Mathematical Notes/Comments

Most of these notes keep promises made in previous chapters. "Square circles" is in a light-hearted vein. No one need be offended. David Hume was my favorite philosopher in college.

Except for their use of algebra, the articles on "How Imaginary Becomes Reality" could have come earlier. They show natural, inevitable ways that mathematics grows, with no mystery about invention vs. discovery.

"Calculus refresher" is included because it isn't possible to talk sense about mathematics without an acquaintance with or recollection of calculus. By omitting exercises and formal computations, I present a semester of calculus in a few easy pages.[1]

The last article is a wonderful piece of mathematical artistry. The late George Boolos proves Gödel's great incompleteness theorem in three simple pages![2]

Arithmetic

[1] It's taken, with minor improvements, from the Teacher's Guide which accompanies the study edition of *The Mathematical Experience*, co-authored with Philip J. Davis and Elena Anne Marchisotto, Birkhauser Boston, 1995.

[2] It's reproduced from the *Notices* of the American Mathematical Society to make it available to a larger readership.

Logic

Sets

Geometry

How Imaginary Becomes Reality

Calculus

More Logic

ARITHMETIC

What Are the Dedekind-Peano Axioms?

Instead of constructing the natural numbers out of sets à la Frege-Russell, we can take them as basic, and describe them by axioms from which their other properties can be derived.

The axioms should be consistent, of course; chaos could ensue if they were contradictory (see below). We would like them to be minimal—not include any redundant axioms.

The standard axioms for the natural numbers were given by Richard Dedekind, inventor of the Dedekind cut. Following the usual rule of misattribution in mathematical nomenclature, they are called "Peano's postulates." The undefined terms are "1" and "successor of."

1. 1 is a number.
2. 1 isn't the successor of a number.
3. The successor of any number is a number.
4. No two numbers have the same successor.
5. (Postulate of mathematical induction) If a set contains 1, and if the successor of any number in the set also belongs to the set, then every number belongs to the set.

How to Add 1's

We show that Dedekind's axioms imply

$$2 + 2 = 4.$$

None of the symbols in this equation appears in the axioms, so we must define all four of them.

"=" is defined by the rule that for any x and y, if x = y, then in any formula y may be replaced by x and vice versa. This rule is called "substitution."

For present purposes, we need only define addition by 1 and 2, for all n. Let S stand for the successor operation.

Define 2 as S(1), 3 as S(2), 4 as S(3).
Define "n + 1" as S(n) and "n + 2" as S(S(n)).

Then by substitution

(A) 2 + 2 = S(S(2)).

Again by substitution,

(B) 4 = S(S(2)).
Voilà! 2 + 2 = 4.

To define n + k for all n and k would take more work, using recursion on both n and k.

Is 2231 prime?

I know how to find out. 2231 is prime if it's not divisible by any number between 1 and 2231. I could just divide 2231 by all the numbers from 2 to 2230.

This labor can be cut down a lot. If 2231 is factorable, it factors into two numbers, one larger, one smaller or both equal to each other. It's sufficient to find the smaller. Since

$47^2 = 2249$,

the smaller factor has to be less than 47. Moreover, it's not necessary to divide by any composite number, because if 2231 has a composite factor, that composite factor has prime factors that also factor 2231.

So we only have to check the prime numbers less than 47—2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43.

Now 2231 is odd, so 2 isn't a factor. The sum of the digits is 8, which is not divisible by 3, so 3 isn't a factor. It doesn't end in 5 or 0, so 5 isn't a factor. The alternating sum

+ 2 − 2 + 3 − 1 isn't 0,

so 11 isn't a factor. Get your calculator and divide 2231 by 7, 13, 17, 19, 23, 29, 31, 37, 41, and 43. If none of them divides 2231 without remainder, 2231 is prime.

Logic

Zermelo-Fraenkel, Axiom of Choice, and the Unbelievable Banach-Tarski Theorem

"Given any collection of nonempty sets, it is possible to form a set that contains exactly one element from each set in the collection."

Surely a harmless-sounding assumption to make about finite collections of finite sets, and even countably infinite collections of countably infinite sets. But

when it's applied to collections and sets of arbitrarily great uncountable cardinality, trouble comes! Consequences follow, which many mathematicians would rather not believe. Zermelo proved, using the axiom of choice, that any set—for instance, the uncountable set of real numbers—can be rearranged to be well-ordered. (But no one can actually do it, and no one expects anyone to be able to do it.) Stefan Banach and Alfred Tarski proved, using the axiom of choice, that it's possible to divide a pea (or a grape or a marshmallow) into 5 pieces such that the pieces can be moved around (translated and rotated) to have volume greater than the sun (see Wagon). As mentioned in Chapter 4 on proof, a transitory movement to avoid the axiom of choice has long been given up.

1/0 Doesn't Work (0 into 1 Doesn't Go)

Division by 0 is not allowed. Why not? If it's allowed to introduce a symbol i and say it's the square root of -1 *which doesn't have a square root,* why not introduce some symbol, say Q, for $1/0$?

We introduce new numbers, whether negative, fractional, irrational, or complex, to preserve and extend our calculating power. We relax one rule, but preserve the others. After we bring in i, for example, we still add, subtract, multiply, and divide as before. I now show that there's no way to define $(1/0) \times 0$ that preserves the rules of arithmetic.

One basic rule is,

$0 \times$ (any number) $= 0$.

(Formula I) So $0 \times (1/0) = 0$.

Another basic rule is

$(x) \times (1/x) = 1$, provided x isn't zero. (But if we want $1/0$ to be a number, this proviso becomes obsolete.)

(Formula II) So $0 \times (1/0) = 1$

Putting Formulas I and II together,

$1 = 0$.

Addition gives

$2 = 0$, $3 = 0$, and so on, $n = 0$

for every integer n.

Since all numbers equal zero, all numbers equal each other.

There's only one number—0.

The supposition that $1/0$ exists and satisfies the laws of arithmetic leads to collapse of the number system. Nothing is left, except—nothing.

What Is Modus Ponens?

In scholastic (medieval Aristotelian) logic, Latin names were given to the different permutations of Aristotle's syllogisms. Modus ponens is the simple argument: "If

A implies B, and A is true, then B is true." In modern formal logic the other syllogistic arguments can be eliminated. Modus ponens turns out to be sufficient.

Formalizable

A statement is "formalized" when it's translated into a formal language. Computer languages like Basic, Pascal, Lisp, C, and others—are formal languages. The notion of a formal language goes back to Peano, Frege, Russell, and Leibniz. A formal language has a vocabulary specified in advance—x_1, x_2, $+$, $\times$, $=$, etc. It has an explicit grammar, which prescribes the admissible permutations of the vocabulary. Whether a sentence in natural language is formalizable depends on the formal language under consideration. To be formalizable a sentence is supposed to be unambiguous, and to mention only objects that have names in the formal language.

How One Contradiction Makes Total Chaos

Suppose some sentence A and its negation "not-A" are both true.

We claim that (A and not-A) together implies B, no matter what B says. First, notice that:

(I) not-(A and not-a) means the same thing as (not-A or A), which is a "tautology"—it's true, no matter what A says. Also, notice that a tautology is implied by any sentence at all, because an implication is false only when the antecedent is true and the conclusion is false; if the conclusion is a tautology, it can't be false. Therefore

(II) not-B implies the tautology (not-A or A)

Now, by the definition of "implies," if a sentence P implies a sentence Q, then not-Q implies not-P. This deduction rule is called "contrapositive."

So, applying contrapositive to II,

(III) not-(not-A or A) implies not-(not-B)

(IV) But not-(not-B) is the same as B (double negative.) So, substituting (IV) into (III),

(V) not-(not-A or A) implies B.

Now applying negation ("not") to both sides of (I), we get

(VI) (A and not-a) means the same thing as not-(not-A or A)

Combining (V) and (VI), we have, as claimed,

(A and not-a) implies B, for any B.

The way (A and not-A) makes the whole logical universe collapse is rather like the way $1/0$ makes the whole number system collapse. Is there a connection?

