TURNING A SURFACE INSIDE OUT

Normally a sphere can be turned inside out only if it has been torn. In differential topology one assumes that the surface can be pushed through itself, but then the problem is to avoid forming a "crease"

by Anthony Phillips

The great mathematician David Hilbert once said that a mathematical theory should not be considered perfect until it could be explained clearly to the first man one met in the street. Hilbert's successors have generally despaired of living up to this standard. As mathematics becomes more specialized it is difficult for a mathematician to describe, even to his colleagues, the nature of the problems he studies. From time to time, however, research on an advanced and inaccessible mathematical topic leads to a discovery that is intuitively attractive and can be explained without oversimplification. A striking example is Stephen Smale's theorem concerning regular maps of the sphere, published in 1959.

The field in which Smale was then working—differential topology, which combines concepts from topology and calculus—is one of the more abstract domains of modern mathematics. Nevertheless, a visualization devised by the late Arnold Shapiro of Brandeis University enables us to depict a startling consequence of Smale's theorem: It is possible, from the topologist's point of view, to turn a surface such as a sphere inside out.

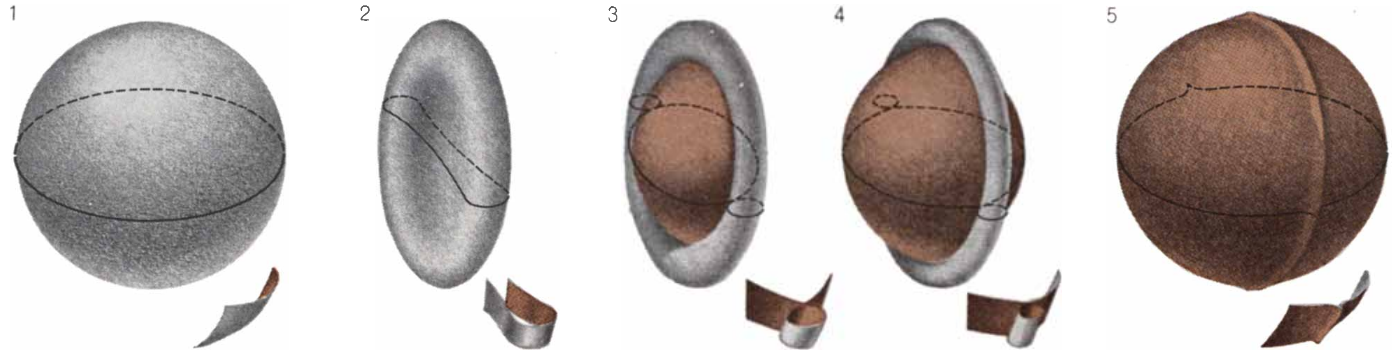
How is this accomplished? It is intuitively clear that unless a sphere is somehow torn it must remain right-side out, no matter how one is allowed to deform and displace it. If we are mentally allowed to move the surface through itself, however, so that two points on it can occupy the same point in space (this is permissible in differential topology), then a solution suggests itself. It is the deformation in which two regions of the sphere are pushed toward the center from opposite sides until they pass through each other. The original inner surface begins to protrude in two places, which are then pulled apart until the knurl—the remaining portion of the outside—vanishes. In the process, unfortunately, the knurl forms a tight loop that must be pulled through itself [*see upper illustration on opposite page*]. This results in a "crease" that is displeasing to differential topologists, whose discipline is limited to smooth surfaces.

The problem for the differential topologist is how to turn the sphere inside out without introducing a crease in getting rid of the knurl. Here too intuition indicates that the problem cannot be solved, and when Smale first announced he could prove that a solution existed, even his thesis adviser at the University of Michigan warned him that there was an "obvious counterexample" to his claim. Intuition was wrong; no fault could be found with the logic of Smale's proof. In fact, it was theoretically possible to follow the proof step by step and to discover an explicit description of the deformation that turns the sphere inside out. The argument was so complicated, however, that the actual task seemed hopeless. For some time after Smale's discovery it was known that the sphere could be turned inside out without a crease, but no one had the slightest idea how to do it.

This problem, concerning such a simple object as the sphere, was extremely mysterious and challenged the minds of many mathematicians. As far as I know the only ones who eventually worked it out were Nicolaas Kuiper, who is now at the University of Amsterdam, and Shapiro. Shapiro's idea of the deformation provides the basis for the illustration that begins at the bottom of the opposite page and ends on page 117. The deformation of the gray sphere begins with the pushing of two regions on opposite sides of the sphere toward the center and through each other so that the colored interior protrudes. The surface is then stretched, pinched and twisted through several intermediate stages too complicated to depict in their entirety. We can follow the changes between successive stages by watching what happens to ribbon-like cross sections of the surface as it is turned inside out. It is possible to understand the entire deformation by interpolating the missing parts of the surface at each stage and checking that the changes in the ribbons depicting various sections fit together coherently.

Turning a sphere inside out is only one example of the deformations shown to be possible by Smale's theorem. To explain the full implications of the theorem we must introduce mathematical definitions of the objects and deformations with which we are concerned. Let us begin with some precise descriptions of curves. The circle is defined to be the set of points in a two-dimensional plane (located by an x axis and a y axis) at distance 1 from the origin (the intersection of the two axes). Having established the radius as 1, we can describe any point P on the circle by the angle θ between the x axis and a line from the origin to P [*see upper illustration on page 114*]. In other words, any point on the circle can be described by a number from 0 to 360 giving the number of degrees in the corresponding angle θ.

A curve in the plane will be defined as a map from the circle into the plane. We shall consider several examples of such maps; each of them is a rule assigning to each point of the circle (or, equivalently, to each number θ between 0 and 360) a point in the plane. If c is such a map, then the point it assigns to θ is denoted $c(\theta)$ and is called "the image of θ under c." Naturally we require that $c(0)$ equal $c(360)$,

