSN 1048-8820) is published bimonthly by the National Science Teachers Association in cooperation with Springer-Verlag New York Inc. Volume 3 (6 issues) will be published in 1992–1993. *Quantum* contains authorized English-language translations from *Kvant*, a physics and mathematics magazine published by the Russian Academy of Sciences and the Russian Academy of Pedagogical Sciences, as well as original material in English. **Editorial offices:** NSTA, 3140 North Washington Boulevard, 2nd Floor, Arlington, VA 22201. **Production offices:** Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010. Production Editor: Madeline R. Kraner

Advertising:

Advertising Representatives: (Washington) Paul Kuntzler (202) 328-5800; (New York) Thomas Heitzman (212) 460-1675; and G. Probst, Springer-Verlag, Heidelberger Platz 3, 1000 Berlin 33, FRG, telephone (0) 30-82 07-1, fax 30-82 07-1.

Second-class postage pending at New York, NY, and additional mailing ofices. **Postmaster:** send address changes to: *Quantum* Springer-Verlag New York, Inc., Journal Fulfillment Services Department, 44 Hartz Way, Secaucus, NJ 07096-2491.

Printed in U.S.A.

Subscription Information: North America

Student rate: \$14; Personal rate (nonstudent): \$18; Institutional rate: \$30; Single issue price: \$5. Rates include postage and handling. (Canadian customers please add 7% GST to subscription price. Springer-Verlag GST registration number is 123394918.) Subscriptions begin with next published issue (backstarts may be requested). Bulk rates for students are available. Send orders to *Quantum*, Springer-Verlag New York, Inc., P.O. Box 2485, Secaucus, NJ 07096; or call 1-800-SPRINGER (in New York, call (201) 348-4033).

All Countries Outside North America

Subscription rates in U.S. currency as above (all rates calculated in DM at the exchange rate current at the time of purchase) plus postage and handling. SAL (Surface Airmail Listed) is mandatory for Japan, India, Australia, and New Zealand. Customers should ask for the appropriate price list. Air mail delivery to all other countries is available upon request. Orders may be placed through your bookseller or directly through Springer-Verlag, Heidelberger Platz 3, 1000 Berlin 33, FRG.

Topology and the lay of the land

A mathematician on the topographer's turf

by Mikhail Shubin

Art by Leonid Tishkov

THE TITLE OF THIS ARTICLE IS SOMEWHAT controversial: topology (unlike topography) doesn't have to do with relief maps of the Earth's surface. It's not concerned with terrain at all. This branch of mathematics explores properties of geometric figures that are usually regarded as the most fundamental—the properties preserved under very general kinds of transformation. These transformations, called *topological transformations*, or *homeomorphisms*, are defined as one-to-one mappings that, together with their inverses, are continuous. (To get a good grip on this notion, as well as other important topological ideas, see "Flexible in the Face of Adversity" in the September/October 1990 issue of *Quantum*.)

Objects that look quite different on the outside can have the same topological properties, if they can be turned into one another by way of a continuous deformation. Martin Gardner once described topologists as mathematicians who can't tell a coffee cup from a doughnut. Sometimes topological properties and values arise as combinations of quite simple geometrical properties and values. This just happens to be the case with Euler's famous theorem on polyhedrons (Theorem 2 below). But I'll begin with another example of this sort—a theorem from "Morse theory."[1] And here is where terrain—the "lay of the land"—comes into play.

On geographical maps the relief of the landscape is usually shown by means of

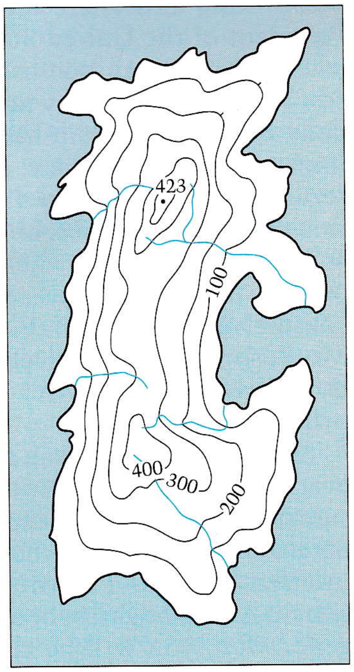


Figure 1

[1]Harold Marston Morse (1892–1977), an outstanding American mathematician, created a theory that later took his name and represents an important branch of modern topology.

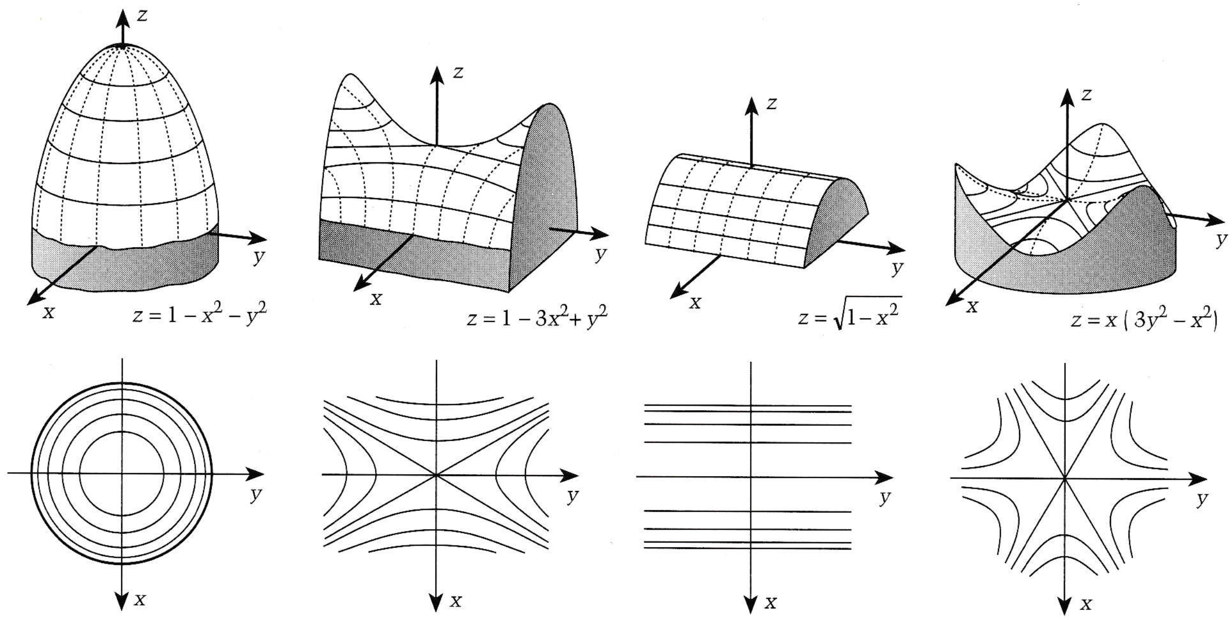


Figure 2 a b c d
Examples of geographic surfaces.

altitude lines—that is, lines connecting points on the map that represent points on the Earth at the same altitude above sea level (fig. 1). The mathematical model for a bounded region of the Earth's surface is a geometrical surface in space defined by an equation of the form $z = f(x, y)$. Examples of such surfaces are shown in figure 2. Solid lines on each surface are sets of points lying at the given height z = constant; the altitude lines are their projections onto the (x, y)-plane. (Dotted lines on the surfaces are the lines of steepest descent and are perpendicular to the altitude lines.) To sum up: an altitude line is a set of points (x, y) satisfying an equation $f(x, y)$ = constant; the constants in these equations in effect "enumerate" the respective altitude lines. In figure 2 the families of altitude lines are placed under the corresponding surfaces.

To formulate our first theorem, we must get acquainted with different types of points on a surface with respect to changes in the surface height near the points. The first three types—summits, basins, and mountain passes (figures 3a through 3c)—are "critical" or "equilibrium" points in our "landscape": if you set a ball at any such point, it will remain there forever (though the equilibrium is stable only for basins). For each kind of point I'll give both a graphical description and a more rigorous definition in terms of altitude lines.

A *summit* is shown in figure 3a. It might seem that a summit can be defined simply as the highest point in its neighborhood. However, the structure of the surface around such a maximum point can be much more complicated than that in figure 3a. For instance, one can imagine a maximum point that is the accumulation point of a sequence of other maximum points with increasing altitudes (that is, a limiting case of figure 4, with infinitely many smaller peaks): any neighborhood of such a point contains other maximum points. To

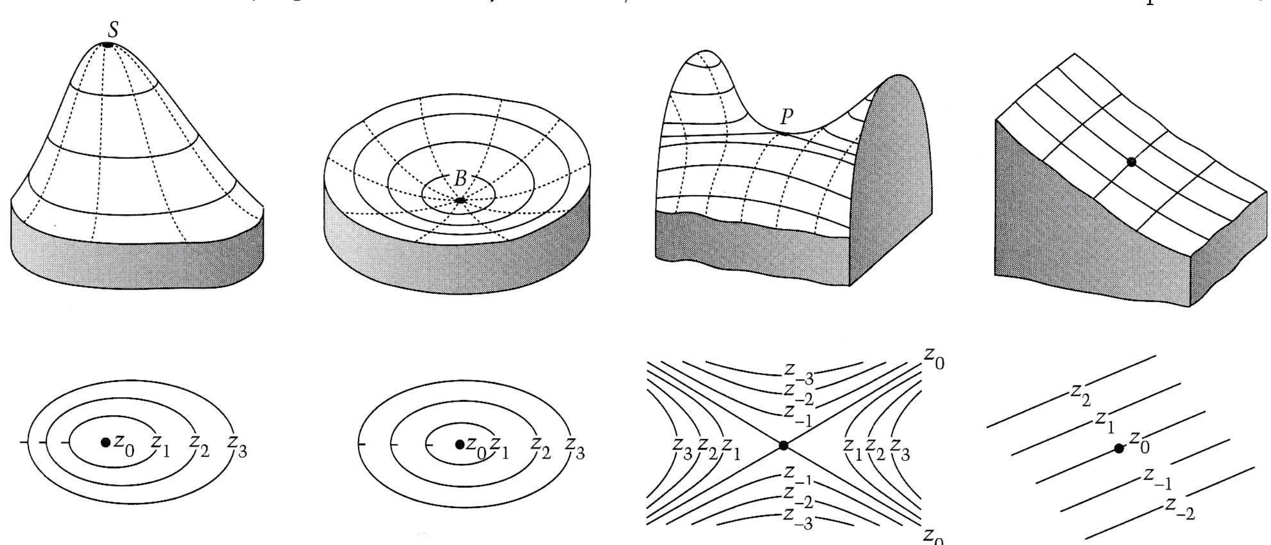


Figure 3 a b c d

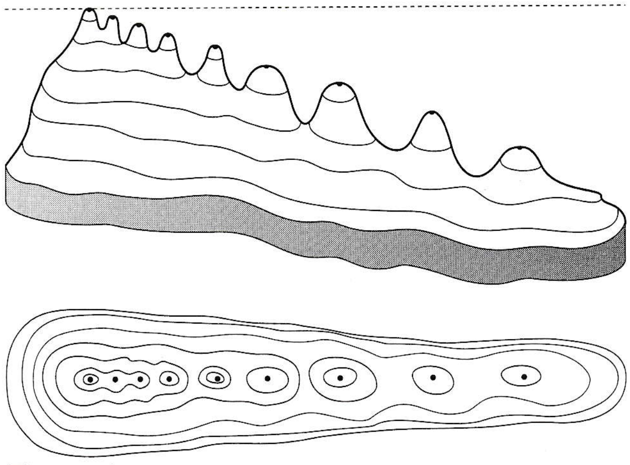


Figure 4

exclude such situations, we define a summit as a point that is the highest one in its vicinity, and in addition, the altitude lines around it are arranged as in figure 3a—that is, they are closed non-self-intersecting nested curves embracing the point, their respective heights decreasing as we recede from it: $z_0 > z_1 > z_2 > \ldots$.

A *basin* (fig. 3b) is an upside-down summit—it becomes a summit when the sign of the function $f(x, y)$ is changed. So it's a point such that the neighboring altitude lines look as in figure 3b (which is the same as in figure 3a), but their heights increase as they move away from the point: $z_0 < z_1 < z_2 < \ldots$.

A mountain *pass* (fig. 3c) can be described as a point surrounded by terrain that looks like a saddle.[2] Whereas the altitude lines passing through a summit or basin degenerate into a single point, the altitude lines through a pass P consist of two intersecting lines. These lines (or rather their tangents at P) divide the neighborhood into two pairs of vertical angles. Moving away from P in any direction lying in one of these pairs of angles, we ascend; moving along directions that fill up the other pair of angles, we descend: in figure 3c, … $z_{-2} < z_{-1} < z_0 < z_1 < z_2 < \ldots$.

The fourth type of point is the most common. Unlike what happens at critical points, a ball set at such a point will roll down, since the surface at this point is sloping. We call such a point a *slope point*. The structure of altitude lines near a slope point is seen in figure 3d. In particular, the altitude line passing through the point itself consists of one piece, and the height of the neighboring altitude lines change monotonically as we move them along a path through the slope point: in figure 3d, … $z_{-2} < z_{-1} < z_0 < z_1 < z_2 < \ldots$.[3]

As a rule, almost all points on a map are slope points. For example, all basins, summits, and passes are isolated and surrounded by slope points.

Of course, there are other types of points in addition to the four types defined above. For instance, some maps have plateaus (entire regions whose points are all at the same altitude) or ridges (a line, rather than a point, of maximum altitude—see figure 2c), or even "trickier" points, like the triple pass in figure 2d or the string of descending summits in figure 4. But all points that are not slope points (other than summits, basins, and passes) can be removed by changing the terrain ever so slightly: the "bad" points either disappear or turn into a number of summits, basins, and passes.[4] So it's reasonable to restrict ourselves to only these three types of critical points in addition to slope points.

Now we can formulate the main theorem in this article—one of the simplest theorems in Morse theory.

Theorem 1. *Figure 5 depicts an island whose every point is either a slope point, a summit, a basin, or a mountain pass, and in addition all the points of its coastline are slope points. If S, B, and P are the numbers of summits, basins, and passes, respectively, then*

$$S + B - P = 1.$$

Proof. To make the proof clearer, I'll divide it into three steps.

1. *Reconstructing the terrain.* Let's change the terrain of the island without changing the numbers S, B, and P so as to satisfy the following conditions:

(a) All the summits are the same height (equal to the altitude of Mount Everest, the highest mountain on Earth);
(b) All the basins are at sea level;
(c) All the passes are at different altitudes.

Condition (a) is the easiest to satisfy: we simply add some earth on top of our hills and mountains to make

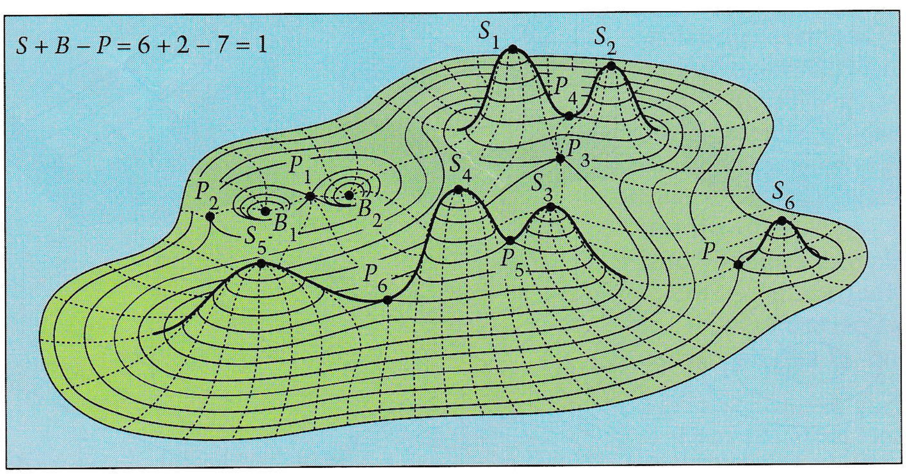


Figure 5

[2]Topologists sometimes call this a saddle point.—*Ed.*

[3]A more formal description can be given in terms of topological transformations: the family of altitude lines near a slope point can be obtained by way of such transformation from a family of parallel lines; similarly, altitude lines near a summit or basin can be obtained from a family of concentric circles and, near a mountain pass, from a family of hyperbolas with common asymptotes.

[4]This fact is rigorously formulated and proved in Morse theory, but we won't need that theorem here.

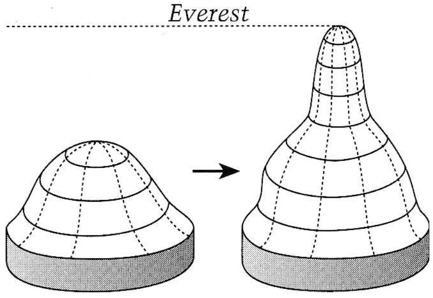


Figure 6

them reach the necessary Everestian altitude (fig. 6). At the same time let's dig a pit at the bottom of each basin to reach the depth of the Earth's deepest trough (the Marianas trench in the Pacific Ocean). This will be needed to ensure that condition (b) is satisfied, which is not that easy: trying to fill up the basins (to raise their level), we may stumble upon a pass. But there's a simple way out: drain the sea around the island. We can assume that the island has steep underwater cliffs that go as deep as the Marianas trench. Then the sea can simply be drained to this depth, and all the basins are at the (new) sea level.

To satisfy condition (c) we must be able to slightly lift or depress the neighborhood of a given mountain pass without touching other critical points or creating new summits, basins, or passes. To this effect, let's draw two circles of radii r and $2r$ with small enough r and centered at the pass. Points outside the bigger circle will be left intact; part of the surface inside the inner circle will be lifted or depressed as required; and the ring between the circles is a sort of connective tissue (see figure 7).

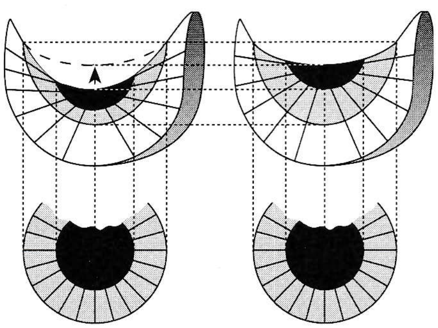


Figure 7
Lifting a mountain pass. The black part of the surface is lifted slightly as a rigid curved plate; the gray part (think of it as being made of a thin rubber sheet) is stretched slightly from the fixed outer boundary.

Thus, we can suppose conditions (a), (b), and (c) to be true.

2. *The big flood.* Imagine that rain gradually inundates the island so that the water level in the basins and in the sea uniformly rises from the initial "sea level" to the height of Everest. Lakes appear in the basins in the island's interior, and the island is divided into smaller islands by the rising water. Let's watch the change in the number of islands and lakes (naturally, we won't consider the sea a lake).

Right after the flood begins, a lake is formed in every basin. So the initial number of lakes is B. By the end of the flood all the lakes become connected with the sea, and only the tops of all the mountains are still towering above the water (that's why we raised them up!). We can draw up the following table.

	Lakes	Islands
Beginning of flood	B	1
Before full inundation	0	S

3. *Flooding a mountain pass.* When do the numbers of lakes and islands change? Clearly, not as long as only slope points are flooded. What happens when a mountain pass is flooded (fig. 8)? Notice that only one pass can be flooded at a time, since all these points are at different altitudes. So there are two possibilities:

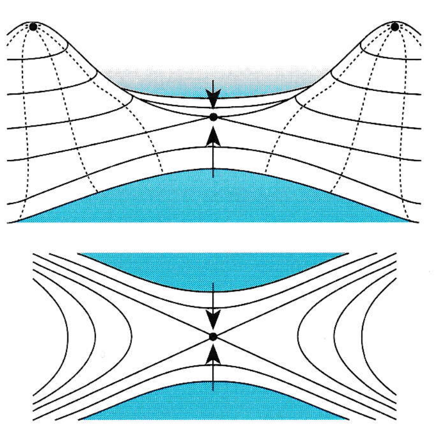


Figure 8

(a) Water from a single lake, or from the sea, flows together from two directions; then the number of lakes doesn't change, but a new island appears (fig. 9a);

(b) Two different lakes flow together (fig. 9b); then the number of lakes decreases by one, while the number of islands remains the same.

Thus, when water comes to a mountain pass, either the number of lakes decreases by one or the number of islands increases by one. Since the total change in these numbers is B and $S - 1$, respectively, we have $P = B + (S - 1)$—that is, $S + B - P = 1$, completing the proof.

Corollary. *If every point on the Earth is either a slope point, a summit, a basin, or a mountain pass, then*

$$S + B - P = 2.$$

Proof. Drain the sea and all the lakes and pour some water into the bottom of the deepest basin to make a single lake. We can regard this lake as the sea and all the land as one island. Then the entire Earth has as many summits and mountain passes as the island and has one more basin than the island. Now use Theorem 1.

An important example of terrain to which this corollary can be applied is any polyhedron containing a point O whose perpendicular projections onto each face and each edge lie inside this face or edge (and not on their extensions). If the altitude at a given point of the polyhedron is defined as the distance from O to this point,

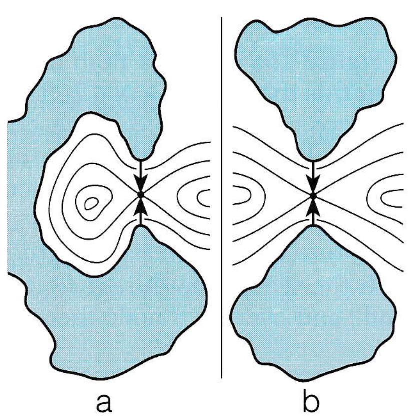


Figure 9

then the vertices of the polyhedron are the summits of this terrain, the projections of point O onto the faces are basins, and the projections of O onto the edges are mountain passes. (Check it yourself!) This observation yields the following theorem.

THEOREM 2 (Euler's theorem on polyhedrons). *Let F, V, and E be the numbers of faces, vertices, and edges of a convex polyhedron. Then*

$$F + V - E = 2.$$

So far we've proved this theorem only for a special class of polyhedrons that satisfy the additional condition above. The proof for the general case is given below; it is easily derived from the next theorem.

THEOREM 3 (Euler's theorem on maps). *Let there be a political map*[5] *of an island bounded by a closed curve and satisfying the following conditions:*

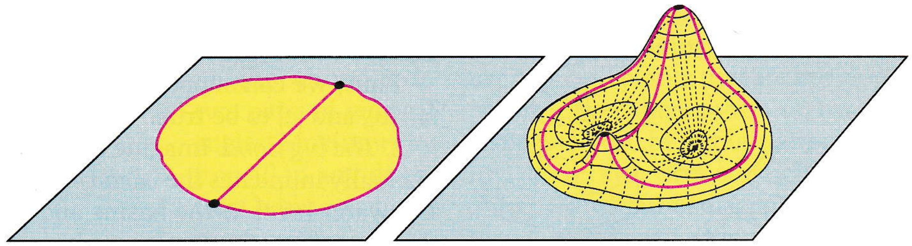
(a) *The number of countries is greater than or equal to 2;*
(b) *Each country is bounded by a closed curve without self-intersections;*
(c) *None of the countries lies inside another (like Monaco or the Vatican).*

If C is the number of countries, N is the number of nodes (junction points of three or more countries or the sea), and B is the number of borders (that is, segments of borders between two nodes, counting the borders with the sea), then

$$C + N - B = 1.$$

Figure 10a shows a map drawn from this theorem ($C = N = 2$, $B = 3$).

PROOF OF THEOREM 3. Let's construct a landscape on the map such that each country contains exactly one basin, each border contains exactly one mountain pass (borders with the sea can be shifted a bit inland), and over each node there is a summit. This can be done as follows: erect a watchtower shaped like the Eiffel tower over each node; build ramparts along the borders descending toward the midpoint of each border, gently sloping down on both sides; and dig a basin inside each country, its slopes smoothly merging with the slopes of ramparts. (Such a landscape for the map in figure 10a is shown in figure 10b.) Now we apply Theorem 1 and we're done.



a b

Figure 10

COROLLARY. *If a political map drawn on a sphere has not less than 3 countries and satisfies conditions (b) and (c) of Theorem 3, then for this map*

$$C + N - B = 2.$$

PROOF. Let's think of one of the countries as a sea and the rest of the globe as one island. Then Theorem 3 is applicable and quickly leads to the desired result.

Now we can go back and prove the previous theorem.

PROOF OF THEOREM 2 (for an arbitrary convex polyhedron). Take a point O inside the given polyhedron and a large sphere whose center is at O. Project the polyhedron onto the sphere from point O (the projection of some point X is defined as the point where the ray OX meets the sphere). In this way we obtain a spherical political map whose countries and borders are the projections of the faces and edges of the polyhedron. To complete the proof, apply the above corollary to this map.

Problems

1. State and prove assertions analogous to the corollaries to theorems 1 and 3 for the case of a torus (the surface of a doughnut shape—see figure 11).

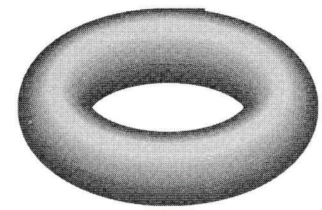


Figure 11

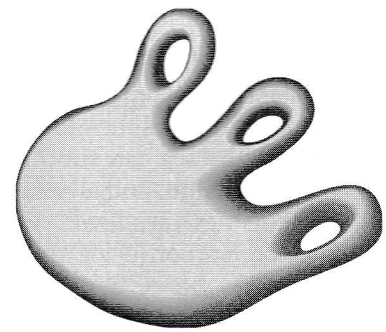


Figure 12

2. Do the same for a "sphere with p handles" (figure 12 shows a sphere with three handles).

[5] That is, a map showing the countries in different colors (as opposed to a topographic map).—*Ed.*

Dinosaurs in the haystack

Does it matter whether a world ends with a bang or a whimper?

by Stephen Jay Gould

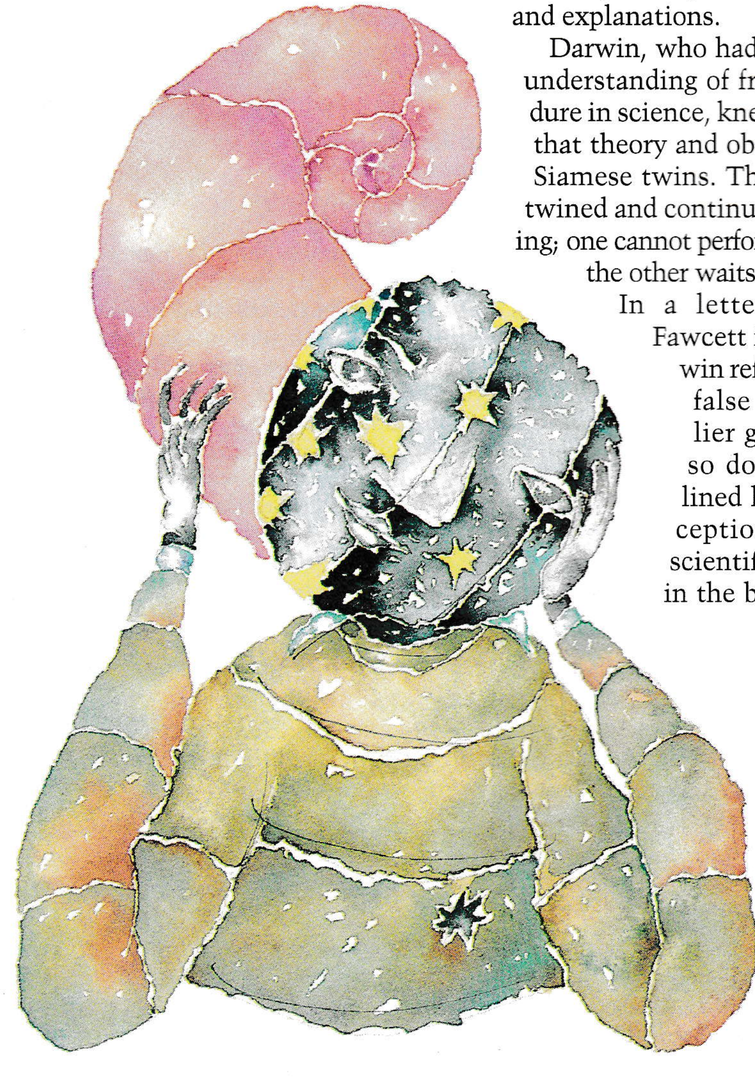
THE FASHION INDUSTRY thrives on our need to proclaim an identity from our most personal space. For academics, who by stereotype (although not always in actuality) scorn the sartorial mode, office doors serve the same function. Professorial entranceways are festooned with testimonies of deepest beliefs and strongest commitments. We may, as a profession, have a deserved reputation for lengthy and tendentious proclamation, but our office doors feature the gentler approach of humor or epigram. The staples of this genre are cartoons (with Gary Larson as the unchallenged *número uno* for scientific doors) and quotations from gurus of the profession.

Somehow, I have never been able to put someone else's cleverness so close to my soul. I wear white T-shirts, and although I wrote the preface to one of Gary Larson's *Far Side* collections, I would never identify my portal with his brilliance. But I do have a favorite quotation—one fit for shouting from the housetops (if not for inscription on the doorway).

My favorite line, from Darwin of course, requires a little explication. Geology, in the late eighteenth century, had been deluged with a rash of comprehensive, but mostly fatuous, "theories of the earth"—extended speculations about everything, generated largely from armchairs. When the Geological Society of London was inaugurated in the early nineteenth century, the founding members overreacted to this admitted blight by essentially banning all theoretical discussion from their proceedings. Geologists, they ruled, should first establish the facts of our planet's history by direct observation, and then, at some future time when the bulk of accumulated information becomes sufficiently dense, move to theories and explanations.

Darwin, who had such a keen understanding of fruitful procedure in science, knew in his guts that theory and observation are Siamese twins. They are intertwined and continually interacting; one cannot perform first while the other waits in the wings.

In a letter to Henry Fawcett in 1861, Darwin reflected on the false view of earlier geologists. In so doing, he outlined his own conception of proper scientific procedure in the best one-liner



Art by Sergey Ivanov

ever penned. The last sentence is indelibly impressed on the portal to my psyche.

> About thirty years ago there was much talk that geologists ought only to observe and not theorize; and I well remember someone saying that at this rate a man might as well go into a gravel-pit and count the pebbles and describe the colors. How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service!

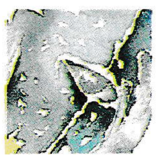
The point should be obvious. Immanuel Kant, in a famous quip, said that concepts without percepts are empty, whereas percepts without concepts are blind. The world is so complex; why should we strive to comprehend with only half our tools? Let our minds play with ideas, let our senses gather information, and let the rich interaction proceed as it must (for the mind processes what the senses gather, while a disembodied brain, devoid of all external input, would be a sorry instrument indeed).

Yet scientists have a peculiar stake in emphasizing fact over theory, percept over concept—and Darwin wrote to Fawcett to counteract this odd but effective mythology. Scientists often strive for special status by claiming a unique form of objectivity inherent in a supposedly universal procedure called *the* scientific method. We attain this objectivity by clearing the mind of all preconception and then simply seeing, in a pure and unfettered way, what nature presents. This image may be beguiling, but the claim is chimerical and ultimately haughty and divisive. For the myth of pure perception raises scientists to a pinnacle above all other struggling intellectuals, who must remain mired in constraints of culture and psyche.

But followers of the myth are ultimately hurt and limited, for the immense complexity of the world cannot be grasped or ordered without concepts. "All observation must be for or against some view if it is to be of any service!" Objectivity is not an unobtainable emptying of mind but a willingness to abandon a set of preferences when the world seems to work in a contrary way.

This Darwinian theme of necessary interaction between theory and observation gains strong support from a scientist's standard "take" on the value of original theories. Sure, we love them for the usual "big" reasons—because they change our interpretation of the world and lead us to order things differently. But ask any practicing scientist, and you will probably get a different primary answer, for we are hung up on the details and rhythms of our daily work, and we don't think about ultimates very often. We love original theories because they suggest new, different, and tractable ways to make observations. By posing new questions, they expand our range of ordinary activity. Theories drive us to seek new information that only becomes relevant as data either for or against a hot idea. Data adjudicate theory, but theory also drives and inspires data. Both Kant and Darwin were right.