Sets

The Natural Numbers Come Out of the Empty Set

I will describe the sequence of "constructions" by which we "create" the real number system out of "nothing." It has philosophical interest, and it's ingenious.

In practice, however, we think of numbers in terms of their behavior in calculation, not in terms of this "construction."

Start with the empty set. We define it as "the set of all objects not equal to themselves," since there are no such objects. All empty sets have the same members—no members at all! Therefore, as sets they're identical, by definition of identity of sets. In other words, there's only one empty set. This unique empty set is our building block. Next comes the set whose only member is—the empty set. This set is *not* empty. Think of a hat sitting on a table—an empty hat. An example of an empty set. Then put the hat into a box. The hat is still empty. The box containing one thing—an empty hat—is an example of a set whose single member is an empty set. We say the contents of the box has cardinality 1. We have so far two entities, the empty hat and the box containing the empty hat. Now put box and another hat together into a bigger box. The contents of the bigger box has cardinality 2. The interested reader can now construct sets with cardinality three, four, and so on. From an empty set we construct the natural number system!

How the Rational Numbers Are Dense but Countable

The natural numbers are discrete—each is separated from its two nearest neighbors by steps of size 1. On the other hand, the rational numbers (fractions) are "dense." Between any two you can find a third—the average of the two. Repeat the argument, and you see that between any two rationals there are infinitely many. (A fact intensely irritating to Ludwig Wittgenstein. He called it "a dangerous illusion." See Chapter 11.)

This seems to mean there are many more rationals than naturals. But that's not true. There are just as many!

Georg Cantor thought of a simple way to associate the rationals to the naturals. To each natural a rational, to each rational a natural.

Arrange the rational numbers in rows according to denominators. In the first row, all the fractions with denominator 1, numerators in increasing order:

$$\frac{0}{1}, \quad \frac{1}{1}, \quad \frac{2}{1}, \quad \text{and so on.}$$

In the second row, the fractions with denominator 2, numerators in increasing order:

$$\frac{0}{2}, \quad \frac{1}{2}, \quad \frac{2}{2}, \quad \text{and so on.}$$

Each row is endless, and the succession of rows is also endless.

Starting in the upper left corner at 0/1, draw a zigzag line: go down one step, then go diagonally up and to the right to the top row (with ones in the denominator). Go one step to the right, then go diagonally down and to the left to the first column (the fractions with 0 in the numerator). Go another step

down, and diagonally up and right again, and so on and on. This jagged line passes exactly once through every fraction in the doubly infinite array. That means you've arranged the fractions in linear order. There's now a first, a second, a third, and so on. Every rational number appears many times in this array (only once in lowest terms), so we have a mapping of the rational numbers onto a subset of the natural numbers. We describe this relationship by saying the rationals are countable.

Yet the real numbers, obtained by filling in the gaps in the rationals, are uncountable!

How the Real Numbers Are Uncountable

The basic infinite set is N, the natural or counting numbers. Many other sets can be matched one to one with N—for example, the even or odd numbers, the squares, cubes, or any other power, the positive and negative integers, and even the rationals, as explained in the previous article. Therefore it comes as a shock that the *real* numbers can't be put in one-to-one correspondence with the naturals. Any attempt to make a list of the real numbers is bound to leave some out!

The proof is simple.

Any real number can be written as an infinite decimal, like

3.14159 . . .

From any list of real numbers written as infinite decimals, Cantor found a way to produce another number *not on the list.* It doesn't matter how the list was constructed. So all the real numbers can never be written in a list.

How does Cantor produce his unlisted number? Step by step. It is an infinite decimal, constructed one digit at a time.

Look at the *first* real number on the list. Look at its *first* digit. Choose some other number from 0 to 9—any other number. That's the first digit in your new, unlisted number. Now go to the *second* real number on the list. Look at its *second* digit. Choose any other number from 0 to 9. That's the second digit of your new, unlisted number. And so on. The n'th digit of your new unlisted number is obtained by looking at the n'th digit of the nth real number on the list, and picking some *other* number for the nth digit in your new, unlisted number.

This construction doesn't terminate. But in calculus a number is well-defined if you can approximate it with *arbitrarily high* accuracy. By going out far enough in its decimal expansion, you approximate the unlisted number as accurately as you wish.

How do you know the new number isn't on the original list? It can't be the first number on the list, because they differ in the first digit. It can't be the second number on the list, because they differ in the second digit. No matter what n you choose, your new number isn't the n'th number on the list, because they differ in the n'th digit. The new number can't be the same as any number on the list! It's not on the list!

Geometry

What's "Between"? What's "Straight"?

What is the "straightness" of the straight line? There's more in this notion than we know, more than we can state in words or formulas. Here's an instance of this "more."

a, b, c, d are points on a line. b is between a and c. c is between b and d.

What about a, b, and d? How are they arranged?

It won't take you long to see that *b has to be between a and d.*

This simple conclusion, amazingly, can't be proved from Euclid's axioms! It needs to be added, an additional axiom in Euclidean plane geometry. This oversight by Euclid wasn't noticed until 1882 (by Moritz Pasch). A gap in Euclid's proof was overlooked for 2000 years!*

Some theorems in Euclid require Pasch's axiom. Without it, the proof is incomplete. The intuitive notion of the line segment wasn't completely described by the axioms meant to describe it.

More recently, the Norwegian logician Thorolf Skolem discovered mathematical structures that satisfy the axioms of arithmetic, but are much larger and more complicated than the system of natural numbers. These nonstandard arithmetics include *infinitely large integers.* In reasoning about the natural numbers, we rely on our mental picture to exclude infinities. Skolem's discovery shows that there's more in that picture than is stated in the Dedekind-Peano axioms. In the same way, in reasoning about plane geometry, mathematicians used intuitions that were not fully captured by Euclid's axioms.

The conclusion that b is between a and d is trivial. You see it must be so by just drawing a little picture. Arrange the dots according to directions, and you see b has to be between a and d. You're using a pencil line on paper to find a property of the ideal line, the mathematical line. What could be simpler?

But there are difficulties. The mathematical line isn't quite the same as your pencil line. Your pencil line has thickness, color, weight not shared by the mathematical line. In using the pencil line to reason about the mathematical line, how can you be sure you're using *only* those properties of the pencil line that the mathematical line shares?

In the figure for Pasch's axiom, we put a, b, c, and d *somewhere* and get our picture. What if we put the dots in other positions? How can we be sure the answer would be the same, "b is between a and d"? We draw one picture, and we believe it represents all possible pictures. What makes us think so?

The answer has to do with our sharing a definite intuitive notion, about which we have reliable knowledge. But our knowledge of this intuitive notion

* H. Guggenheimer showed that another version of Pasch's axiom can be derived as a theorem using Euclid's fifth postulate.

isn't complete—not even implicitly, in the sense of a base from which we could derive complete information.

A few simple questions to ponder while shaving or when stuck in traffic:

Is "straight line" a mathematical concept?

When you walk a straight line are you doing math?

When you *think* about a straight line, are you doing math?

Appletown, Beantown, and Crabtown are situated on a north-south straight line.

Must one be between the other two? Can more than one of the three be between two others? How do you know? Can you prove it?

Dogtown, Eggtown, and Flytown are on a *circle*, center at Grubtown. On that circle, must one be between the other two? Can more than one be between two others? How do you know? Could it be proved?

Is a straight line something you know from observation? From a definition in a book? Or how?

Is it something in your head?

Is the straight line in your head the same as the one in my head? Could we find out?

Is Euclid's straight line the same as Einstein's?

Is the straight line of a great-grandma in the interior of New Guinea the same as Hillary Rodham Clinton's? If Hillary Clinton visited her and they had a common language, could she find out?