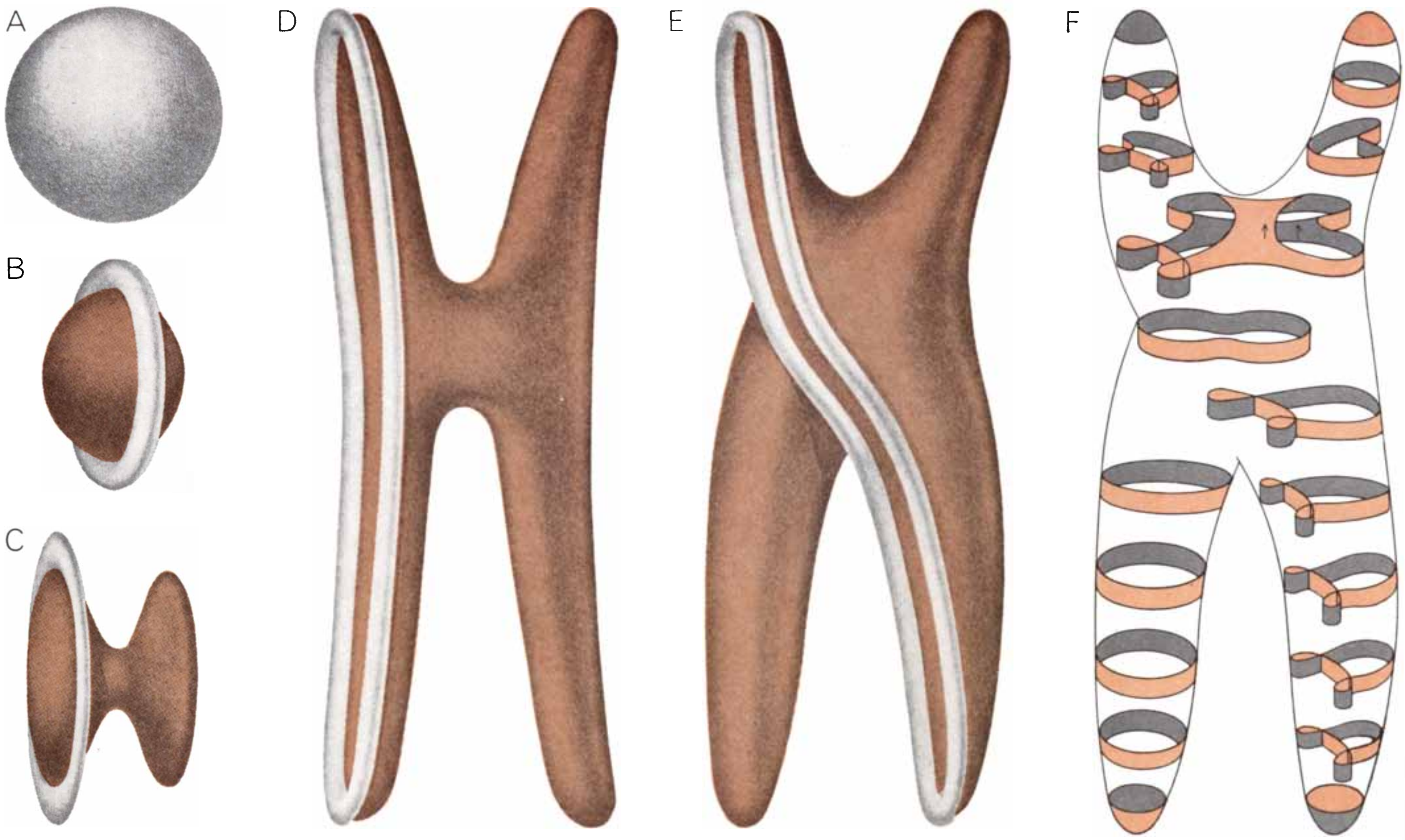


INADMISSIBLE PROCEDURE for turning a sphere inside out entails pushing regions on opposite sides toward the center (*2*) and through each other. The original interior (*color*) begins to protrude on two sides (*3*); these two sides are pulled out to form a sphere (*4 and 5*). When the looped portion of the original surface is pulled through itself, a "crease" is introduced in the surface; this violates a law of differential topology, a discipline of mathematics that is concerned only with smooth surfaces. In this discipline moving a surface through itself is permissible. The ribbons at bottom depict a section of the surface during stages of deformation.

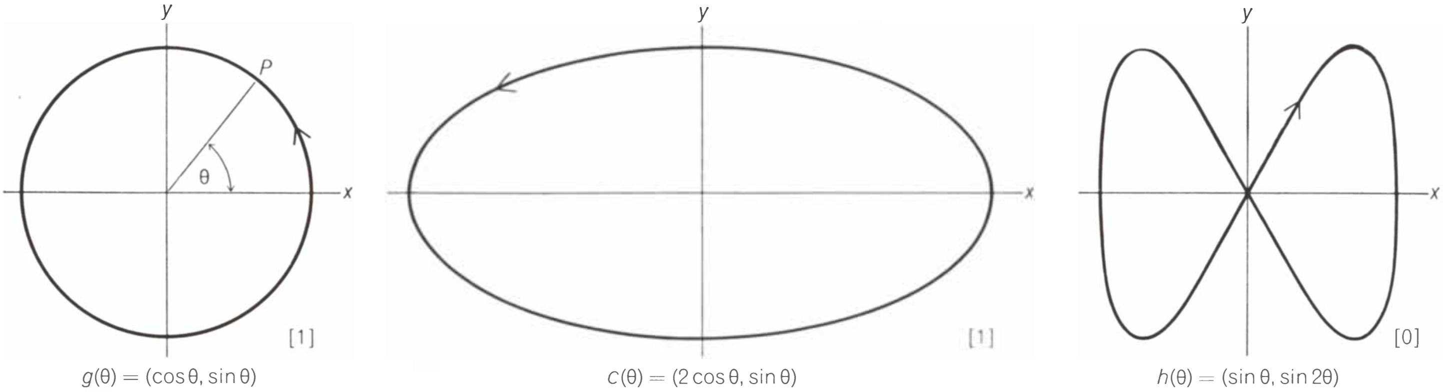
since 0 and 360 describe the same point on the circle.

Why should one give such abstract and complicated meanings to such simple terms as "curve"? There are two important reasons. First, precise definitions are essential for a sound mathematical theory. Second, a good definition will suggest by analogy how a theory can be extended to encompass new material. For example, to get a picture of a curve in the plane it is helpful to think of the path of a moving particle. This could be made into a definition more easily grasped by the intuition than "a map from the circle into the plane," but the more abstract definition will enable us to generalize our study of curves to that of two-dimensional surfaces, our main interest. A simple example of a curve is provided by the map c that assigns to each θ the point in the plane with an x coordinate of $2 \cos \theta$ and a y coordinate of $\sin \theta$. (If θ is taken as an acute angle of a right triangle, then $\sin \theta$, the sine of θ, is the ratio of the length of the opposite leg to the length of the hypotenuse; $\cos \theta$, the cosine of θ, is the ratio of the length of the adjacent leg to the length of the hypotenuse.) This description can then be abbreviated to $c(\theta) = (2 \cos \theta, \sin \theta)$.

The curve defined by $g(\theta) = (\cos \theta, \sin \theta)$ is even simpler. If the point P on the circle corresponds to the angle θ,



HOW TO TURN A SPHERE INSIDE OUT by steps that conform to the laws of differential topology is depicted in this illustration and those at bottom of the next four pages. The deformation begins with the pushing of opposite sides toward the center and through each other so that the interior (*color*) protrudes in two regions (*B*). One part of the original interior is then distended (*C*) to give the surface resembling a saddle on two legs (*D*). The two legs are then twisted counterclockwise to give surface *E*. This surface is shown again (*F*) with ribbons depicting it in cross section on different levels. Thin black line indicates missing parts of surface.