The problem of mass extinctions



I bring up this personal favorite among quotations because my profession of paleontology has recently witnessed a fine example of theory confirmed by data that no one ever thought of collecting before the theory itself demanded such a test. (Please note the fundamental difference between demanding a test and guaranteeing the result. The test might just as well have failed, thus dooming the theory. Good theories invite a challenge but do not bias the outcome. In this case, the test succeeded twice, and the theory has gained strength.) Ironically, this particular new theory would have been anathema to Darwin himself, but such a genial and generous man would, I am sure, have gladly taken his immediate lumps in exchange for such a fine example of his generality about theory and observation, and for the excitement of any idea so full of juicy implications.

We have known since the dawn of modern paleontology that short stretches of geological time feature extinctions of substantial percentages of life—up to 96 percent of marine invertebrate species in the granddaddy of all such events, the late Permian debacle some 225 million years ago. These "mass extinctions" were originally explained, in a literal and common-sense sort of way, as products of catastrophic events, and therefore truly sudden. As Darwin's idea of gradualistic evolution replaced this earlier catastrophism, paleontologists sought to mitigate the evidence of mass dying with a reading more congenial to Darwin's preference for slow and steady. The periods of enhanced extinction were not denied—how could they be in the face of such evidence?—but they were reinterpreted as more spread out in time and less intense in effect: in short, as intensifications of ordinary processes, rather than impositions of true and rare catastrophes.

In the *Origin of Species* (1859), Darwin rejected "the old notion of all the inhabitants of the earth having been swept away at successive periods by catastrophes," as well he might, given the extreme view of total annihilation, with its antievolutionary implication of a new creation to start life again. But Darwin's preferences for gradualism were also extreme and false:

> We have every reason to believe . . . that species and groups of species gradually disappear, one after another, first from one spot, then from another, and finally from the world.

Darwin himself had to admit the apparent exceptions:

> In some cases, however, the extermination of whole groups of beings, as of the ammonites towards the close of the secondary period, has been wonderfully sudden.

We now come to the central irony that inspired this essay. So long as Darwin's gradualistic view of mass extinction prevailed, paleontological data, read literally, could not refute the basic premise of gradualism—the "spreading out" of extinctions over a good stretch of time before the boundary, rather than a sharp concentration of disappearances right at the boundary itself. For the geological record is highly imperfect and only a tiny fraction of living creatures ever become fossils. As a consequence of this imperfection, even a truly sudden and simultaneous extinction of numerous species will be recorded as a more gradual decline in the fossil record. This claim may sound paradoxical, but consider the following argument and circumstance.

Some species are very common and easily preserved as fossils; we may, on average, find specimens in every inch of strata. But other species will be rare and poorly preserved, and we might encounter their fossils only once every 100 feet or so. Now suppose that all these species died suddenly at the same time, after 400 feet of sediment had been deposited in an ocean basin. Would we expect to find the most direct evidence for mass extinction—that is, fossils of all species through all 400 feet of strata right up to the very top of the sequence? Of course not.

Common species would pervade the strata to the top, for we expect to find their fossils in every inch of sediment. But even if rare species live right to the end, they only contribute a fossil every 100 feet or so. In other words, a rare species may have lived through 400 feet, but its last fossil may be entombed 100 feet below the upper boundary. We might then falsely assume that this rare species died out after three-fourths of the total time had elapsed.

Generalizing this argument, we may assert that the rarer the species, the more likely that its last fossil appears in older sediments even if the species actually lived to the upper boundary. If all species died at once, we will still find a graded and apparently gradualistic sequence of disappearances, the rare species going first and the common forms persisting as fossils right to the upper boundary. This phenomenon—a classic example of the old principle that things are seldom what they seem and that literal appearances often obscure reality—even has a name: the Signor–Lipps effect, to honor two of my paleontological buddies, Phil Signor and Jere Lipps, who first worked out the mathematical details of how a literal petering-out might represent a truly sudden and simultaneous disappearance.

The Alvarez hypothesis

We can now sense the power of Darwin's argument about needing theories to guide observations. We say, in our mythology, that old theories die when new observations derail them. But too often, indeed I would say usually, theories act as straitjackets to channel observations toward their support and to forestall data that might refute them. Such theories cannot be rejected from within, for we will not conceptualize the potentially refuting observations. If we accept Darwinian gradualism in mass extinctions, and therefore never realize that a graded series of fossil disappearances might, by the Signor–Lipps effect, actually represent a sudden wipeout, how will we ever come to consider the catastrophic alternative? For we will be smugly satisfied that we have "hard" data to prove gradualistic decline in species numbers.

New theories are to this conceptual block what Harry Houdini was to straitjackets. We escape by importing a new theory and by making the different kinds of observations that any novel outlook must suggest. For "all observation must be for against some view," and a new view can therefore engender original observations. I am not making an abstract point or waving arms for my favorite Darwinian motto. Recently, two lovely examples with the same message have been published by a pair of my closest colleagues: studies of ammonites and dinosaurs through the last great extinction.

Anyone who keeps up with press reports on hot items in science knows that a new catastrophic theory of mass extinction has illuminated the paleontological world (and graced the cover of *Time* magazine) during the past decade. In 1980, the father–son (and physicist–geologist) team of Luis and Walter Alvarez published, with colleagues Frank Asaro and Helen Michel, their argument and supporting data for extraterrestrial impact of an asteroid or comet as the cause of the Cretaceous–Tertiary extinction, most recent of the great mass dyings and the time of extinction for dinosaurs along with some 50 percent of marine invertebrate species.

This proposal unleashed a furious debate that cannot be summarized in a page, much less an entire essay, or even a book. Yet I think it fair to say that the idea of extraterrestrial impact has weathered this storm splendidly and continually increased in strength and supporting evidence. At this point, very few scientists deny that an impact occurred, and debate has largely shifted to whether the impact caused the extinction in toto (or only acted as a *coup de grâce* for a process already in the works), and whether other mass extinctions may have had a similar cause.

New forms of supporting evidence are reported on a monthly basis in almost every issue of major journals. In the past few weeks we have learned about minute diamonds in sediments from the impact boundary. Diamonds are a form of pure carbon produced under immense pressures that impacts, and no other known process active at the earth's surface, can generate. This discovery may represent the literal fulfillment of that Beatles classic about psychic hallucinations, "Lucy in the Sky with Diamonds" (and its obvious acronym). Lucy is, literally, light—and the impact was quite a blast. Moreover, the smoking gun may now have been located as a massive crater in the Gulf of Mexico, off the Yucatán Peninsula.

Paleontologists, with very few exceptions, reacted negatively, to say the least, and Luis Alvarez, a virtual model for the stereotype of the self-assured physicist, was fit to be tied. Luis, in retrospect, was also mostly right, so I forgive his fulminations against my profession. I, if I may toot my horn, was among his few initial supporters, but not for the right reason of better insight into the evidence. Catastrophic extinction simply matched my idiosyncratic preference for rapidity, born of the debate over punctuated equilibrium (see my essay of August 1991 [in *Natural History*]). After all, my colleagues had been supporting Darwinian gradualism for a century, and the fossil record, read literally, seemed to indicate a petering-out of most groups before the boundary. How could an impact cause the extinction if most species were already dead? But the extraterrestrial impact theory soon proved its mettle in the most sublime way of all: by Darwin's criterion of provoking new observations that no one had thought of making under old views. The theory, in short, engendered its own test and broke the straitjacket of previous certainty.

Tearing apart the haystack

My colleagues may have disliked the Alvarez hypothesis with unconcealed vigor, but we are an honorable lot, and as debate intensified and favorable evidence accumulated, paleontologists had to take another look at their previous convictions. Many new kinds of observations can be made, but let us focus on the simplest, most obvious, and most literal example. In the light of new prestige for impact and sudden termination, the Signor–Lipps argument began to sink in, and paleontologists realized that catastrophic wipeouts might be recorded as gradual declines in the fossil record.

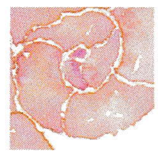
How then to break the impasse of this indecisive broad appearance of petering-out? Many procedures, some rather subtle and mathematical, have been proposed and pursued, but why not start with the most direct approach. If rare species actually lived right to the impact boundary but have not yet been recorded from the uppermost strata, why not look a whole lot harder. The obvious analogy to the usual cliché suggests itself. If I search for a single needle in the haystack by sampling ten handfuls of hay, I have very little chance of locating it. But if I take apart the stack, straw by straw, I will recover the needle. Similarly, if I really search every inch of sediment in every known locality, I might eventually find even the rarest species right near the boundary—if it truly survived.

This all seems rather obvious. I cannot possibly argue that such an approach was unthinkable before the Alvarez hypothesis. I cannot claim that conceptual blinders of gradualism made it impossible even to imagine pulling apart the haystack rather than sampling it. But this example becomes so appealing precisely through its entirely pedestrian character. I could cite many fancy cases of original theories that open entirely new worlds of observation: think of Galileo's telescope and all the impossible phenomena thus revealed. In this case, the Alvarez theory suggested little more than hard work.

So why wasn't the effort expended before? Paleontologists are an industrious lot; we have faults aplenty, but laziness in the field is not among them. We do love to find fossils; this is why most of us entered the profession in the first place. We didn't scrutinize every inch of sediment for the most basic of all scientific reasons. Life is short and the world is immense; you can't spend your career on a single cliff face. The essence of science is intelligent sampling, not sitting in a single place and trying to get every last one. Under Darwinian gradualism, intelligent sampling followed the usual handful-from-the-haystack method. The results obtained matched the expectations of theory, and conceptual satisfaction (in retrospect, one might say "sloth") set in. No impetus existed for the much more laborious dismember-the-entire-haystack method, a quite unusual approach in science. We could have worked this way, but we didn't because we had no reason to do so. The Alvarez theory made this unusual approach necessary. It forced us to look in a different way. "All observation must be for or against some view if it is to be of any service!"

Consider two premier examples—the best-known marine terrestrial groups to disappear in the Cretaceous–Tertiary extinction: ammonites and dinosaurs. Both had been prominently cited as support for gradual extinction toward the boundary. In each case, the Alvarez hypothesis inspired a closer look via the dismember-the-haystack method, and in each case, this greater scrutiny yielded evidence of persistence to the boundary and potentially catastrophic death.

Ammonites at Armageddon . . .



Ammonites are cephalopods (mollusks classified in the same group as squids and octopuses) with coiled external shells closely resembling those of their nearest living relative, the chambered nautilus. They were a prominent, and often dominant, group of marine predators, and their beautiful fossil shells have always been prized by collectors. They arose in mid-Paleozoic times and had nearly become extinct twice before—in two other mass dyings at the end of the Permian and of the Triassic periods. But a lineage or two had scraped by each time. At the Cretaceous–Tertiary boundary, however, all lineages succumbed, and to cite Wordsworth from another context, there "passed away a glory from the earth."

My friend and colleague Peter Ward, paleontologist from the University of Washington, is one of the world's experts on ammonite extinction; a vigorous, committed man who adores fieldwork and could never be accused of laziness on the outcrop. Peter didn't care much for Alvarez at first, largely because his ammonites seemed to peter out and disappear entirely some thirty feet below the boundary at his favorite site, the cliffs of Zumaya on the Bay of Biscay in Spain. In 1983, Peter wrote an article for *Scientific American* entitled "The Extinction of the Ammonites." He stated his opposition to the Alvarez theory, then so new and controversial, at least as an explanation for the death of ammonites:

The fossil record suggests, however, that the extinction of the ammonites was a

> consequence not of this catastrophe but of sweeping changes in the late Cretaceous marine ecosystem. . . . Studies of the fossils from the stratigraphic sections at Zumaya in Spain suggest they became extinct long before the proposed impact of the meteoritic body.

But Peter, as one of the smartest and most honorable men I know, also acknowledged the limits of such "negative evidence." A conclusion based on *not* finding something has the great virtue of unambiguous potential refutation. Peter wrote: "This evidence is negative and could be overturned by the finding of a single new ammonite specimen."

Without the impact hypothesis, Peter would have had no impetus to search these upper thirty feet of section with any more care. Extinctions were supposed to be gradual, and thirty feet of missing ammonites made perfect sense, so why look any further. But the impact hypothesis, with its clear prediction of ammonite survival right up to the boundary itself, demanded more intense scrutiny of the thirty-foot haystack. In 1986, Peter was still touting sequential disappearance: "Ammonites . . . appear to have become extinct in this basin well before the K/T [Cretaceous–Tertiary] boundary, supporting a more gradualistic view of the K/T extinctions" (*Palaios*, vol. 1, pp. 87–92).

But Peter and his field partners, inspired by Alvarez (if only by a hope of disproving the impact hypothesis), worked on through the haystack: "The remaining part of the Cretaceous section was well exposed and vigorously searched and quarried." Finally, later in 1986, they found a single specimen just three feet below the boundary. It was crushed, and they couldn't tell for certain whether it was an ammonite or a nautiloid, but this specimen did proclaim a need for even more careful search. (Since nautiloids obviously survived the extinction—the chambered nautilus still lives today—such a fossil right at the boundary would occasion no surprise.)

Peter started a much more intense search in 1987, and the ammonites began to turn up—mostly lousy specimens and very rare, but clearly present right up to the boundary. Peter writes in a book just published:

> Finally, on a rainy day, I found a fragment of an ammonite within inches of the clay layer marking the boundary. Slowly, over the years, several more were found in the highest levels of Cretaceous strata at Zumaya. Ammonites appeared to have been present for Armageddon after all.

Peter then took the obvious next step: look elsewhere. Zumaya contained ammonites right up to the end, but not copiously, perhaps for reasons of local habitat rather than global abundance. Peter had looked in sections west of Zumaya and found no latest Cretaceous ammonites (another reason for his earlier acceptance of gradual extinction). But now, he extended his fieldwork to the east, toward the border of Spain and France. (Again, these eastern sections were known and had always been available for study, but Peter needed the impetus of Alvarez to ask the right questions and to develop a need for making these further observations.) Peter studied two new sections, at Hendaye on the Spanish–French border and right on the yuppie beaches of Biarritz in France. He found an abundance of ammonites just below the boundary line of the great extinction. He writes in his new book:

> After my experience at Zumaya, where years of searching yielded only the slightest evidence . . . near the Cretaceous–Tertiary boundary, I was overjoyed to find a source of ammonites within the last meter of Cretaceous rock during the first hour at Hendaye.

. . . and dinosaurs, too

We professionals may care more about ammonites, but dinosaurs fire the popular imagination. No argument against Alvarez has therefore been more prominent, or more persuasive, than the persistent claim by most (but not all) dinosaur specialists that the great beasts, with the possible exception of a straggler or two, were gone long before the supposed impact.

I well remember the dinosaur men advancing their supposed smoking gun of a "three-meter gap"—the barren strata between the last-known dinosaur bone and the impact boundary. And I recall Luis Alvarez exploding in rage and with ample justice (for I felt a bit ashamed of my paleontological colleagues and their very bad argument). The last bone, after all, is not the last animal, but rather a sample from which we might be able to estimate the probable later survival of creatures not yet found as fossils. After all, if my colleague throws a thousand bottles overboard and I later pick up one on an island fifty miles away, I do not assume that he only tossed a single bottle. But if I know the time of his throw and the pattern of currents, I might be able to make a rough estimate of how many he originally dropped overboard. The chance of any animal becoming a fossil is surely much smaller than the probability of my finding even one bottle. All science is intelligent inference: excessive literalism is a delusion, not a humble bowing to evidence.

Again, as with Peter Ward and the ammonites, the best empirical approach would order a stop to the shouting and organize a massive effort to dismember the haystack by looking for dinosaur bones in every inch of latest Cretaceous rocks. Peter means "rock" in Latin, so maybe men of this name are predisposed to a paleontological career. Another Peter, my friend and colleague Peter Sheehan of the Milwaukee Public Museum, has been guiding such a project for years. Just last month (I write this essay in December 1991), he published his much awaited results (see "Sudden Extinction of the Dinosaurs: Latest Cretaceous, Upper Great Plains, U.S.A.," by P. M. Sheehan, D. E. Fastovsky, R. G. Hoffmann, C. D. Berghaus, and D. L. Gabriel, *Science*, November 8, 1991, pp. 835–39).

Dinosaurs are almost always rarer than marine creatures, and this haystack really has to be pulled apart fragment by fragment and over a broad area. The National Science Foundation and other funding agencies simply do not supply grant money at such a scale for projects that lack experimental glamor, whatever their importance. So Peter (Sheehan this time) availed himself of a wonderful resource that mere ammonites could never command. I will tell this story in his words:

> We co-opted the longstanding volunteer-based "Dig-a-Dinosaur" program at the Milwaukee Public Museum. Sixteen to 25 carefully trained and closely supervised volunteers and 10 to 12 staff members were present during each of 7 two-week field sessions during three summers. The primary objective of each volunteer was to search a predetermined area for all bone visible on the surface. The volunteers were arrayed in "search party" fashion across exposures so that all outcrops were surveyed systematically. Associated with the field parties were geologists whose function was to measure stratigraphic sections and identify facies.

I cannot think of a more efficient and effective way to tackle a geological haystack. Peter's personnel logged 15,000 hours of fieldwork and have provided our first adequate sampling of dinosaur fossils in uppermost Cretaceous rocks. They worked in the Hell Creek Formation in Montana and North Dakota, the classic strata for latest Cretaceous dinosaurs. They studied each environment separately, with best evidence available from stream channels and floodplains. They divided the entire section into thirds, with the upper third extending right up to the impact boundary, and asked whether a steady decline occurred through the three units, leaving an impoverished fauna when the asteroid struck. Again, I will let their terse conclusion, summarizing so much intense effort, speak for itself:

> Because there is no significant change between the lower, middle, and upper thirds of the formation, we reject the hypothesis that the dinosaurian part of the ecosystem was deteriorating during the latest Cretaceous. These findings are consistent with an abrupt extinction scenario.

You can always say, "So what; T. S. Eliot was wrong; some worlds at least end with a bang, not a whimper." But this distinction makes all the difference, for bangs and whimpers have such divergent consequences. Peter Ward sets the right theme in his final statement on the nonnecessary demise of ammonites:

> Their history was one of such uncommon and clever adaptation that they should have survived, somewhere, at some great depth. The nautiloids did. It is my prejudice that the ammonites would have, save for a catastrophe that changed the rules 66 million years ago. In their long history they survived everything else the earth threw at them. Perhaps it was something from outer space, not the earth, that finally brought them down.

The true philistine may still say, "So what; no impact and we still have both ammonites and nautiloids. What do I care. I had never even heard of nautiloids before reading this essay." Think about dinosaurs and start caring. No impact to terminate their still-vigorous diversity and perhaps they survive to the present. (Why not? They had done well for more than 100 million years, and the earth has only added another 66 million since then.) If they survive, mammals almost surely remain as small and insignificant as they were during the entire 100 million years of dinosaurian domination. And if mammals stay so small, restricted, and unendowed with consciousness, then surely no humans emerge to proclaim their indifference. Or to name their boys Peter. Or to ponder the nature of science and the proper interaction between fact and theory. Too dumb to try and too busy scrounging for the next meal and hiding from that nasty *Velociraptor*.

Stephen Jay Gould *teaches biology, geology, and the history of science at Harvard University. He is the author of* The Panda's Thumb, The Mismeasure of Man, *and many other books.*

HOW DO YOU FIGURE?

Challenges in physics and math

Math

M61

Repeating sums of digits. Prove that in any infinite arithmetic sequence of positive integers, one can always find two numbers with equal sums of digits. (S. Genkin)

M62

This is some year!! Prove that 1992!! – 1991!! is divisible by 1993 (n!! denotes the product of all positive integers not greater than n and of the same parity as n: $n!! = n \cdot (n - 2) \cdot (n - 4) \cdot \ldots$). (V. Proizvolov)

M63

All you need is l . . . The diameter AB of a semicircle is arbitrarily divided into two parts, AC and CB, on which two other semicircles are constructed (fig. 1). Find the diameter of a circle inscribed in a curvilinear triangle formed by the three semicircles, given only the distance l from this circle's center to line AB.

M64

Ordering roots. Let a, b, and c be the roots of the equations

$$\cos a = a,$$
$$\sin \cos b = b,$$
$$\cos \sin c = c,$$

lying in the interval $(0, \pi/2)$. Without any calculations, arrange them in increasing order. (S. Gessen)

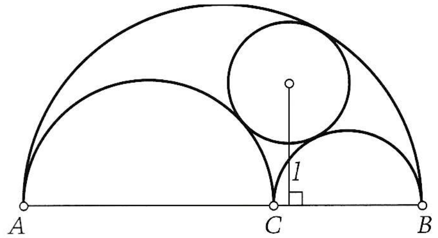


Figure 1

M65

Circles encircled by circles. Prove that (a) it is possible to plot a finite number of nonintersecting circles on the plane so that each of them touches (externally) exactly five of the others; (b) you cannot replace the number five in part (a) with six. (D. Fomin)

Physics

P61

Rubber cord. Two athletes are standing at points A and B holding a rubber cord. At a signal, runner A moves eastward with a velocity v_0 = 1 m/s, and runner B moves southward with a constant acceleration. Determine the acceleration if it is known that a knot C tied on the cord passed point D (the scale is given in figure 2). (S. Krotov)

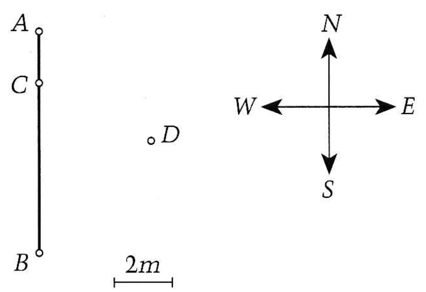


Figure 2

P62

Flying top. A conical top is rapidly rotating about its axis on a smooth table (fig. 3). At what velocity v of its forward motion does the top not hit the table as it falls off the edge? The top's axis remains vertical; its dimensions are shown in the figure. (A. Zilberman)

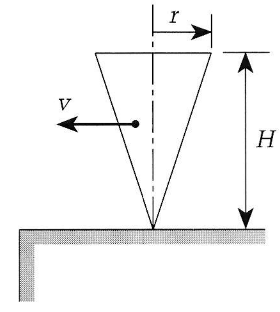


Figure 3

P63

Low-pressure retort. A spherical retort with a volume of 1 liter was pumped out and sealed. A monomolecular layer of air remained on the retort's walls. What is the approximate pressure within the retort when heated to 300°C if it is known that at such a temperature the retort walls are totally free of gas? (A. Mitrofanov)

P64

Battery charger. To charge a storage battery with emf $\mathcal{E}$ = 12 V from a power source with a voltage V = 5 V, a circuit was constructed by using a coil with inductance L = 1 henry, a diode D, and a switch K that periodically opens and closes at equal time intervals $\tau_1 = \tau_2$ = 0.01 s (fig.4). Determine the average current I_{av} charging the battery. (A. Zilberman)

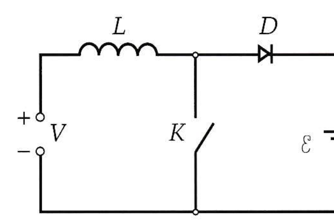


Figure 4

P65

Blowin' in the wind. Why is it hard to hear words shouted into the wind? (A. Buzdin)

ANSWERS, HINTS & SOLUTIONS ON PAGE 57

SMILES

Doing it the hard way

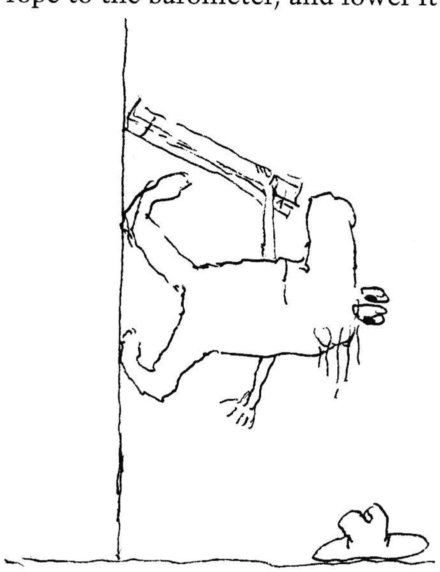
"Among twenty towering buildings
The only moving thing
Was the plummeting barometer."
—Steven Wallace, "Eight Ways of Looking at a Blackboard"

by M. Tulchinsky

THIS ARTICLE OFFERS SEVeral (eight) methods of solving a problem of great theoretical as well as practical importance: how to measure the height of a multistory building with a sufficiently long rope and a 1-meter-long mercury barometer.

Some of these methods are applicable to a wide range of analogous problems (measuring the height of the Eiffel Tower, the Empire State Building, Mt. Everest, etc.).

Method 1 (a trivial one). Go up to the roof of the building, tie your rope to the barometer, and lower it to ground level. Then lift it up and measure the length of the rope.



Art by Pavel Chernusky

Method 2 (a straightforward one). Holding the barometer vertically, go up the stairs and mark the length of the case on the wall. Counting the number of marks, you will obtain the height of the building.

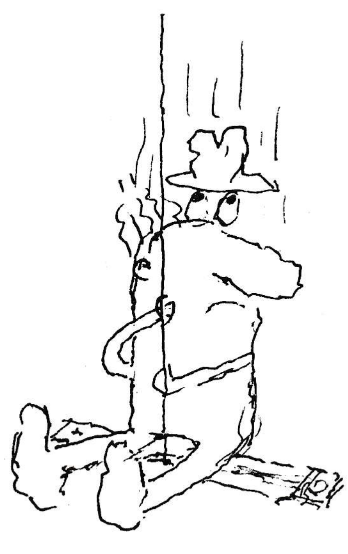
Note: If you measure the height of one floor in this way and multiply its height by the number of floors, the error will be too large.

Method 3 (an aerostatic one). Measure the atmospheric pressure at ground level and at roof level. Figure out the height of the building according to the change in the barometer readings.

Method 4 (a geometrical one). Take the barometer outside on a sunny day. Stand it up vertically. Measure the length of its shadow and the shadow of the building. Using the similarity of triangles, calculate the height of the building.

Method 5 (a sociological one). Ask all the tenants to estimate the height. Find the arithmetical mean. Promise to give your barometer as a prize.

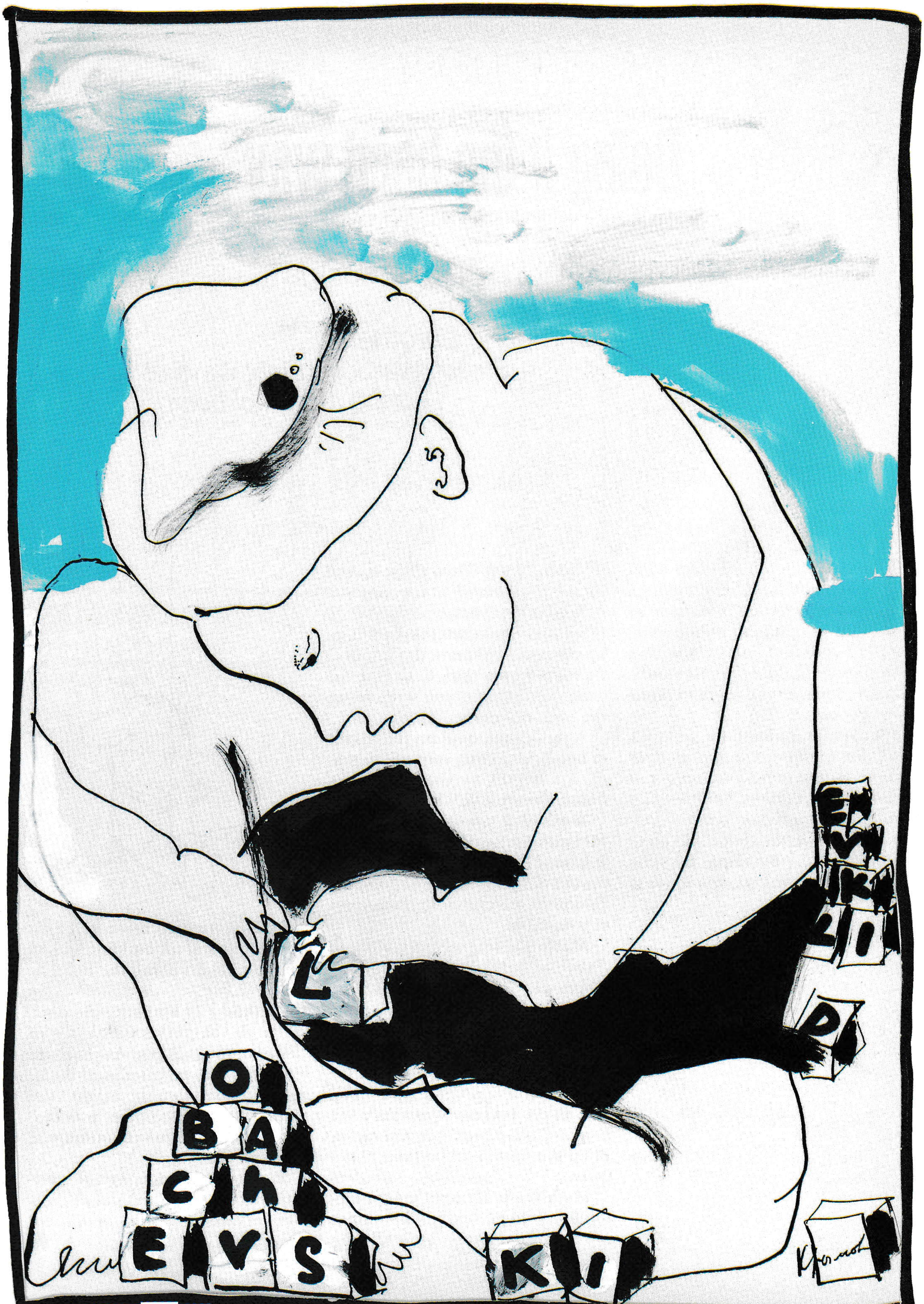
Method 6 (a kinematic one). Determine your pulse rate. Then count the number of heartbeats while the



vertically oriented barometer falls from the roof of the building. Calculate the height using the formula $h = gt^2/2$.

Method 7 (a bureaucratic one). Write to the persons who are in charge of the building. Ask them to look into the architectural documents and tell you the height. (Remember, your barometer was broken in your "kinematic" attempt at measurement.)

Method 8 (a pedagogical one). Perhaps our readers will devise their own methods and send us descriptions . . .



Turning the incredible into the obvious

How many geometries do you know?

by Vladimir Boltyansky

EDITOR'S NOTE: THIS IS THE first of a series of articles dedicated to the bicentenary of the great Russian mathematician Nikolay Lobachevsky (1792–1856), one of the creators of non-Euclidean geometry. It will introduce you to the problem of testing an axiomatic system (in this case, non-Euclidean geometry) by constructing a model of it. Upcoming issues of *Quantum* will bring you the historical and mathematical details of Lobachevsky's discovery, its physical interpretations, and applications. (See also the Anthology department in the last issue.)

An "odd" association

In a certain town an association was founded that united philatelists, numismatists, collectors of the eyes of needles and all other things worth collecting. It was decided that the association's administration must consist of an odd number of people (which is convenient for voting). Also, it was decided that the administration should be accommodated in one building and every member of the administration should have a separate intercom unit. A switchboard was considered an unjustifiable luxury, but to keep the intercom system from becoming too complicated, each member was supposed to be connected directly to three others.

"Well," one of the founders said, "I think this arrangement is convenient enough."

"What I like best about it," said another founder, a famous collector of mathematical curios, "is that the proposed scheme can be described by a set of abstract axioms. Look, we have three primary notions—'member,' 'line,' 'connect'—and we want the following three *axioms* to hold:

1. The number of members is odd;
2. Each line connects exactly two members;
3. Each member is connected to exactly three members."

"But is this set of axioms very useful? Is it possible, for instance, to deduce from it any theorems that would make a 'theory' that reflects the operation of our administration?"

"Sure, here's a theorem for you! By axiom 1 there exists at least one member (zero is an *even* number!). Then, axiom 3 ensures there are three more members, making the total not less than four. Finally, applying axiom 1 once again we conclude that *the total number of members is not less than five*. And this is the first theorem of our 'theory.' "

"I see. This statement can indeed be considered a theorem, since it was proven—that is, logically derived from axioms."