Euclid's Alternate Angle Theorem

This is the first part of theorem 29, Book 1 of Euclid. "A straight line falling on parallel straight lines makes the alternate angles equal to one another." *Proof*: Let AB and CD be parallel. Let EF cross them, intersecting AB at G and CD at H. We claim the alternate angles AGH and GHD are equal, for they are both supplementary to angle CHG (adding to two right angles). For by construction

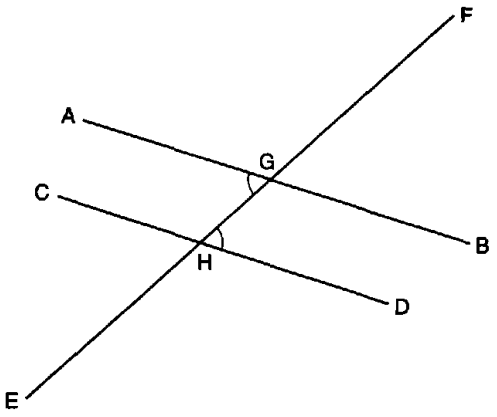

Figure 3. Alternate angles.

CHG and DHG add up to a straight angle, or two right angles. And by Euclid's fifth postulate (his definition of parallel lines) AB and CD parallel means the interior angles AGH and CHG add to two right angles.

$$\text{CHG} + \text{DHG} = 2\text{R} \qquad \text{CHG} + \text{AGH} = 2\text{R}$$
$$\text{DHG} = 2\text{R} - \text{CHG} = \text{AGH}$$

The proof is complete.

Euclid's Angle Sum Theorem

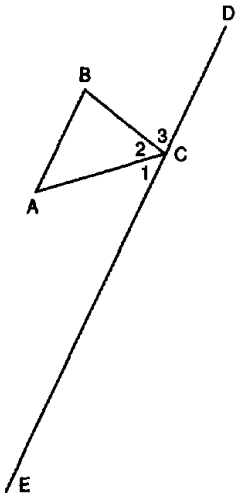

Figure 4.

In an arbitrary triangle ABC, choose a vertex, say C. Through C draw a line DCE parallel to AB. At C the three angles 1, 2, 3 add up to the sum of two right angles (180 degrees). Angle 2 is the same as angle C in triangle ABC. Angle 1 equals angle A, since they are alternate angles between two parallel lines (using Euclid's alternate angle theorem proved above). Similarly, angle 3 equals angle B. Adding,

Angle A + Angle B + Angle C =
Angle 1 + Angle 2 + Angle 3
= two right angles, q.e.d.

The Triangle Inequality

Here's an inequality valid for any six real numbers a, b, c, d, e, f:

$$\sqrt{(a-c)^2 + (b-d)^2} \leq \sqrt{(a-e)^2 + (b-f)^2} + \sqrt{(c-e)^2 + (d-f)^2}.$$

This algebraic inequality has a geometric name—the "triangle inequality."

Why?

Let the three pairs (a,b), (c,d), (e,f) be rectangular coordinates of three points P, Q, and R in the plane. Then this inequality says the distance from P to Q is less or equal the distance from P to R plus the distance from R to Q.

If P, Q, and R are vertices of a triangle, the last statement says any side of a triangle is shorter or equal to the sum of the other sides. This is the triangle inequality—than which nothing could be visually more obvious. "A straight line is the shortest distance between two points."

The complicated-looking formula is the translation into algebra of this simple geometric fact. The geometric fact "motivates" the algebraic formula. One can, with effort, give an algebraic proof of the algebraic formula, and thereby give a complicated proof of a very simple geometric fact.

But you could just as well turn the procedure around. The triangle inequality is geometrically evident. Therefore its complicated-looking algebraic statement is also true. To prove the messy algebraic inequality, use its geometric interpretation with its simple visual proof.

What's Non-Euclidean Geometry?

(See the sections on Certainty, Chapter 4, and Kant, Chapter 7.) The fifth axiom of Euclid's *Elements*, the parallel postulate, was long considered a stain on the fair cheek of geometry. This postulate says: "If a line A crossing two lines B and C makes the sum of the interior angles on one side of A less than two right angles, then B and C meet on that side."

The usual version in geometry textbooks is credited to an English mathematician named Playfair: "Through any point P not on a given line L there passes exactly one line parallel to L." This is equivalent, and easier to understand.

This parallel postulate was true, everybody agreed. Yet it wasn't as self-evident as the other axioms. Euclid's version says that something happens, but perhaps very far away, where our intuition isn't as clear as nearby. From Ptolemy to Legendre, mathematicians tried to prove the parallel postulate. No one succeeded.

Many so-called "proofs" were found. But each "proof" depended on some "obvious" principle which was only a disguised version of the parallel postulate. Posidonius and Geminus assumed there is a pair of coplanar lines everywhere equally distant from each other. Lambert and Clairaut assumed that if in a quadrilateral three angles are right angles, the fourth angle is also a right angle. Gauss assumed that there are triangles of arbitrarily large area. Each of these different-sounding hypotheses is equivalent to the fifth postulate

In the early nineteenth century Gauss, Lobachevsky, and Bolyai all had the same idea: Suppose the fifth postulate is *false*!

Euclid's axiom can be replaced in two different ways. Either "Through P pass ***more than one*** line parallel to L" or "Through P pass ***no*** lines parallel to L." The

first is called "the postulate of the acute angle." The second is "the postulate of the obtuse angle." These postulates generate two different non-Euclidean geometries, called "hyperbolic" and "elliptic." The hyperbolic was studied first, and is often referred to as just "non-Euclidean geometry."

An elegant contrast between the three geometries mentioned already in Chapter 4 is the sum of the angles in a triangle. In Euclidean geometry, as we proved above, the sum equals two right angles. In elliptic geometry, the sum of the angles of every triangle is *more* than two right angles. And in hyperbolic geometry it's *less.*

Gauss was the earliest of the three discoverers. As I mentioned earlier, in the section on Kant, he didn't publish his work, to avoid "howls from the Boeotians." In classical Athens, "Boeotions" meant "ignorant hicks." To Gauss, it meant perhaps "followers of Kant." They would say non-Euclidean geometry is nonsense, since Kant proved there can be no geometry but Euclid's.

Gauss's fear was justified. When non-Euclidean geometry became public, Kantian philosophers did say it wasn't really geometry. One of them was Gottlob Frege, the founder of modern logic.

Before Gauss, deep penetrations into the problem had been made by the Italian Jesuit priest Saccheri, by Lagrange, and by Johann Heinrich Lambert (1728–1777), a leading German mathematician who was an acquaintance or friend of Kant. Decades before Gauss, Lambert wrote:

> Under the (hypothesis of the acute angle) we would have an absolute measure of length for every line, of area for every surface and of volume for every physical space. . . . There is something exquisite about this consequence, something that makes one wish that the third hypothesis be true! In spite of this gain I would not want it to be so, for this would result in countless inconveniences. Trigonometric tables would be infinitely large, similarity and proportionality of figures would be entirely absent, no figure could be imagined in any but its absolute magnitude, astronomers would have a hard time, and so on. But all these are arguments dictated by love and hate, which must have no place either in geometry or in science as a whole. . . . I should almost conclude that the third hypothesis holds on some imaginary sphere. At least there must be something that accounts for the fact that, unlike the second hypothesis (of the obtuse angle), it has for so long resisted refutation on planes.

It's astonishing that Lambert actually gives the acute angle hypothesis a fair chance. The issue is to be decided by mathematical reasoning, not by "universal intuition." He honestly contemplates the possibility that a non-Euclidean geometry may be valid. His very ability to do so refutes Kant's universal innate Euclidean intuition.

In the end, Lambert slips into the same ditch as Legendre and Saccheri. He "proves" the Euclidean postulate by getting a "contradiction" out of the acute angle postulate. Laptev exposes Lambert's fallacy.

Beltrami, Klein, and Poincaré constructed models that showed that Euclidean and non-Euclidean geometry are "equiconsistent." If either one is consistent, so is the other. Since no one doubts that Euclidean geometry is consistent, non-Euclidean is also believed to be consistent.