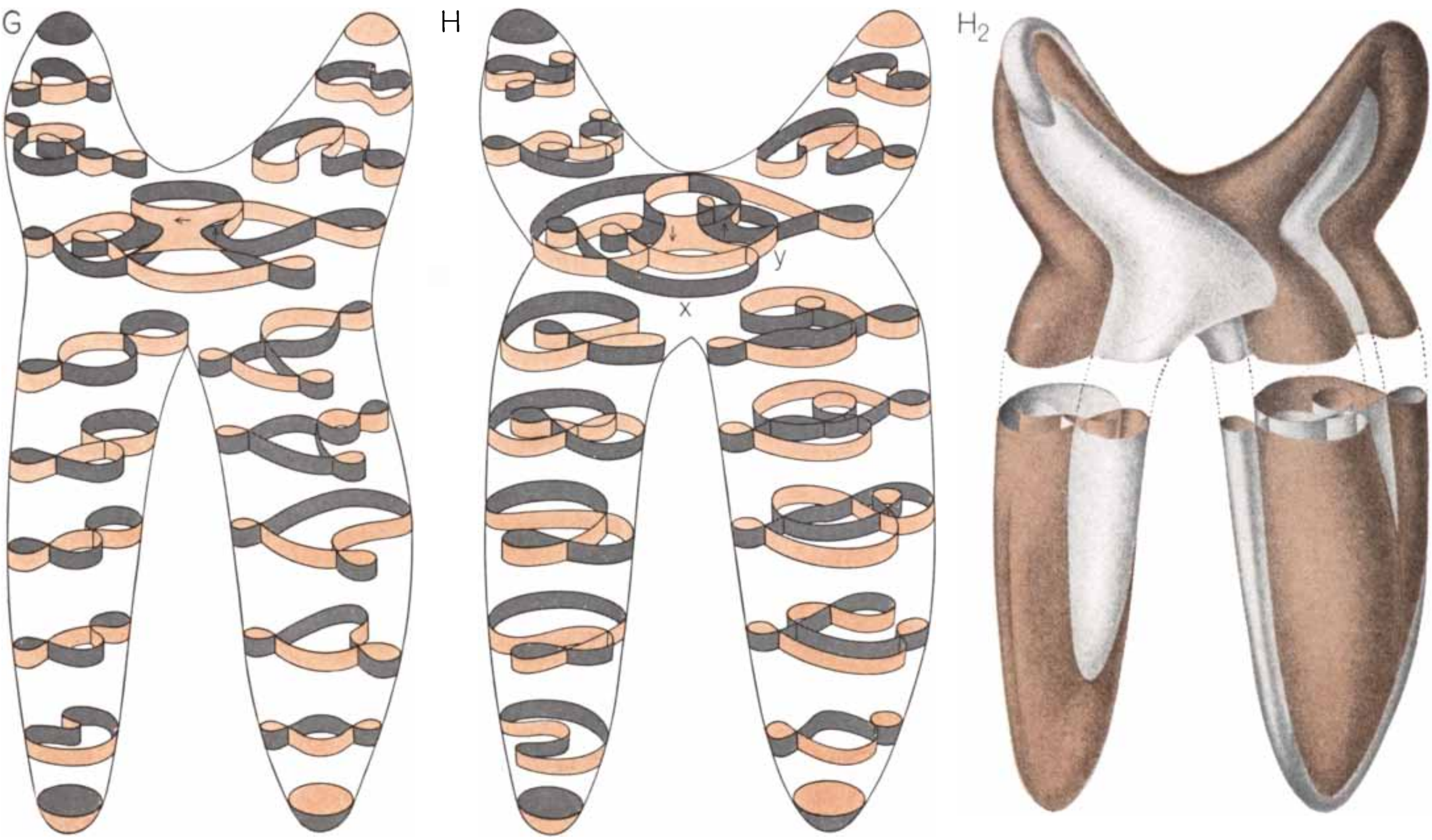


CURVES IN THE PLANE are defined and exemplified. The circle is defined as the set of points in a two-dimensional plane at distance 1 from the origin (the intersection of the x axis and the y axis). A point P on the circle is described by the angle θ between the x axis and the line from the origin to P. A closed curve in the plane can be given by a map that assigns to each number θ a point in the plane. The map c that assigns to each θ the point whose x coordinate is given by two times the cosine of θ and whose y co-

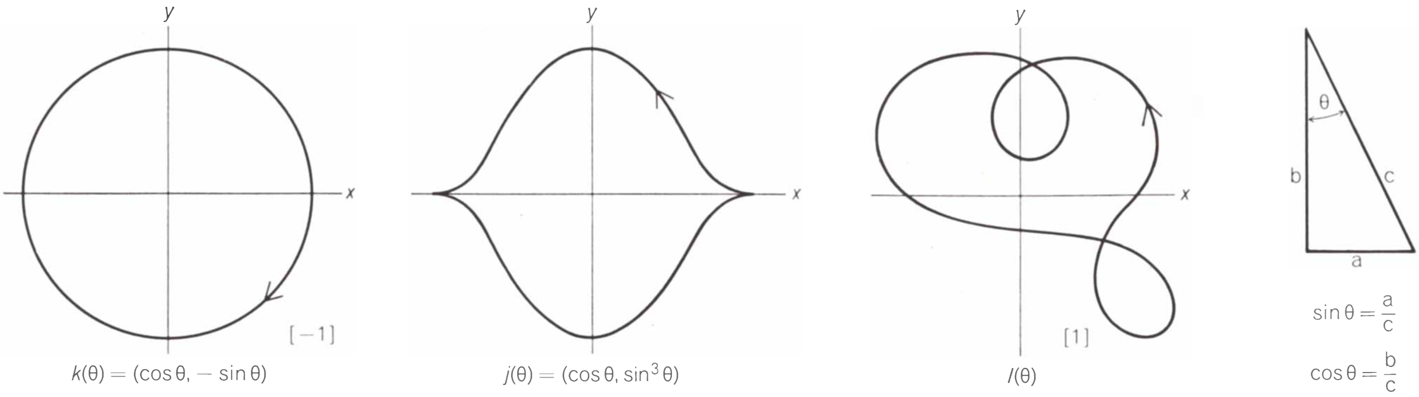
then the coordinates of P itself are cos θ and sin θ. In other words, the map g assigns to each point P the point P once again. The map g is called the standard embedding of the circle in the plane.

The curves given by the maps c and g are both examples of regular curves in the plane. A curve in the plane will be called regular if, as a point runs around the circle at constant speed, its image moves smoothly and with a velocity other than zero in the plane. If one represents a curve by tracing it, the motion of the pencil on the page is equivalent to the motion of the image of θ as θ runs from 0 to 360. Thus the map l, the image of which appears in the upper illustration on these two pages, is seen to be a regular curve but the map j is not, since the pencil cannot move smoothly past a pointed end (a "cusp") without stopping.