"Here's another example of a very simple theorem: *among any five members there are always two who are not connected by a line.*"

"Well, that's pretty obvious—otherwise, there will be at least four lines leading to each of these five members."

"Exactly! By the way, how did you put it? A line 'leads' to a member? Within the framework of our set of axioms we can give an accurate definition of this situation: if A is a member and line l connects this member to some other member, the pair (A, l) will be called a *lead*."

"So now we can say that every member A belongs to exactly three leads (A, l_1), (A, l_2), (A, l_3). And that's what we mean when we say that there are three lines l_1, l_2, l_3 leading to A (axiom 3)."

"Wonderful!" The collector of mathematical curios (the Collector, for short) grinned contentedly.

"Continuing your line of reasoning, I can formulate the following theorem: *the total number of leads is odd*. After all, the number of members is odd (axiom 1), and each of them belongs to exactly three leads."

All of a sudden the Collector turned gloomy and said in a downcast tone, "Unfortunately, I can also prove that *the number of leads is even*: each line (connecting, for instance, A and B) belongs to *two* leads—(A, l) and (B, l)—so the number of leads is *twice* the number of lines."

"But that contradicts the previous theorem!"

"That's the problem! Our axioms turn out to be *inconsistent*, because we can deduce two mutually contradictory theorems from them. Too bad! Our theory falls to pieces: it's

Art by Dmitry Krymov

simply impossible to create the communication system described by axioms 1, 2, and 3."

Changing axioms

"But what should we do? Allow the number of administrators to be even? But we tried so hard to avoid that!"

"There's another way out," the Collector answered. "Let's leave axioms 1 and 2 as they are, but replace the third axiom with

3′. Each member is connected to exactly four other members.

"Of course, we'll have to lay a greater number of lines, but in return we'll make our set of axioms consistent, and three of the theorems—with a slight modification—will remain valid. As before, the total number of members is not less than five, two of any *six* (not five!) members are not connected with a line, and the total number of leads will be even (without any contradictions!). We can go on and prove other theorems as well."

"But, with your permission, how can you be sure there are no contradictions whatsoever? True, the number of leads is no longer a problem; but, perhaps, proving new theorems over and over again we'll nevertheless come across a contradiction *sometime*. You're not saying you know in advance *all* the theorems that can ever be deduced from the new axioms, are you? And if you're not, who can guarantee the absence of contradictions?"

"Oh, I'm absolutely certain about this! I'll tell you why I'm so confident. You don't doubt the 'correctness' of arithmetic, do you?"

"Of course not. But what does arithmetic have to do with our association?"

"Just this. I'm going to construct a *model*, as mathematicians say, of our set of axioms from the 'material' of arithmetic. By the way, how many members, approximately, will the administration include?"

"Not less than 30, I think."

"Great! I suggest we consider the numbers 1, 2, ..., 37 as the 'members'; for convenience I'll plot them on a circle (fig. 1), so that 37 is followed by 1. Now I define a 'line' as a pair of number such that one of them is next to, or one person away from, the other along this circle. For instance, the pairs (1, 3), (36, 37), (37, 2) are 'lines,' while (3, 6) isn't a 'line'—3 and 6 stand too far apart."

"I think I see what you're getting at. This model, as you call it, comprises 37 'members'—an odd number (as required by axiom 1). Each 'line' connects exactly two members (axiom 2) because a pair consists of two members. And it's clear that exactly four 'lines' lead to every 'member' (axiom 3′) as illustrated in figure 2. Thus, all three axioms 1, 2, and 3 hold for this model. But why does this guarantee the consistency of the axioms in question?"

"Just because the model is made of numbers! If two contradictory theorems could be derived from our axioms, this contradiction would come out in our model, too. It would mean we could obtain a contradiction while reasoning about *numbers*. But you don't doubt the infallibility of arithmetic, do you?"

"Well, now your arguments are clear to me. And so, let me, please, repeat them in a general form, so to speak. You consider two theories, *P* and *Q*. Theory *P*—in our case, arithmetic—doesn't raise any doubts, and can be regarded as unshakable and infallible; theory *Q*, on the other hand, is a new one, defined by a list of axioms. We want to obtain guarantees that *Q* is a consistent theory. For this purpose the following technique is applied: using notions from theory *P* as 'building material,' we try to construct a model of theory *Q*—that is, choose some notions of *P* that represent the *primary* notions of *Q*, so that all the axioms of *Q* hold for them. If we're able to do it, the consistency of theory *Q* is proven."

A strange geometry

"We've digressed a bit from discussing questions about our association. But because I collect mathematical curios, I'd like to offer you one of them—a rather unusual model of geometry."

"I'll be glad to take a look at it."

"Imagine that one point, *O*, is removed (mathematicians would say *deleted*) from the plane. (Of course, it actually stays in place but we pretend it's not a 'point.') All the other points (distinct from *O*) are considered 'points' in our model."

"And this is the whole *point* of your model?!"

"Wait a minute. Every circle passing through the deleted point *O* will be now called a 'straight line' (including 'circles of infinite radius'—that is, regular straight lines passing through *O*). Some of the new 'straight lines' are illustrated in figure 3."

"This is really unusual. We take deliberately *curved* lines and declare them '*straight*.' But what for?"

"Let's try, for example, to draw a 'straight line' through two points *A* and *B*. How can we do it?"

"You want the 'line' to pass through *A* and *B*. In addition, since it's a 'line,' it must contain *O*. This means we have to draw a circle

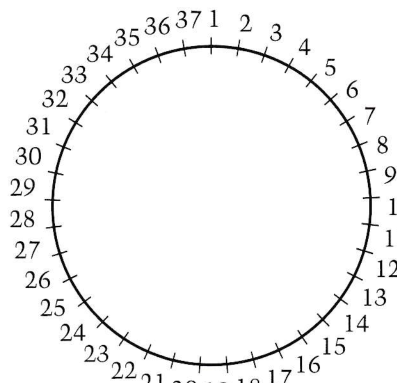


Figure 1

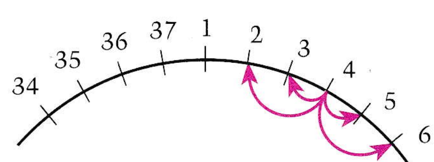


Figure 2

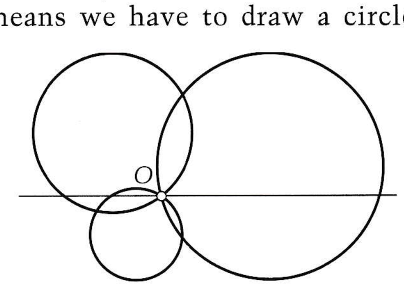


Figure 3

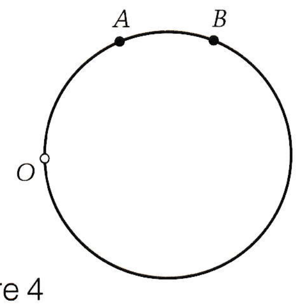


Figure 4

through three points A, B, O (fig. 4). A child could do it!"

"What's important is that such a 'line' always *exists* and is *unique*. Right?"

"Right. Three points uniquely determine a circle as long as they don't belong to a regular straight line; but if they do, the 'line' we need is simply this regular line, a 'circle of infinite radius.' "

"And so this model satisfies the following axiom: *there is one and only one line passing through any two different points.*"

"I think I'm getting interested in your curio."

"Now look at figure 5, which shows a *triangle ABC*. Each of its angles is equal to the corresponding angle at point O. So their sum equals the sum of the marked angles at point O. But *they add up to a straight angle*! Which means that the sum is 180 degrees."

"Amazing!"

"Notice also that in this model there's only one 'straight line' that doesn't intersect a given 'line' a and passes through a given point A outside a. It's the circle touching the given 'line' a at point O (fig. 6): point O is counted out, so the two 'lines' in the figure have no common 'points.' "

"This means that the parallel postulate is valid in your model too?"

"Absolutely. And not only this postulate: all the axioms (and, consequently, all the theorems) of the geometry studied at school—Euclidean geometry, named after the thinker who was the first to lay it out systematically more than 2,000 years ago. And so all the statements of Euclidean geometry hold in this model, too. But I haven't finished describing it. It wasn't specified how to measure distances, what triangles should be considered isosceles, and so on. Besides, we should have added a special 'point at infinity' instead of the deleted point O to make the model complete. But the essence of the matter must be clear enough: I've constructed a model of the usual (Euclidean) geometry."

"May I say then that what you've done was to take the usual geometry as the original theory P and construct from its own 'material' a model of this very theory?"

"Exactly. And the reason I told you about this model was just to show it's not so unnatural to call curved lines 'straight' in some models."[1]

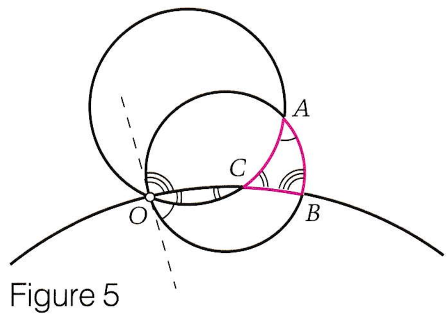


Figure 5

Lobachevsky's tragedy

"And now one more curio from my collection. It's a geometry discovered by the great Russian mathematician Nikolay Lobachevsky."[2]

"As far as I know he constructed a geometry in which there is more than one line through a point outside a given line parallel to it (fig. 7)."

"Right. But although all his arguments were logically irreproachable, during his lifetime he failed to gain recognition for his ideas. His conclusions seemed too weird to his colleagues (to say nothing of laypersons)."

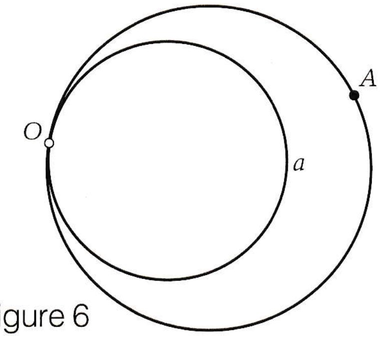


Figure 6

"Can you give an example?"

"Well, for example, parallel lines approach each other (fig. 8)—one can even construct a triangle with parallel sides (fig. 9). And though the lines in figure 9 are curved, there are *straight* lines in Lobachevskian geometry situated in just this way. In addition, the sum of the triangle's angles is always *less* than 180 degrees, and so on, and so forth."

"But that's not the way it is *in reality*, now, is it?"

"There! There! That's just what Lobachevsky's contemporaries said. But he was aware that what matters is whether the geometry obtained from a set of axioms laid at its basis is consistent or not. Besides, any argument like "this isn't the way it is in reality" is based entirely on the fact that our observations cover only a small part of space in which there is simply no marked difference between Lobachevsky's and Euclid's geometries (possible differences would lie beyond the limits of measuring accuracy.) By the way, modern physics continues to confirm it."

"That's surprising! But if I understood you correctly, you were going to talk about a purely mathematical confirmation that Lobachevskian geometry is correct."

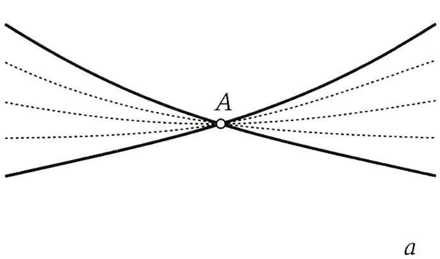


Figure 7

[1]In fact, this unusual model of the usual geometry is simply the image of the Euclidean plane after *inversion*, a remarkable transformation discussed in the article that begins on page 40.

[2]It was also discovered by Gauss and Bólyai, but Lobachevsky was the first to publish his results. That's why this geometry, usually called hyperbolic geometry, is sometimes—for instance, in this article—referred to as "Lobachevskian geometry."

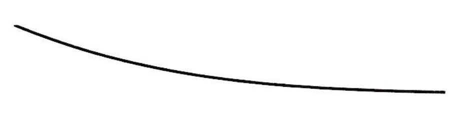


Figure 8

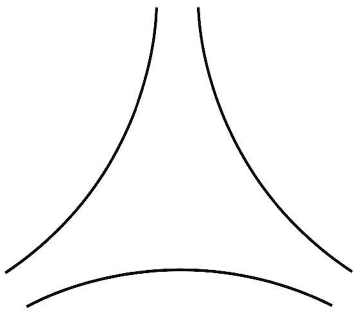


Figure 9

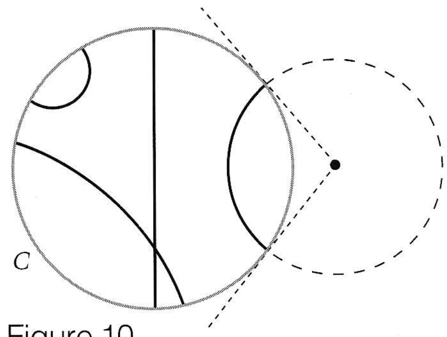


Figure 10

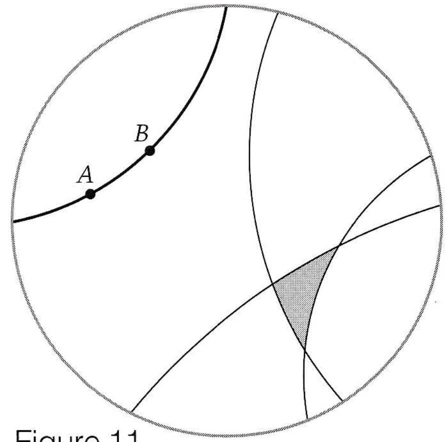


Figure 11

"More precisely, that it's *consistent*."

"Oh, I see! Apparently, you want to say one can build a model of Lobachevskian geometry from the 'material' of some other theory P whose correctness we don't doubt."

"Exactly! And this theory is Euclidean geometry!"

"Well, did Lobachevsky know about that?"

"No, he didn't. But he built another remarkable model. From the 'material' of his geometry he managed to construct a model of Euclidean geometry."

"So, if people had been sure of the impeccability of Lobachevskian geometry, but doubted the validity of Euclidean geometry, this model could have convinced them that Euclidean geometry is consistent too! But it would have been better the other way around!"

Justice delayed

"Now you see that Lobachevsky was fully aware of what a model is and how the consistency of his geometry could be proven. Taking Euclidean geometry as the original theory P, one had to construct from its 'material' a model of theory Q, Lobachevskian geometry. But Lobachevsky failed to do it. Such models were found only after his death."

"Who did it?"

"Beltrami, Cayley, Klein, Poincaré, and others."

"Are these models very complicated?"

"Not really. For instance, take the model discovered by the outstanding French mathematician Henri Poincaré (1854–1912). Its points are the interior points of some circle C. 'Lines' are defined as follows: take a circle (or a straight line) perpendicular to circle C; then its arc (or segment) intercepted by C is called a 'line' (see figure 10, showing several 'lines')."

"This model is similar to the model of Euclidean geometry discussed earlier. 'Lines' there were also circles."

"And not only that. It can also be proven for this model that there is a single 'line' through any two points A and B (fig. 11). One can consider triangles, and figure 12 shows there are *infinitely many* 'lines' through a point A outside a given line a and not intersecting it. Two of them (those touching 'line' a at points on the circumference of C (in a certain sense, at infinity) are said to be 'parallel' to a. And figure 13 illustrates an 'infinite triangle' whose sides are parallel to each other. In short, this is indeed a model of Lobachevskian geometry. Of course, my description is far from complete—I ought to explain how to measure lengths, define the congruence of figures, and so on.[3] But the general idea of the model was presented correctly."

"What about the other mathematicians you named—were their models the same?"

"Not at all! Nowadays, many different models of Lobachevskian geometry are known. For instance, Cayley and Klein created a model that also involved only interior points of a circle, but with truly 'straight' lines—all *chords* of the

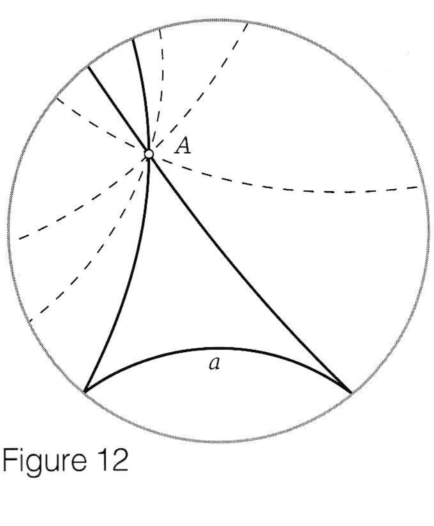


Figure 12

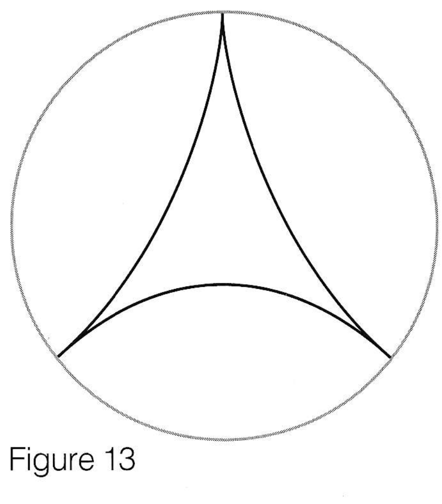


Figure 13

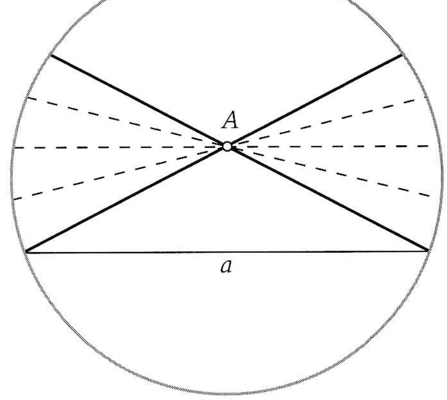


Figure 14

[3]For these and other definitions, see the article "Inversion."

circle (without their endpoints). In this model, it's even easier to see that there are infinitely many 'lines' through a point outside a given 'line' *a* not intersecting *a* (fig. 14). But it's more difficult to measure angles in this model than in Poincaré's."

"Your 'mathematical curios' are very curious indeed. Now I understand the significance of Lobachevsky's discovery. He not only constructed a very unusual geometry, he originated the search for other new 'geometries.' I've heard that mathematicians today study all kinds of different spaces and geometries, which are applied in physics and other branches of science. I guess in your collection, too, there are many wonderful spaces, geometries, and mathematical 'worlds.' "

"There certainly are! But we'll talk about them some other time."

IN THE LAB

Can white be blacker than black?

Here's a simple way to find out

by V. V. Mayer

LET'S START WITH A VERY simple observation. Take sheets of white and black paper, put them near each other, and make the room completely dark. Then you won't see them, because both sheets will be equally black.

It would seem that there are no conditions in which the white sheet of paper can be blacker than the black one. And yet this isn't the case. Try to think it through and devise an experiment in which white turns out to be blacker than black. But first read the following passage.

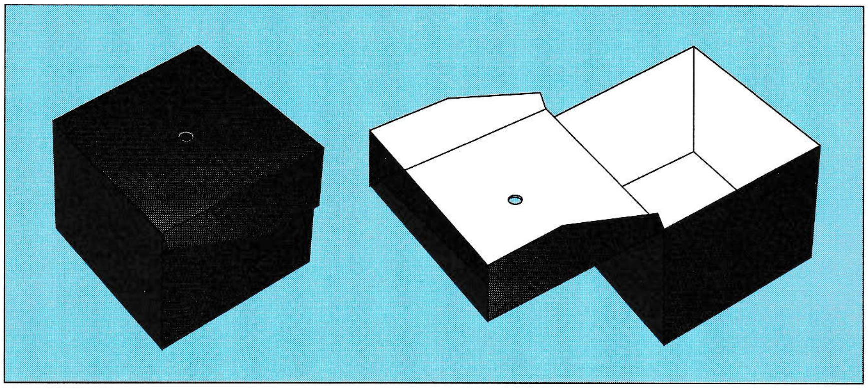
A body that completely absorbs incident radiation at any frequency and any temperature is called a *black body*. We understand that this is an idealization—that absolutely black bodies do not exist in nature. Bodies that we usually call black (soot, black paper, black velvet, and so on) are actually gray—that is, they partially absorb and partially scatter incident light. So to answer the question we've posed, we could, for example, take a sheet of white paper and make a body that is closer to being a black body than black paper is. Now the solution is almost self-evident.

A spherical cavity with a small opening turns out to be a good model of a black body. If the opening is no larger than 1/10 of the diameter of the cavity, then (according to the corresponding calculation) a light beam entering the hole would be able to exit only after repeated scattering and reflection. But every time the beam of light touches a wall, light energy is partially absorbed, so the amount of light that escapes the box is insignificant. So we can say that the hole in the cavity absorbs light of any frequency almost completely, just like a black body.

You can construct such an experimental device. Glue together pieces of cardboard to form a box with approximate dimensions 100 mm × 100 mm × 100 mm, and construct a lid that flips open. Cover the inside with white paper and paint the outside black (with India ink or gouache) or cover it with the black paper that comes with photographic paper to protect it from light. Make a hole with a diameter of not more than 10 mm in the lid. Below is a rendering of such a box.

To perform the demonstration, shine a table lamp on the lid. The hole will seem to be blacker than the black lid. Open the lid, and everybody will see that behind the hole is—white paper, which really was "blacker than black" in the experiment!

There is a simpler way to observe this phenomenon. Take a white porcelain cup and cover it with a black paper lid that has a small hole in it—the effect will be practically the same.



PHYSICS CONTEST

The tip of the iceberg

"And ice, mast-high, came floating by,
As green as emerald."
—Samuel Taylor Coleridge, "The Rime of the Ancient Mariner"

by Arthur Eisenkraft and Larry D. Kirkpatrick

CLOUDS FLOAT IN THE SKY. Rocks plunge to the bottom of lakes. Children gaze at the sky as a helium balloon rises and rises, wondering what its future will be. Physicists also wonder about such things. And as they wonder, they think about gravity, buoyancy, sinking, and floating.[1]

One of the first people to ponder floating and sinking was Archimedes. It's difficult to believe that Archimedes was once so consumed by his king's challenge of determining the constituency of the royal crown that in his burst of insight, he jumped out of the bathtub and ran through the town screaming "Eureka!" If you heard such a commotion outside and then observed a naked man running down the street screaming a Greek word or even its English equivalent ("I found it!"), what would you do? Your first thought probably involves locking doors or calling police. When you found out that this was the reaction of a wise man after figuring out whether the king's crown was pure gold or an alloy, I can't imagine many people unlocking their doors or hanging up the phone: "It's okay, dear—just a wise man discovering a new law of physics."

Archimedes's law was a great achievement. Everybody knew that an object dropped in water made the water level rise (that is, it displaced some water). But Archimedes was the first to recognize that the amount of water displaced is related to the object's volume in a very precise way!

You can try to "discover" Archimedes's law, and we promise you won't have to take a bath or run through your town. Gather 100 pennies. Fill a glass with water. Carefully place the pennies in the water. What happens when 50 pennies are placed in water? What happens when 100 pennies are placed in water? Does the water rise? Does the water spill out of the glass? How much water? Can you find the relationship between the water and the pennies? That was part of Archimedes's challenge.

We know that when you try to submerge a table tennis ball in water, it will find its way to the surface. A well-designed boat will float. The same boat, with lots of cargo, may sink. Why is it that some people find floating in water easier than others?

One helpful way of analyzing floating and sinking is to note that in a container of water, the water at each level of the container is, in effect, floating at that level. Physicists would say that the water at each depth is in equilibrium. The buoyant force on any piece of water pushing it up must be exactly equal to the gravitational force or weight of the piece of water pulling it down. We then conclude that the pressure (force per area) of the water increases with depth. The bottom of a submerged object experiences a higher pressure than its top. The difference in pressure pushes the object up. Gravity is pulling the object down. Since the piece of water moves neither up nor down, the force due to the pressure differences must exactly equal the weight.

The pressure at the top of the piece of water is equal to p and the pressure at the bottom of the piece of water is equal to $p + \Delta p$. This difference in pressure provides the upward force that must be equal to the weight of the piece of water:

$$(p + \Delta p)A - pA = m_w g = \rho_w V g.$$

If we assume that the piece of water is a slab of length l, width w, and height Δh, then

$$(p + \Delta p)A - pA = \rho_w A \Delta h g,$$
$$\Delta p = \rho_w g \Delta h.$$

Since the density of the water and the acceleration due to gravity are constant, we can conclude that the pressure change is proportional to the change in depth of the water.

If we now replace our piece of water with an identically shaped piece of wood, we know that the upward force on the wood is identical to the upward force on the piece of water. This is due to the pressure difference between the top of the wood and the bottom of the wood. The weight of the wood may be different than the

[1] See "Boy-oh-buoyancy!" in the September/October 1990 issue of *Quantum*.

weight of the piece of water. If the wood weighs less than the equivalent volume of water, the buoyant force prevails and pushes the wood to the surface. If the wood weighs more than the equivalent volume of water, the force of gravity prevails and the wood sinks to the bottom.

And so we have theoretically derived Archimedes's principle: *A body in water (or any fluid) will have a buoyant force equal to the weight of the fluid that it displaces.* When Archimedes sat in his bathtub and displaced that bath water, he reached this same conclusion. Anybody feel like shouting "Eureka"?

Have you ever wondered how much of an ice cube is beneath the water? Or, maybe, how much of an iceberg is below the surface? The density of ice is 0.92 g/cm³. The density of ocean water is 1.04 g/cm³. Assume that the weight of the iceberg is W_i, its volume is V_i, and its density is ρ_i; the volume of the displaced water is V_w; and the density of the displaced water is ρ_w. Then

$$W_i = \rho_i V_i g.$$

The buoyant force equals

$$F_B = \rho_w V_w g.$$

Since the buoyant force equals the weight,

$$\rho_i V_i g = \rho_w V_w g,$$

$$\frac{V_w}{V_i} = \frac{\rho_i}{\rho_w}.$$

Assuming the densities of ice and water given above,

$$\frac{V_w}{V_i} = \frac{0.92}{1.04} = 0.88.$$

We conclude that 88% of the iceberg is below the surface, or that we see only the "tip of the iceberg."

Our contest problem relates to buoyancy and the tendency of a tall object floating on the surface to tip over. If the mast of a ship tilts due to a strong wind or a wave, the submerged part of the ship will no longer be directly below the center of mass. The buoyant force and the weight will exert torques on the ship that could cause it to capsize.

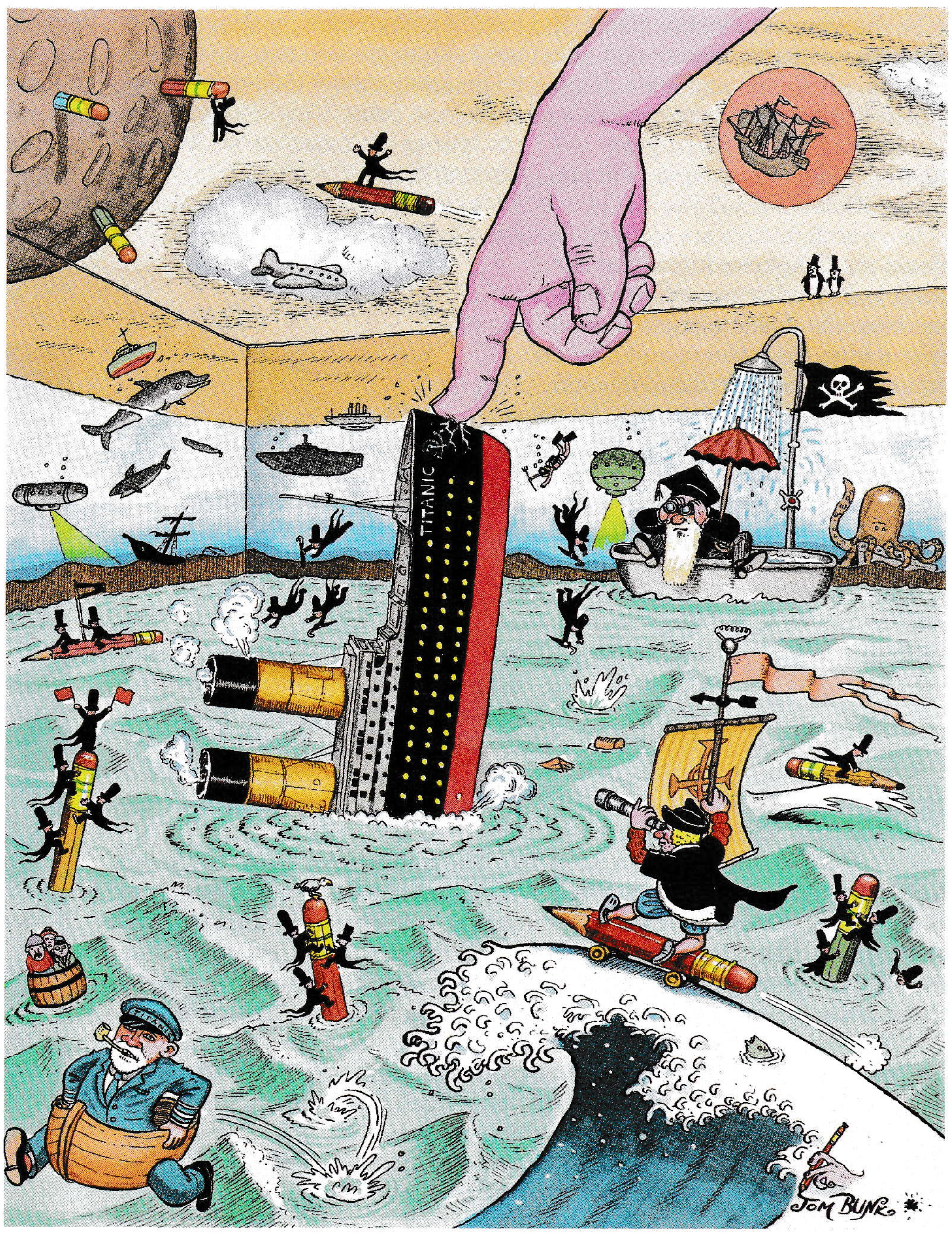
We can simulate this tilting ship by considering what happens if you hold a pencil at its tip and lower the other end into a large pail of water. At first we expect that the pencil will remain vertical. As more of the pencil enters the water, the pencil will have a tendency to tilt in the water. If the pivot is at a given height above the water, the pencil will choose a specific stable angle. Don't believe us—try it!

The contest problem has three parts:

A. What is the relationship between the angle of the pencil and the height of the pivot above the water?

B. Why is this position stable?

C. What happens to the fraction of



Art by Tomas Bunk

the submerged pencil as the pivot is lowered?

Please send your solutions to *Quantum*, 3140 North Washington Boulevard, Arlington, VA 22201 within a month after receipt of this issue. The best solutions will be noted in this space and their authors will receive special certificates from *Quantum*.

The clamshell mirrors

In the March/April issue we asked you to find the separation of two spherical mirrors that would produce a real image in a hole in the upper mirror. Furthermore, the image was to be the same size as the object. If you want to experiment with the mirrors, they are available from Edmonds Scientific Company (101 E. Gloucester Pike, Barrington, NJ 08007–1380) as the Optic Mirage (item A72,381) for $42.95.

Since both mirrors are required to obtain the image, we know that the image is formed by two reflections—one from the upper mirror followed by one from the lower mirror. Since the only dimension in the problem is the focal length of the mirrors, let's choose to express all other distances in units of the focal length. If we assume that the mirrors are separated by nf, the object distance s_1 for the image produced by the upper mirror (mirror 1) is also nf. The image distance s'_1 can be found from the mirror formula:

$$\frac{1}{s_1}+\frac{1}{s_1'}=\frac{1}{f}.$$

Therefore,

$$s_1'=\frac{nf}{n-1}.$$

For $n > 1$, the image will be real and located below the surface of the lower mirror (mirror 2). The image acts as an object for the lower mirror with an object distance

$$s_2=nf-s_1'=\frac{n(n-2)}{n-1}f.$$

Using the mirror formula a second time, we locate the image formed by the lower mirror:

$$s_2'=f\frac{n(n-2)}{n^2-3n+1}.$$

However, the conditions of the problem require that

$$s'_2=nf.$$

This yields a quadratic equation in n with the roots $n = 1$ and $n = 3$.