Kant said that only one geometry is thinkable (see Chapter 7). But the establishment of non-Euclidean geometry offers a choice between several geometries. Which works best in physics? The choice must be empirical, to be settled by observation.

It's tempting to simply declare that "obviously" or "intuitively" Euclid is correct. This was not believed by Gauss. There's a legend that he tried to settle the question by measuring angles of a gigantic triangle whose vertices were three mountain tops. (The larger the triangle, the likelier that there would be a measurable deviation from Euclideanness.) Supposedly the measurement was inconclusive. Perhaps the triangle wasn't big enough.

What's a Rotation Group?

A "group" is a closed collection of reversible actions. For instance, multiplication by the positive real numbers is a group, since the product or quotient of two positive real numbers is a positive real number. The set of rotations in 3 dimensions is a group. Motions of your arm can be thought of as rotations around your shoulder and your elbow. So awareness of how your arm moves is an intuitive acquaintance with the 3-dimensional rotation group.

The Four-Color Theorem

In political maps, countries that share a border of positive length (not just some isolated points) are required to have different colors. It turns out that four colors always suffice to meet this requirement. This was stated as a mathematical conjecture in 1852. It was first proved by Haken and Appel in 1976. They broke the problem into a great many arduous calculations, which were performed on a computer. There followed discussion and dispute on whether this way of proving was new and different in mathematics. (See article on proof, Chapter 4.)

Two Bizarre Curves

A function has a curve as its graph; and a curve (subject to mild restrictions) is the graph of a function. Today we teach the function as primary. The graph is derived from the function. Until a hundred years ago or so, it was the other way around. As a geometric object, the curve was part of the best understood branch of mathematics. Functions leaned on geometry. Mathematicians were upset when, late in the nineteenth century, they learned of functions with wild graphs

impossible to visualize. Example 1 is the Riemann-Weierstrass curve. It's continuous, but at every point it has no direction! Example 2 is the Peano-Hilbert curve. It fills a two-dimensional region—actually passes through *every* point of a square.

I'll give brief sketches of these monsters.

For example 1, van der Waerden's construction is simpler than Riemann or Weierstrass. Start with two connected line segments in the x-y plane. The piece on the left has slope 1 and rises from the x-axis to height 1. There it meets the second piece, which descends back down to the x-axis with slope −1. At the corner where the two segments meet, the slope is undefined.

From this first step, define a second with two peaks, having slope twice that of the first, but height or "amplitude" only half as great. It oscillates twice as fast as the first, but rises only half as high.

In this manner define a sequence of graphs made of connected line segments, each half as high and twice as steep as the previous one, with corners half as far apart, or twice as frequent.

Then—add them up!

The sum converges, because the terms are getting smaller in a ratio of 1:2. As you add more and more terms, the corners get closer and closer, and the slope in between gets bigger and bigger. In the limit, the corners are dense, and the slope in between is infinite. There is no direction.

In example two, start with a square. Cut it into four subsquares, then 16 subsquares, then $256 = 16^2$ subsquares, and so on. At each stage, draw a broken line (polygonal curve) connecting the centers of all the subsquares. You obtain a polygonal line through the centers of many small squares that cover the whole original square. This sequence of polygonal curves converges to a limit curve, which actually passes through every point of the original square.

Square Circles

> Mad Mathesis alone was unconfin'd,
> Too mad for mere material chains to bind,
> Now to pure Space lifts her ecstatic stare,
> Now running round the circle finds it square.
>
> —Alexander Pope, *The Dunciad, Book IV*

This article is an imaginary conversation with David Hume, who rashly presumed there could be no such thing as a square circle (see Chapter 10).

We are not concerned here with the classic Greek problem, proved in modern times to be unsolvable, of constructing with ruler and compass a square with area equal to that of a given circle. By "circle" we mean, as usual, a plane figure in which every point has a fixed distance (the radius) from a fixed point (the center.)

Suppose I live in a flattened, building-less war zone. Transportation is by taxi. Taxis charge a dollar a mile. There are no buildings, so they can run anywhere, but for safety, they're required to stick to the four principal directions: east, west, north, and south.

People measure distance by taxi fare. If two points are on the same east-west or north-south line, the fare in dollars equals the straight-line distance in miles. Otherwise, the fare in dollars equals the shortest distance in miles, traveling only east-west and north-south.

The taxi company has a map showing the points where you can go for $1. These points form a square, with corners a mile north, south, east, and west of the taxi office. In the taxicab metric, *this square is a circle*—it's the set of points $1 from the center.

Yes, a square circle! Inconceivable, yet here it is!

But Hume says, "You can't just change the meaning of 'distance' that way. You know I mean the regular Euclidean distance."

"Very well, David. What's a square?"

"A quadrilateral with equal sides and equal angles."

"Fine. Take your regular Euclidean circle, and inscribe four equally spaced points on it. Then the circumference is divided into four equal sides, and they all meet at the same angle, 180 degrees. Another square circle!"

"No, no!" cries Hume in exasperation. "I mean a quadrilateral with straight sides, not curved sides!"

"O.K. Let the regular Euclidean circle be the equator of the earth. That's a great circle, as straight as a line can be, here on earth. Doesn't *that* have four equal *straight* sides and four equal angles?"

"No, no, no! The four equal angles have to be *right* angles!" shouts Hume.

"That wasn't part of your definition."

"Any way," says he, "call the equator straight if you like, but it isn't! It's a circle! No line on the surface of the earth is straight."

"What's this? You say that what the Mind can conceive is possible, and what the Mind can't conceive is impossible. Now you tell me there's no straight line on the surface of the earth! I grant your mind conceives that, but most minds can't conceive it. Either give up geometry, or give up your notion that what you can't conceive is impossible."

To carry the argument a step further, I leave David Hume behind, and introduce the equation

$$(x^p) + (y^p) = 1.$$

Here x and y are the usual rectangular "Cartesian" coordinates. To avoid irrelevant complications, take x and y in their absolute values, so the graphs are symmetric in the x- and y-axes. p is an arbitrary positive real number, a parameter.

For each different value of p we have a different equation and a different graph in the x-y plane.

$$\text{If } p = 2, \quad x^2 + y^2 = 1$$

The graph is the familiar Euclidean circle. For any p bigger than 1 the graph is a smooth convex curve passing through four special points:

$$(1, 1), (-1, 1), (1, -1), (-1, -1).$$

This curve is the unit circle in a new metric, where distance from the origin to the point (x,y) is defined as

$$\text{the } p\text{'th root of } (x^p + y^p).$$

Two cases are especially simple:

$p = 1$ and p infinite.
For $p = 1$,

the graph is exactly the unit square of the "taxicab metric" defined above!

For p infinite, the graph is a larger square, with horizontal and vertical sides: the four lines

$$x = +1, \quad x = -1, \quad y = +1, \quad y = -1.$$

The square for $p = 1$ is inscribed in the square for p infinite. Its corners are at the midpoints of the sides of the larger square.

Call the small square inner, and the large one outer. If you let p increase, starting with $p = 1$, the graph on your computer's monitor will expand smoothly from the inner square through a family of smooth convex curves to become the outer square. A student who watched this transformation could inform Hume, "We have infinitely many unit circles of various shapes. The first and the last are square!"

Embedded Minimal Surfaces

Soap bubbles and soap films are constrained by a physical force called "surface tension." This force makes the surface area as small as possible, subject to appropriate side conditions. In a bubble, the side condition is the volume occupied by the air inside. In a soap film, the side condition is where the film is attached to something—a bubble pipe, another bubble, or the fingers of a child.

It's easy to guess or observe that for a soap bubble the minimal surface is a sphere. For a soap film, with endless different possible boundaries, the problem is more complicated.

One basic fact is clear. In order to have minimal area "in the large" (globally) the film must have minimal area "in the small" (locally). That is, if you mentally mark out any small simple closed curve in the soap film, the area of film

enclosed by that curve must be the smallest area that can be enclosed by that curve.

Because of this "local" property, the soap film surface satisfies a certain complicated-looking partial differential equation discovered by Joseph-Louis Lagrange in 1760.