Two regular curves are said to be regularly homotopic if one can be deformed into the other through a series of regular curves. This implies that between the two original curves there is a family of regular curves, each representing the shape of the curve at a given stage of the deformation. It can be



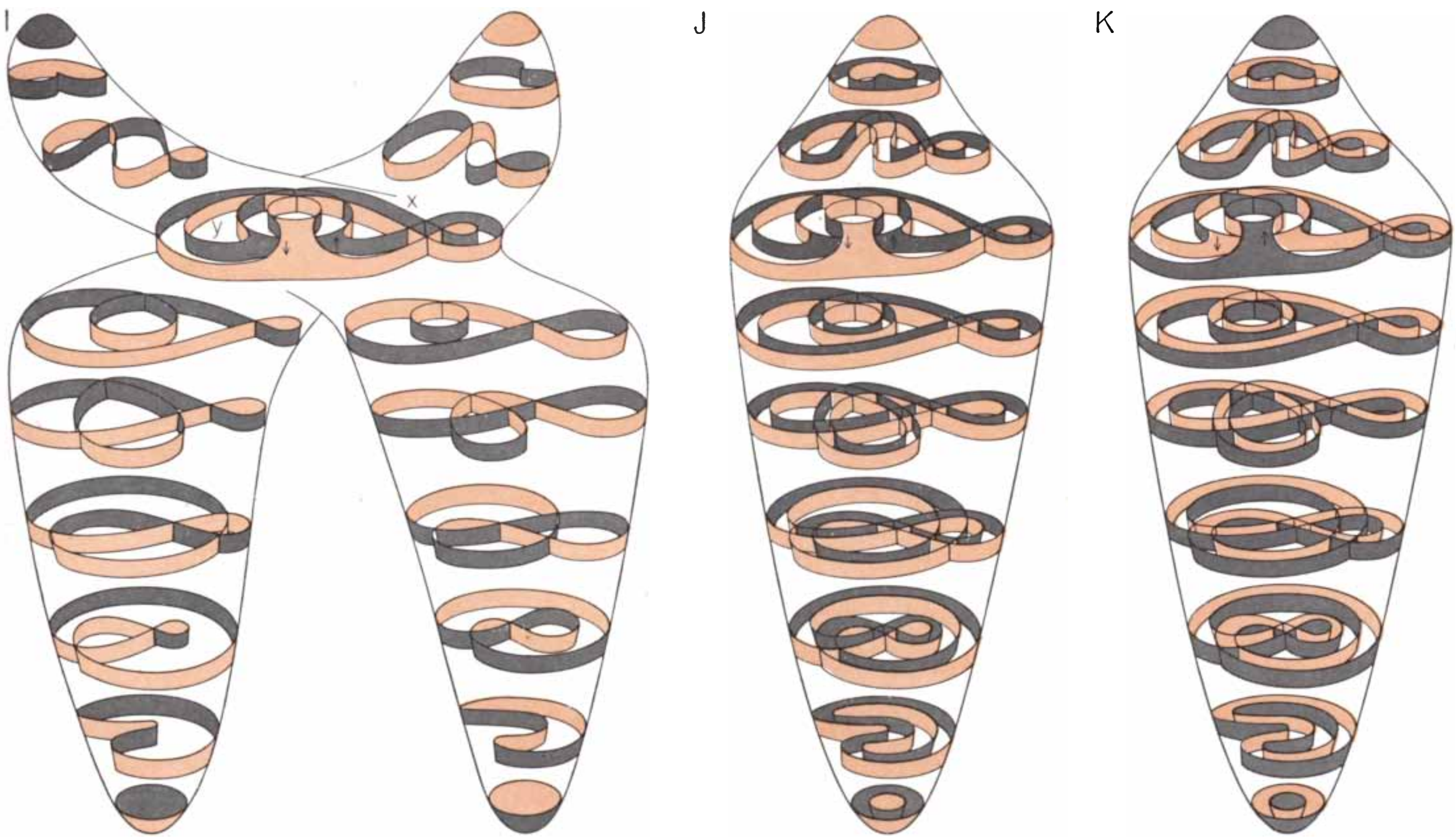
DEFORMATION OF SPHERE IS CONTINUED here and at bottom of opposite page. Depicting the entire surface at each stage would not clarify the process; the reader must interpolate the missing parts of the surface at each stage by considering the 10 cross sections and checking that the changes at all levels fit together coherently. One stage (H_2) is depicted schematically so that the overall view of the surface can be borne in mind. Surface G was formed by pinching and rotating the saddle of surface F 90 degrees.



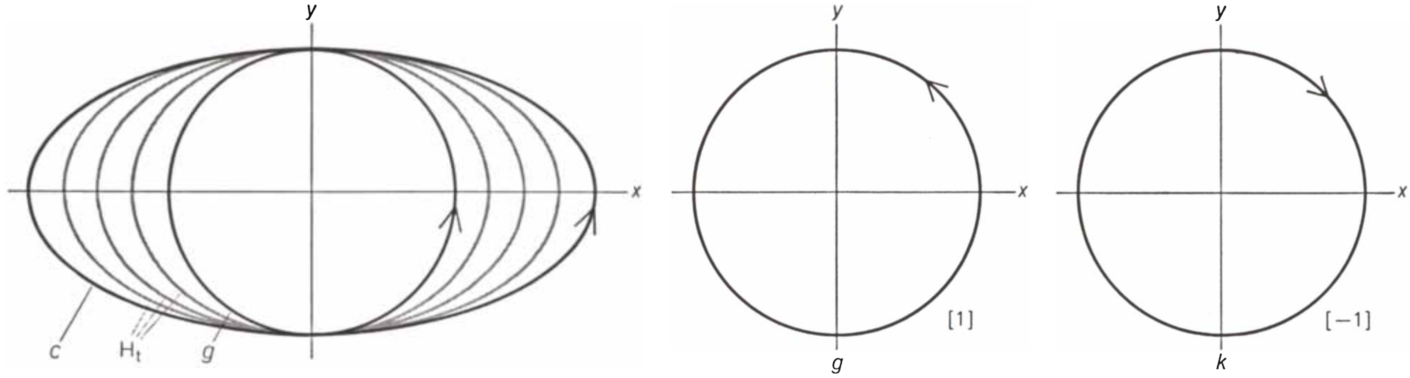
ordinate is given by the sine of θ is written c $(\theta) = (2 \cos \theta, \sin \theta)$. (Key at far right defines the trigonometric functions sine and cosine.) A closed curve in the plane is called regular if, as a pencil is traced around it, there is no "cusp," or spot at which the pencil must come to a momentary stop. (All curves in this illustration except *j* are regular.) The "winding number" listed in brackets beneath each curve gives the total number of counterclockwise turns made by the curve. Clockwise turning is counted as negative.

demonstrated that the curves given by the maps c and g are regularly homotopic [*see upper illustration on next two pages*]. Here the family of curves H_t, or the deformation this family represents, is called a regular homotopy between c and g.

One can show that two curves are regularly homotopic without finding the specific homotopy between them, by using the concept of "winding number." The winding number of a regular curve is the total number of counterclockwise turns the curve makes. (Clockwise turning is counted as negative, so the winding number of a curve can be negative.) The upper illustration on these two pages shows that the winding numbers of curves g, h and k are respectively 1, 0 and -1. It is plausible and not very difficult to prove that the winding number of a regular curve must remain fixed during a regular homotopy, so two regularly homotopic curves must have the same winding number. (It follows that no two of g, h and k can be regularly homotopic.) In 1937 Hassler Whitney, then at Harvard University, proved the more difficult converse statement: Any two regular curves with the



REVERSAL OF COLORS, indicating a reversal of the original interior and exterior, is achieved in subsequent steps. Between stages *H* and *I* the parts of the surface marked *x* and *y* (*middle level*) move to the rear. Between *I* and *J* the two similarly shaped legs move through each other. At each level of the surface at stage *J* the cross-sectional ribbon has two gray sides facing each other. Between stages *J* and *K* the inner layer expands and the outer layer contracts, giving surface *K*, which is *J* with the colors reversed.