The solution $n = 1$ is the one actually used in the Optic Mirage. If we go back and look at our mathematics for this case, we find that the image produced by the upper mirror is located at infinity. Therefore, the rays forming the image leave the upper mirror parallel to each other, as shown in figure 1. These parallel rays are focused by the lower mirror to form an image at its focal point, which is located in the hole in the upper mirror. Notice the symmetry of the problem. Interchanging the image and the object has no effect.

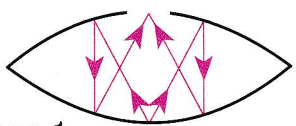


Figure 1

If we take a few liberties with infinities, we can use the formula for the magnification of the image

$$m=\frac{-s'}{s}$$

to find the magnification at each stage and then multiply them together to get the overall magnification. For this case, we obtain $m = -1$, indicating that the image is inverted. Observation of the image shows that this is correct.

The solution $n = 3$ is a surprise to many students. The ray diagram for this case is shown in figure 2. The upper mirror forms a real image midway between the mirrors that has a magnification of $-1/2$. The lower mirror then forms an image in the hole with a magnification of -2. Therefore, the overall magnification is $+1$ and the image is erect. Notice once again the symmetry of the problem. The upper half of figure 2 is just a mirror reflection of the lower half. This

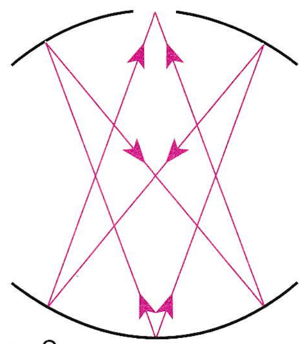


Figure 2

must be the case for all solutions to this problem.

Additional solutions can be found by letting the light reflect two, three, four, or more times from each mirror. In each case, symmetry requires that the rays be parallel, forming an image at infinity, or they form an image midway between the mirrors. For more information about these solutions, we refer you to the article by Andrzej Sieradzan in *The Physics Teacher* (November 1990, page 534).

IN YOUR HEAD

The game of bop

Wacky wordplay for the mathematically minded

by Sheldon Lee Glashow

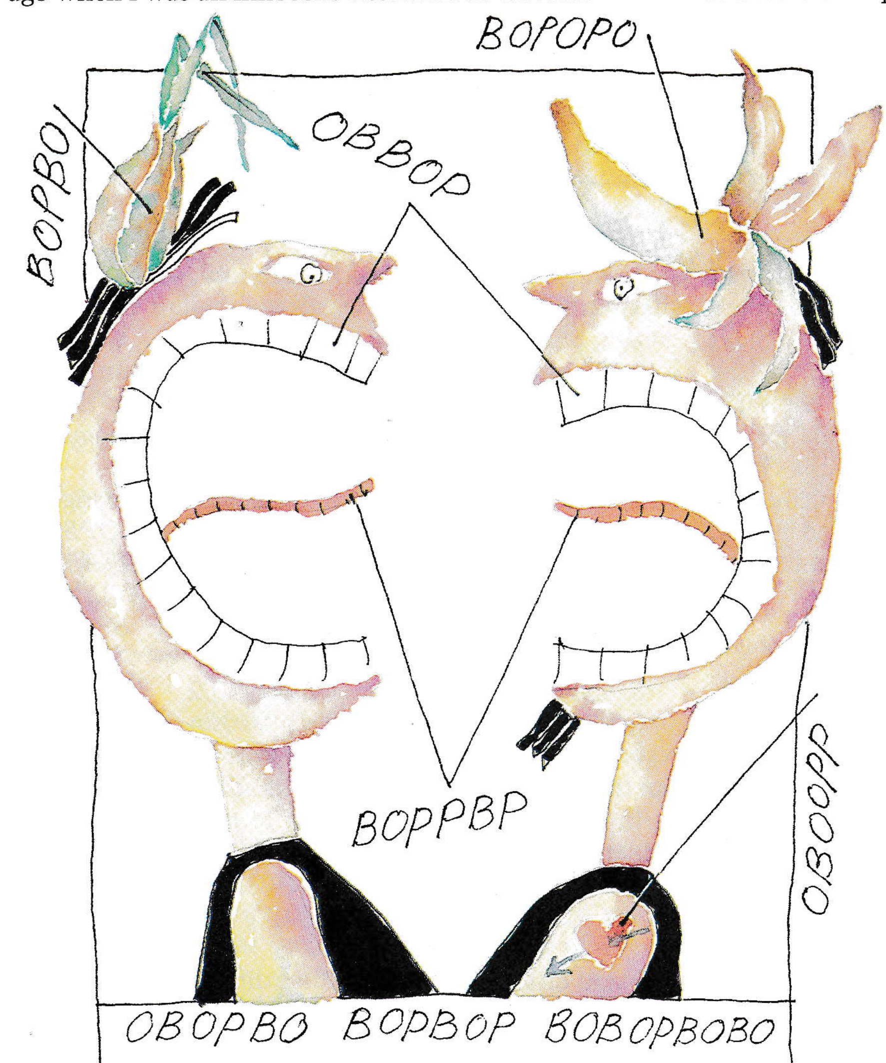
MATHEMATICS IS A POWERFUL TOOL, A WAY of thinking, and a variety of high art, but there's no such thing as a *language of mathematics*. No formula says "dinner is served" or "a lion is loose!" On the other hand, the *mathematics of language* is serious stuff to linguists and can also be an amusing and instructive form of wordplay. The fun began forty years ago when I was an innocent freshman at Cornell.

"BEBOPBOP?" asked a fellow nerd.

"Precisely what might you have in mind?" I responded (perhaps more colloquially than this).

"It's a common English word in a substitution cipher," he explained.

"NONSENSE!" said I in a trice—and so began the saga of the game of bop and this tongue-in-cheek study of the structure and spelling of words.



Art by Sergey Ivanov

Human speech consists of a sequence of sounds that phonetic languages transcribe as a string of symbols chosen from a commonly agreed upon alphabet or syllabary. Grouped by spaces and punctuation marks, they form words and sentences. Symbols vary in size and substance from language to language. For example, Hebrew letters are consonants, while Sanskrit signs signify syllables. English uses 26 letters; other alphabets have more or fewer letters:

Italian	21	Russian	33
Hebrew	22	Japanese	46
Korean	24	Sanskrit	48

We don't need so many letters. The book of life, as writ in our genes, makes do with four chemical signs. Computer jocks and radio hams can "say" anything in a binary system (like the Morse code) with just two symbols. Speech, however, involves lots of different sounds. Phonetic tongues use many letters because they try, with varying degrees of success, to link each sound to its own sign. A language with only a few letters would be ponderous, aphonetic, anathema to our poetic sensibility, and useless for the game of bop.

Bop is not so much a contest as a voyage of discovery in an utterly deconstructed literary sea. It deals solely with the letter sequences of words and not at all with how they sound or what they mean. We call an arbitrary series of p letters a *pword*. Those found in books are simply *words*. The properties of words and pwords relevant to the game of bop are those left intact by arbitrary permutations of the letters of the alphabet. Let me explain what I mean.

A substitution cipher is an easily broken kind of code. Every letter of a message is switched for another according to a fixed rule. Identical (or different) letters of the original become identical (or different) letters of the secret script. Since English has over a million words, we simplify our analysis of their letter sequences by defining two words to be bopwise equivalent if they're turned into one another by such a coding process, as in

bebopbop = *nonsense*,
mammal = *pepper*,
Harvard = *warfare*.

Humanists who are compelled to compose essays using only equivalent words could scarce do worse than this:

Avast, idiot! Enemy ozone icing papal cache again. Every ninth paper bible awash,

wherein each word is equivalent to *BEBOP*.

Our relation between pwords shares three vital properties with ordinary numerical equality. If x, y, and z are pwords, it's easy to see that

$x = x$ (reflexivity),
$x = y$ implies $y = x$ (symmetry),
$x = y$ and $y = z$ implies $x = z$ (transitivity).

These define what mathematicians call an *equivalence relation*. It groups pwords into disjoint sets such that members of the same set are equivalent to each other, while those of different sets are not. We call each such set of equivalent pwords a p-letter *form*, or *pform*.

Words of one letter, like *I* and *a*, are all of the same form. All two-letter words are equivalent to *me* except for *oo*,[1] which once signified the last letter of the Greek alphabet (as you'll recall from "The Greek Alphabet" in the March/April issue of *Quantum*). Onward and upward to three-letter words, four of whose forms are exhibited in

See the eel, Pop!

The fifth and last 3-form is a tripled letter like *ooo*. There's no such word in my dictionary. Indeed, I assert that *there is no English word of any length (excluding arabic numbers) with a triple letter!* We declare this to be an axiom and hereby exclude from consideration pwords and forms involving three consecutive identical letters.[2]

Let $N(p, m)$ be the number of distinct pforms (subject to triple-letter exclusion) of pwords using m different letters. N is straightforward to work out for small values of p and m. We've already done so for $p = 2$ and $p = 3$. Table 1 shows $N(p, m)$ for $p < 7$.

Table 1: Values of $N(p, m)$ for $p < 7$

	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$
$p = 2$	1	–	–	–	–
$p = 3$	3	1	–	–	–
$p = 4$	5	6	1	–	–
$p = 5$	8	22	10	1	–
$p = 6$	13	69	61	15	1

Every four-letter word belongs to one of twelve forms, of which five have two distinct letters and one has four. Any four-letter word with *three* different letters belongs to one of the six remaining 4-forms. As their exemplars, we introduce the "bop-forms" *BBOP, BOPB, BOPO, BOPP, OBOP* and the "boop-form" *BOOP*. It's a cinch to find words equivalent to each of these 4-forms. The sentence

Baby errs that eels sell here

uses every one.

Before we hunker down to the study of bigger and better words, here are some puzzles to ponder. If you find them worth your effort, read on!

BOProblems

1. Find a four-letter word in each form with only two distinct letters—that is, words equivalent to *BOOB, BOBB, BOBO*, and *BBOB*. (I give up on *BBOO*.)

2. What is the longest word you can find containing precisely two different letters?

3. What is the longest word you can find that contains no string of three distinct abutting letters?

4. Find a word equivalent to each of the following sixteen pwords: *BOPBOPB, BLOPBOP, PLOPBOP, MBOPBOP, BOPABOPA, BOPABOP, BOPABOB, BOPAPOB, MBOPABOPA, BOPSTOP, BOPBOX, POPUPBOP, BOPANDBOP, BOOPPS, BOPEEP, BOOBOOPIO.*

5. Show that $N(m + 1, m) = m(m + 1)/2$.

6. Show that $N(p + 2, 2) = N(p + 1, 2) + N(p, 2)$ for all $p > 2$.

7. Consider a word built from more than three different letters. Let L be the length of its shortest consecutive string containing four different letters. For example, $L(inane) = 5$ and $L(asinine) = 4$. Find a word with the largest L value you can. (Par is 8.)

* * * * * * * *

Focus on words using just three different letters. Most 3-letter and many 4-letter words are like this, and you can easily find familiar words in any chosen form. We won't

[1]Also possible is *'ee*, a colloquial contraction of *ye*.
[2]The German snow owl, or *sneeeule*, rarely seen in *Kaaawa* (rhymes with "cut the power"), Hawaii, proves that my rule has exceptions.

have such resounding success later on, when we deal with larger words of this category.

We assign a unique *canonical form* to every word, however long, that uses three distinct but never tripled letters. A word is of *bop form* if its three distinct letters abut one another somewhere in the word. Identification of the first occurrence of that sequence as *bop* defines a substitution code for the entire word and specifies its canonical form.

If the three distinct letters never abut, the word is of *boop form*. The first occurrence of a string equivalent to *boop* defines the code. Every pword made from three or more letters contains a string equivalent to either *bop* (whereupon it's of bop form) or *boop* (whereupon it's of boop form, unless it also has a bop). A few examples may clarify these assignments. Bopless *reeve* is of boop form *BOOPO*, while boopish *testees* is of bop form *BOPBOOP*. *Seems* seems to have a boop but is of bop form *PBBOP*. Finally, *murmur* is equivalent to *PBOPBO* and *OPBOPB*, but its form is *BOPBOP*.

We can't easily and elegantly classify words using more letters because the shortest string containing *four* different letters can be equivalent to *bopa, boopa, bopoppa, bopooppopa*, and so on *ad infinitum*. This is why we stick to words using three letters.

Let's hunt for words corresponding to 5-forms using three letters. We know from the table above that $N(5, 3) = 22$. Here are my offerings for all but one of the 18 bop and 4 boop forms. Get ready! Here come the bops.

Table 2: Bop 5-forms and their words

BBOPB = eerie	BOPBP = tests	OBOPB = cocao
BBOPO = llama	BOPOB = kayak	OBOPO = rarer
BBOPP = ?	BOPOO = melee	OBOPP = amass
BOBOP = cocoa	BOPOP = queue	OBBOP = teeth
BOPBB = fluff	BOPPB = sells	OOBOP = eeled
BOPBO = onion	BOPPO = freer	PBBOP = hooch

And now for the boops.

Table 3: Boop 5-forms and their words

BBOOP = eette	BOOPP = année	BOOPO = error
OBOOP = sassy		

Cocao is an alternate spelling of cocoa (along with cacao and cocco). *Année* is a French year and accented to boot. *Eette*, says the Oxford English Dictionary, is an antique form of eat.

Six-letter words begin to get out of hand. Of the sixty-nine 6-forms using three letters, 58 are bop and 11 are boop. After an inordinate waste of time—which would have been far better spent seeking a Theory of Everything—I managed a weak D. Table 4 constitutes my 64% solution to the Great Bop Problem.

Table 4: Bop 6-forms and their words

BBOBP = ?	BOPBPO = tertre	OBOBOP = ananas
BBOPBB = ookroo	BOPBPP = ?	OBOPBB = ?
BBOPBO = eerier	BOPOBB = ?	OBOPBO = Sashas
BBOPBP = ?	BOPOBO = revere	OBOPBP = sestet
BBOPOB = ?	BOPOBP = natant	OBOPOB = kokako
BBOPOO = ?	BOPOOB = ?	OBOPOO = elevée
BBOPOP = llenen	BOPOOP = ?	OBOPOP = emeses
BBOPPB = ?	BOPOPB = tenent	OBOPPB = ?
BBOPPO = ?	BOPOPO = banana	OBOPPO = anatta
BOBBOP = mammal	BOPOPP = ?	OOBBOP = ?
BOBOPB = acacia	BOPPBB = ?	OOBOPB = ?
BOBOPO = papaya	BOPPBO = sheesh	OOBOPO = ?
BOBOPP = ?	BOPPBP = easses	OOBOPP = eeless
BOOBOP = inning	BOPPOB = redder	PBBOPB = eddied
BOPBBO = amraam	BOPPOO = settee	PBBOPO = ?
BOPBBP = gregge	BOPPOP = powwow	PBBOPP = ?
BOPBOB = senses	BPBBOP = caccia	POOBOP = reefer
BOPBOO = testee	OBBOPB = teethe	PPBBOP = ?
BOPBOP = murmur	OBBOPO = doodad	
BOPBPB = andada	OBBOPP = appall	

Some of the words in table 4 are of dubious or distant provenance. *Amraam* is acronymic for Advanced Medium-Range Air-to-Air Missile; *caccia* is Italian for hunt; *andada* and *llenen* are forms of the Spanish verbs for walk and fill. Englishmen once said *gregge* and *tenent* for aggravate and tenet in the good old days when *easses* were earthworms. *Sheesh* is a variant of shish, as in shish kabob; *Sashas* are Russian Alexanders familiarly known; *kokakos* fly the New Zealand skies; *elevée* is raised Frenchwise. *Emeses* is when several people vomit; a *tertre* is a small French hill; *anatta* is a red-orange dye. Unlucky eelers return home *eeless* (not to say eelless) to eette *ookroo*, an oddly spelled (if no better tasting) variety of okra.

Let's not neglect the boops. Many gaps remain to be filled by *Quantum* wordsmiths.

Table 5: Boop 6-forms and their words

BBOOPO = ?	BOOPOP = seeded	OBOOPO = sasses
BBOOPP = ?	BOOPPB = ?	OBOOPP = tattoo
BOBOOP = cocoon	BOOPPO = ?	OOBOOP = ?
BOOPOO = assess	OBBOOP = ?	

Table 6: Some big bops and boops

BBOPBBO = eel-weel	BOPOBBOB = referrer
BOBBOOP = coccoon	BOPOBOO = referee
BOBBOPB = sissies	BOPOPPOP = benennen
BOBBOPOB = pepperer	BOPPBOPP = greegree
BOBOPBOBO = tête-à-tête	BOPPOBO = mottomo
BOBOPOB = rerefer	BOPPOOB = settees
BOBOPPOB = susurrus	BOPPOOP = lessees
BOOBOOP = muumuus	BOPPOPB = ginning
BOOBOPOP = teeterer	OBBOPOB = Essenes
BOOPOOPO = assesses	OBBOPPO = efferre
BOPBBOO = feoffee	OBBOPPOBO = attaccata
BOPBBOP = checché	OBOOBOP = nennend
BOPBOBO = tastata	OBOPBPO = sestets
BOPBOOP = testees	OOBBOPP = eelless
BOPBOPB = entente	PBBOPBB = essless
BOPBOPBOP = cha-cha-cha	PBBOPBBO = pooh-pooh
BOPBOPO = Barbara	POOBOPPOP = seeresses

Who would dare make a systematic study of larger words made of three letters? Most of the many forms are wordless, except perhaps in Polynesian tongues. Table 6 is my hard-won gathering of big bop words with a few boops thrown in for zest. A growing international contingent includes the French *tête-à-tête* and *entente*, Japanese *mottomo* (much), and both *nennend* and *benennen* from the German *nennen* (name). The *cha-cha-cha* is a Spanish–Caribbean dance. To Italians, *attaccata* is devoted, *tastata* is a feel, and *checche* is whatever.

Efferre means to emit, *coccoon* is a documented variant spelling, while *Barbara* and her far from *essless Essenes* are proper. *Rerefer* and *teeterer* are weak, but better than unlisted *rereferer*, *rereferee* and *lesseeless*. Some African charms are *greegrees* just as *eel-weels* are misspelt wheels with which ingenious eelers avert eellessness.

* * * * * * * *

My essay is done, but not the game of bop. Many wonderful words remain to be found. Someone more skillful than I may figure out a general formula for *N*. Readers finding bigger bop words or better bop games should send them to me c/o *Quantum*.

ANSWERS ON PAGE 61

BRAINTEASERS

Just for the fun of it!

B61

Zoo at home. Mademoiselle Dubois loves pets. All her pets but two are dogs, all but two are cats, and all but two are parrots. Those that are not dogs, cats, or parrots are cockroaches. She has more than two pets. How many pets of each kind does she own? (A. Rudinskaya)

B62

To make a parallelogram. A convex quadrilateral is cut along a diagonal, a congruent quadrilateral is cut along the other diagonal. Put the four pieces together to make a parallelogram. (V. Proizvolov)

B63

Cubes in water. You have two small cube-shaped plastic blocks of the same dimensions. The first block, floating in water, is submerged 2 cm; the second block is submerged only 1 cm. How deep will the lower block be submerged if the first block is placed on the second? What about the reverse (the second block is placed on the first)? (N. Dolbilin)

B64

Monochromatic vertices. All the points of a circle are arbitrarily painted in two colors. Prove there is an isosceles triangle inscribed in the circle whose vertices are all the same color. (I. Tonov [Bulgaria])

B65

Squares in a row. Arrange the integers 1 through 15 in a row so that the sum of any two adjacent integers is a perfect square. How many ways can this be done? (B. Recamán [Swaziland])

ANSWERS, HINTS & SOLUTIONS ON PAGE 60

Art by Dmitry Krymov

Playing with the ordinary

"The experiment—Life's eternal teacher."
—Johann von Goethe

WELL, YOU'RE BACK in school, and no doubt your labs are well stocked with equipment and materials. But you can also perform a great many experiments on your own with just a ruler, a magnet, or any number of household items. It's just a matter of looking at a familiar phenomenon from the physical point of view—to "play" with it, recalling some little fact from your physics textbooks. Many seemingly abstract concepts will take on a new life. You'll not only understand them better, it will be as if they've become a part of you.

We hope the selection below will help you conduct a few experiments with materials at hand and maybe even learn a few things you might never have thought about.

Questions and problems

1. A cylindrical pot is filled to the brim with water. How can you pour off exactly half of the water?

2. Two pieces of glass that were left out in the rain have stuck together so that it's very difficult to pry them apart. Why? What can you do to separate them with practically no effort?

3. You happen to know that there is a big air bubble inside an aluminum ball. Can you can find out where the bubble is—in the center of the ball or closer to the surface?

4. In order to detect a puncture in an air mattress, you can put a pile of books on it to exert additional pressure. Will the air escape faster if you put the books on the mattress in two stacks?

5. How can you determine the diameter of a large ball with a wooden ruler?

6. Fill a clean bowl with boiled water and scatter several wooden matches over the surface. If you touch the water between the matches with a sugar cube, the

the surface. If you touch the water between the matches with a sugar cube, the matches will stick together; but if you touch the water with a piece of soap, the matches will scatter in different directions. Why?

7. A one-color flag fluttering above a building seems to be striped at night when it is illuminated at a certain angle, and the stripes are moving all the time. How can you explain that?

8. Is it possible to determine the direction of rotation of the rotor in a coffee grinder without taking it apart?

9. Why does a driver at night see a puddle on the asphalt as a dark spot?

10. How can you use a strong magnet to determine what kind of current—DC or AC—is flowing through a light bulb?

11. If you pour some hot tea from a thermos in cold weather and put the cork back in the thermos, you may find that the cork pops out after a while. Why?

ANSWERS, HINTS & SOLUTIONS ON PAGE 60

Art by Pavel Chernusky

The story of a dewdrop

"I dream of Jeanie with the light brown hair,
Floating, like a vapor, on the soft summer air."
—Stephen Foster

by A. A. Abrikosov

WE COME ACROSS THE processes of evaporation and condensation literally at every step we take. And though most of us probably think right away of a boiling kettle, the most important example of these processes is the water cycle in nature: without it life on Earth would be, at the very least, different.

The heat of the Sun evaporates moisture at the Earth's surface, and then diffusion and convection currents carry the water vapor to the upper layers of the atmosphere. As the vapor travels upward, the air temperature decreases, the vapor condenses, and clouds are formed. Drops inside the clouds coalesce and grow, which leads to rain—and that completes the cycle.

We'll be looking at one of the most important links in the chain: the formation of drops in cooling water vapor. But first we need to recall one important notion.

You know that saturated vapor is the vapor that is in equilibrium with the liquid phase. There is a definite value of this saturated vapor pressure $p°(T)$ for every temperature T (fig. 1). The line $p°(T)$ divides the p-T plane into regions corresponding to the liquid and gaseous states.

The equilibrium of liquid and gas provides an example of *dynamic equilibrium*. Molecules are constantly exchanged between the two phases at the liquid–vapor interface—it's as if the processes of evaporation and condensation are running counter to each other. If the vapor is saturated, the flow of particles is the same in both directions and the amount of the substance in each phase remains constant.

So the concentration of saturated vapor is determined by the rate of evaporation per unit area of the liquid—that is, the saturated vapor density is a measure of the *volatility* of the substance.

"Telecannibalism" among drops: its qualitative aspect

Now we can move on to the main topic of this article: the influence of surface shape on the equilibrium of the liquid and gas phases. Surface deformation, as it turns out, leads to a change in the saturated vapor pressure. Let's look at an example.

Imagine that there are two practically spherical droplets of the same size under a bell jar. We already know that the apparent invariability of this system is deceptive. In reality there is vapor under the bell jar, and molecules, after leaving one drop, can easily reach the other. Is dynamic equilibrium stable in this case?

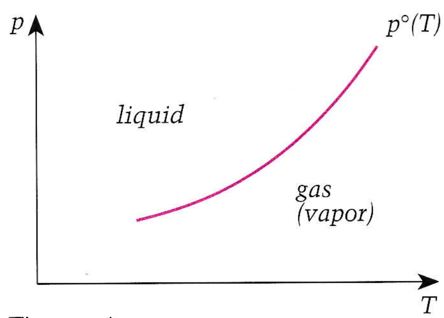


Figure 1
Phase diagram for the liquid–gas system.

First, let's bring energy considerations to bear on the situation. The free surface of the drop has the energy

$$E_S(r) = \sigma S = 4\pi r^2\sigma,$$

where σ is the coefficient of surface tension and $S = 4\pi r^2$ is the surface area for a sphere with radius r. It's easy to show that the minimum energy would be that of a single drop with doubled mass, and in our case the total surface area is $\sqrt[3]{2}$ times larger.

Theoretically this could mean that the equilibrium is unstable. But is there a mechanism that could lead to a directed transfer of the substance if, because of fluctuation, the symmetry is broken and the droplets become even slightly different?

Yes, there is such a mechanism. We're going to prove that the rate of evaporation depends on the shape of the evaporating surface. On a highly curved surface, which is the case with a small drop, this process is more intensive. The counterflow of molecules (condensation) is in turn determined only by the vapor density

At right: "Falling from the leaf, the dewdrop cries out in terror."

Art by Leonid Tishkov


Падая с листа
росинка кричит
от страха
3 мая 1992 Тишков

under the bell jar and is the same for both drops. If the radii of the drops are different, the drops cannot simultaneously be in equilibrium with the vapor. The larger drop will be more stable and will be able to gradually devour the smaller one, even without touching it. Such long-distance feasting might be called "telecannibalism."

Mathematical basis: Kelvin's formula

The law of conservation of energy permits us to calculate a correction to the vapor pressure necessitated by the surface deformation. Let's analyze the machine depicted in figure 2. It is a "perpetual-motion machine," and its only shortcoming (so the inventor thought) is its low power.

The Parisian Academy of Sciences has rejected such devices for more than two centuries (since 1775). The law of conservation of energy guarantees that they won't work. But sometimes it's useful to find where the inventor made an error—the concrete physical principle that wasn't taken into account. Methodologically this resembles an indirect proof, which is common in mathematics.

The general idea of this perpetual-motion machine is the following. A capillary tube made of a nonwettable material is immersed in a container partially filled with a volatile liquid. The level of the liquid in the capillary tube will be lower than in the container by

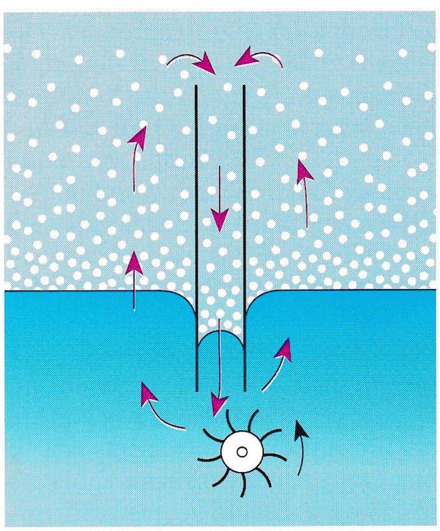


Figure 2
Proposed perpetual-motion machine.

$$h = \frac{2\sigma}{\rho_l g r},$$

where ρ_l is the density of the liquid, r is the radius of the capillary tube, and g is the acceleration due to gravity. The vapor above the surface of the liquid has the pressure $p°(T)$, while over the meniscus in the capillary tube its pressure will be greater by

$$\Delta p = \rho° g h,$$

where $\rho°$ is the saturated vapor density (according to the Clausius–Clapeyron–Mendeleyev equation $\rho° = p°(T)M/RT$, where M is the molar mass of the liquid and R is the universal gas constant). According to the inventor's conception, the excess vapor will cause condensation in the capillary tube, and so the liquid will begin to circulate in the system.

This line of thought has only one weak spot: the inventor didn't take into account the effect mentioned in the previous section. Let's compare molecules near the surface of the liquid in the vessel and in the capillary tube (fig. 3). It's evident that in case (a) any molecule has more neighbors, which are linked with it by intermolecular forces. And this means that in case (b) it is easier for a molecule to leave the liquid, the rate of evaporation is greater, and in order to compensate for it, a greater counterflow of particles is necessary.

The law of conservation of energy asserts that there is no directed circulation in the system. We come to the conclusion that the equilibrium pressure of the vapor above the convex surface of the liquid is

$$\begin{aligned} p_{\frown}(r) &= p°(T) + \Delta p \\ &= p°(T) + \frac{2\sigma}{r}\frac{\rho°(T)}{\rho_l}. \qquad (1) \end{aligned}$$

(Notice that gravitational acceleration g has dropped out of the final answer, although it was present in intermediate calculations.)

Analysis of the perpetual-motion machine was the shortest path to our goal. But the formula obtained is universal in nature: for a given liquid, the va-

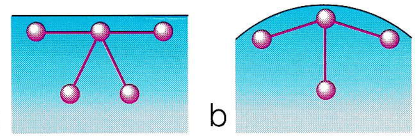


Figure 3
A molecule at a flat surface (a) is linked with more molecules than at a convex surface (b).

por pressure depends not only on the temperature but also on the curvature of the surface.

Now let's suppose that the interface is concave. After placing a wettable capillary tube in our perpetual-motion machine, we immediately obtain the result in which the added element Δp has the opposite sign:

$$p_{\smile}(r) = p°(T) - \frac{2\sigma}{r}\frac{\rho°(T)}{\rho_l}.$$

If the radii of curvature are not too small, our equations are equivalent to the more exact equation

$$p(r) = p°(T)e^{\frac{2\sigma}{r}\frac{M}{\rho_l RT}},$$

obtained by Lord Kelvin in 1871. The radius of curvature is taken with a "+" sign for a convex surface and with a "−" sign for a concave surface.

For the given liquid we can obtain the characteristic value of the radius of curvature at which a correction to the pressure becomes comparable to the vapor pressure above a flat surface:

$$r_0(T) = \frac{2\sigma}{p°(T)}\frac{\rho°}{\rho_l} = \frac{2\sigma M}{\rho_l RT} = \frac{2\sigma V_l}{RT}$$

(here V_l is the molar volume of the liquid phase). The magnitude of r_0 depends on temperature only, since σ and V_l are functions of temperature.

The first numerical estimates

Let's discuss the magnitude of the effect discovered and determine whether it plays any noticeable role in real life. The values of σ, ρ_l, $\rho°$, $p°$, and r_0 for water at different temperatures are given in the table on the next page. At first it seems that the radius r_0 is discouragingly small—it is simply unobservable (it's less than the wavelength of visible light

Table

T (K)	$\sigma \cdot 10^5$ (J/m^2)	$\rho_l \cdot 10^{-3}$ (kg/m^3)	$\rho^\circ \cdot 10^3$ (kg/m^3)	$p^\circ \cdot 10^{-3}$ (Pa)	$r_0 \cdot 10^9$ (m)
273	75.5	1.00	4.88	0.611	1.2
293	72.5	1.00	17.3	2.33	1.1
373	58.8	0.96	598	101.3	0.7

by a factor of several hundred). But a ball of this size contains

$$N(r_0) = \frac{4}{3}\pi r_0^3 \rho_l \frac{N_A}{\mu}$$

$$\sim 10^2 \text{ molecules of water}$$

($N_A = 6.0 \cdot 10^{23}$ mole^{-1} is Avogadro's number and $\mu = 0.018$ kg is the molar mass of water). We can't help rejoicing at this result, and here's why.

While deriving the basic formula, we implicitly assumed that the system was macroscopic. We can talk about surface tension and vapor pressure only if the deformed portion of the surface contains a sufficient number of molecules. The value of $N(r_0)$ obtained proves that it is true for practically all $r > r_0$.

But where could the properties of deformed menisci be revealed? It's certainly pointless to try to manufacture a single capillary tube of radius r_0. But we can manage here with less exotic means.

First of all, it isn't necessary to have a proper capillary net, so we can try using a porous substance. Silica gel—a wettable material consisting of sintered granules of SiO_2—can serve as an example. The size of the granules can vary from $2 \cdot 10^{-7}$ to 10^{-6} cm, and the spaces between them are even smaller—of the order of r_0 (fig. 4). If water vapor can penetrate the spaces between granules, the vapor pressure above this substance must decrease significantly. So by means of silica gel one can create a dry atmosphere for storing hygroscopic substances, say, or to protect metals from corrosion. Silica gel can also be used to remove gases and other impurities from the air in industrial plants as well as in the home.