This suggests the famous "Plateau's problem": Given a curve in 3-space, construct a soap film (a minimal surface) having that curve as boundary. J. A. F. Plateau was a blind Belgian physicist who made the problem known to the world in 1873. Jesse Douglas, a New York mathematician, won a Fields Medal in 1936 for his solution.

In the late nineteenth-century Karl Weierstrass, Bernhard Riemann, Hermann Amandus Schwarz, and A. Enneper discovered a number of interesting new minimal surfaces. For years the corridors of university mathematics departments were lined with glass-fronted cabinets displaying plaster models of these surfaces.

It turned out that a mathematical minimal surface need not be a physical one. The mathematical conditions permit the surface to intersect itself, which soap film doesn't ordinarily do. A surface is called "embedded" if it has no self-intersection.

The use of computer graphics, starting some fifteen years ago, has revitalized the subject by making possible the visualization of complicated surfaces formerly described only by equations. The computer pictures often reveal instantly whether a surface has self-intersections, and may show other properties of the surface that can be used to provide rigorous proofs. An infinite number of new complete embedded minimal surfaces have been found in this way. David Hoffman and Jim Hoffman, whose computer-generated video is the basis of our dust-jacket picture, have been leaders in this work.

How Imaginary Becomes Reality

In these notes we have already "constructed" the natural numbers from the empty set. Now we go the rest of the way. Step by step, we construct the integers (positive and negative whole numbers); then the fractions or rational numbers; then the real numbers, rational and irrational; and at last the complex numbers. I show five different ways to construct the complex numbers! And then we go even further, to exotic creatures called quaternions. These extensions of number systems show where the axiom-theorem model is misleading. We don't just obey axioms, we modify them.

Creating the Integers

From the natural numbers we wish to construct the integers—the natural numbers *plus* zero *plus* the negative whole numbers. We can subtract one natural

number from another—for instance, 3 from 7—if the former isn't bigger than the latter. But with the natural numbers, we can't subtract a larger from a smaller. "Seven from three you can't take," say the first-graders.

Mathematicians have two ways to extend a mathematical system. One is brute force—create a needed object by fiat. For instance, there's no natural number x such that

$$x + 1 = 0.$$

But we might need such a number. (For instance, to keep track of money we owe.) No problem. Just make up a symbol: -1 and state *as a definition* that

$$-1 + 1 = 0.$$

That's how it's done in school.

But can we really create what doesn't exist, just by definition? Who gave us a license for such presumption?

The fact is, it's not necessary to "create" anything. We can use a more sophisticated approach known as "equivalence classes." Since we want -1 to equal $(0 - 1)$ and $(1 - 2)$ and $(2 - 3)$ and so on, we just collect all these ordered pairs into a great big "equivalence class":

$$\{ (0 , 1), (1, 2), (2, 3) \ldots \}$$

It includes infinitely many elements; each element is an ordered pair of natural numbers. Yet it deserves and has the name you expect: -1.

My definition of equivalence class was vague. I just wrote down three members of this infinite class "to give you the idea," and then wrote three dots. . . . I can't possibly write the complete list. We need a membership rule for the class.

We find this by elementary-school arithmetic. When does

$$a - b = c - d$$

without using minus signs? Of course, when

$$a + d = b + c.$$

So we make a precise definition of equivalence:

$(a - b)$ is equivalent to $(c - d)$ if and only if
$a + d = b + c$.

To make a fresh start, not depending on a fiat, we temporarily give up the expression $a - b$, and instead write $\{a,b\}$—an ordered pair in curly brackets.

$$\{a, b\} = \{c, d\} \text{ if and only if } a + d = b + c.$$

We have to figure out how to add, subtract, and multiply the equivalence class of $\{a,b\}$ and the equivalence class of $\{c,d\}$. (Division isn't generally possible,

since we don't have fractions yet.) Now, from junior high school we know the rules to add, subtract, and multiply positive and negative whole numbers. It's simple to rewrite those rules in terms of ordered pairs in curly brackets. You can do it, if you wish.

One class plays a special role:

$$\{ (0, 0), (1, 1) \ldots \}.$$

It's the additive identity, and has a special name—zero. But that name is already in use, for the whole number before 1. So we allow the same word to have two meanings.

We say two classes are "negatives" of each other, or "additive inverses," if their sum is zero. You can easily check that for any natural numbers a and b, the class of (a, b) is the negative of the class of (b, a). That is to say,

$$\{a, b\} + \{b, a\} = \{0, 0\}.$$

This means, in particular, that each equivalence class has one and only one negative or additive inverse.

In every equivalence class, there's exactly one ordered pair that includes the natural number 0. If that pair is {a, 0}, we call that equivalence class "a." (It's in a sense the "same" as the natural number a.) And if that pair is {0, a}, we call it −a.

We Have Constructed the Integers, Including the Negative Numbers!

"Constructed" out of some infinite sets, to be sure. Rather than ordered into existence "by fiat."

Why $-1 \times -1 = 1$

In extending from positive whole numbers to integers we preserve *all* the rules of arithmetic except *one*. The rule we give up is:

No number comes before 0.

But we still have:

Rule A $\quad a \times 0 = 0 \times a = 0,$ for all a.

So in particular

Rule A' $\quad -1 \times 0 = 0.$

And we still have

Rule B $\quad a \times 1 = 1 \times a = a,$ for all a.

So in particular

Rule B' $\quad -1 \times 1 = 1 \times -1 = -1$

And we still have the distributive law:

Rule C $x \times (y + z) = (x \times y) + (x \times z)$.

Consider

$(-1) \times (-1 + 1)$.

This is -1×0, which is zero, by Rule A'. On the other hand, by Rule C,

$(-1) \times (-1 + 1) = (-1 \times -1) + (-1 \times 1)$.

The last term on the right equals -1, by Rule B'. So

$0 = (-1 \times -1) + (-1)$.

That is to say,

(-1×-1)

is the additive inverse of -1. But -1 has a unique additive inverse: 1.

So -1×-1 is 1, as we claimed.[3]

Creating the Rationals

In the course of human progress people acquired property and money. So they needed fractions. When a baker had a whole loaf and a half, he had to know he had three halves to sell. Anybody can add

$$\frac{1}{2} + \frac{1}{2} = \frac{2}{2} = 1.$$

Confusion arises with improper fractions:

$$\frac{4}{5} + \frac{4}{5} = \frac{8}{5},$$

and still worse with unequal divisors:

$$\frac{1}{3} + \frac{1}{8} = \text{??}.$$

But people did manage to extend addition and multiplication to fractions (both positive and negative.) This enlarged system is called "the rational numbers." ("Rational" meaning, not "reasonable" or "logical," but just *ratio* of whole

[3] Thanks to Howard Gruber for suggesting this example.

numbers.) With these numbers, any problem of addition, subtraction, multiplication, or *division* (except by zero) has a solution.

We pay a penalty for this enlargement. The natural numbers are ordered "discretely"—every one has a unique follower, and all but 1 have a unique predecessor. This beautiful property makes possible a powerful method—"proof by induction." It's no longer true for the rationals.

Ordinarily we write fractions as a/b, but I will temporarily write them as ordered pairs (in *square* brackets, [a,b]).

Since

$$\frac{1}{2} = \frac{2}{4} = \frac{3}{6},$$

the rule now is,

$$[a,b] = [c,d] \text{ (or } a/b = c/d) \text{ if } ad = bc.$$

This defines "equivalence classes of pairs of integers"—what we call "equal fractions" in the fourth grade. The ones we used when we practiced reducing fractions to lowest terms.

With fractions, as with negatives, we need rules for calculating with these new ordered pairs. And again, we just take the known rules for fractions and rewrite them in terms of ordered pairs in square brackets. We have constructed the rational numbers! Jacob Klein shows that to the Greeks, number meant "positive whole number greater or equal to 2." Number 1 wasn't like other numbers. Fractions were a commercial and practical necessity, but they weren't *numbers.* Klein writes that the broadening of "number" to include positive fractions took place only in the late middle ages and early Renaissance, and with difficulty.