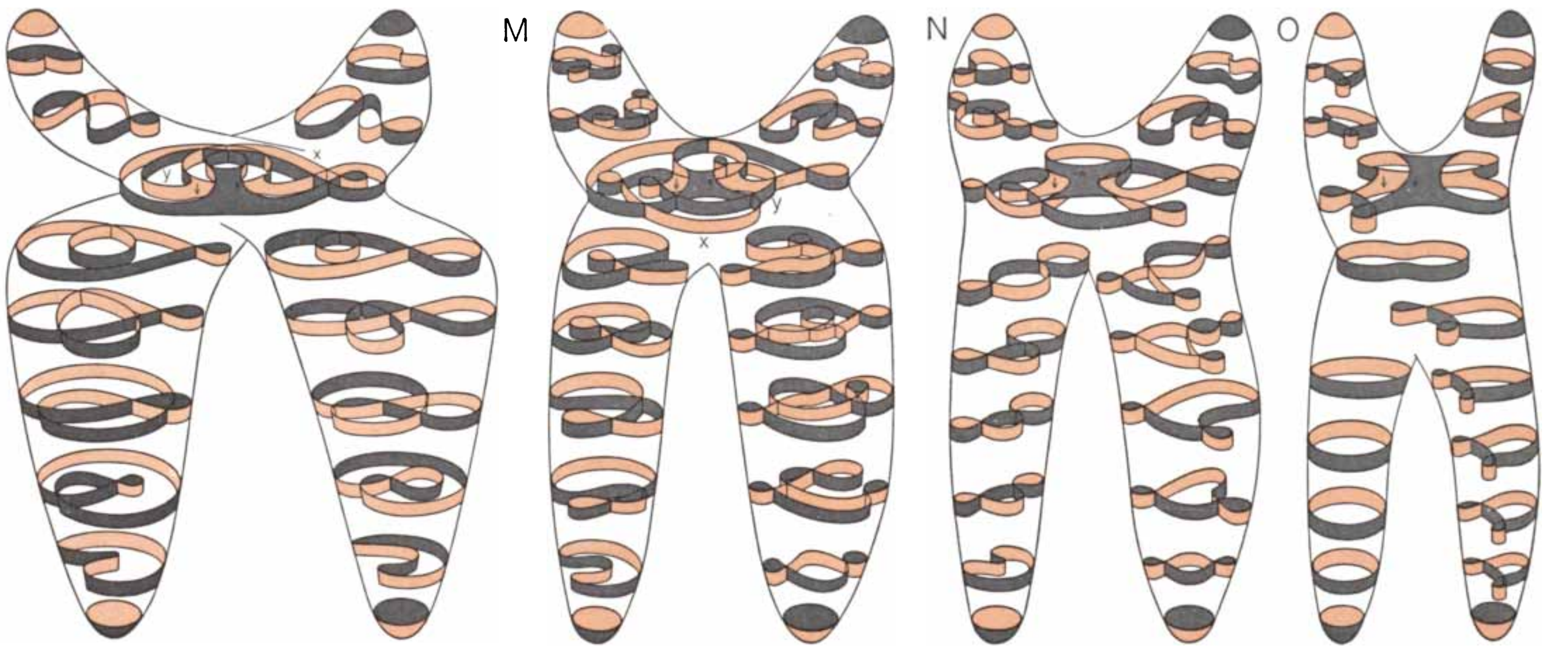
REGULARLY HOMOTOPIC CURVES IN PLANE can be deformed, one to the other, through a family of regular curves between them (*gray*). Thus the curves *g* and *c* are regularly homotopic, and the gray curves (H_t) represent successive stages of the deformation of *g* to *c*. It has been proved that only curves with the same winding number are regularly homotopic (because the wind-

same winding number are regularly homotopic. The reader may find it interesting to devise a regular homotopy between two curves with the same winding number, such as curves *g* and *l* in the upper illustration on the preceding two pages. This can be made easier with a length of chain or string. The problem is to loop the chain or string to form one of the curves, then shift it (without lifting it from the table) to form the other without introducing a cusp at any point.

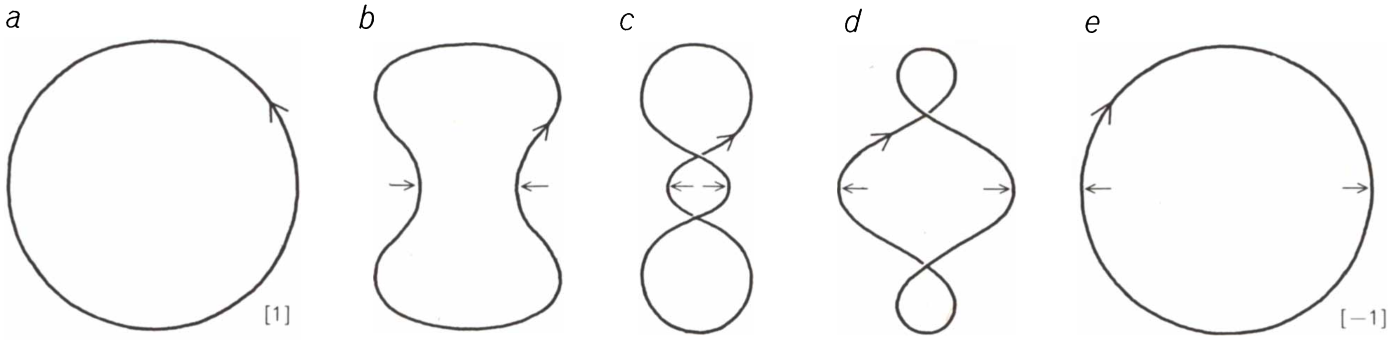
The definitions of a curve on the sphere and of a regular curve on the sphere are analogous to those of curves in the plane, the sphere being the set of points in three-dimensional space at distance 1 from the origin. Some regular curves on a sphere are shown in the top illustration on page 118. It turns out that two such curves are regularly homotopic if the parity of the number of their self-intersections is the same, that is, if they both have an odd number of self-intersections or both have an even number of them. Unlike curves on a plane, regular curves on the sphere can be regularly homotopic even if they do not have the same "winding number."

We began by defining a curve as a map from the circle into the plane. Since the circle and the sphere have analogous definitions, one is led to ask: What happens when everything is moved up one dimension? The analogue to a curve is a map from the sphere into three-dimensional space. Such a map would assign to each point of the sphere some point (its image) in three-space. An example of such a map is the standard embedding, which assigns to a point *P* of the sphere the point *P* considered as a point in three-space. This is entirely analogous to the standard embedding of the circle in the plane. Another example is the antipodal map *A* [*see bottom illustration on page 118*] that assigns to each point *P* of the sphere its diametrically opposite point $A(P)$, considered as a point in three-space. The analogous curve is given by $a(\theta) = (-\cos\theta, -\sin\theta)$.