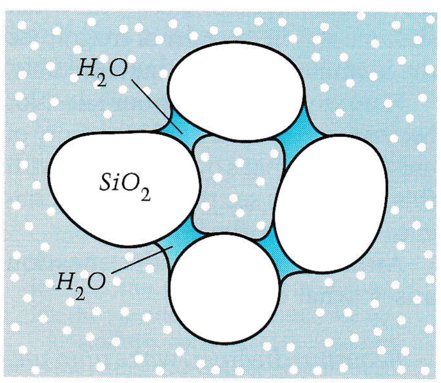


Figure 4
Silica gel (SiO_2) collects moisture in the spaces between granules.

The real importance of the imaginary surface

We meet another example of the influence of the interface on the properties of a liquid–vapor system in a case where this interface is not yet formed—in supersaturated vapor.

If a saturated vapor is in contact with the liquid, cooling or compression causes condensation. The states of this system are described by the points $p^\circ(T)$ in the phase diagram (fig. 1). But if the liquid phase is absent, we can supersaturate vapor by isothermically compressing it or cooling it in a constant volume. Formula (1) helps us understand what this means.

Vapor whose pressure exceeds $p^\circ(T)$ will be saturated with respect to drops of radius

$$r(p) = \frac{2\sigma}{p - p^\circ(T)} \frac{\rho^\circ}{\rho_l}$$

$$= \frac{p^\circ(T)}{p - p^\circ(T)} r_0(T).$$

Such drops are called *critical embryos*. (You'll recall that smaller drops evaporate readily.)

The number of molecules inside the embryo is

$$N(r) = \left(\frac{p^\circ}{p - p^\circ}\right)^3 N(r_0).$$

If the degree of supersaturation is 10%—that is, $(p - p^\circ)/p^\circ = 0.1$, then $N(r)$ is of the order of 10^5 molecules. In the gas phase the same number of molecules occupies a volume several thousands of times larger than in the liquid phase. You see that the formation of a critical embryo is rather problematic, since the vapor molecules are not likely to suddenly want to "crowd together." So supersaturated vapor can stay in its energetically unprofitable, so-called *metastable* state for a long time.

Real condensation, however, proceeds a bit differently. Let's try to estimate the characteristic size of the droplets that lead to fog and dew. Suppose that the humidity the previous evening was 100%—that is, the partial pressure of vapor in the air was $p^\circ(T_e)$, where T_e is the evening temperature. By morning the temperature had decreased by ΔT degrees. At a temperature $T_m = T_e - \Delta T$, vapor at a pressure of $p^\circ(T_e)$ is supersaturated, since the pressure of saturated vapor decreases with temperature and $p^\circ(T_e) > p^\circ(T_m)$.

To determine the size of the embryo under these conditions, we need to know how saturated vapor pressure depends on temperature. This is determined by the Clapeyron–Clausius equation:

$$\frac{dp^\circ(T)}{dT} = \frac{q}{T(V_v - V_l)}.$$

Here q is the molar latent heat of evaporation of the given liquid, V_l and V_v are molar volumes of the liquid and the saturated vapor.

This means that after the decrease in temperature of ΔT, the pressure was greater than the equilibrium pressure by

$$\Delta p = \frac{q}{T_e(V_v - V_l)} \Delta T,$$

and the radius of the critical embryo is

$$r \cong \frac{2\sigma}{q} V_l \left(\frac{T}{\Delta T}\right).$$

For an initial temperature of 293 K ($\sigma = 72.5 \cdot 10^{-5}$ J/m^2, $q = 44.0 \cdot 10^3$ J/mole) with an overnight temperature decrease of 5 degrees, we get $r \cong 3r_0$.

Now we understand that the purely fluctuational formation of an embryo containing $N \cong 6 \cdot 10^3$ molecules is unlikely. So we come to the conclusion that there must be a mechanism that facilitates condensation.

The ancient Greeks teach us

Jason, the fabled hero of the story of the Golden Fleece, sowed the field of Ares with dragon teeth, in accordance with an order from Æetes, king of Colchis. (Ares was the ancient Greek god of war, better known by his Roman name Mars.) The treacherous plan of Æetes came to light as soon as the first shoots appeared: the seeds sprouted armed warriors, ready for battle with anyone they saw. Jason was saved by a ruse suggested to him by the sorceress Medea. A stone thrown into the center of the field attracted the attention of the warriors, and they turned their weapons against one another.

This is actually a pretty good picture of the collapse of the metastable phase. Any real medium contains some inhomogeneities. The laboratory container may have defects or impurities on the walls; the atmosphere contains microscopic specks of dust. During cooling, condensation begins around these centers, which play the part of embryos and stimulate the transition to the stable phase. So metastable states are rarely observed in nature. Painstaking preparation is needed to investigate them in the laboratory, because the further the system is from equilibrium, the smaller the inhomogeneities and defects that become troublesome.

Stepping back

Now let's try to get away from the real-life but too concrete example and generalize our results. They turn out to bear a relation not only to the gas–liquid transition we discussed but to practically all first-order phase transitions.[1] In fact, we've been talking about effects caused by additional energy at the phase interface. For a system only slightly out of equilibrium, the phase transition may start only near alien defects, and the critical size of the embryo is inversely proportional to the deviation from equilibrium. That's why there isn't a speck of dust in the air on a dewy summer morning, and in winter even the smallest twigs are decorated with frost.

This sensitivity of metastable states to the presence of embryos found a practical application. For a long time elementary particles were detected by means of the Wilson cloud chamber, which contained the supercooled vapor of some liquid as its working medium. When a charged particle passed through the chamber, it caused condensation, which made its track visible. Now the Wilson cloud chamber has been replaced by more sensitive bubble chambers. In bubble chambers a superheated liquid (for instance, liquid hydrogen), which is a denser medium, is used for detection. The particle tracks reveal themselves as chains of bubbles (the hydrogen boils).

Another example is the cultivation of high-quality monocrystals. A small seed monocrystal is immersed in a near-equilibrium melt,[2] where it becomes the only center of crystallization. (If there is no seed, the melt could be cooled down to $T \cong 0.5T_{\text{melting}}$ before crystallization starts.) But, alas, you pay for high quality with long growth times. It takes several weeks or even months to grow a large monocrystal.

A conclusion, but not a finale

Before typing the final period in this article, I should clear up one point. Up to now we have discussed how drop formation begins. But when does it end? Why are all dewdrops so much alike in the morning? What stops their growth? We made sure that the larger the radius of a drop becomes, the more greedily it "drinks." So there must be some additional factor that determines the maximum size.

You needn't go far afield for the answer. Up to now, we haven't taken into account the influence of gravity on a dewdrop. But the flattened form of large drops suggests that this isn't always permissible.

Let's try to estimate how long we can neglect the gravitational energy. For a spherical drop the gravitational energy E_g is

$$E_g = \frac{4}{3}\pi r^3 \rho g r$$

and, since it is proportional to r^4, it must inevitably exceed the surface energy $E_s = 4\pi r^2\sigma$ as r increases.

To find the limiting value of the radius, let's equate the energy of a dewdrop to the total energy of two "half-drops" into which it could separate:

$$\begin{aligned} E(1) &= 4\pi r^2\sigma + \frac{4}{3}\pi r^4 \rho g \\ &= 2\left[4\pi\left(\frac{r}{\sqrt[3]{2}}\right)^2 \sigma + \frac{4}{3}\pi\left(\frac{r}{\sqrt[3]{2}}\right)^4 \rho g\right] \\ &= 2E\left(\frac{1}{2}\right) \end{aligned}$$

($r/\sqrt[3]{2}$ is the radius of the half-drop). So

$$r_{\max} = \sqrt{\frac{3\sqrt[3]{2}\sigma}{\rho g}} \cong 0.5 \text{ cm}.$$

At first glance it seems to be too large for a dewdrop. But we haven't taken all the factors into account, of course. In particular, the size of a drop depends on the wettability of the surface where it forms. Besides, very large drops simply can't stay on the leaves and blades of grass—they just slide off. But in its order of magnitude the estimate we made is quite credible.

Here we can interrupt our tale. Not because it's over—it's only the story of one dewdrop that's over. But even with this droplet we see that no step in science can ever be the final one.

After each answer, the question arises: what else?

Beautiful dreamer, wake unto me,
Starlight and dewdrop are waiting
for thee.

—*Stephen Foster*

[1] "First-order" phase transitions are those from one phase state to another in which such characteristics of the substance as density or concentration change abruptly; a definite amount of heat per unit mass is liberated or absorbed during this process. Examples of such transitions are evaporation and condensation, melting and solidification, sublimation and condensation into the solid phase, and so on.

[2] The nearer to equilibrium the melt is, the better the crystals obtained and the longer they will grow.

MATH INVESTIGATIONS

Periodic functions in hiding

Can you root them out?

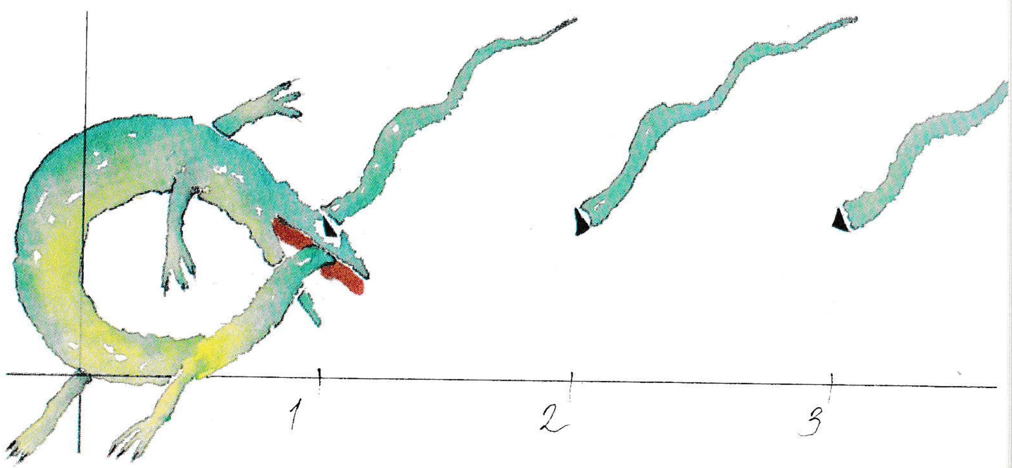
by George Berzsenyi

THE FOLLOWING PROBLEM appeared in my "Competition Corner" in the December 1990 issue of the now defunct *Mathematics Student* journal of the National Council of Teachers of Mathematics. It is offered here as a first challenge to my readers: **Suppose that the function f satisfies the functional equation $f(a, b) = f(a + b, b - a)$ for all real numbers a and b, and define g by $g(x) = f(4^x, 0)$. Show that g is periodic.**

This problem was originally submitted by one of my former contestants, Martin Gelfand, who was a high school student at that time. Since then, Andrew went on to receive a Ph.D. in physics (Cornell, 1989), and he is presently on the faculty of the University of Illinois. Andrew's problem was too good to be discarded after one use, so I went on to pose the following problem in the most recent round of the USA Mathematical Talent Search (USAMTS, featured via *Consortium*). It is offered here as our second challenge: **Prove that if f is a nonconstant real-valued function such that for all real x, $f(x + 1) + f(x - 1) = \sqrt{3}f(x)$, then f is periodic.**

After solving these problems, the reader should also gain some insights into the creation of problems and wonder what other numbers beside $\sqrt{3}$ might lead to periodic behavior. I have done some such investigations prior to posing this problem in the USAMTS, and my colleague Dr. John Rickert (another former student-participant in the "Competition Corner" ever so long ago) did a lot more. But I was happiest for extensions communicated to me by the teacher of a student in less favorable circumstances. This student attends a juvenile court school in a major city in the southwestern US, doing mathematics at a level several grades below others his age in public or private schools. Nevertheless, after solving the original problem, he has gone on (with the help of his teacher) to discover the general pattern behind this curiosity! My next challenge to my readers is to follow his example: **Find all constants k for which the difference equation $f(x + 1) + f(x - 1) = kf(x)$ implies periodic behavior.**

In closing, I wish to commend this bright young man and his teacher for their serious and thoughtful response. I would urge my readers to further their discoveries and **find other difference equations whose solutions are necessarily periodic.** Such investigations may even lead to separate publications. Recently, one of my readers, Michael Filaseta, was led to some wonderful results based on my "Triangles of Differences" (see page 30 in the May/June issue of *Quantum*); while another regular reader, Brian Platt, obtained similarly publishable results related to my "Restricted Distances" (see page 31 of the January/February issue). I hope that many others will follow their examples. ◘



The purpose of this column is to direct the attention of *Quantum*'s readers to interesting problems in the literature that deserve to be generalized and could lead to independent research and/or science projects in mathematics. Students who succeed in unraveling the phenomena presented are encouraged to communicate their results to the author either directly or through *Quantum*, which will distribute among them valuable book prizes and/or free subscriptions.

GETTING TO KNOW ...

Inversion

A most useful kind of transformation

by Vladimir Dubrovsky

FIGURE 1 ILLUSTRATES THE dramatic changes suffered by a plain chessboard when it is inverted in a circle. Although the pitiless transformation of inversion turns the inside of the circle out (and the outside in) and bends straight lines into circles, somehow it contrives to preserve some fundamental features of figures—for instance, the magnitudes of angles between curves; and we can usually recognize an object in its *inverse*—the image under inversion. Owing to its remarkable properties, inversion often simplifies the solutions of rather difficult geometrical problems (like math challenge M63). And *Quantum* has already presented two of its numerous applications: in "Constructions with Compass Alone" (May 1990), Dmitry Fuchs showed how to perform an arbitrary compass-and-ruler geometric construction *without* a ruler; and Yury Solovyov, in "Making the Crooked Straight" (November/December 1990), used inversion to explain the work of some hinge mechanisms that convert circular motion into rectilinear. The reason I invite you to revisit this wonderful transformation now is that it's indispensable for understanding the so-called Poincaré model of non-Euclidean geometry, which will be discussed in detail in the next issue of *Quantum* (see also the article by Vladi-mir Boltyansky on p. 18).

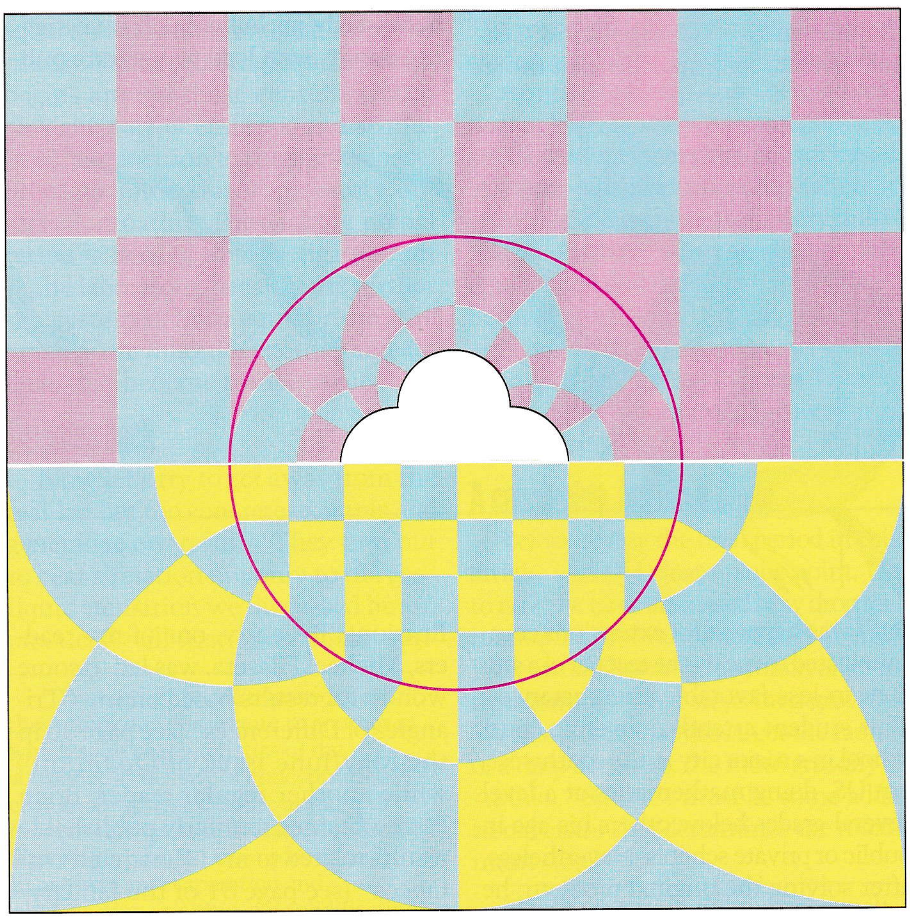


Figure 1

Inverting points

By definition, the *inverse of a point X in a circle* ω with center O and radius R is the point X' on the ray OX such that

$$OX' \cdot OX = R^2.$$

The transformation I_ω (or I_O^R) assigning to every point its inverse is called *inversion*; ω, O, and R are the *circle, center*, and *radius* of inversion.

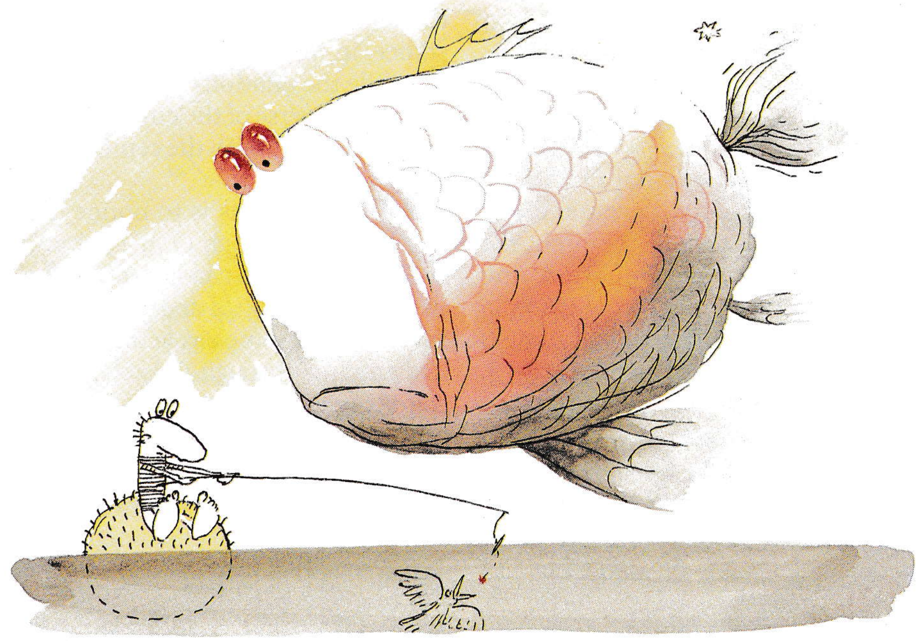
We see at once that inversion keeps all the points of ω in their places; all the other points except the center O can be paired so that one point of each pair lies inside ω and the other outside, and either of them is the other's inverse.

To visualize better what happens to the plane when it is inverted, note that the nearer a point is to the center, the further away its inverse is. As a point approaches the center, its inverse recedes to infinity. It is therefore reasonable—and convenient—to add to the plane a special ideal *point at infinity* P_∞ that will serve as the inverse of the inversion center O. (The inverse of P_∞ is O.)

In a coordinate system whose origin is at the center of inversion I_O^R, the inverse of point (x, y) is given by the formula

$$(x', y') = \left(\frac{R^2x}{x^2+y^2}, \frac{R^2y}{x^2+y^2}\right). \quad (1)$$

Indeed, the coordinates (x', y') are proportional to (x, y) with a positive factor of $R^2/(x^2 + y^2)$, so points (x, y) and (x', y') lie on the same ray from the origin. A little algebra shows that the product of their distances from the origin, $\sqrt{x'^2+y'^2} \cdot \sqrt{x^2+y^2}$, equals R^2. A geometric construction of the inverse X' of a point X outside ω is illustrated in figure 2: we draw a tangent XA to ω and drop the perpendicular AX' from A to OX. This construction is easily reversed to find X given X' inside ω. Another construction, with compass alone, is shown in figure 3: it's even simpler, but it only works with points X such that the circle with radius XO centered at X intersects ω (the compass-alone construction for the other case—$OX \le R/2$—can be found in the aforementioned article by Fuchs).



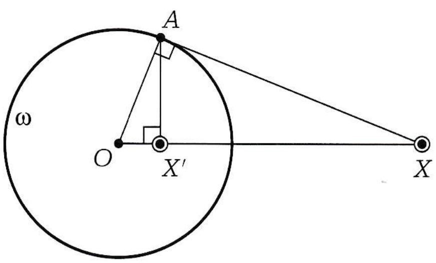


Figure 2

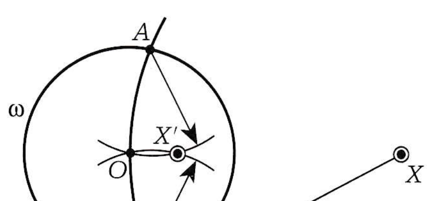


Figure 3

Construction of the inverse X' of point X in circle ω with compass alone. Draw a circle through O centered at X and two more circles through O centered at the points A and B where the first circle meets ω. X' is the second point of intersection of the last two circles.

Exercises

1. Prove that point X' is indeed the inverse of X in ω (a) in figure 2 and (b) in figure 3.

2. Prove that two successive inversions with the same center O and radii R_1 and R_2 result in one dilation of the plane relative to center O by a factor of $(R_2/R_1)^2$.

Inverting circles

One of the most remarkable and useful properties of inversion can be concisely stated as follows: *the inverse of a circle is a circle.* The term "circle" in this short formulation must be understood in the generalized sense to include straight lines, which are thought of as "circles" that pass through the point at infinity P_∞ and have an infinite radius. Since P_∞ is the inverse of the center of inversion, we can make the following statement: *the inverse of a line or a circle through the center of inversion is a line, and the inverse of any other line or circle is a ("finite") circle.*

I'll give a proof using coordinates that is simple, comprehensive, and rather short, though it lacks the beauty of the purely geometric approach.

Consider a circle with center C and radius r. Let the origin of the coordinates be at the center O of inversion, and let the x-axis pass through C, so that C has the coordinates $(d, 0)$ ($d = OC$; see figure 4). To write an equation for the locus described by the inverse $Q(x, y)$ of a point P tracing the given circle, notice that P is the inverse of Q, so the coordinates (x', y') of P are given by formula (1). On the other hand, the equation $CP^2 = r^2$ of the given circle yields

$$(x' - d)^2 + y'^2 = r^2,$$

or, after plugging in formula (1),

$$\left(\frac{R^2x}{x^2+y^2} - d\right)^2 + \left(\frac{R^2y}{x^2+y^2}\right)^2 = r^2.$$

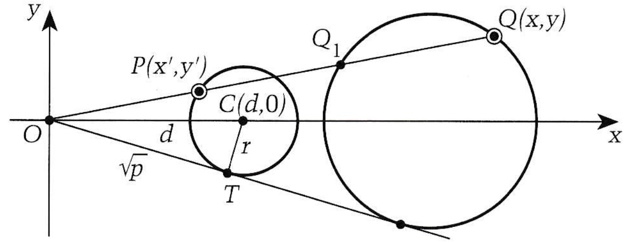


Figure 4

After routine juggling of variables and coefficients, we arrive at the equation

$$(d^2 - r^2)(x^2 + y^2) - 2R^2dx + R^4 = 0. \quad (2)$$

The value $p = d^2 - r^2 = OC^2 - r^2$ is called the *power of point O with respect to the circle* (with center C and radius r). If O is outside the circle, the power p equals the square of the length of a tangent OT from O to the circle (consider the right triangle TOC in figure 4); see also exercise 3.

In the case of a circle passing through O, the power p (of O with respect to the circle) equals 0, so equation (2) takes the form

$$x = \frac{R^2}{2d}$$

and defines a straight line perpendicular to the x-axis—that is, to line OC (fig. 5).

If O is not on the circle, we can divide equation (2) by $p = d^2 - r^2 \neq 0$ and, after completing the square, get the equation

$$\left(x - \frac{R^2d}{p}\right)^2 + y^2 = \frac{R^4d^2}{p^2} - \frac{R^4}{p} = \frac{R^4r^2}{p^2},$$

or

$$(x - kd)^2 + y^2 = (kr)^2, \quad k = \frac{R^2}{p}.$$

This is the equation of the circle with center $(kd, 0)$ and radius $|k|r$. So *the inverse of a circle not passing through the center O of inversion coincides with the dilation of this circle by a factor of* $k = R^2/p$ *relative to the center O.* (But notice that separate points of a circle are mapped differently by an inversion and the corresponding dilation—for instance, in figure 4 the inverse of P is Q, while the dilation takes P into Q_1. Also, the inverse of the circle's center *never* coincides with the center of the circle's inverse.)

All that remains is to notice that any line not through O in a suitable coordinate system can be represented in the form $x = R^2/d$ and therefore is inverted into some circle through O; and any line through O is obviously self-inverse. This completes the proof of the statement at the beginning of this section.

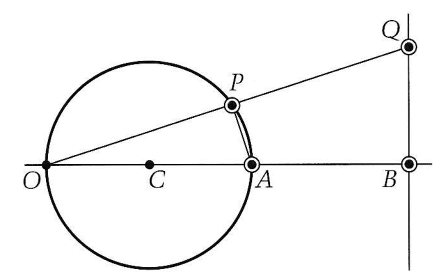


Figure 5
A circle through the center of inversion that is inverted into a line. A direct geometric proof of this fact follows from the similarity of right triangles OAP and OQB. (Complete this proof on your own.)

Exercises

3. An arbitrary line through a given point P cuts a given circle ω in points A and B. Prove that the product $PA \cdot PB$ does not depend on the line and equals the power of P with respect to ω if P lies outside the circle, and the negative of this power if P lies inside the circle. Using this statement and exercise 2, give another proof that circles are inverted into circles.

4. Find the radius of circle ω in figure 6 given the side length a of the square $ABCD$ (AEB is a semicircle, arc AEC is centered at D). (Hint: invert the figure in a circle with center A and radius AB.)

Let's look more closely at the circles inverted onto themselves. Of course, the circle of inversion ω is self-inverse—all its points stay fixed; and this is the only self-inverse circle with center O. In the general case, using the fact that the inversion of a circle can be replaced by its dilation by a factor of $k = R^2/p$, we deduce that a circle ω_1 with center other than O is self-inverse if and only if the respective factor k is equal to 1, or $p = d^2 - r^2 = R^2$. Geometrically, this means that point O lies outside circle ω_1 ($d > r$), and a tangent OT drawn from O to ω_1 is a radius of ω (because $p = OT^2$—see figure 7). The tangents to ω and ω_1 at their point of intersection T (TC and TO in figure 7) are perpendicular; such circles are called *orthogonal* to each other. Thus, *a circle distinct from ω is self-inverse if and only if it is orthogonal to ω.* (By the way, this is true for lines—"circles of infinite radius"—as well; lines through O are orthogonal to ω.)

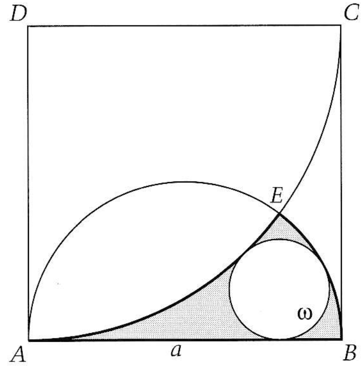


Figure 6

Now we can describe inversion in terms of orthogonal circles: *point Q is the inverse of P in ω if any circle through P and Q is orthogonal to ω.* (The inverse of such a circle will again pass through P and Q and, in addition, intersect ω at the same points as the original circle, which stay fixed.) Notice that point P here can be the point at infinity; then Q is the center of ω, and the circles through P and Q turn out to be straight lines—extended diameters of ω, which, needless to say, are orthogonal to ω. Inversion in a circle and reflection in a line thus share a common property. If points P and Q are images of each other (with respect to either sort of transformation), then any circle through P and Q is orthogonal to the circle of inversion or line of reflection. For this reason, inversion is sometimes called "reflection in a circle," and two points inverse to each other in some circle are said to be "symmetric about this circle."

Exercise

5. Given points A and B and circles ω_1 and ω_2, construct a circle ω (a) through A and B orthogonal to ω_1; (b) through A orthogonal to ω_1 and ω_2; (c) orthogonal to ω_1 and inverting A into B.

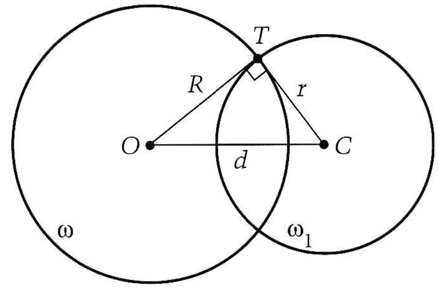


Figure 7

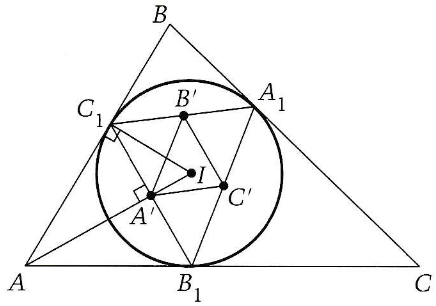


Figure 8

As an example of how inversion is applied to solving geometric problems, let's *express the distance d between the centers O and I of the circumcircle and incircle of a triangle ABC, given their respective radii R and r.*

Let A_1, B_1, C_1 be the points of contact of the incircle with the sides of the given triangle (fig. 8). Comparing triangle IC_1A with triangle OAX in figure 2, we see that the inverse of vertex A in the incircle is the midpoint A' of segment B_1C_1; similarly, the inverses of B and C are the midpoints of C_1A_1 and A_1B_1. It follows that the circumcircle of triangle ABC is inverted into the circumcircle of $A'B'C'$. Since triangle $A'B'C'$ is half as big as triangle $A_1B_1C_1$ (the sides of the former are the midlines of the latter), the circumradius of $A'B'C'$ is half the circumradius of $A_1B_1C_1$—that is, half the inradius of ABC, or $r/2$. On the other hand, by the formulas above for the inverse of a circle, the circumradius of $A'B'C'$ equals the circumradius of ABC times $r^2/|p|$, where $p = d^2 - R^2 < 0$ is the power of the center of inversion I with respect to the circumcircle of ABC. So we arrive at the equation $r/2 = R \cdot r^2/(R^2 - d^2)$, yielding the following formula:

$$d^2 = R^2 - 2Rr,$$

one of the numerous formulas bearing the name of the great Leonhard Euler.

Inverting angles

Consider a circle ω, a point P on it, their respective inverses ω′ and P', and line $l = OP$ through the center of inversion O (fig. 9). It appears from

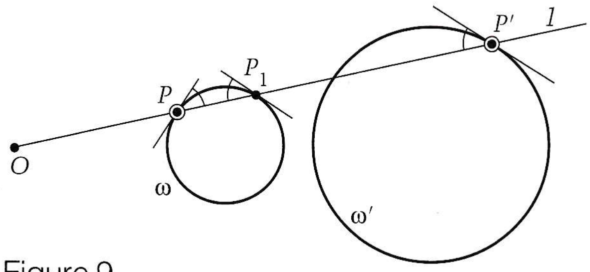


Figure 9

the figure that the angle between ω and l at P—that is, between the tangent to ω to P and l—is equal to the angle between ω′ and l at P'. And this is easy to prove: both these angles are equal to the angle between ω and l at their second point of intersection Q (the angle at P is taken into the angle at Q under the reflection in the diameter of ω perpendicular to l, and the angle at Q is taken into the angle at P' under the dilation relative to O taking ω into ω′, which always exists).

Now consider the angle between any two circles (that is, between their tangents) at one of their points of intersection P. Any such angle can be represented as the sum (or difference) of the angles between line $l = OP$ and the circles (fig. 10). These angles are equal to the corresponding angles between l and the inverses of the circles, so the angle between our circles equals the angle between their inverses. In other words, *inversion preserves angles between circles.*

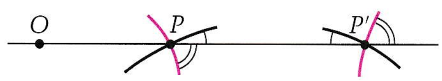


Figure 10

Actually, the same is true for the angle between a circle and a line, or between two lines, or, generally, between any two curves. (We can draw two circles touching the curves at their point of intersection; then the angle between the circles will be equal to the angle between the curves, and the inverses of the circles will touch the inverses of the curves. So the angle between the inverted curves equals the angle between the inverted circles; and since the angle between the circles is preserved, so is the angle between the curves.)