Why $\sqrt{2}$ Is irrational

The most famous theorem in Euclid is the "Pythagorean": "In any right triangle, the sum of the squares of the lengths of the two shorter sides equals the square of the length of the long side" (the "hypotenuse"). You can construct a pair of right triangles by drawing a diagonal in a square of side 1. Then Pythagoras's theorem says the diagonal has length $\sqrt{2}$. On the other hand, the Pythagoreans also discovered that *there is no ratio equal to* $\sqrt{2}$!

Since it doesn't exist, there's nothing to exhibit or construct. All the proof can do is show that the presumption such a ratio exists is absurd. This is called "indirect proof." Suppose that for some pair of numbers p and q,

$$(p/q)^2 = 2.$$

If so, p/q can be put in "lowest terms"—p and q should have no common factor. In particular, they don't both have 2 as a factor—they aren't both even. Multiplying both sides by q^2 gives

$$p^2 = 2\,q^2.$$

A factor 2 is visible on the right side, so the right side of the equation is an even number. Therefore p^2, the left side of the equation, is also even. It's easy to check that the square of an even number is always even, the square of an odd number always odd. Since p^2 is even, p is even. That means p is twice some other whole number. Let's call it r, so $p = 2\,r$. Then

$$p^2 = 4\,r^2.$$

We replace p^2 by $4\,r^2$ in the previous equation, and get

$$4\,r^2 = 2\,q^2,$$

which simplifies to

$$2\,r^2 = q^2.$$

This is just like the equation $p^2 = 2\,q^2$ we started with, but p is replaced by q and q by r. So the same argument as before proves that q is even, as p was proved to be even. But p and q aren't allowed to both be even. CONTRADICTION!

The contradiction shows that the presumption that such a fraction p/q exists is impossible, or absurd.

If we only have whole numbers and ratios, we're stuck with the conclusion that $\sqrt{2}$ doesn't exist. It exists as a line segment, the diagonal of the unit square, but not as a number. The diagonal of the unit square does not have a length! Yet, using operations of Euclidean geometry, we can add it to other line segments, and also subtract, multiply, and divide. Line segments constitute an arithmetical system richer than the system of arithmetical numbers! This impasse suggests we go beyond the rational numbers. We need a theory of irrational numbers.

Creating the Real Numbers—Dedekind's Cut

So we want the "real numbers"—rationals and irrationals together. (The name "real" is in contrast to the imaginary and complex numbers, which we will meet shortly.) We use $\sqrt{2}$ to motivate our construction of the irrationals. No rational number when squared can equal 2 (proof is above). Yet we can approximate $\sqrt{2}$:

1
1.4
1.41
1.414

and so on, as far as our computing budget permits. This sequence converges, but what does it converge *to*? $\sqrt{2}$, naturally. But what *is* $\sqrt{2}$, if it can't be a rational number?

We want mathematics to include $\sqrt{2}$ —and many other irrational numbers, of course. We have to somehow take such "convergent" sequences of rationals, which don't have rational limits, and make them into numbers—"real numbers."

Georg Cantor, Karl Weierstrass and Richard Dedekind each found a way to do this. Dedekind's is especially easy.

Arrange the rational numbers in a row or a line in the usual way, increasing from negative to positive as you go from left to right. By a "cut" Dedekind means a separation of this row into two pieces, one on the left, one on the right. The row can be cut in infinitely many different places. Dedekind regards such a split or "cut" in the rationals as being a new kind of number! He shows in a natural way how to add, subtract, multiply, or divide any two cuts (not dividing by zero, of course). In an equally natural way, he defines the relation "less than" for cuts, and the limit of a sequence of cuts. Once these rules of calculation are laid out, the cuts are established as a number system.

Every rational number x defines an associated cut. The left piece is simply the set of rational numbers less or equal x, and the right piece is the set of rationals greater than x. By this association between cuts and rational numbers, we make the rational numbers a subsystem of the system of cuts. To identify Dedekind cuts as the sought-for "real number system," we must show that they include *all* the rationals and irrationals—all the numbers that can be approximated with arbitrary accuracy by rationals.

I'll be satisfied to show that one particular irrational is included as a Dedekind cut— $\sqrt{2}$. To do so, I must identify a left half-line and right half-line associated with $\sqrt{2}$. What rationals are less than $\sqrt{2}$? Certainly all the negative ones, and also all those whose squares are less than 2. All numbers x such that either $x < 0$ or $x^2 < 2$. That specifies the left piece of the cut, the left half-line associated to $\sqrt{2}$. Its complement is the corresponding right half-line. It's easily verified that when this cut is multiplied by itself, it produces the cut identified with the rational number 2. Among Dedekind cuts 2 does have a square root!

All that's left to prove is that no numbers are missing. Dedekind's cuts provide a limit for every convergent sequence of rationals, but we need more. We need a limit for every convergent sequence of *real* numbers—every convergent sequence of cuts. This property is called completeness. The proof is in every text on real analysis and many texts on advanced calculus. I give the essence of it. Let a_n be a convergent sequence of Dedekind cuts (real numbers.) We want to produce a cut a which is the limit of this sequence. We know that every cut a_n is the limit (in many ways) of a convergent sequence of rational numbers. So we replace each a_n by an approximating rational number, choosing the rational approximation more and more accurately as we go out in the a_n sequence. This is easily shown to be a convergent sequence of rationals, and it's easily shown that its limit cut is the limit of the original sequence of cuts.

These constructions are "existence proofs." If you believe Dedekind cuts exist, you have proved that the real numbers exist.

What's the Square Root of −1?

Does $\sqrt{-1}$ exist? There's no real number that yields −1 when squared. That's the reason we say $\sqrt{-1}$ doesn't exist.

Yet in our next breath we bring it into existence!

I'll show you five different ways to do it.

The simplest way is the high-school way. Just ***define*** i as a "quantity" that obeys the laws of arithmetic and algebra in all respects, except that

$$i^2 = -1.$$

If you wish, instead of a "quantity" you may call it a "symbol," which ***by definition*** satisfies

$$i^2 = -1.$$

This approach is direct. It is clear cut. i is treated algebraically like any "letter" or "indeterminate." It can be added and multiplied. These operations and their inverses obey the same commutative, distributive, and associative laws as the real numbers do. The only difference is

$$i^2 = -1.$$

Real multiples of i, like $2i$ or $-3i$, are called "imaginary" or "pure imaginary." Numbers of the form $z = x + iy$, where x and y are real numbers, are called "complex." x is called "the real part" of z, and y is "the imaginary part."

Either x or y or both can be 0, so the imaginary numbers and the real numbers are among the complex numbers! (0 is the only complex number that is both real and imaginary.) This shouldn't be a shock. The positive and negative whole numbers (the integers) are among the rational numbers (fractions.) When we enlarge a number system, we want the numbers we start with to be included among the numbers we "construct."

But since no real number satisfies $x^2 = -1$, is it legitimate to simply "introduce" the square root of −1? Isn't this cheating?

We've seen that pretending some number equals $1/0$ leads to disaster. If $1/0$ is fatal, how can we be sure $\sqrt{-1}$ is O.K.?

One answer might be that analysis with complex numbers is a powerful theory that has never led to a contradiction. That would be saying, "We never had trouble so far, so we never will have trouble." A dubious defense. To resolve such worries, we renounce "introducing" or "creating" the square root of -1. Instead, we'll ***find*** it, already there! As promised, we'll do it in five different ways.

1. A point in the x-y coordinate plane.
2. An ordered pair of real numbers.
3. A 2-by-2 matrix of real numbers.

4. An equivalence class of real polynomials.
5. In the Grand Universal Super-Structure of Sets.

1. After centuries of skepticism, mathematicians accepted complex numbers when they found them "already there," as points in the x-y plane. The complex number $3 + 4i$, for example, is associated to the point with coordinates $x = 3$, $y = 4$. In this way, every complex number gets a point in the coordinate plane, and every point in the plane gets a complex number.