Suppose *c* is a curve on the sphere; thus *c* assigns to each point of the circle a point on the sphere. Since a map from the sphere into three-space assigns to each point on the sphere a point in three-space, such a map will transform *c* into a curve in three-space. This remark is the basis for the definition of the two-dimensional analogue of a regular curve: a regular map from the sphere into three-space is a map



"BACKWARD" DEFORMATION of the surface shown in stage *K* results in a sphere having the colored original interior as its exterior. Intermediate surface *L* corresponds to surface *I* with the colors reversed; surface *M* corresponds to surface *H*, *N* to *G*, *O* to *F*, O_2 to *E* and so on. The colored sphere (*surface S*) corresponds of course to the gray sphere shown as surface *A* on page 113. The en-



ing number of a curve remains fixed during a regular homotopy). Thus curves *g* and *k* are not regularly homotopic. The deformation of one into the other (which is the equivalent in the plane of turning a sphere inside out) creates a change in winding number. If this deformation is duplicated with some string ("*a*" *through* "*e*") on a table, one will find it cannot be done without introducing a cusp.

that transforms each regular curve on the sphere into a regular curve in three-space.

The standard embedding is obviously regular, and the bottom illustration on the next page shows that the antipodal map *A* transforms a regular curve into a regular curve, and that *A* is thus also a regular map. On the other hand, a sphere with a crease cannot be the image of a regular map; any curve perpendicular to the crease has a cusp.

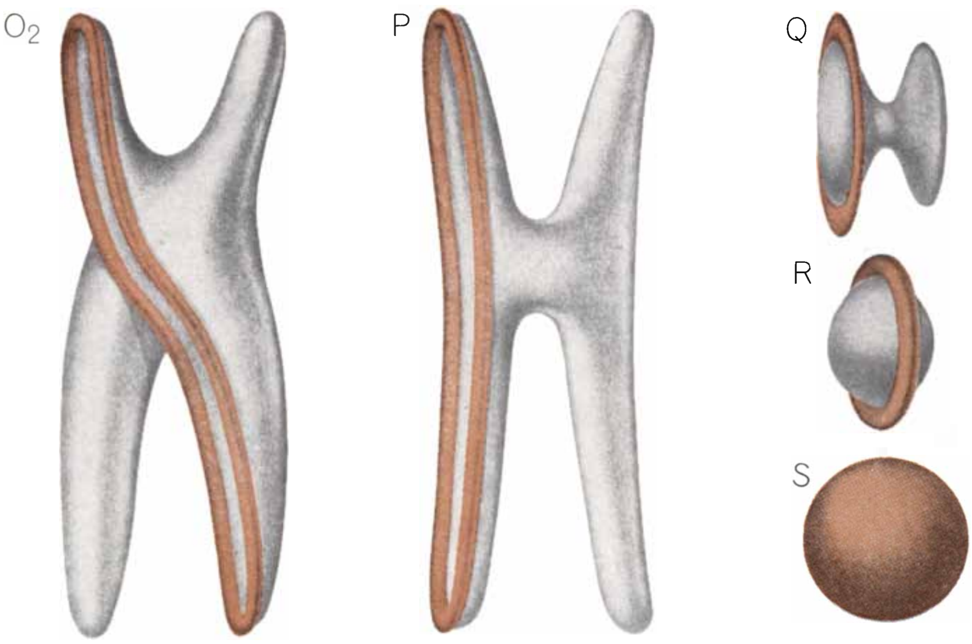
Pursuing the analogy, we define two regular maps from the sphere into three-space as regularly homotopic if there is a family of regular maps (a regular homotopy) joining them; in other words, if one regular map can be deformed into the other through a series of regular maps. Now, we already know that the standard embedding is regularly homotopic to the antipodal map. In fact, Shapiro's visualization is of just such a regular homotopy. Since we had not yet defined regular maps when we first discussed turning the sphere inside out, we described the homotopy by showing how the images of the maps could be deformed one into the other. In retrospect it is clear that each of the surfaces in the illustrations at the bottom of pages 113 through 117 is the image of a regular map from the sphere into three-space, and that these maps can be chosen to vary smoothly and in such a way as to provide a regular homotopy between the standard embedding and the antipodal map. In particular every point on the surface designated *J* in the bottom illustration on page 115 lies opposite its antipodal point, and the deformation of *J* to *K* exchanges these points.

One big advantage of considering a homotopy of maps rather than a deformation of surfaces is that the status of self-intersections becomes logically clear. It is no longer necessary to speak of two points on the surface "occupying" the same point in space. A map from the sphere into three-space has a self-intersection when it sends two points of the sphere to the same point in space.

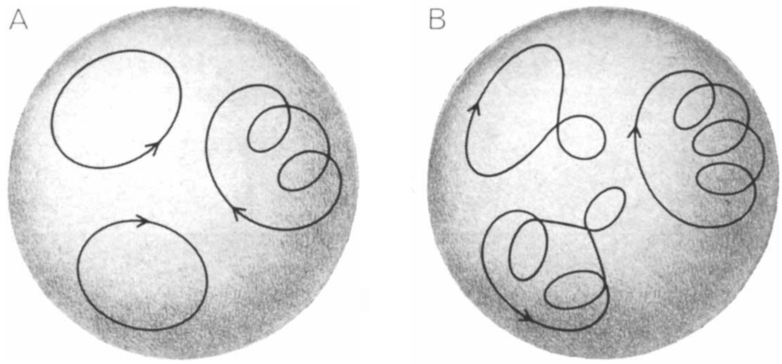
The regular homotopy between the standard embedding of the sphere and the antipodal map can be used to provide a regular homotopy that turns other simple surfaces, such as the torus, inside out. The torus is the doughnut-like surface shown in the illustration on page 119. The homotopy depicted can be described as follows. Extrude a small sphere from the surface of the torus and turn the sphere inside out. Expand the reversed sphere until it swallows the torus. A tube now leads from the outside of the reversed sphere to the inside of the torus. Enlarging this hole if necessary, pull the torus out by the inside. Shrink what remains of the sphere.

There are regular maps from the torus into three-space that are not regularly homotopic. The illustration on page 120 shows the images of four such maps, no two of which are regularly homotopic. How many such maps can there be?