The preservation of angles implies the preservation of orthogonality. Now recall that if two points P and P' are symmetric about a circle ω, then any circle through them is orthogonal to ω (fig. 11a). Inverting the entire configuration in an arbitrary circle, we obtain points Q and Q' such that any circle through them is orthogonal to the inverse of ω—that is, points symmetric about the inverse of ω. In particular, choosing the center of inversion on ω, we turn ω into a straight line and points P and P' into Q and Q' symmetric to each other about this line (fig. 11b). In this sense, *inversion can be inverted into line reflection.*

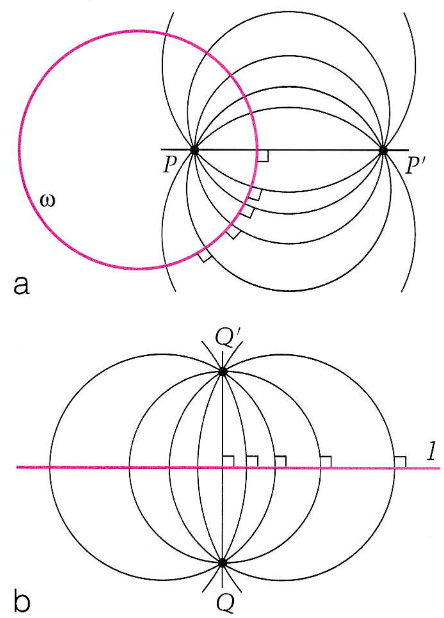


Figure 11

I'll use this connection between the two kinds of transformation to explain how to draw a circle ω through three given points A, B, C with compass alone.

It suffices to construct the center O of the required circle ω (fig. 12).

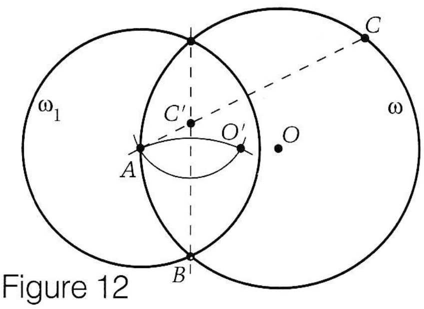


Figure 12

Note that the inverse of O in ω is the point at infinity P_∞. Therefore, the inverses of O and P_∞ in an arbitrary circle ω_1 are symmetric about the inverse of ω in ω_1. If ω_1 is the circle through B with center A, then $I_{\omega_1}(P_\infty) = A$, $I_{\omega_1}(B) = B$, $I_{\omega_1}(\omega)$ is the line through B (in figure 12), and $I_{\omega_1}(C)$ is a point on this line, here labeled C'. So $O' = I_{\omega_1}(O)$ is the *reflection of A in line BC′*. To summarize, the order of construction is as follows: draw circle ω_1, construct the inverse C' of C in ω_1 (by the method in figure 3), then the reflection O' of A about line BC' (as the second intersection of the circles through A centered at B and C'), and finally the inverse of O' in ω_1, which is the desired center O.

This construction, together with that of figure 3, enables us to translate an arbitrary compass-and-ruler construction into a compass-alone construction with only one reservation: we'll consider a straight line as "having been constructed" if any two of its points have been constructed. Suppose, for instance, we must find the intersection point P of two straight lines AB and CD, given only the four points A, B, C, D. We draw an arbitrary circle ω with some center O, construct inverses A', B', C', D' of the given points in this circle (by the method in figure 3), draw two circles—through O, A', B' and through O, C', D'—by the method presented above, and finally invert their second point of intersection P' in circle ω. The inverse of P' is the desired point P, because ω inverts lines AB and CD into circles $OA'B'$ and $OC'D'$, and consequently point P into P'. Similarly, one can construct intersection points of a circle and a line through two given points. So we can replace every application of a ruler in a given sequence of compass-and-ruler constructions by a compass-alone construction yielding the same new points. In the end we'll get exactly the same configuration as if we used a ruler except that straight lines will be represented by pairs of their points (in accordance with our agreement).

More on constructions with compass alone can be found in the article by Fuchs mentioned at the beginning of this article.

Inverting distances

Of course, distances—unlike angles—are not preserved under inversion. What's more, the distance between the inverses A' and B' of points A and B cannot be expressed in terms of the distance AB only; it depends also on how far points A and B are from the center O of inversion. But the formula for $A'B'$ is easily deduced from figure 13. Triangles OAB and $OB'A'$ in this figure are similar (they have a common angle, and $OA/OB' = OB/OA'$ because $OA \cdot OA' = OB \cdot OB'$). So $A'B'/AB = OB'/OA = R^2/(OB \cdot OA)$, where R is the radius of inversion. Finally,

$$A'B' = \frac{R^2}{OA \cdot OB} \cdot AB. \qquad (3)$$

Exercise

6. Prove that for any four points A, B, C, D in the plane

$$AB \cdot CD + AD \cdot BC \geq AC \cdot BD,$$

the inequality turns into an equality if and only if $ABCD$ is a quadrilateral inscribed in a circle (the last statement is Ptolemy's theorem). (Hint: invert in a circle with center A and apply the formula for the distance between inverses.)

I'll use formula (3) to prove that a certain combination of the distances between four points is preserved under inversion. This fact, which seems rather artificial at first glance, turns to be very helpful later.

Let's choose four points A, B, C, D and invert them in some circle ω, with center O. Let's call their inverses A', B', C', D', respectively. By formula (3),

$$A'B' = \frac{R^2}{OA \cdot OB} \cdot AB,$$

$$B'C' = \frac{R^2}{OB \cdot OC} \cdot BC.$$

Dividing, we find

$$A'B'/B'C' = (OC/OA) \cdot (AB/BC).$$

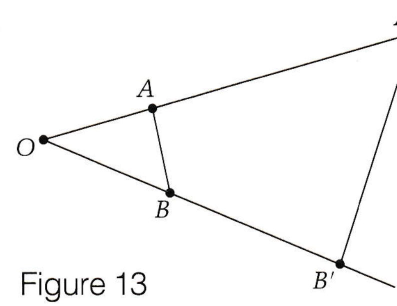


Figure 13

In other words, the ratio AB/BC after inversion is multiplied by a factor that does not depend on B. Therefore, if we divide this ratio by the like ratio for points A, C, and D (instead of B), this factor will cancel out, so the "ratio of two ratios"

$$\frac{AB}{BC} : \frac{AD}{DC}$$

will remain invariant under inversion (the skeptical reader is invited to do the algebra, which will bear this out). This expression, which we denote by $\{AC, BD\}$, is called the *double ratio* or *cross ratio* of points A, C, B, D, in that order.

Inverting space

The definition of inversion can be extended to space without the slightest change. We can repeat our calculations with coordinates (including, where necessary, the third coordinate) to prove that *the inverse of a sphere is a sphere if it does not pass through the center of inversion, and a plane if it does*. A simpler proof is obtained by just rotating figures 4 and 5 about the line through O and C, the center of inversion and the center of the circle. Then the circle describes a sphere, and its inverse describes the inverse of this sphere, which is either a plane or a sphere, depending on whether the first sphere passes through O.

As to circles and straight lines themselves, their inverses in a sphere conform to literally the same rule as in the planar case. But this must be proven separately, since the center of inversion may not lie in the same plane as the inverted circle. The proof, though, presents no problem. A circle (or line) can be thought of as the intersection of two spheres (planes), and the inverse of an intersection is the intersection of the respective inverses. So the inverse of a circle or a line is the intersection of two spheres, or a sphere and a plane, or two planes (depending on the relative position of the inversion center)—that is, a circle or a line.

I leave it to the reader to prove that inversion in space preserves angles between curves. The proof is analogous to that in the two-dimensional case.

Now consider a sphere passing through the center N of a larger sphere and touching it from the inside at point S (points N and S may be thought of as the North and South Poles of the first sphere). When the smaller sphere is inverted in the larger one, it is mapped onto the plane touching both spheres at S (see figure 14, in which the larger sphere—the sphere of inversion—is not shown). This mapping, it turns out, can be described without referring to inversion, simply as the central projection of the smaller sphere onto the plane from point N: each point P of the sphere other than N is mapped onto the point Q where line NP meets the plane. Such a projection of a sphere onto its tangent plane from the point of the sphere diametrically opposite to the point of contact is called a *stereographic projection*.

It follows from the properties of inversion that the stereographic projection maps circles on the sphere into circles or lines on the plane and preserves angles between curves. This makes stereographic projection helpful both with "pure" geometrical problems (like problem M65 in this issue) and with problems of a more

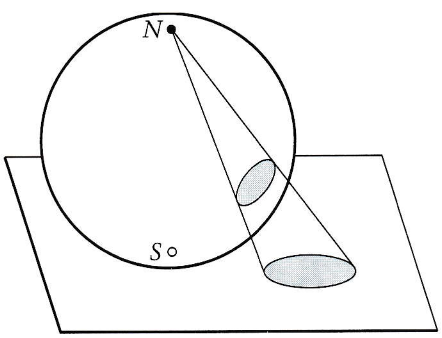


Figure 14

practical nature—say, for drawing geographic maps.

Two points of a sphere are called *symmetric about a given circle* ω *on the sphere* if any circle through them is orthogonal to ω. By the preservation of angles, stereographic projections of such points are symmetric about the projection of ω (in the plane). So an "inversion *on* the sphere" is projected into an inversion or line reflection on the plane, and vice versa. To make this correspondence complete, we must assume the stereographic projection of the pole N to be the point at infinity, which is only natural. Thus, the sphere is actually a more suitable "ground" for considering inversions than the plane, because on the sphere the inversion is a one-to-one mapping without any additional ideal points. We see also why, from the "inversive" point of view, the point at infinity is absolutely equivalent to any "good finite" point.

Exercise

7. Let ω be the "equator" corresponding to the poles N and S, and ω_1 its stereographic projection from N. Prove that the reflection of the sphere about the plane through ω is transformed under this projection into inversion of the plane in ω_1.

Inverting the parallel postulate

The intrinsic kinship of inversion and line reflection is most dramatically displayed in the model of non-Euclidean (hyperbolic) geometry, discovered in the early 1880s by the great French mathematician Henri Poincaré when he was studying a certain class of functions of a complex variable. In this model, a certain family of inversions plays exactly the same role as line reflections in the regular, Euclidean geometry. The Poincaré model provides a geometric system with its own concepts of lines, distances, and angles. These are different from the ones we're used to, but nevertheless satisfy all the usual axioms of Euclidean geometry with the one (albeit critical) exception of the parallel postulate, which is "inverted," so to speak. Instead of the uniqueness of a line passing through a given point and parallel to a given line, it is assumed that *through a point not on a given line l there is more than one line parallel to l.* I'll skip deliberations about why it's so important to have a model for a system of axioms—in particular, for that of non-Euclidean geometry. This is thoroughly discussed in Boltyansky's article (page 18), where the Poincaré model has been introduced. Now that we've gotten to know inversion, Poincaré's construction can be described in more detail.

It would be quite appropriate and helpful to think of it as a "stage production" of a rather bizarre "play" entitled "Non-Euclidean Geometry." So, to begin with, we must choose a stage for the show—the counterpart of the entire plane in Euclidean geometry. For the model in question it's the *interior of some circle* α. (We'll see in the end that the stage for the Poincaré model is, in fact, transformable and may turn into a half-plane.) The role of *line reflections* is assigned to inversions that map α onto itself—that is, to the *inversions in circles orthogonal to* α (which naturally include conventional reflections in diameters of α). This is the decisive choice that determines the rest of the casting. For instance, it implies that the part of straight lines must be played by the circles orthogonal to α (or rather, to their arcs intercepted by α) and to diameters of α. All of these will be called *p-lines*, and we'll use the tag *p-* to label the counterparts of various geometric objects in the Poincaré model—for instance, "*p*-plane" for the interior of α, and so on.

We can immediately test the validity of Euclidean axioms for *p*-lines. And most of them prove to be true. For example, by the solution to exercise 5a, there is a unique *p*-line through any two points. Also, a point on a *p*-line evidently divides it into two arcs (*p-rays*), and a *p*-line divides the *p*-plane into two areas (*p-half-planes*) such that two points lie in the same area (on one side of the *p*-line) if and only if the *p-segment* joining them (define it yourself!) does not cross the *p*-line. Non-Euclideanism shows up as soon as we touch on parallelism. As shown in figure 15, one

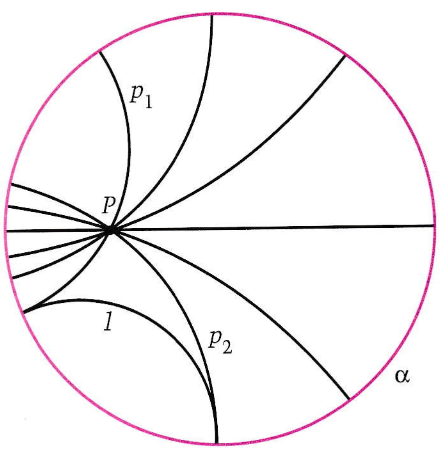


Figure 15

can always draw two p-lines through each of the endpoints of a p-line l and a point P not on it. The "endpoints" in fact do not belong to the p-plane, so the two p-lines actually never meet l. These p-lines (labeled p_1 and p_2 in figure 15) are called *parallel* to l, while the p-lines that have no common points with l, even taking into account the circumference of α, are called *superparallel* to l. One way or the other, we see that the parallel postulate is violated. So our "staging" appears to be correct.

Now we're coming to distances, angles, and other notions connected with measurement. The definitions of p-distances and p-angles are suggested by the fact that line reflection preserves Euclidean distances and angles. As for p-reflection (that is, inversion), we've seen that it, too, preserves the Euclidean measure of angles between curves. So it's only reasonable (and even inevitable) simply to set the *p-measure of an angle* between intersecting p-lines to be equal to its Euclidean measure.

The definition of the *p-distance* is trickier because Euclidean distances are changed by inversion. What we need is a function $d(A, B)$ of pairs of p-points that satisfies at least the following two requirements:

(1) It must remain unchanged by p-reflections (inversions);
(2) $d(A, C) + d(C, B)$ must be equal to $d(A, B)$ whenever C lies on a p-segment AB.

Now it's time to make use of the cross ratio we've been holding in reserve. Let A_0 and B_0 be the "endpoints" of a p-line AB, and let A_0 be further from A than from B (fig. 16). Then the cross ratio of A, B, A_0, B_0,

$$R(A, B) = \{AB, A_0B_0\} = \frac{AA_0}{A_0B} : \frac{AB_0}{B_0B},$$

depends only on A and B and matches condition (1), because if A' and B' are the p-reflections of A and B in some p-line (fig. 16), then the "endpoints" of the p-line through them will be the p-reflections of A_0 and B_0, respectively, in the same p-line (why?); and since the cross ratio is preserved, $R(A', B') = R(A, B)$.

But condition (2) fails for $R(A, B)$: to compute $R(A, B)$ given $R(A, C)$ and $R(C, B)$ for C on a p-segment AB, we have to multiply the given values rather than add them up:

$$R(A, B) = R(A, C) \cdot R(C, B)$$

(the "endpoints" are the same for all three pairs AB, AC, CB, and "single ratios" are multiplied: $AA_0/A_0B = (AA_0/A_0C) \cdot (CA_0/A_0B)$). Fortunately, God has granted us a standard way of converting multiplication into addition. Taking the logarithm of R, we derive a function

$$d(A, B) = \log R(A, B)$$

that meets both conditions (1) and (2), since $\log xy = \log x + \log y$ for $x, y > 0$. The base of the logarithm makes no difference—it merely fixes the unit of p-length. The only constraint is that the base must be greater than 1 to ensure that p-distances are positive, since the definition of R implies $R(A, B) > 1$ for $A \neq B$). By the way, with this definition of the p-distance, it turns out that bounded arcs representing straight lines in the Poincaré model are of infinite length. Why? When point B in figure 16 recedes from A to A_0, $R(A, B)$ varies directly as B_0B/A_0B and so increases indefinitely, making $d(A, B)$ do the same.

At this point I'll stop. All the leading roles in our "play" have been cast, so the way is clear for you to make your own explorations; the exercises

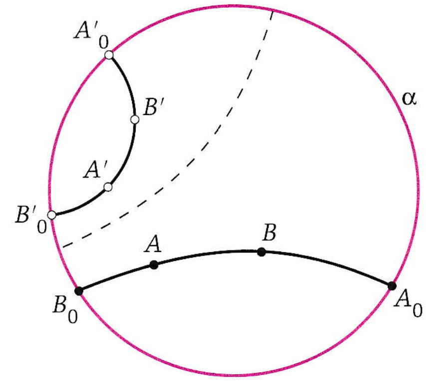


Figure 16

below may help you get started. In the next issue of *Quantum* you'll meet the Poincaré model once again—this time in a half-plane instead of a circle. You'll see that it has a neat physical interpretation. Geometrically, the two versions are absolutely equivalent: one is obtained from the other by way of inversion that turns circle α into a half-plane, and vice versa.

Exercises

8. Prove that $d(A, B) = d(B, A)$, and that $d(A, B) < d(A, C) + d(C, B)$ when C does not lie on p-segment AB.

9. Prove that the angles of a p-triangle always add up to a value less than 180°.

10. Two figures in the p-plane are *p-congruent* if one of them can be mapped onto the other by a number of successive p-reflections. Prove the side–angle–side test for p-congruence of p-triangles.

ANSWERS, HINTS & SOLUTIONS IN THE NEXT ISSUE

AT THE BLACKBOARD

A trio of topics

Questions of mass, electricity, and time

What is the center of mass?

IF YOU THROW A STONE AT AN ANGLE TO THE horizon, it's going to fly along a parabola, as you well know. But what if you throw a long stick, giving it a good flick to make it twirl as it flies? Different points of the stick will move differently in rather intricate trajectories, of course; but the flight of the stick as a whole will pretty much be like the flight of the stone: ascent, maximum elevation, descent. Not only that, if you neglect air resistance, there is a point on the stick that moves exactly the same as the stone flying freely along a parabola. This point is the stick's center of mass.

A center of mass exists for every body, and for any system of bodies as well. It has some very interesting properties, some of which we'll look at in this article.

Let's begin with determining the position of the center of mass. Consider a system of material points with masses $m_1, \ldots, m_n$. If we know their coordinates, how can we determine the coordinates of this "most important point"? Here's what the answer looks like:

$$x_{\mathrm{cm}} = \frac{m_1x_1 + m_2x_2 + \cdots + m_nx_n}{m_1 + m_2 + \cdots + m_n}, \qquad (1)$$

Art by Pavel Chernusky

with analogous expressions for the y_{cm}- and z_{cm}-coordinates. You'll see why the center of mass is defined in this particular way later, when we consider its dynamic properties. For now, let's get used to expression (1) by discussing some questions related to it.

(a) In the case of two point masses m_1 and m_2, their center of mass lies along the line connecting the points and is closer to the point with greater mass (see figure 1); the ratio of the distances from the center of mass to the two points is the inverse ratio of their masses (you can check all this yourself). It's clear that in the general case the center of mass is to be found somewhere between the points of a system, and its position depends on the spatial distribution of the masses.

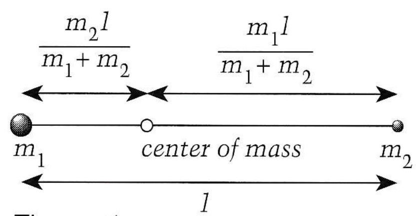


Figure 1

(b) All the material points have "equal rights" in determining the position of the center of mass. If the masses are distributed symmetrically with respect to some point in space, it is this point that is the center of mass. For example, the center of mass of a uniform sphere coincides with its center (the same is true for a cylinder, a cube, and so on).

(c) One more comment. It turns out (try and prove this fact on your own) that the position of the center of mass remains unchanged if we single out a part of the system and concentrate all its mass at one point—its center of mass. For example, the center of mass of a wire triangle coincides with the center of mass of the system consisting of three points situated at the midpoints of the triangle's sides (the masses of the points being equal to the masses of the corresponding sides).

Now let's move on to what is most important—an examination of the physical properties of the center of mass.

Let the points be displaced through distances $\mathbf{s}_1, \ldots, \mathbf{s}_n$ in a small time interval Δt. From equation (1) the displacement of the center of mass is seen to be equal to

$$\mathbf{s}_{\mathrm{cm}} = \frac{m_1\mathbf{s}_1 + m_2\mathbf{s}_2 + \cdots + m_n\mathbf{s}_n}{m_1 + m_2 + \cdots + m_n}$$

(recall that projections of the displacement vector are equal to the changes in the corresponding coordinates). Dividing the displacement by the time interval Δt, we find the velocity of the center of mass:

$$\mathbf{v}_{cm} = \frac{m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + \cdots + m_n\mathbf{v}_n}{m_1 + m_2 + \cdots + m_n}. \qquad (2)$$

Notice what's sitting there in the numerator: the system's total momentum **P**. Expression (2) can therefore be rewritten as

$$\mathbf{P} = (m_1 + m_2 + \ldots + m_n)\mathbf{v}_{cm}. \qquad (3)$$

So the first property is: if the entire mass of the system is to be mentally concentrated at its center of mass, the momentum of this imaginary point will be equal to the total momentum of the system. What does this mean? We know, for instance, that the momentum of a closed system is conserved. This means that if the system is closed, the velocity of its center of mass $\mathbf{v}_{cm}$ remains constant.

Let's look at an example. A uniform thin rod of length l stands vertically on a smooth floor (fig. 2). Then it is released and falls down flat. How can one calculate the distance through which the bottom end of the rod will be displaced at the moment the rod hits the floor? The rod by itself does not form a closed system; but since it is subjected only to vertical forces, the horizontal component of its momentum does not change—in our case it remains zero. So the center of mass does not move in the horizontal direction—that is, the center of the rod will hit the floor exactly at the point where the bottom end was when the rod was released, and the bottom end will be displaced the distance $l/2$.

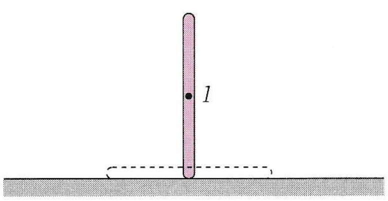


Figure 2

Let's look at other properties of the center of mass. Consider a system of two material points with masses m_1 and m_2. Let this system be open—that is, let's assume that each point of the system experiences the action of both internal and external forces. Starting from Newton's second law, let's write the change in the momentum of each point in time Δt:

$$m_1\Delta\mathbf{v}_1 = (\mathbf{F}_1 + \mathbf{F}_{12})\Delta t,$$
$$m_2\Delta\mathbf{v}_2 = (\mathbf{F}_2 + \mathbf{F}_{21})\Delta t,$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are external forces, $\mathbf{F}_{12}$ is the force exerted by the second point on the first, and $\mathbf{F}_{21}$ is the opposite.

According to Newton's third law, $\mathbf{F}_{12} = -\mathbf{F}_{21}$, so the change in the total momentum of the system equals

$$\Delta\mathbf{P} = m_1\Delta\mathbf{v}_1 + m_2\Delta\mathbf{v}_2 = (\mathbf{F}_1 + \mathbf{F}_2)\Delta t$$

—that is, only external forces can change the system's momentum. On the other hand, expression (3) implies

$$\Delta\mathbf{P} = (m_1 + m_2)\Delta\mathbf{v}_{cm}.$$

From this an expression for the center of mass that is analogous to Newton's second law can be obtained:

$$(m_1 + m_2)\Delta\mathbf{v}_{cm} = (\mathbf{F}_1 + \mathbf{F}_2)\Delta t,$$

which can be rewritten in the more familiar form

$$\mathbf{F}_1 + \mathbf{F}_2 = (m_1 + m_2)\mathbf{a}_{cm}. \qquad (4)$$

This is our most important result: the center of mass moves as if all the system's mass were concentrated at it and all the external forces were applied to it. Take note: not *all* forces, only *external* forces. The internal forces do not influence the motion of the center of mass. This is why in many cases the motion of the center of mass turns out to be so simple.

This property of the center of mass finds many applications. For example, now you probably understand why the center of the stick thrown at an angle to the horizon moves—just as a stone does—along a parabola. Neglecting the force of friction between the stick and the air, the only external force is that of gravity $m\mathbf{g}$, and so the acceleration of both the stick and the stone equals $\mathbf{g}$, regardless of the stick's rotation.

Let's look at an example of this. A construction crane is positioning a heavy slab on the construction site. To turn the slab properly, two workers push the slab at points A and B with equal (in magnitude) forces (fig. 3). What is the point about which the slab begins to rotate? We guarantee that many of you will immediately answer, "Point C, of course, which is halfway between points A and B." Not so fast! The correct answer is: point O—the slab's center of mass. Look at equation (4). Since the sum of the external forces equals zero, the acceleration of the center of mass should also be zero—it is this point that will remain at rest.

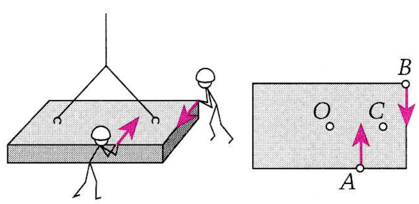


Figure 3

In conclusion, I'll offer one other "convenience" of the center of mass. As we saw from equation (3), the total momentum of a system of bodies in the reference system associated with its center of mass equals zero. The motion in such a reference system naturally looks simpler, since the system as a whole in this case is at rest. This method is especially convenient when the system is closed. In this case the acceleration of the center of mass equals zero (see equation (4)) and the reference system associated with it is inertial. For example, a head-on collision between two elastic balls in such a system looks so simple that we can immediately guess the answer: after the collision the balls will fly apart with the velocities they had before the collision. Can you figure out why?

—A. I. Chernutsan

How does electric current flow in a metal?

THIS QUESTION USUALLY PRESENTS NO DIFFIculty for students. How does it flow? It's simple. If you create a difference in potential between two ends of a conductor (for instance, a metallic one), an electric field arises in the conductor. This field acts upon free electrons in the metal and gives them an acceleration directed toward the end whose potential is greater (the charge on electrons is negative). So a motion of the charges arise, and this is what we call electric current.

We can't say that this answer is *wrong*. The words are all true enough. Yet this answer, which seems exhaustive at first glance, gives rise to a pack of other questions and objections. Let's try to straighten things out.

How do electrons in a conductor move when a difference in potential is created across its ends? It would seem that they are accelerated, since they are constantly under the influence of the force $\mathbf{F} = e\mathbf{E}$ ($\mathbf{E}$ is the strength of the electric field in the conductor). But on the other hand, if this is actually the case, the current through any cross section should increase with time, which contradicts Ohm's law: the current induced in a conductor by a constant difference of potentials is constant and equals $I = V/R$. Now what? Let's recall what we know of the internal structure of a metal.

The valence electrons of atoms in a metal are weakly coupled with the atoms they belong to. So when the crystal lattice is formed, they are easily torn away to form a rather dense electronic gas (even if each atom gives up only one electron, their concentration in such a gas would be of the order of $n \sim 10^{29}$ per cubic meter, which you can easily check). When we spoke above about a current flowing through metal, we considered the electrons to be free. In a certain sense this is true, but we mustn't forget the ion crystal lattice surrounding the electrons.

The classical electronic theory of resistance in metals, formulated in the late nineteenth and early twentieth centuries, says that electrons induced to move by an electric field undergo collisions with the ions of the crystal lattice. In some of these collisions, electrons give all of the kinetic energy acquired from the electric field to the lattice. It is these collisions—called *effective*—that are responsible for resistance in metals. The other collisions are not essential for understanding the flow of current through metals (they change only the direction of the electron's velocity but not its magnitude).

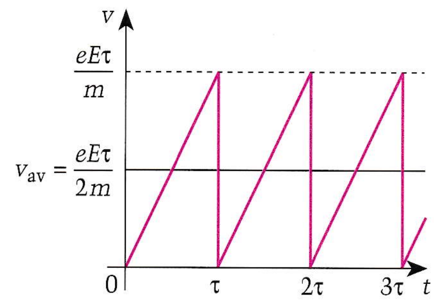


Figure 4

Let the average time between collisions be τ. Then we can imagine the following model of the motion of an electron in a metal in which an electric field is created. During the time interval between 0 and τ, the electron moves with acceleration $\mathbf{a} = e\mathbf{E}/m$, and so the projection of its velocity opposite to the electric field $\mathbf{E}$ increases linearly with time: $v = at = eEt/m$. At the moment τ, the electron collides with an ion and gives all its kinetic energy to the lattice. Then it is accelerated again by the electric field and the whole process is repeated. The plot of the time dependence of the velocity of this ordered motion is given in figure 4. Such motion is, in fact, equivalent to a uniform drift of electrons in the direction opposite to the field with velocity $v_{av} = eE\tau/(2m)$. Let's calculate the strength of the current associated with this motion.

The number of electrons passing through a cross section S in time Δt is $\Delta N = nSv_{av}\Delta t$. These electrons transfer charge $\Delta q = e\Delta N = neSv_{av}\Delta t$. So the current flowing through the conductor is equal to

$$I = \frac{\Delta q}{\Delta t} = nev_{av}S = \frac{ne^2\tau}{2m}SE.$$

The quantity

$$j = \frac{I}{S} = \frac{ne^2\tau}{2m}E$$

is called the current density.

The coefficient of the field strength $\mathbf{E}$, which is composed only of the microscopic characteristics of the metal, is simply the inverse of the metal's resistivity.

So some things seem clearer now. Some questions remain, however. For example, let's estimate the average drift velocity of the electrons. Let a current $I = 10$ A flow through a copper conductor with a cross section of 10 mm^2, and let the concentration of electrons $n = 1.67 \cdot 10^{29}$ per cubic meter. Then the average velocity is

$$v_{av} = \frac{I}{neS} \cong 0.04 \text{ mm/s}.$$

If we determine the time between effective collisions on the basis of the experimentally measured resistivity $\rho = 1.7 \cdot 10^{-8}\ \Omega \cdot \text{m}$, we get $\tau \sim 10^{-14}$ s. And if we assume that the mean free path between effective collisions is covered with an average velocity $v_{av} \sim 0.1$ mm/s, we arrive at the absurd conclusion that the distance between two successive collisions of an electron is $l = v_{av}\tau \sim 10^{-18}$ m, which is many orders of magnitude smaller than the distance between the ions in the lattice. So once again we have overlooked something. And what we failed to take into account is that electron-gas particles in a metal, like the particles of an ideal gas in a vessel, are in continuous chaotic motion. But even if we make use of this analogy and substitute the value of the mean thermal velocity $v_t = \sqrt{3kT/m}$ for v_{av} in the expression for l, it will still be insufficient to obtain a value consistent with experimental data (prove this on your own).

We have exhausted all the possibilities of classical physics. In fact, a consistent theory of the resistance of metals was constructed only in the middle of the twentieth century by means of ideas from quantum physics. It turns out that electrons in a metal move with huge velocities $v_e \sim 0.01c$ (where c is the speed of light in a vacuum). This chaotic motion of the particles of the electron-gas is purely of quantum rather than thermal origin—it does not stop even at absolute zero. Even with such enormous velocities of chaotic electron motion, though, the average charge transfer through any cross section is still equal to zero when no electric field is present. When an electric field is activated, the ordered drift of electrons in the direction opposite the field is superimposed on this chaotic motion as described above. It is the large velocity of chaotic motion that accounts for the distance between two successive collisions. This distance for the copper conductor we chose is several dozen (maybe even several hundred) interatomic distances, which seems reasonable.

And one last surprise. According to the laws of quantum mechanics, an electron in an ideal periodic crystal lattice moves such that it *never collides with the ions forming it*. Well—what do we do now with all our previous mental constructs? How *do* electrons transfer their energy to the lattice?

It turns out that at low temperatures the electrons collide with the atoms of impurities and other defects that are always present in a crystal lattice. If these defects are removed, the resistivity of a crystalline metal can be reduced to infinitesimally small values. At room temperature, electrons are basically scattered on *oscillations of the lattice*. In an ideal (immobile) lattice, electrons might somehow arrange their behavior so as to bypass all periodically situated ions; but if the ions are also involved in thermal oscillations, there is no way for the electrons to keep track of the chaotic movement of the ions, and so the electrons inevitably collide with ions.