Addition and multiplication of complex numbers turn out to be elementary geometric operations! Addition is just shifting. Adding $3 + 4i$, for example, shifts any complex number 3 units to the right and 4 units up.

Multiplying is stretching and turning. To see this, use polar coordinates. The "polar distance r" of a point $x + iy$ is its distance from the origin. For

$$3 + 4i,$$

the Pythagorean theorem gives $r = 5$.

The "polar angle Q" of a point is the angle between the positive x-axis and the ray from the origin to that point. Multiplying by $3 + 4i$ then turns out to be simply ***multiplying*** distance by $r = 5$ and ***increasing*** polar angle by Q.

For $i = 0 + 1i$, evidently $x = 0$ and $y = 1$. The point corresponding to i is on the (vertical) y-axis. So we call the y-axis the "imaginary axis." The "imaginary unit" i is there, one unit above the origin. The (horizontal) x-axis is the "real axis."

For the point i, polar distance r is 1, and polar angle Q is a right angle, 90 degrees.

Multiplying $i \times i$ results in ***squaring*** r and ***doubling*** Q.

Since $r = 1$, $r^2 = 1$.

Since Q is a right angle, 90 degrees, its double is two right angles—180 degrees.

This means that i^2 is on the x-axis (the real axis) one unit *left* of the origin. It has coordinates $(-1,0)$. Its complex number is $-1 + 0i$, or simply -1. We have demonstrated geometrically that

$$i^2 = -1.$$

That is, the point i or $0 + i$ is a square root of -1!

Since classical times geometry was the most venerated part of mathematics. Identifying the complex numbers with plane geometry made them respectable.

2. From a more critical viewpoint, something is still missing. The complex numbers are defined by laws of arithmetical operations. They're an independent ***algebraic*** system, defined prior to their geometric interpretation. We should give an ***algebraic*** proof of consistency. This was done by Ireland's greatest mathematician, William Rowan Hamilton (remembered also for quaternions, the Hamilton-Jacobi equations, and Hamiltonian systems of differential equations).

To construct the complex numbers, Hamilton creates from the real numbers a simple new kind of thing: an ***ordered pair*** of real numbers. This will look a lot like how we constructed the integers and the rationals—but historically Hamilton's construction of the complex numbers came first!

He defines equality of his ordered pairs:

$$(a, b) = (c, d) \text{ if and only if both } a = c \text{ and } b = d.$$

He defines addition in a very natural way:

$$(a, b) + (c, d) = (a + c, b + d)$$

Multiplication is more complicated:

$$(a, b) \times (c, d) = (ac - bd, ad + bc).$$

Hamilton didn't pull this multiplication rule out of thin air. He just translated the known multiplication of $(a + bi)$ times $(c + di)$ into his notation of ordered pairs. Seen this way, the whole performance looks trivial. But it gets rid of the suspicious i, and replaces it by the innocent (0, 1). Please check that its square is $(-1, 0)$, which is $-1 + 0i$, which is -1.

One should verify the arithmetical laws that complex numbers share with real numbers: commutative laws of addition and multiplication, associative laws of addition and multiplication, and distributive law of multiplication over addition. These verifications are straightforward calculations that the interested reader can carry out.

Notice that ordered pairs whose second component is 0 behave just like real numbers. The zero in the second place never "gets in the way." The multiplicative identity is (1,0); it's algebraically "the same" as 1, the multiplicative identity of the reals.

The pair $(-1, 0)$ is algebraically "the same" as the real number -1. The additive identity is (0, 0); it's algebraically "the same" as the real number 0.

It's straightforward to define subtraction:

$$-(3, 4) = (-3, -4) \qquad -(a,b) = (-a, -b)$$

and to check that

$$(a, b) + (-(a, b)) = (0, 0).$$

Division is trickier. I'll save time by just telling you how to do it—you can check that it works. First, for the special example (3,4),

$$\frac{1}{(3, 4)} = \frac{(3, -4)}{(3^2 + 4^2)} = \left(\frac{3}{25}, \frac{-4}{25} \right).$$

And in general,

$$\frac{1}{(a, b)} = \left(\frac{a}{[a^2 + b^2]} \quad , \quad \frac{-b}{[a^2 + b^2]}\right).$$

which you can check by multiplying

$$(a,b) \times \frac{1}{(a, b)}$$

using the multiplication rule given above. The answer is (1 , 0), or simply 1.

If the definitions of multiplication and division seem baffling, go back to the geometric interpretation of complex numbers to make them intuitively clear.

From a strict formal point of view, one oughtn't to write

$$a + 0i = a.$$

That's "equating apples and oranges." A single real number a just isn't the same as the pair of real numbers (a, 0). Instead of "=" one could say "is isomorphic to."

3. Another way to construct complex numbers uses 2 × 2 matrices of real numbers instead of ordered pairs. The complex number $a + bi$ corresponds to the matrix

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

If you know how to multiply 2 × 2 matrices, you can check that the usual rules of matrix algebra correspond to the usual rules of addition and multiplication of complex numbers. The number −1 corresponds to the matrix

$$\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

The matrix

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

gives −1 when squared. We are entitled to call this matrix "i"!

We found a square root of −1 by interpreting −1 as a 2 × 2 matrix. What does this say about existence of $\sqrt{-1}$? It exists if you interpret −1 the right way!

4. A fourth way of finding $\sqrt{-1}$ is inspired by a branch of modern algebra called Galois theory, after Evariste Galois. He was a student, killed in a duel in 1838 at age 21, before being recognized as a precocious genius.

Instead of matrices or ordered pairs we use "polynomials with real coefficients." For instance,

$$5x^4 + 3x^3 + 7x^2 - x^1 + 5.$$

We divide all our polynomials by $x^2+ 1$. Why? Because the thing we're after, i, is a root of $x^2 + 1$!

As in division of numbers, so in division of these polynomials, we get a quotient and a remainder. And the remainder has ***degree*** less than the ***degree*** of the divisor. We're dividing by $x^2 + 1$—a second degree polynomial—so the remainder has degree 1 or 0. There might be ***zero*** remainder, or a constant remainder different from zero (a zero-degree polynomial), or a first-degree remainder—a polynomial of the simple form $ax + b$.

Two polynomials, whatever their degree, are ***equivalent*** if they have the ***same remainder*** on division by $x^2 + 1$. This equivalence splits the polynomials into equivalence classes—sets of polynomials having the same remainder on division by $x^2 + 1$.

An equivalence class is a sack. We're putting polynomials into sacks. All the polynomials in any sack have the same remainder, which is some polynomial of degree 1 or 0. The polynomials that are multiples of $x^2 + 1$, including the number 0, all have zero remainder, so that sack, or if you will that equivalence class, is the zero class, the zero of this algebra.

It's straightforward to define operations between classes or sacks—multiplication, addition, subtraction, division, additive inverse, multiplicative inverse. Everything is done

"mod $(x^2 + 1)$."

Meaning: "Whenever $x^2 + 1$ shows up, throw it away." It's equivalent to zero, because the remainder of $(x^2 + 1)$ on division by $(x^2 + 1)$ is 0.

The multiplicative inverse of the sack of polynomials with remainder $(ax + b)$ is the sack of polynomials with remainder

$$\frac{(-ax + b)}{(a^2 + b^2)}$$

Why? When multiplied together, they yield a polynomial whose remainder is 1, which, naturally, is the multiplicative unit in this algebraic structure.

And of course its additive inverse, which we denote by -1, is the sack of polynomials that leaves the remainder -1.

Now the big question. What about a square root of -1? The answer is so easy, it feels like swindle.

Since $x^2 + 1$ is equivalent to 0, or as an equation, $x^2 + 1 = 0$,
then subtracting 1 from both sides,

$$x^2 = -1.$$

Hey, that's it! We've found the thing that when squared equals minus one! It's the equivalence class containing the simple special polynomial x. If you prefer, its the class with remainder $ax + b$, where $a = 1$, $b = 0$.