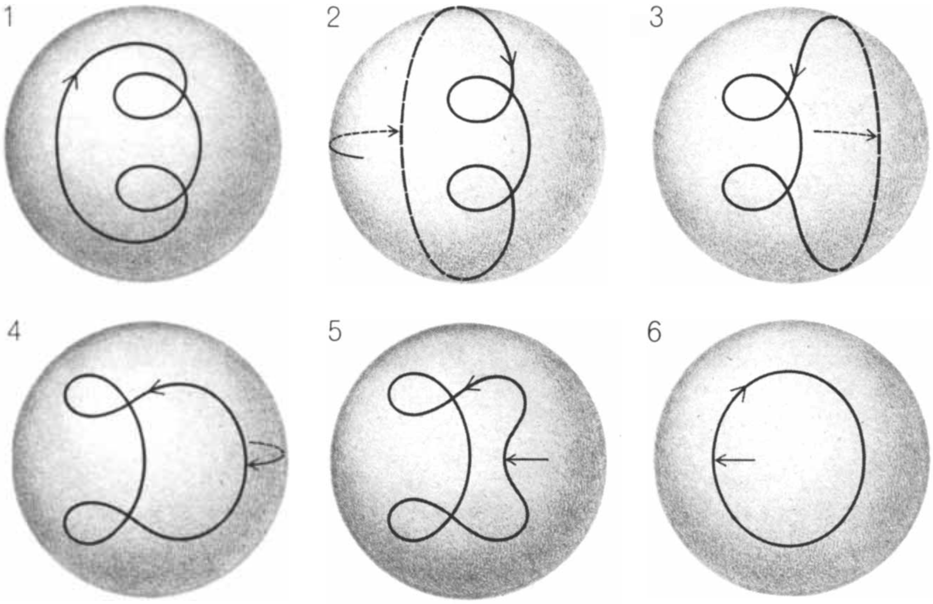
We have seen that there is an infinite collection of regular curves in the plane (one for each winding number), no two of which are regularly homotopic. We have also remarked that two regular curves on the sphere are regularly homotopic if and only if the number of their self-intersections is odd in each case or even in each case. It is therefore possible to divide all the regular curves on the sphere into two sets (those with an even number of self-intersections and those with an odd number of self-intersections) such that



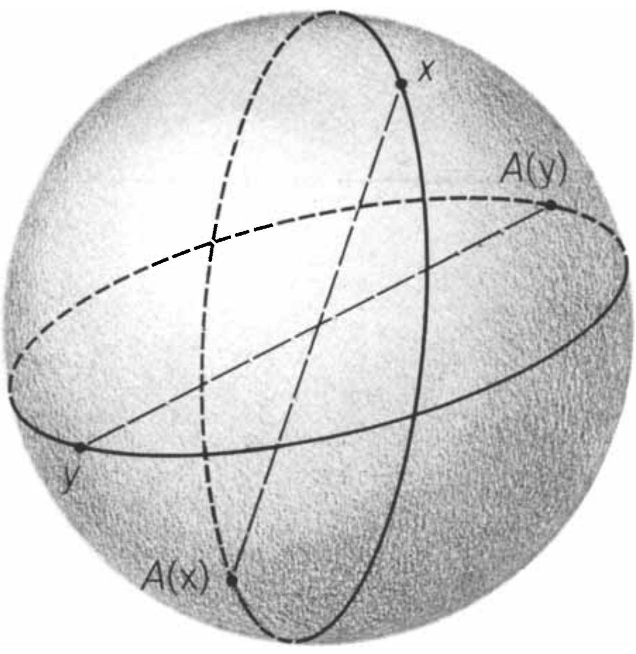
tire deformation is accomplished without introducing a crease in the surface. The feat was first proved possible by Stephen Smale, then at the University of Michigan. The intermediate steps of the deformation were first imagined by Arnold Shapiro of Brandeis University.



REGULAR CURVES ON THE SPHERE can be shifted "around the back" and are not governed by rules for curves in a plane. Two regular curves on the sphere are regularly homotopic if they both have an odd or both have an even number of self-intersections. Curves on sphere A belong to same "regular homotopy class." Curves on sphere B belong to a different regular homotopy class. No curve on A is regularly homotopic to one on B.



REGULAR HOMOTOPY ON THE SPHERE is illustrated for two curves on sphere A of illustration at top. Broken segment of the curve has been shifted around the back of sphere.



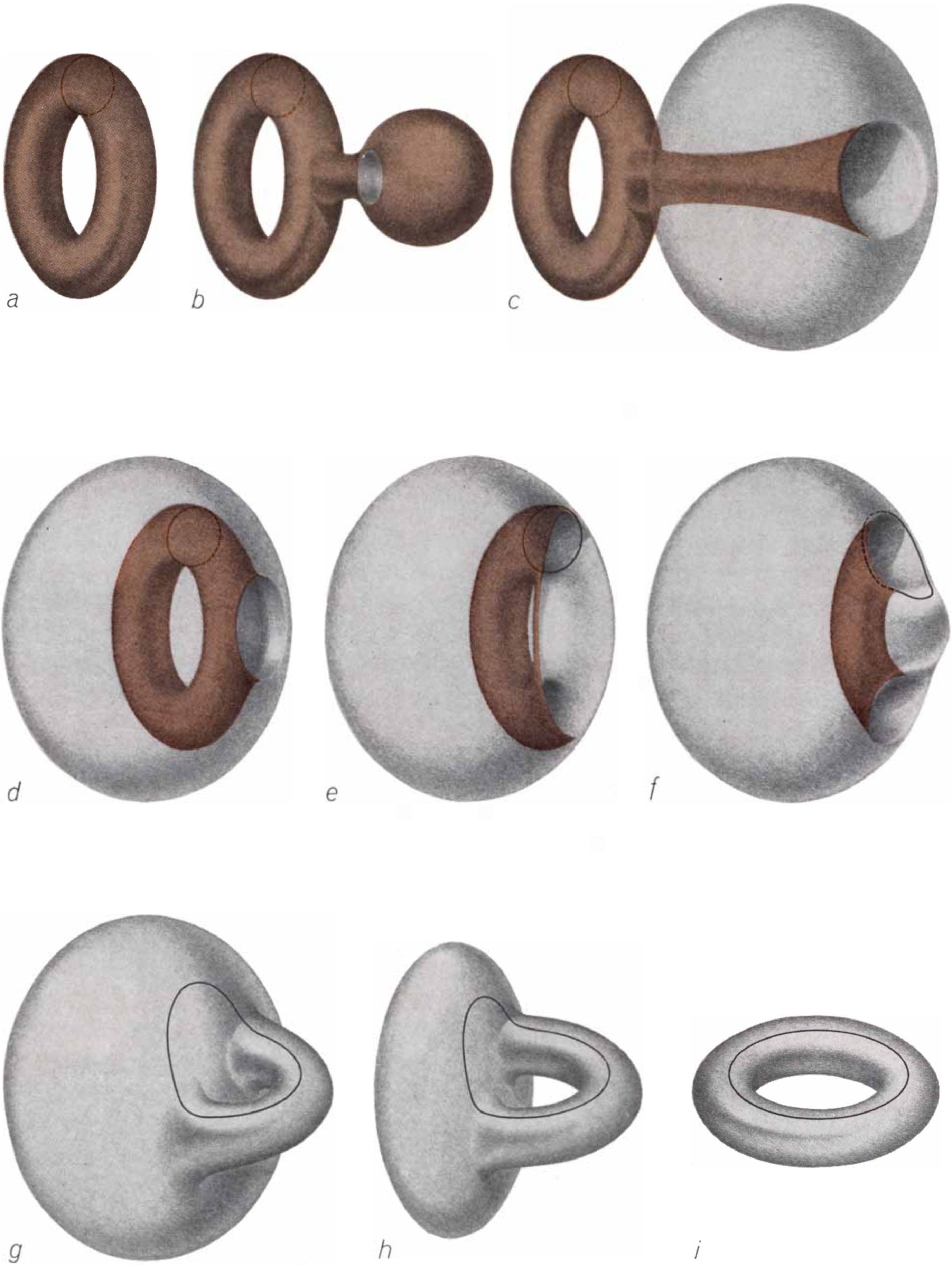
ANTIPODAL MAP assigns to each point on sphere its diametrically opposite point. Discovery of a regular homotopy between sphere and antipodal map proved it could be everted.

any two curves in the same set are regularly homotopic, but no curve in one set is homotopic to a curve in the other set. If we agree to call sets with these two properties "regular homotopy classes," we can state that there are infinitely many regular homotopy classes of regular curves in the plane, but only two regular homotopy classes of regular curves on the sphere.