Such are the reefs we encounter when we look more closely at questions that seem so simple.

—Andrey Varlamov

Time machines and the theory of relativity

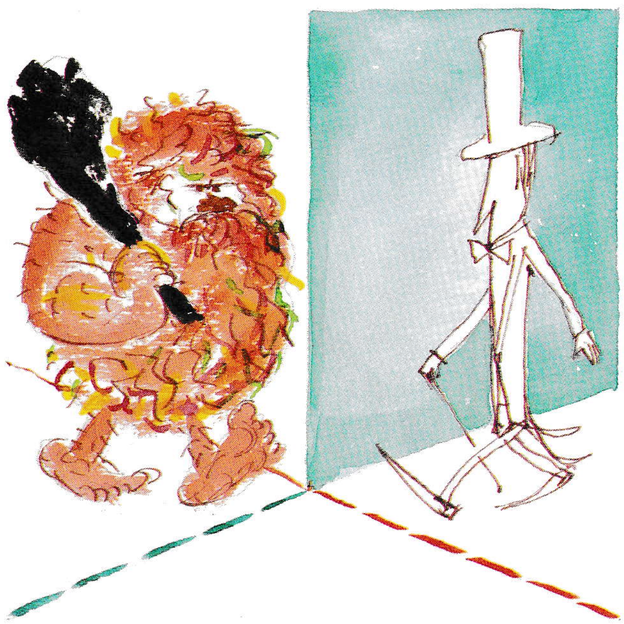
ONE OF THE MOST IMPORTANT STATEMENTS of the special theory of relativity says that the simultaneity of spatially separated events is relative—that is, it depends on the reference system of the observer. This assertion, as well as many others about the theory of relativity, seems strange, even paradoxical, and in any case it flies in the face of common sense. Why?

From the very beginning, science fiction's favorite device has been "time travel." Many authors use it to create a beautiful "time paradox"—a chain of events that absurdly loops back on itself. For example, the heroine in a story travels to the previous century and gets acquainted with a nice boy who, as she knows, will eventually make great discoveries. And, sure enough, it all happens, but only because the girl, who has a good memory, tells the boy all the details of each discovery, which she already knows from "future" textbooks and monographs. As you well understand, nobody discovered anything. Because of time travel, an effect (the heroine knowing the details of a discovery) changes places with its cause (the discovery itself), so that the effect precedes the cause—a paradox.

In order to avoid such paradoxes, common sense suggests, the past and the future must not change places.

"Hey, wait a minute!" you say. "In this case something is wrong with the theory of relativity. After all, the relativity of simultaneity implies that the sequence of events in time can be different for various observers."

Let's make this clearer with a concrete example. Consider a very long immobile platform A, with two photodetectors at its ends, and two trains B and C that move to the right and to the left, respectively. At some moment a light flashes exactly in the middle of the platform. We'll call it event No. 1. Then event No. 2 is the detection of the flash by the left detector, and event No. 3 is the detection by the right one. It's clear that in the reference system centered on the platform, events No. 2 and No. 3



are *simultaneous*. In the system associated with train B moving to the right, however, event No. 3 takes place *earlier* than event No. 2, since in this system the right detector moves toward the light ray and the left detector moves away from it. On the other hand, in the system associated with train C moving to the left, event No. 3 occurs *later* than event No. 2. The point is, of course, that for all three observers, the light moves at the same speed (according to the second postulate of the special theory of relativity).

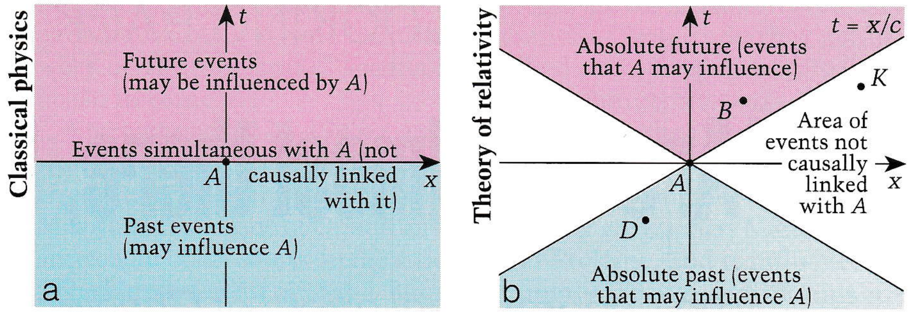


Figure 5
Diagram of the absolute (independent of reference system) classification of events. For simplicity only one spatial coordinate has been kept—that is, each event corresponds to a point on the "space (x)–time (t)" plane.

So according to the theory of relativity, the very notion of "earlier–later" is relative. An event that occurred "earlier" in one reference system may turn out to have occurred "later" in another.

"This all looks very strange," you say. "We've already seen that switching places in time is fraught with dire consequences."

Don't worry, it's not as bad as all that. Despite the fact that the special theory of relativity discards the absolute aspect of the notions "earlier–later" and "simultaneously," it never breaks the causal relationships between events. If one event results from another, then in any reference system the resulting event will always occur "later." Take note that in all three reference systems (A, B, and C) event No. 1 (the emission of light) takes place earlier than events No. 2 and No. 3 (the detection of this light). It occurs earlier from the standpoint of any observer—in no reference system can the light be first detected by a detector and then emitted by a source. The conclusion is clear: it would be wrong to assert that *any* two events can, by changing the reference system, switch places in time.

Well, when is it possible and when isn't it? Let's work from the opposite direction and establish when it is that a cause-and-effect relationship exists between events.

In classical (pre-Einstein) physics, this question had a simple answer: if one event takes place later than another, that event may be the consequence of the earlier one regardless of where the events occur. There was no reason to think that the transmission speed of information (signals) is limited. So even if event B occurred a great distance from event A, it is possible to send information about this event to point B at a speed sufficient for it to arrive before event B. It turns out that any disruption of the "earlier–later" sequencing in time could lead to a disruption of causal relationships. So it's natural that the concepts "earlier–later" and "simultaneously" in classical physics were absolute—that is, they could not depend on the reference system (see figure 5a).

In accordance with the theory of relativity, no signal can propagate at a speed greater than the speed of light (c). This statement implies that the condition of a possible causal link between events changes. If a light beam from point A arrives at point B before the event takes place, event A can influence event B. Precisely in such a case event B is considered causally linked with event A. Let's write down this condition as follows:

$$t_B - t_A \geq \frac{r_{AB}}{c}, \tag{I}$$

where r_{AB} is the distance between the points where events A and B occur. The condition

$$t_A - t_D \geq \frac{r_{AD}}{c} \tag{II}$$

implies that event A may be caused by event D—that is, event D may be causally linked with event A, but in reverse order. In the case

$$|t_A - t_K| < \frac{r_{AK}}{c}, \tag{III}$$

events A and K, even if they are not simultaneous, are completely independent of each other, since no information about the event can reach the location of the other one before it begins.

All these relations are illustrated schematically in figure 5b. For example, if event B is linked with event A by relationship (I), the same relationship binds these events in any other reference system—the condition of causal relationship cannot be changed. So the domain of events B is called the "absolute future" with respect to event A. Similarly, an event D that can affect event A occurs earlier than A for any observer. But event K, which is not linked with event A by any causal relationship, can be made simultaneous with event A (try to prove this on your own) or can even change places in time with it, and this will not lead to any paradoxes.

And so we've had a chance to convince ourselves that the cited statement of the special theory of relativity doesn't lead to any contradiction with common sense when we analyze it carefully. Of course, we had to modify common sense itself a little bit.

—A. I. Chernutsan

"Are we almost there, Captain?"

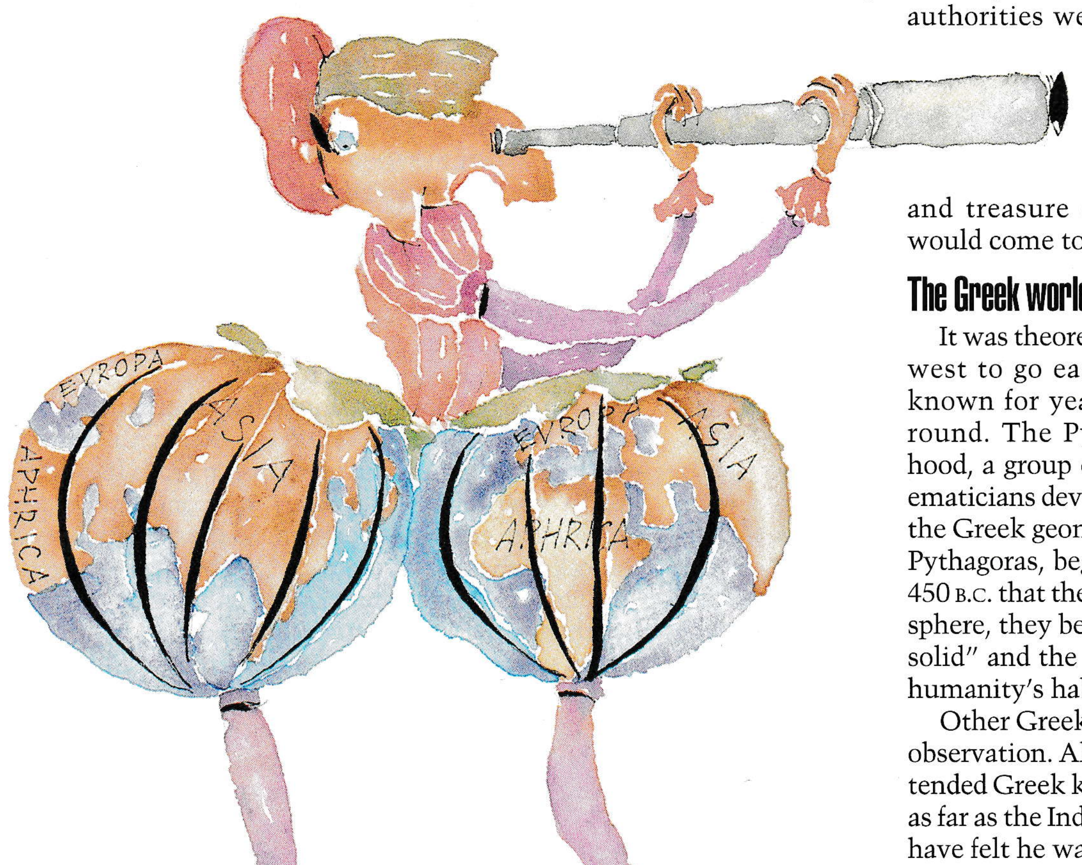
The geographical errors of Christopher Columbus

by Glenn M. Edwards

PEOPLE BELIEVE A LOT OF things about Columbus that just aren't true. For instance, that he proved the Earth is round and that people need have no fear of falling off the end of the world. Another is the legend that he convinced Queen Isabella of Spain to bankroll his project by balancing an egg on its point. Most people assume, at any rate, that Columbus knew where he was going in the fateful year of 1492. But did he?

Perhaps the greatest event of the fifteenth century before the Europeans discovered the Americas was the Turkish conquest of the Middle East. After the great city of Constantinople fell to them in 1453, the Turks were sitting squarely on the only east–west trade route known to the Europeans. The Portuguese were trying to break the Turkish monopoly on trade with the East as they slowly felt their way down the western coast of Africa, hoping to circumnavigate the vast continent and reach Asia.

It was a risky venture. No one in Portugal, or anywhere else in Europe, knew much about the world even close to home. Geographers in ancient Greece and Rome had written that there was no great southern ocean, but that southern Africa extended eastward and linked up with Asia. The Indian Ocean would then be a large inland sea. If these ancient authorities were right (and no one was in a position to say they were wrong), the route around Africa was a dead end. The time and treasure spent on this quest would come to nothing.



Art by Sergey Ivanov

The Greek world

It was theoretically possible to sail west to go east, since people had known for years that the world is round. The Pythagorean Brotherhood, a group of mystics and mathematicians devoted to the memory of the Greek geometer and philosopher Pythagoras, began to believe around 450 B.C. that the Earth is a sphere. The sphere, they believed, is the "perfect solid" and the only proper shape for humanity's habitat.

Other Greeks put more weight on observation. Alexander the Great extended Greek knowledge of the globe as far as the Indus River, and he must have felt he was very near the end of the inhabited world when he got there. Of the many cities he founded and named after himself, here he created Uttermost Alexandria.

Not that Alexander and his troops ever thought they were in danger of falling off the edge of the world. Alexander's teacher, Aristotle (probably the greatest scientist of the ancient world), knew that the Pythagoreans had postulated a globular Earth on philosophical grounds. But he had better reasons for agreeing with them: the shadow on the surface of the Moon during an eclipse is curved. Aristotle knew that only a sphere would always cast such a shadow. He also noted that new constellations become visible as a traveler moves south.

Aristotle had other reasons for believing in a round Earth. For instance, his theory of physics held that everything tends toward the center of the universe—which, for him, corresponded to the center of the Earth. And he was not alone among the ancients in his belief.[1] So, if the Earth was thought almost certainly to be a sphere, the question almost automatically arises: how big a sphere is it?

Aristotle had said that the distance from Spain to India was not large, though we don't know why he made that claim. (This idea influenced the medieval thinkers Roger Bacon and Pierre D'Ailly, who in turn influenced Columbus.) In his treatise *De Caelo* ("On the Heavens"), Aristotle wrote that certain mathematicians had calculated the circumference of the Earth as 400,000 stadia (one "stadium" was generally taken to equal 600 Greek feet, or about 188 meters).

Aristotle and Alexander bring us to Egypt, where the conqueror founded the most celebrated of his Alexandrias. Under the leadership of the last Egyptian dynasty, the Ptolemies, it became the greatest center of learning the ancient world ever knew. And it was here that work on the problems of geography began in earnest.

Eratosthenes (276–194 B.C.) was working there when he computed the circumference of the Earth. See the May/June 1992 issue of *Quantum* for an account of his efforts (in the article "The Universe Discovered"). But errors in his initial data were responsible for a large error in his final result. After arriving at a circumference of 250,000 stadia, he then rounded it up to 252,000. Why? Because it is evenly divided by 360—the number of degrees in a circle. His calculations gave him the result that each degree of the Earth's circumference is 700 stadia long. In fact, it is closer to 594.

Degrees of uncertainty

What does all this have to do with Columbus? Just this: it was now possible to determine one's latitude—that is, how far north of the equator one is. (Figure 1 illustrates how latitude is related to the equator and the elevation of the pole above the horizon.) The invention of an astronomical computer called the astrolabe (by Hipparchus in 140 B.C.) made it possible to determine latitude simply and with a fair degree of precision. This device remained in use until the seventeenth century and in the Islamic world for some two centuries more. In the northern hemisphere, where the Greeks lived, one could easily locate the north pole because a star, Polaris, happens to be very close to it.

If you want to know where you are on the globe, however, you need to know your longitude—that is, how far east or west of any given point you are. And if you want to do that, you have to know what time it is someplace else. The ancients theorized (correctly) that observers could see the same eclipse and note the difference in local time. For each hour of difference, the observers would be located 15 degrees apart (the 360 degrees in a circle divided by the 24 hours in a day). As a practical matter, though, this is almost impossible without mechanical clocks or instantaneous communication. So the ancients never made any measurements of longitude that could be called close, let alone precise.[2]

Posidonius measured the Earth less exactly than Eratosthenes did—he came up with a figure of 180,000 stadia. (This number was taken up by the later geographers Marinus of Tyre and Ptolemy, and it influenced Columbus greatly.) Posidonius measured the angle of a certain star above the horizon as seen in different cities, but as with Hipparchus, substantial errors in his initial data led to seriously flawed results (in hindsight). Nevertheless, Marinus of Tyre was sufficiently convinced by the work of Posidonius to use his figure for the arc distance between Alexandria and Rhodes—1/48 of the circumference,

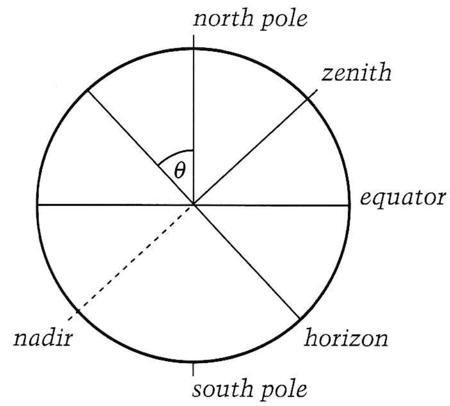


Figure 1
To all appearances, the Earth sits in the center of a huge sphere that rotates around us. Observers assume this to be a celestial sphere whose axis of rotation passes through the sphere at diametrically opposed points—the north and south poles that lie above the corresponding poles of the terrestrial globe. The celestial sphere, like our globe, is divided in half by an equator, which lies above the terrestrial equator. But for an observer there is a horizontal plane, called the horizon, that intersects the celestial sphere. There is also a vertical line that transfixes the observer and intersects the celestial sphere at a point directly overhead—the zenith—and directly under the observer's feet—the nadir. These are the poles of the horizon circle. (The angle θ is the terrestrial latitude in degrees.)

[1]Adrastias of Aphrodisias noted that celestial bodies rise sooner in the east, and that a lunar eclipse can be seen later in the evening by an observer in the east. If the Earth were flat, he pointed out, all observers would see the eclipse at the same time. Pliny the Elder, the Roman naturalist, observed that sailors can see the tops of mountains before they see the towns at their bases, and that sailors atop a mast can see these things before sailors on deck can.

[2]Hipparchus had noted that an eclipse seen at Carthage at one hour was seen on the Tigris River three hours later, local time. This would mean the two places are 45 degrees apart, whereas the true figure is closer to 30.

or about 7.5 degrees (as opposed to the more accurate 1/60, or 5.25 degrees).

Early attempts at mapmaking

This Marinus is an important figure in our story, though he was not as influential as his brilliant successor Ptolemy. Marinus made a point of interviewing visitors to Alexandria to learn what he could of the lands they had traveled in. He was interested in making more accurate maps, and he was the first after Hipparchus to divide the Earth into squares of longitude and latitude.

Marinus and Ptolemy understood that one problem with their map project was that the lines on it were straight—which could not, of course, correspond to the realities of a globe. Ptolemy devised a system in which latitude lines were parallel and longitude lines were curved to more accurately reflect the curvature of the Earth.

There was much else to improve in their map system. When Ptolemy gave the coordinates for a place, he based them not on observation but on travelers' reports of ground distances. (Remember, an accurate method of determining longitude wasn't devised until the eighteenth century.) And in general Ptolemy followed Posidonius's error of making each degree equal 500 stadia (rather than 594). So when he heard that two places were 500 stadia apart, he would make them one degree apart on his map. Ptolemy also overestimated the length of the inhabited world—180 degrees from the Canary Islands (his starting point, or prime meridian) to the great Chinese trading city of Sera. (Marinus had estimated the distance at 225 degrees; the correct figure is closer to 125 degrees.)

Columbus used as many sources as he could in estimating the Earth's circumference: Aristotle, Hipparchus, Eratosthenes, Ptolemy, and the Greek geographer Strabo, whose work had been recovered in the Renaissance and was very popular. He also learned from the Moslems of North Africa and the Middle East, who had learned, absorbed, and continued the work of the ancient

Be more clever than Chris!

by Yakov Perelman

"CHRISTOPHER COLUMBUS was a great man," the student wrote in an essay. "He discovered America and made an egg stand up." The student was equally amazed by both achievements. On the other hand, Mark Twain didn't see anything remarkable in the fact that Columbus discovered America: "It would have been surprising if he *hadn't* found it."

But I would venture to say that the second feat of the great mariner isn't such a big deal. Do you know how Columbus made the egg stand on end? He just pushed it against the table, breaking the shell on the bottom. So he obviously changed the shape of the egg. But how can one stand an egg up without changing its shape? The intrepid voyager didn't solve *that* problem.

And yet this is incomparably easier than discovering America, or even a tiny island, for that matter. I'll show you three ways of doing it: one for cooked eggs, one for raw eggs, and one for either kind.

To make a cooked egg stand up, all you have to do is spin it like a top: the egg will rotate upright and will stay in that position as long as it is spinning. After two or three tries, it's pretty easy to do the trick.

You can't use this method to make a raw egg stand up. A raw egg, as you've no doubt noticed, doesn't spin too well. In fact, this gives us a foolproof way of determining whether an egg is raw or cooked without breaking the shell. The liquid contents of a raw egg resist being drawn into the rapid spinning motion of the shell, and so they brake the shell, so to speak. We have to find another method of standing the egg up. And here it is. Shake the egg vigorously several times. The yolk will tear its delicate sack and will spread inside the egg. Then hold the egg upright, blunt end down, for a little while—the yolk, which is denser than the white, will drop to the bottom of the egg and collect there. Because of this, the egg's center of gravity is lowered and the egg is more stable than it was before this operation.

Finally, the third method of standing an egg on end. Put a cork back in a wine bottle, put the egg on the cork, and on top of that put another cork with two forks stuck in it from opposite sides. This entire "system" (as a physicist would say) is rather stable and preserves its equilibrium even if you carefully tip the bottle over. Why don't the egg and the cork fall? For the same reason a pencil can stand on its point if you stick a penknife in it just the right way. "The system's center of gravity is below its point of support"—that's how a physicist would explain it. This means that the point at which the force of gravity is applied is located below the place where the system is being held up. You can test this law of equilibrium with many objects, combining them in such a way that the heavy parts are lower than the point of support. The objects will stay put stably in the most startling positions.

And so you now have three advantages over Columbus in the art of making an egg stand up. As for discovering new continents, he has only one advantage over you: merely the fact that he discovered America.

From the book *For Young Physicists*, published in 1929.—*Ed.*

Greeks and Romans at a time when much of this culture was lost in Europe. For instance, the great caliph al-Mamun in the ninth century organized several attempts to determine the length of a degree, arriving at a result of $56\frac{2}{3}$ miles.

Columbus made use of later medieval sources as well. His own copies of two works, *Imago Mundi* ("Image of the World") by Pierre D'Ailly and *Historia Rerum Ubique Gestarum* ("History of Deeds") by Pope Pius II, were filled with marginal notes written by Columbus and his brothers. The *Historia* was a somewhat critical digest of Ptolemy's *Geography* and also incorporated information from Marco Polo's travels.

To sail or not to sail

Was it feasible to sail west to get to the east? Was Asia close enough? To find out, Columbus had to determine how far eastward Asia extended. Subtracting that distance from an estimate of the Earth's circumference, he would know how far he had to sail to reach Asia from the other direction. Let's see how Columbus dealt with a factor that played a crucial role in all his calculations.

Columbus took the length of a degree to be $56\frac{2}{3}$ miles (an error that goes back to al-Mamun). He noted in the margin of *Imago Mundi* that he had verified this measurement himself—a statement many writers have considered a lie because the value is wrong. The great German naturalist Alexander von Humboldt said that Columbus's measurements agreed with the Arab estimate because he "knew in advance what he wanted to find." But given the resources of his own time, Columbus could indeed have made his own measurements and come to the same false conclusion.

Columbus wrote that he had sailed south from Lisbon toward Guinea several times. He "noted with care the route followed, according to the customs of pilots and navigators, and took the elevation of the sun many times with a quadrant and other instruments, and I found agreement with Alfraganus [his source for al-Mamun's figure (al-Fargani in Arabic)—*Ed.*]. That is to say, each degree corresponds to $56\frac{2}{3}$ miles, wherefore credence should be given to this measurement."

Columbus had used the same method for determining the Earth's circumference as Eratosthenes and al-Mamun's geographers had. In each case they determined their location astronomically, traveled a certain distance along what they thought was the same longitude, and then determined their new location astronomically. It's a matter of easy arithmetic to compute the length of a degree. Al-Mamun's workers apparently took several measurements, getting a slightly different result each time—from 56 to $57\frac{1}{3}$ miles. The figure they decided to accept falls right in the middle of this range.

One advantage Columbus had over these earlier geographers, though, was that he had an interval of 40 degrees of arc to work with. Eratosthenes had only seven. Columbus made his measurements along the line from Lisbon to the islands of Los Idolos. Ptolemy had estimated that Lisbon was 40 degrees 15 minutes north latitude, and a famous map made by Martin Behaim in 1492 had placed it at slightly above 40 degrees. It is actually at 38 degrees 42 minutes (a minute is 1/60 of a degree). The Portuguese had estimated that Los Idolos was at 1 degree 5 minutes north, whereas it is actually at 9 degrees 30 minutes. So Columbus thought that the difference between these two points is 39 degrees 10 minutes, when it is closer to 29 degrees 12 minutes. This was the first of his errors.

The second came from the way he estimated distance, which probably came from the dead reckoning of Portuguese sailors. Columbus thought that the distance from Lisbon to Los Idolos is 2,192 miles. And since the

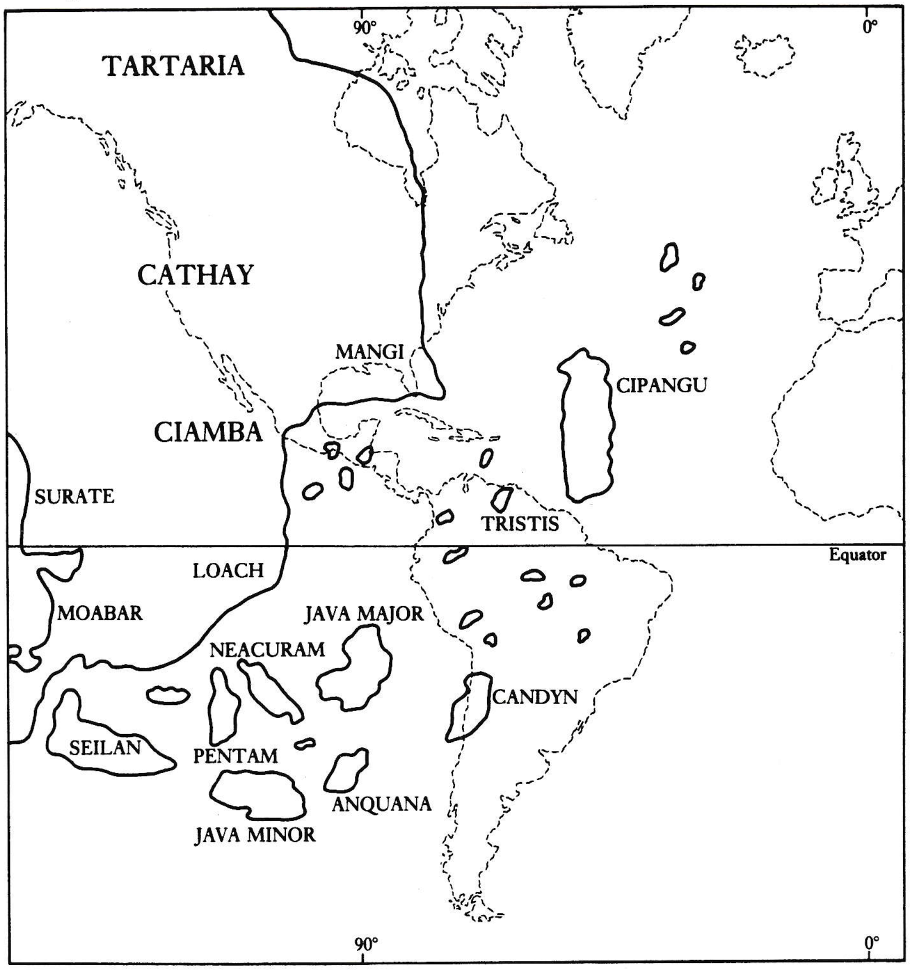


Figure 2

Martin Behaim's conception of the Atlantic Ocean. From Columbus *by Felipe Fernandez-Armesto, published by Oxford University Press, 1991. Reproduced by permission of Oxford University Press.*

angular distance between them was erroneously believed to be 39 degrees 10 minutes, he got a result of about 56 miles to a degree. This was close enough to al-Mamun's figure to effectively confirm it. So the circumference of the globe was determined to be about 20,400 miles ($56\frac{2}{3}$ miles/degree × 360 degrees) rather than 25,000.

Columbus based his estimate of the length of Asia on the reports of Marco Polo and others, and the question hangs on the distance one assigns to a degree. The Bartholomew Columbus map of 1503 notes that Christopher Columbus and Marinus of Tyre both gave the distance from Cape St. Vincent to Cattigara as 15 hours, or 225 degrees. Ptolemy had said 12 hours, or 180 degrees, but Columbus (using his own calculation for the length of a degree) had decided that Marinus was correct. Martin Behaim had followed Ptolemy in putting Cattigara on the 180th meridian, but he also estimated the further distance to Mangi as 60 degrees; thus, the known world extended some 240 degrees from west to east. At $66\frac{2}{3}$ miles to the degree (the figure Behaim used), the breadth of the known world was 16,000 miles at the equator. Columbus took this distance, but divided it by his own measurement of $56\frac{2}{3}$ miles to the degree, which gave him 283 degrees as the distance from the west coast of Spain eastward to the coast of China.

To this figure Columbus added Marco Polo's inflated estimate of the distance from the Chinese coast to Cipangu (as Japan was then called by Europeans) and determined that he would have to sail only some 2,500 miles before he reached Japan. This was well within the technical capabilities of fifteenth-century sailors. (These distances are reflected in figure 2, which gives a portion of Martin Behaim's map.)

But Columbus—a man of the Middle Ages—was also something of a mystic. He noted in the margin of *Imago Mundi*, and repeated in a letter in 1503, that the world is six parts dry land to only one part water (citing the fourth chapter of the apocryphal Book of Esdras). For religious reasons, then, he expected the watery distance from Spain to Asia to be only one seventh the circumference of the Earth, which he reckoned at 17,500 miles at the latitude he would be sailing at. This would put Asia about 2,500 miles west of Spain. So, for Columbus, science and religion neatly supported each other and made his venture plausible.

The scholar G. E. Nunn wrote that Columbus was "painstaking in his inquiries" and used the best information available to him. He also happens to have been wrong: Japan is actually about 10,600 miles west of Spain. But fate intervened, placing the New World between him and his goal. The rest, as they say, is history.

Glenn M. Edwards *wrote his doctoral dissertation on a medieval astronomical treatise and has taught at the university level ever since receiving his Ph.D. in history from the University of Southern California. Dr. Edwards also works as a supervising production editor for Springer-Verlag New York.*

ANSWERS, HINTS & SOLUTIONS

Math

M61

Let a be the first term and d the difference of the arithmetic sequence. Evidently $d \geq 0$. Then all the terms of the form $a + 10^n d$ with large enough n (such that $10^n > a$) have the same sum of digits equal to the sum of all digits of a and d.

M62

Since $1992!! = (1993 - 1) \cdot (1993 - 3) \cdot (1993 - 5) \cdot \ldots \cdot (1993 - 1991)$, this number has the same remainder when divided by 1993 as the product $(-1)(-3)(-5)\ldots(-1991) = (-1)^{1992/2} \cdot 1991!! = 1991!!$. Therefore, $1992!! - 1991!!$ has a remainder of zero when divided by 1993.

The number 1992 here can be replaced with any multiple of 4. For all the other even numbers n (of the form $4k + 2$) we must change the sign: $(n - 1)!! + n!!$ is divisible by $n + 1$.

M63

If you tried to solve this problem by way of calculations, you'll certainly appreciate the solution below. It's based on the transformation called *inversion* (see page 40 in this issue) and it is perhaps one of the most beautiful applications of this transformation.

Let's invert the diagram to the problem (fig. 1) in some circle with center C (we'll choose its radius later on). The inverses of semicircles AC and BC are the rays starting at points A' and B' on line AB (inverses of A and B) and perpendicular to this line, and the inverse of semicircle AB is the semicircle with diameter $A'B'$ (fig. 2). The circle in question, ω, inverts into a circle ω′ touching the two rays and the semicircle $A'B'$. It's clear that the diameter of ω′ is equal to $A'B'$, and so it's equal to the distance from the center of ω′ to $A'B'$. But we can choose the radius of the circle of inversion so that ω inverts into itself! (It suffices to make it equal to the length of the tangent from C to ω—see the article mentioned above). Then ω′ coincides with ω, and so the diameter of ω equals l. (V. Dubrovsky)

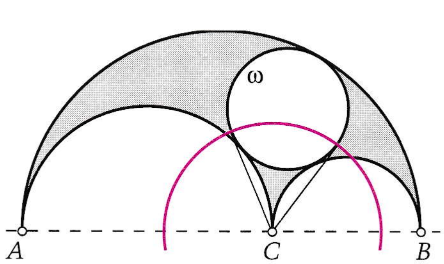


Figure 1

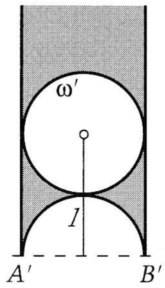


Figure 2

M64

The answer is $b < a < c$. To prove it, recall that $\sin x < x$, and $\cos x$ monotonically decreases for all x in the interval $(0, \pi/2)$. Plugging $\cos x$ instead of x into the above inequality, and taking the cosine of both its parts, we get

$$\sin \cos x < \cos x,$$
$$\cos \sin x > \cos x,$$

respectively. It follows that

$$b = \sin \cos b < \cos b,$$
$$c = \cos \sin c > \cos c,$$

or

$$\cos b - b > 0 = \cos a - a > \cos c - c.$$

But the function $y = \cos x - x$ decreases on the interval $(0, \pi/2)$, so $b < a < c$.

M65

(a) The required arrangement of circles is shown in figure 3. But a figure is not a proof, and we have to provide rigorous arguments confirming the existence of such an arrangement—that is, to compute the radii and distances between the centers of the circles; or, at least, to prove that the equations defining them are solvable. This is far from easy to do with the traditional methods of school geometry. The solution given below is based on ideas presented in the articles "Off into Space" (January/February 1992) and "Inversion" (this issue).

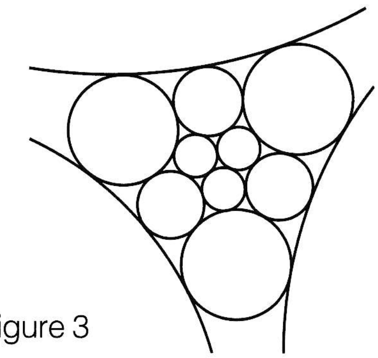


Figure 3

Take a regular dodecahedron with circles inscribed in each of its faces (fig. 4). This set of 12 circles evidently solves the problem except that they don't lie in the plane. They do, however, lie on the sphere touching all the edges of the dodecahedron, and a sphere can be converted into a plane (so that circles remain circles) by way of stereographic projection (see page 45). So the desired configuration can be obtained from our set of the 12

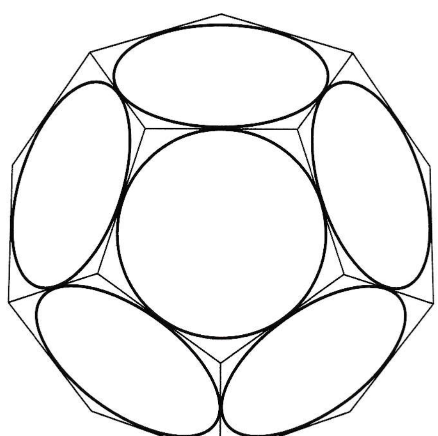


Figure 4

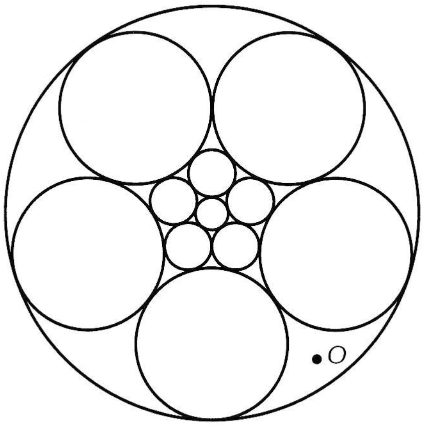


Figure 5

circles as its stereographic projection from any point on the sphere that does not lie inside any of the circles. In particular, when the center of projection lies on the line through the sphere's center and any vertex of the dodecahedron, the projection yields figure 3. If we project from the point lying on the line joining the centers of the sphere and one of the circles, we get figure 4. Incidentally, the existence of this configuration is easy to prove strictly without "going off into space." But it doesn't immediately solve the problem because the biggest circle contains all the rest. But inverting the entire picture with respect to the appropriate center (for instance, point O in figure 5) will produce what we want.

(b) To begin with, suppose the required arrangement exists and not all of the circles are congruent to one another. Then we can find a circle of the smallest radius r that touches at least one circle of another radius r_1 ($r_1 > r$). Let O and O_1 be their centers,

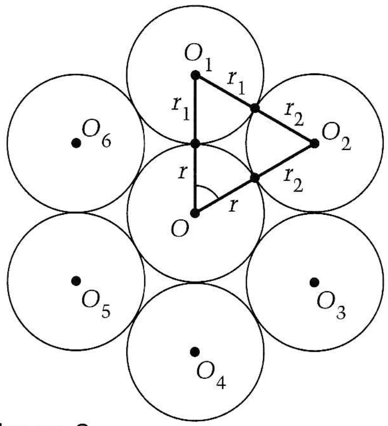


Figure 6

and let $O_2, \ldots, O_6$ be the centers of all the other circles touching circle O in clockwise order (fig. 6). Then $O_1O_2 = r_1 + r_2$, $O_2O = r_2 + r$, $OO_1 = r + r_1$, where r_2 is the radius of circle O_2. Since $r_2 > r$, O_1O_2 is the longest side in triangle OO_1O_2 and the opposite angle O_1OO_2 is its largest angle. It follows that $\angle O_1OO_2 > 60°$. Similarly, angles O_2OO_3, O_3OO_4, …, O_6OO_1 are not less than 60°. Then the sum of all six angles is greater than 360°, which is impossible.

So all the circles must be of the same radius. But in this case each of them must be surrounded by the other six, like circle O in figure 6, and this is possible only with an infinite number of circles; otherwise, we'd be able to find a "border" circle that is not completely surrounded.

Thus, assuming the existence of a solution, we always arrive at a contradiction. (V. Dubrovsky, D. Fomin)

Another approach to this problem comes from graph theory and uses Euler's well-known formula $V + F = E + 2$, where V, E, and F are the number of vertices, edges, and faces (including the infinite face) of a finite graph.

Suppose that a finite configuration of n circles exists, each of which touches exactly six of the others externally. If we connect the center of each circle to the center of each of its six neighbors, we get a graph. We now use a standard technique in problems of this kind: that of counting the edges of the graph in two different ways. On the one hand, each vertex is connected to exactly six other vertices. This observation would give $6V$ as the number of edges, except that it counts each edge twice: once for each endpoint. Thus $E = 3V$, or $V = E/3$. Now we count the edges again, using the faces of the graph. Since each face has at least three sides, E would be at least $3F$—except that we have again counted each edge twice. So E is at least $3F/2$, or F is at most $2E/3$.

So we have found that $V + F \le E/3 + 2E/3 = E$, which contradicts Euler's formula. We can have no such configuration. (Our thanks to Noam Elkies of Harvard University for contributing this alternate solution.)

Physics

P61

At any moment while it is moving the cord is strained uniformly; thus, the ratio of the distances between knot C and the ends of the cord will not change over time. Figure 7 shows that initially this ratio is

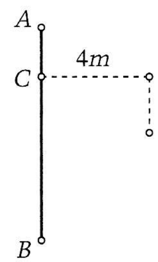


Figure 7

$$|AC| : |CD| = 1 : 4.$$

It's evident that the displacement Δx of the knot eastward is determined by the displacement ΔS_x of runner A, and at any moment it equals $\frac{4}{5}$ of this displacement—that is,

$$\Delta x = \frac{4}{5}\Delta S_x = \frac{4}{5}v_0 t.$$

Using the scale given in the figure, we see that point D has moved from point C a distance $\Delta x = 4$ m eastward. Thus, the knot passed point D at time

$$t = \frac{5}{4}\frac{\Delta x}{v_0} = 5 \text{ s}$$

after the runners started moving.

The displacement Δy of the knot southward is determined by the displacement ΔS_y of runner B, and at any moment $\Delta y = \frac{1}{5}\Delta S_y$. Using the scale, we see that during the time $t = 5$ s the knot moved a distance $\Delta y = 2$ m southward from the initial position (point C). Thus, moving with acceleration a during $t = 5$ s, runner B covered the distance $\Delta S_y = 5\Delta y = 10$ m—that is,

$$\frac{1}{2}at^2 = \Delta S_y,$$

and so

$$a = 2\frac{\Delta S_y}{t^2} = \frac{20}{25} \text{ m/s}^2 = 0.8 \text{ m/s}^2.$$

P62

When the top falls off the table, it moves along a parabola (its rotation merely stabilized the vertical position of the top's axis) (fig. 8). The horizontal projection v_x of the top's velocity will be equal to the velocity v of its movement along the table; the vertical position of the top as a function of time is given by $y = gt^2/2$. The top will not hit the edge of the table if during the time τ, during which it falls a distance H, its horizontal displacement $v_x\tau$ is greater than or equal to r. (You should convince yourself that this is correct.) We'll write this condition as follows:

$$v_x\tau = v\tau \geq r, \quad \tau = \sqrt{\frac{2H}{g}}.$$

Thus, the velocity of the top's horizontal movement should satisfy the following condition:

$$v \geq \sqrt{\frac{r^2 g}{2H}}.$$

P63

The number of molecules that remained on the walls is approximately equal to $N \sim 4\pi R^2/d^2$, where $R \cong 0.06$ m is the radius of the retort and $d \sim 10^{-10}$ m is the diameter of a gas molecule. After the gas is removed from the walls, the concentration of molecules in the retort will be

$$n = \frac{N}{V} \sim \frac{3}{d^2 R},$$

and the unknown pressure will be

$$p = nkT \sim \frac{3kT}{d^2 R} \cong 40 \ \text{N/m}^2.$$

Here $k = 1.38 \cdot 10^{-23}$ J/K is Boltzmann's constant and $T \cong 600$ K.

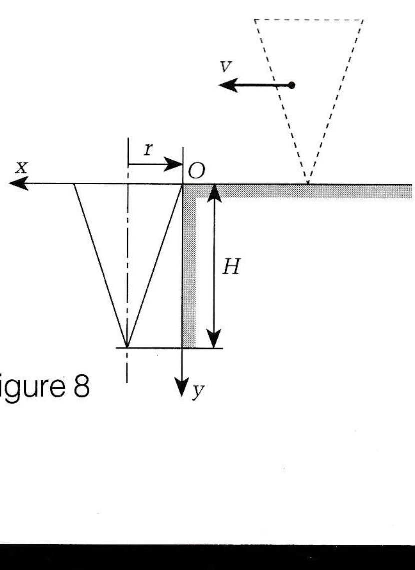


Figure 8

Let's compare this pressure with the normal atmospheric pressure $p_0 \cong 10^5$ N/m^2:

$$\frac{p}{p_0} \cong \frac{40}{10^5} \sim 10^{-4}.$$

This simple example shows how much a "vacuum" deteriorates in a sealed vessel if the gas is not removed from its walls beforehand.

P64

To determine the average current charging the battery, it's necessary to determine the charge sent to the battery during one cycle.

When the switch is closed, the coil is directly connected to the power source, and the emf across the inductor is equal to V. This means that the current through the coil increases linearly with time. Because there is no current in the coil at the moment the switch is closed, we get

$$I = \frac{V}{L}t.$$

By the time the switch is opened, the current is

$$I_0 = \frac{V}{L}\tau_1.$$

After the switch is opened, the diode allows current to flow to the battery. The emf across the inductor is equal to the difference between the voltage across the power source and that of the battery: $V - \mathcal{E}$. Since the emf is now in the opposite direction, current in the coil decreases linearly with time according to

$$I = I_0 - \frac{\mathcal{E} - V}{L}t.$$

Because the rate of decrease of the current (when the switch is open) is greater than the rate of increase (when the switch is closed), the current will drop to zero before the switch is closed and the diode switches off. Since the battery is charging as long as $I > 0$, we can find the time of charging τ_3 from the condition

$$\frac{\mathcal{E} - V}{L}\tau_3 = I_0 = \frac{V}{L}\tau_1,$$

from which it follows that

$$\tau_3 = \frac{V}{\mathcal{E} - V}\tau_1.$$

The charge passing through the battery during this time is

$$\Delta q = I_{av}\tau_3 = \frac{1}{2}I_0\frac{V}{\mathcal{E} - V}\tau_1 = \frac{1}{2}\frac{V^2\tau_1^2}{L(\mathcal{E} - V)}.$$

So the average current charging the battery is

$$I_{av} = \frac{\Delta q}{\tau_1 + \tau_2} = \frac{V^2\tau_1^2}{2L(\mathcal{E} - V)(\tau_1 + \tau_2)} \cong 8.9 \text{ mA}.$$

P65

This problem has a long history—the phenomenon evaded a correct explanation for many, many years. First off, it's worth noting that the velocity of even a strong wind (say, 20 m/s) is much less than the speed of sound (330 m/s). The fact that sound travels a bit faster with the wind than against it is of no importance.

The correct answer was obtained by the English physicist G. G. Stokes in 1857. The gist of the matter is that the wind speed changes with altitude—it is slower near the surface of the Earth and increases with distance above the ground. (This is due to the friction between the layers of air and the ground, as well as the friction within the air itself.) Let's see how this influences the propagation of a vertical wave front of the sound moving against the wind (see figure 9).

The upper part of the wave, being further from the Earth's surface,

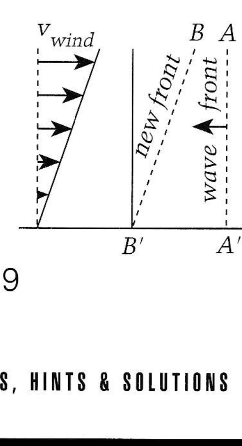


Figure 9

propagates more slowly than the lower part. As a result, the upper part of the wave front is tilted backward.

Because the direction of the sound wave is always perpendicular to its wave front, the sound waves that would normally propagate along the Earth's surface will be deflected upward when the wind is blowing (fig. 9).

Stokes's explanation was confirmed convincingly by the experiments of another famous English physicist, Osborne Reynolds. As the sound source Reynolds used an electric bell, which could be raised or lowered. He discovered that the bell could be heard against the wind at a much greater distance if it was raised high above the ground.

Another scientist who extensively studied the influence of the wind on the propagation of sound was John Tyndall, a brilliant lecturer and popularizer of science. In his experiments he also used a bell (though not an electric one). When he was some distance upwind of the bell, he was able to hear the bell only when he climbed a ladder. Tyndall's experiments clearly demonstrated that sound waves really are diverted upward when they travel into the wind.

Brainteasers

B61

Mlle Dubois has one dog, one cat, one parrot, and no cockroaches.

B62

If $ABCD$ is one of the given quadrilaterals (see figure 10), then translating triangle ABD by vector AC into triangle CBD, we get the required parallelogram BB_1D_1D: its parts BB_1C and DD_1C are congruent to the pieces $C'A'B'$ and $C'A'D'$, respectively, of the second quadrilateral $A'B'C'D'$.

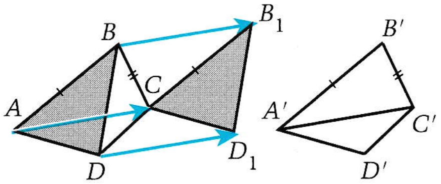


Figure 10

B63

The stacked blocks will be submerged 3 cm in both cases.

B64

Consider an inscribed regular pentagon (fig. 11). At least three of its vertices are the same color (because there are only two colors). But any three of its vertices form an isosceles triangle.

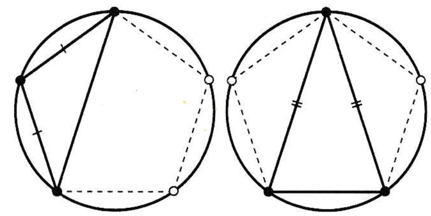


Figure 11

B65

Since we will be adding numbers bigger than 1 and less than 16, the only perfect squares we can get are 4, 9, 16, and 25. If a, b, and c are three numbers in our row, then both $b + a$ and $b + c$ must be perfect squares, and a cannot equal c. A quick check shows that b cannot equal 8 or 9: for every other choice of b, there are two numbers in the required range that can be added to b to get a perfect square.

Therefore, the numbers 8 and 9 must be at either end of the row. This determines the order of the integers:

8 1 15 10 6 3 13 12 4 5 11 14 2 7 9

The only other way to get the required result is to read this solution backwards.

Kaleidoscope

1. See figure 12.

2. The water in the gap between the pieces of glass that are stuck together moistens the glass, and the free (lateral) surface of the water along the edges of the glass is concave. Surface tension prevents one from pulling the pieces apart. If you immerse the pieces of glass in water, the concave lateral surface of the water layer will disappear, and with it the constraining forces of surface tension.

3. Place the ball in some water. If the air bubble is shifted relative to the center, the ball will always turn over so that the cavity is in the highest possible position.

4. The speed of the escaping air is higher when the pressure in the air mattress is greater. But the pressure is lower in the second case.

5. Moisten the ball and roll it along the floor. Measure the length l of the track after one revolution. The ball's diameter $d = l/\pi$.

6. A sugar solution has a greater surface tension than pure water, and the surface tension pulls the matches toward one another. With a soap solution, the surface tension decreases.

7. The brightness of the fabric depends on the luminosity—that is, on the angle of incidence. This angle varies for different parts of the fluttering flag, and so stripes appear.

8. Holding the coffee grinder in your hand, you feel a little kick when you turn it on that tends to turn the coffee grinder in the direction opposite to the direction of rotation of the rotor.

9. The surface of the puddle reflects light like a mirror, which is why the light from the headlights is directed almost completely away from the driver. Asphalt, on the other hand, diffuses the light, and a portion of that light reaches the driver's eyes.

10. With alternating current, a magnet brought up to the bulb causes the filament to oscillate and its outline will become indistinct. With direct current, the filament

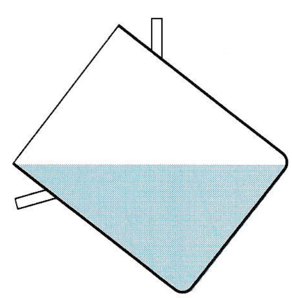


Figure 12

will be seen distinctly because it is displaced only slightly from its original position.

11. While the thermos is uncorked, the freed-up portion of its volume fills with cold air; when the thermos is recorked, the air inside heats up, its pressure increases, and this internal pressure pushes the cork out.

Bop

1. *Deed, sass, mama,* and *eefe,* an obsolete synonym for easy, which I never said these problems were.

2. *Deeded* and *muumuu* are good, but *deedeed* (the euphemistic *d– – – –d* sometimes standing for damned) and the lovely town of *Orrororo* in South Australia are even better.

3. The longest I know of is *assesses.*

4. Sixteen everyday words, as promised: *alfalfa, Iranian, hashish, singing, couscous, derider, decided, repaper, tableable, angling, onions, nonunion, blendable, coolly, piazza, assassins.*

Beriberi or *chercher* are the equals of *couscous;* the *alfalfa* alternative to *entente* was suggested to me by Isaac Asimov.

5. Think of a word with $m + 1$ letters, m of them distinct. Its form is specified by i and j, the first and second positions of the repeated letter. The number of forms is the number of pairs satisfying $1 \le i < j \le m + 1$.

6. Let $A(p)$ [$B(p)$] be the number of pforms using one or two letters that do [do not] begin with a double letter. Evidently $A(1) = 0$ and $A(2) = B(1) = B(2) = 1$. Proceed by induction. Adding one letter to the beginning of a pword, we find $A(p + 1) = B(p)$ and $B(p + 1) = A(p) + B(p)$. These results yield

$$A(p+2) + B(p+2) = [A(p+1) + B(p+1)] + [A(p) + B(p)].$$

For $p > 2$, there are no pwords using only one letter and $A(p) + B(p) = N(p, 2)$.

7. *L(assessed)* = 8, but *Mississippi* beats it by one.

BULLETIN BOARD

Imagine standing in the vortex of a 70-foot-high tornado or crawling through a pitch-black, 100-foot tunnel with only your sense of touch as a guide. These are only two of the attractions at the new **Liberty Science Center** just outside New York City. The four-story structure encompasses three exhibit areas featuring an expanding geodesic sphere, a 170-foot observation tower, and the world's largest Omnimax theater. The center also offers outreach programs to schools nationwide (for example, students can become "pen pals" with corporate scientists). For information, write to Liberty Science Center, Liberty State Park, 251 Phillip St., Jersey City, NJ 07304-4629, or call 201 451-0006. . . . For hands-on, high-energy exhibits in northern Illinois, check out the new **Commonwealth Edison Power House**, adjacent to the Zion Nuclear Generating Station on the shore of Lake Michigan. The 30,000-square-foot education center boasts a working geyser demonstrating the power of thermal energy, exhibits on how humans have harnessed energy through the ages, and much more. The Power House also features a resource center, conference rooms, and an auditorium for presentations to school groups. For information, write to Steve Solomon, Commonwealth Edison, Northern Division, 1000 Skokie Blvd., Northbrook, IL 60062-4103, or call 708 746-7080. . . . There's a fascinating story underfoot, and geologist James Sadd aims to spill the beans in "**Earth Revealed**," a new public television series and college course. Professor Sadd will examine how Earth's dynamic interacting systems relate in the development of life and how they help us unlock the secrets of our environment. The 26 half-hour programs will cover such topics as plate tectonics, the rock cycle, the forces that shape the Earth's features, the pace of geologic change, and the importance of preserving Earth's natural resources. Check your newspaper for information on local broadcast. For information on purchasing the series on videocassette, or taking the course for credit, call 1 800 LEARNER.

Index of Advertisers

The mystery of figure No. 51

What would Pythagoras have thought?

by Alexey Panov

I'M SURE YOU RECOGNIZE the pattern in figure 1. A square cut into seven pieces as shown in this figure becomes a popular puzzle, the *tangram*, thought to have been invented in China about 4,000 years ago. (In fact, according to recent research, it originated in Europe and is only a few centuries old.) The tangram usually comes with a set of silhouettes, the most interesting of which are of people or animals, that the player must construct from all seven pieces. One of my favorite shapes, "woman with a fan," is seen in figure 2. Owing to its universal popularity, the tangram has generated many spinoffs, and one of them—under the not very apt name of *pythagoras*—has become part of my collection of puzzles.[1]

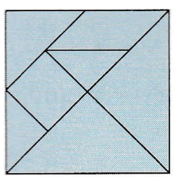


Figure 1

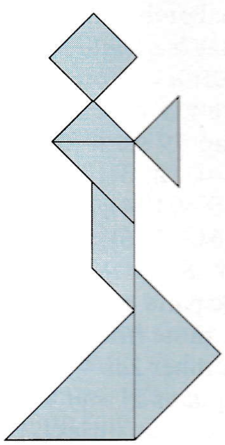


Figure 2

[1]This puzzle was manufactured in Russia, and I don't know if it's available in the US, but in any case it's very easy to make yourself.

Pythagoras (the game) is also a square cut into seven pieces, but in a different way (fig. 3). In the list of shapes to be assembled from the seven pythagoras pieces, one item—No. 51 (fig. 4)—proved to be a real challenge. I spent a lot of time trying to put it together, but the best I could manage was the shape in figure 5. In the end I began to doubt whether it could be done at all.

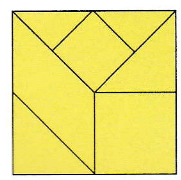


Figure 3

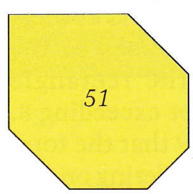


Figure 4

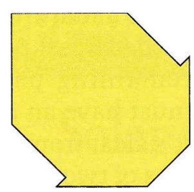


Figure 5

Eventually I found help in one of Martin Gardner's books, where I read about two Chinese mathematicians who in 1942 published a list of all *convex polygons* that can be constructed from the seven tangram pieces. Their idea was to subdivide all the pieces into isosceles right triangles congruent with the two smallest triangles of the tangram (fig. 6a). By a subtle argument the authors showed that from the 16 triangles thus obtained one can make 20 different (noncongruent) convex polygons. Then it's not difficult to examine these polygons one by one to find that only 13 of them can be tiled by the tangram pieces.

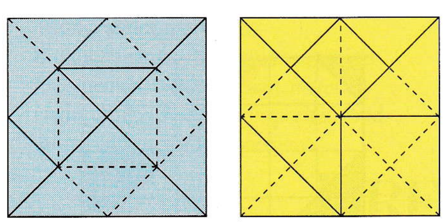
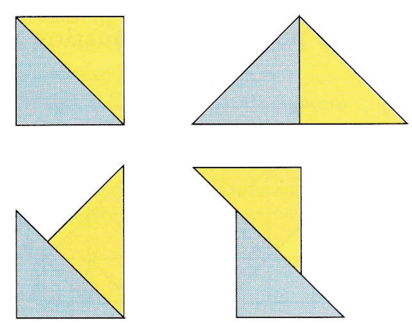

a b

Figure 6

The pieces of pythagoras can be divided into the same 16 triangles (fig. 6b). Figure No. 51 is a convex polygon, so if it doesn't occur among the 20 polygons mentioned above, it will be all the more impossible to create it out of the pythagoras pieces.

Let's make a complete list of these 20 polygons.

An experiment will help us. Prepare a large enough number of congruent 45°–45° right triangles and try to make polygons out of them. You'll


a

b

Figure 7

discover that convex polygons appear only when any two neighboring triangles have a whole common side (as in figure 7a). If a leg of one triangle borders on the hypotenuse of the other, or if the adjacent sides of the two triangles are congruent but do not exactly fit each other (fig. 7b), a convex polygon won't emerge. This "tiling rule" is a clue to a description of the polygons in question. But it's rather difficult to prove, and we'll accept it just as an experimental fact.

Now lay an arbitrary convex polygon tiled by our triangles over a grid of squares whose sides are equal in length to the legs of the triangles so that the vertices of one of the triangles match some three nodes of the grid (fig. 8). Then by the tiling rule the vertices of all the other triangles will also fall on nodes of the grid. Since all the angles of the polygon are composed of the angles of 45°–45° right triangles, they can measure only 45°, 90°, or 135° (because the angles of a convex polygon are always less than 180°). Therefore, the polygon can be represented as a rectangle whose sides lie on the grid lines and whose corners are truncated, the cuts making a 45° angle with the grid lines (fig. 8). If the neighboring lines of the grid are assumed to be one unit apart, the side lengths of the rectangle will be some integers a and b. To completely define the polygon one has to specify four more nonnegative integers x, y, z, and w—the lengths of the legs of the truncated corners. The area A of the polygon is then easily expressed in terms of these six numbers:

$$S = ab - \tfrac{1}{2}(x^2 + y^2 + z^2 + w^2).$$

For the polygons in question the area equals the total area of the 16 isosceles right triangles with a unit leg: $A = \tfrac{1}{2} \cdot 16 = 8$. So our problem underwent a surprising transformation from a purely geometric problem of tiling a polygon to an arithmetic problem of finding nonnegative integer roots of the equation

$$2ab - 16 = x^2 + y^2 + z^2 + w^2. \quad (1)$$

There is, of course, one more condition to be met: each side of the rectangle must be not less than the sum of the two segments (legs of truncated corners) cut off from it. This condition yields four additional inequalities:

$$\begin{aligned} x + y &\le a, \\ y + z &\le b, \\ z + w &\le a, \\ w + x &\le b. \end{aligned} \quad (2)$$

From here on, geometry and arithmetic will walk hand in hand.

First let's figure out the possible dimensions of our rectangle. On the one hand, its area must be not less than that of the polygon in question:

$$ab \ge 8. \quad (3)$$

(This inequality can also be directly deduced from equation (1).) On the other hand, after cutting off corners of the largest possible total area, the remaining part of the rectangle must have an area not exceeding 8. It's clear from figure 9 that the total area of two corners bordering on one side of the rectangle is not greater than the area of the 45°–45° right triangle whose legs are congruent with this side. So to obtain the smallest (in area) nonempty remainder, one must truncate the rectangle as in figure 10a for $a < b$, and as in figure 10b for $a = b$. In the first case, the remainder is a parallelogram of area $a(b - a)$; in the second case, it is a trapezoid of area $a - \tfrac{1}{2}$. So the side lengths a and b of the rectangle must satisfy the inequalities

$$a(b - a) \le 8 \quad (4)$$

for $a < b$, and

$$a - 1/2 \le 8 \quad (5)$$

for $a = b$. (Try to derive (4) and (5) from (1) and (2).)

All the pairs of positive integers (a, b) satisfying inequalities (3) and (4) or (3) and (5) can be found easily by a simple search. For instance, for $a = 1$, we'll find that $b \ge 8$ and $b - 1 \le 8$, so $b = 8$ or 9; for $a = 2$, the possible values of b are 4, 5, or 6; and so on. In all, there are 19 such pairs: (1, 8), (1, 9), (2, 4), (2, 5), (2, 6), (3, 3), (3, 4), (3, 5), (4, 4), (4, 5), (4, 6), (5, 5), (5, 6), (6, 6), (6, 7), (7, 7), (7, 8), (8, 8), (8, 9).

For each of these pairs, according to equation (1), we must find all the instances of the number $2ab - 16$ as the sum of four squares $x^2 + y^2 + z^2 + w^2$ and then check inequalities (2), or draw a rectangle measuring $a \times b$ and find all truncations of its corners that leave a polygon of area 8. (By the way, for some of the 19 pairs (a, b) it is impossible to come up with the appropriate x, y, z, and w.) The results of this search are summarized in the table, which contains a list of all 20 convex polygons that can be assembled from 16 con-

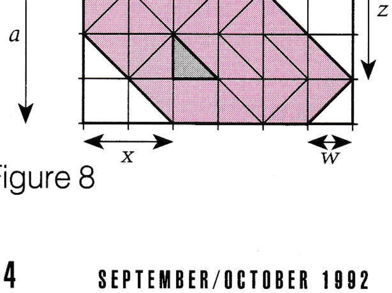


Figure 8

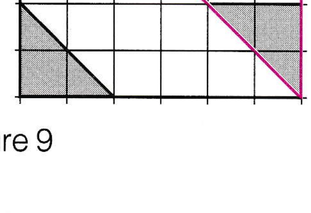


Figure 9

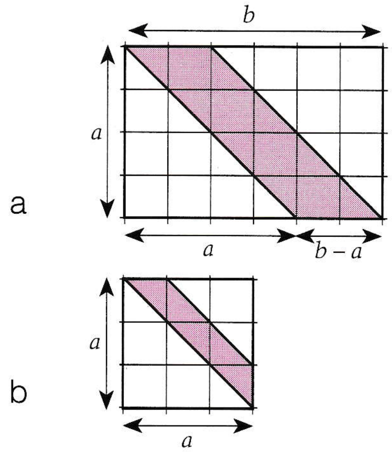


Figure 10

Table

a	b	x	y	z	w	T	P
1	8	0	0	0	0		
1	9	0	1	0	1		
1	9	0	1	1	0		
2	4	0	0	0	0	✓	
2	5	0	0	0	2	✓	✓
2	5	1	1	1	1	✓	✓
2	6	0	2	0	2	✓	✓
2	6	0	2	2	0	✓	✓
3	3	0	0	1	1	✓	
3	3	0	1	0	1	✓	✓
3	4	0	2	0	2	✓	
3	4	0	2	2	0	✓	✓
3	5	0	3	1	2	✓	
3	5	0	3	2	1	✓	✓
4	4	0	0	0	4	✓	✓
4	4	2	2	2	2	✓	✓
4	6	0	4	0	4		✓
5	5	0	3	0	5		✓
5	5	1	4	1	4		✓
8	9	0	8	0	8		

gruent 45°–45° right triangles. The numbers determining the dimensions of the rectangles corresponding to our polygons and of the triangles cut off from them are given in the first six columns. The polygons that can be tiled by the pieces of the tangram or pythagoras are checkmarked in the last two columns. (Try to prove that these two columns are correct.)

We don't even need the last two columns, though, because none of the 20 polygons has exactly seven sides like the mysterious figure No. 51! (Draw them and you'll see.)

But my story has a happy ending. Before me is a recently purchased, brand-new version of pythagoras. Aha! In the new instructions the drawing for figure No. 51 is exactly like my figure 5! But I look on the bright side: if the old instructions hadn't been in error, this article would never have been written. ◘