It has been ascertained that the number of regular homotopy classes of regular maps from the torus into three-space is four. The general problem of determining the number of regular homotopy classes of regular maps from an arbitrary surface into three-space was explicitly solved only recently. The solution came as part of the extensive study of regular maps that was stimulated by Smale's research. Morris Hirsch, while a graduate student at the University of Chicago, showed in his doctoral thesis how Smale's work could be extended to regular maps of an arbitrary surface. Hirsch's work was used in turn by Ioan James of the University of Oxford and Emery Thomas of the University of California at Berkeley, who showed that the number of regular homotopy classes of regular maps from a surface into three-space depends only on a number known as the Euler characteristic of the surface, named for the great Swiss mathematician Leonhard Euler. To calculate the Euler characteristic X of a surface, divide the surface into polygons. Then X is given by the equation $X = P - E + V$, where P is the total number of polygons, E is the total number of distinct edges (each edge will belong to two polygons but should be counted only once) and V is the total number of distinct vertices. It is a significant topological fact that the number thus obtained depends only on the surface and not on how the surface was divided into polygons.

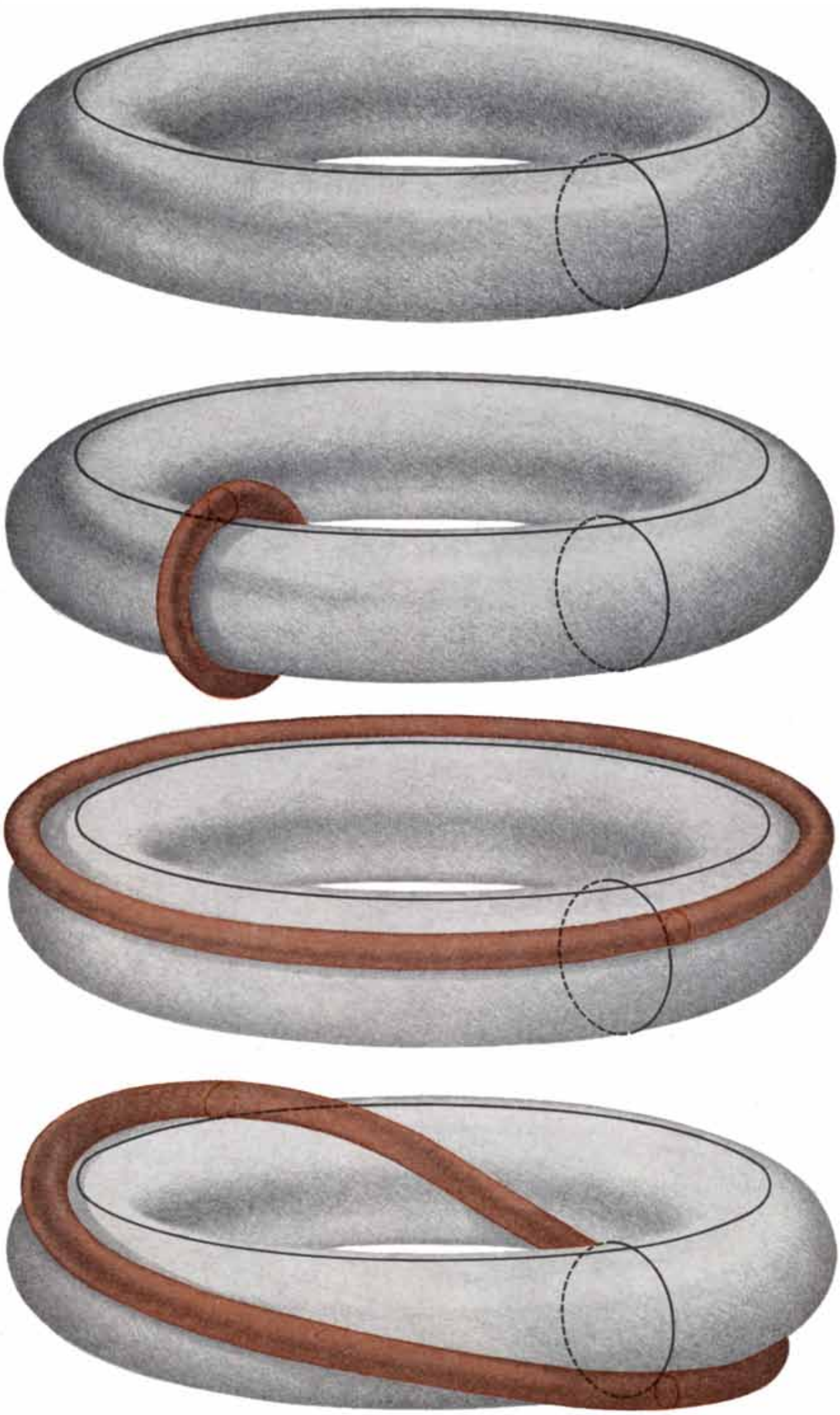
James and Thomas showed that if the Euler characteristic of a surface is X, then the number of regular homotopy classes of regular maps from that surface into three-space is $2^{(2-X)}$. Thus the torus (with an Euler characteristic of $X = 0$) has four distinct regular homotopy classes of maps. The sphere ($X = 2$) has only one. This is the complete statement of Smale's theorem: *Any two regular maps from the sphere into three-space are regularly homotopic.* The existence of a regular homotopy between the standard embedding and the antipodal map was only a special case.

Of course Smale's proof does not consist of drawing pictures of regular maps.



TURNING A TORUS INSIDE OUT involves the regular homotopy for everting the sphere. The torus is depicted with a circle designating its meridian (*a*). A small sphere is extruded from the torus (*b*) and everted (*c*). Then the inside-out sphere is enlarged until it engulfs the torus (*d*). Next the tube leading from the outside of the everted sphere to the inside of the torus is enlarged and the torus is pulled through it ("*e*," "*f*" *and* "*g*"). Finally sphere is shrunk ("*h*" *and* "*i*"). In the process the meridian has become a latitude.

In fact, Smale's paper contains no pictures at all. The intricacy of the pictures, which were in a sense implicit in Smale's abstract and analytical mathematics, is amazing. Perhaps even more amazing is the ability of mathematicians to convey these ideas to one another without relying on pictures. This ability is strikingly brought out by the history of Shapiro's description of how to turn a sphere inside out. I learned of its construction from the French topologist René Thom, who learned of it from his colleague Bernard Morin, who learned of it from Arnold Shapiro himself. Bernard Morin is blind.



NONREGULARLY HOMOTOPIC REGULAR MAPS from the torus into three-dimensional space fall into four distinct classes. This illustration shows the image of one map from each class. No two of these maps could be deformed one to the other through regular maps.