In a fussier notation, $x^2 = -1$ modulo $(x^2 + 1)$. So x^2 is "congruent" (equivalent) to -1, and x is equivalent to $\sqrt{-1}$.

I'll say it once more. x^2 is equivalent to -1 because both give the same remainder on division by $x^2 + 1$. (Or, put even more simply, adding 1 to either gives 0.) If x^2 is equivalent to -1, that means x is equivalent to the square root of -1. So x in our algebra of polynomial equivalence classes is "the same" as the complex number i! Polynomials with remainder 1 are equivalent to the real number 1. A combination of x and 1, say, $ax + b$, corresponds to the complex number $ai + b$. All our equivalence classes correspond to remainders of the form $ax + b$, so the equivalence classes and the complex numbers are in a one-to-one correspondence. They're "isomorphic." These sacks correspond precisely to the complex numbers!

Why go through all this when we can just adjoin i? Because adjoining something new and prescribing rules for it to follow is a leap in the dark. In using equivalence classes, on the other hand, we add nothing and risk nothing. We just notice what's there. The step from real numbers to real polynomials involves bringing in x, but we don't require x to satisfy any weird conditions (like $x^2 = -1$.) We just divide by the polynomial

$$x^2 + 1$$

and look at the remainder. Given two polynomials, we can find their remainders on division by

$$x^2 + 1$$

and see if they're the same or different. This relation automatically sorts the polynomials into classes. Then behold! These equivalence classes are the complex numbers!

Let's compare the three constructions—by ordered pairs, by 2×2 matrices, and by polynomials mod $(x^2 + 1)$.

The construction by ordered pairs uses an algebraic structure created specifically for constructing the complex numbers. Conceptually and computationally, it's the simplest.

The construction by matrices uses something already available—the algebra of 2×2 matrices. It isolates a special subset of them—those whose diagonal elements are equal, and whose off-diagonal elements are equal in absolute value but opposite in sign. One checks that this matrix algebra is closed—sums, products, and inverses of matrices of this type are again of this type. Then, since we know that the identity element is

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

we know that -1 corresponds to

$$\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

and simply check that

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

squared is

$$\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

Just call

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

"i," and you have

$$i^2 = -1$$

You could say Hamilton "constructed" the complex numbers with his algebra of ordered pairs. In the matrix approach, you can't say anything has been *constructed*—the matrices are here already. You might say we "isolated" or "discovered" the complex numbers embedded in the algebra of 2×2 matrices.

What about the method of polynomials mod $(x^2 + 1)$? Here the objects that correspond to Hamilton's ordered pairs or to 2×2 matrices are equivalence classes of polynomials mod $(x^2 + 1)$. (See section below on "Equivalence Classes.") We take all the polynomials that have the same remainder, say $2x + 1$, and throw them into the same sack. We think of the sackful—the whole class of mutually equivalent polynomials—as a single object, which can be added to or multiplied by any other equivalence class. The multiplicative identity is the class of all polynomials with remainder 1. -1 is the class of all polynomials with remainder -1. x is a square root of -1, because the remainder when x^2 is divided by $(x^2 + 1)$ is -1.

Proof: $x^2 = [1 \times (x^2 + 1)] - 1$.

5. Finally, let's see how the complex numbers might be regarded by some anonymous set theorist.

There are two approaches to set theory. One is axiomatic. If something satisfies the 12 axioms of Zermelo and Frankel, it exists. The other way is constructive. Start with the empty set, and step by step, using axiomatically authorized set-theoretic operations, construct ever bigger uncountable sets. Everything you can get by iterating uncountably infinitely often the set-theoretic operations of enlargement is thought to have *already* existed, in advance. Modern set theory is a fascinating and difficult study.

The most famous example of constructing by means of equivalence classes was Frege's "construction" of the natural numbers as equivalence classes of sets. He ended his career resigned to the failure of his set-theoretic foundation. Yet those ideas continue to permeate philosophical logic and set theory.

Where does this put the complex numbers? In the number system they're at the top of the heap, but in the grand set-theoretic structure they're near the bottom. Whether there's a number whose square is -1 is of little set-theoretic interest. But if by chance you want a square root of -1, you have to look in the set-theoretical structure. There isn't any place else!

Recall Hamilton's ordered pairs. Forming ordered pairs is licensed by one of Zermelo's axioms, so all ordered pairs always existed, whether Hamilton knew it or not. Hamilton gave his ordered pairs an algebraic structure. How is that algebraic structure understood set-theoretically? To explain, let's go back to multiplication of natural numbers. We get a natural number as product. The formula $a \times b = c$ describes a *function*, the "times" function, which operates on the pair a, b and yields the value c. So the formula is "really" a set of ordered triples, a, b, c. Therefore, it's a subset of the set of *all* ordered triples. Since it's a subset of a set, it's a set—it exists, by another of Zermelo's axions. This "proves," in a certain strange sense, that multiplication of natural numbers exists. If we go from the natural numbers to ordered pairs (rational numbers) we get the set of all ordered triples of pairs—sextuples. A certain subset of that set represents multiplication of rational numbers. It wouldn't be essentially different to treat a *pair* of operations, like "plus" and "times." Proceeding further, we would find that the real numbers, the complex numbers, and all their operations, are already there in the grand set-theoretic super-universe. The problem is to find them. That means showing that certain sets have certain required properties. To do that requires the same checking we've been doing.

Are more representations of the complex numbers waiting to be discovered? If you look in the right math book, you'll find a theorem, "There's only one system of complex numbers." If we line up our representations carefully, they look like merely verbal variants of each other. In Hamilton's ordered pairs, (1,0) is the multiplicative unit, and (0,1) is the imaginary unit. In the matrix representation of complex numbers,

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

is the multiplicative unit, and

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

is the imaginary unit i. The correspondence is obvious.

In the polynomials mod $(x^2 + 1)$, the class of polynomials having remainder $1 + 0x$ is the multiplicative unit, the class having remainder $0 + 1x$ is the imaginary unit. So the standard complex number $a + bi$, the ordered pair (a, b), the matrix

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

and the equivalence class of a + bx are four names for the same thing.

These correspondences between algebraic systems are called "isomorphisms." They are one-to-one invertible mappings, which preserve algebraic structure. In mathematics teaching an impression is often given that isomorphic systems should be regarded as the same.

The difference between (a, b), the equivalence class of a + bx and

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

is regarded as trivial or meaningless. Or their difference is mere notation, like the difference between x^2 as a function of x and t^2 as a function of t.This would be an error. The mathematician describes this situation by saying that the same structure has several representations. The structure is the abstract thing that each representation represents, in a particular language and from a particular viewpoint. An investigation often is possible only by means of some concrete representation. It can be advantageous to have several representations. Several famous theorems are representation theorems—the Riesz theorems, the Radon-Nikodym theorem, the spectral theorems. An attitude that structure is all, representations are trivial, is a serious misrepresentation.

This question comes up in use of coordinates in geometry. Geometric results should be independent of coordinates; therefore, the story goes, they should be proved without coordinates. Yet the coordinate proof may be more accessible to find and to teach.

Anything that claims to be a new representation must be *substantially new*. Somebody could report a new representation for complex numbers by using 3×3 matrices instead of 2×2. He could simply augment the 2×2 representation by one more row and one more column, all zeroes. This would be new formally but not substantially. Such a change from 2×2 to 3×3 would be uninteresting and obvious. What we think interesting today isn't always what Euler thought interesting, nor what geniuses in 2997 will consider interesting. This question is esthetic. Esthetic questions play a small part in deciding what's correct, a major part in deciding what's interesting. Esthetic considerations are spared little space in the journals, but they're crucial for understanding the development of mathematics.

At present our number system looks stable, although Abraham Robinson's nonstandard real numbers have proved their worth, and John Conway's "surreal numbers" may have a future.

What Are Quaternions?

Hamilton's passion was to find a number system to do for 3 dimensions what the complex numbers do for 2. To define an algebraic structure, each element of which could be identified with a point of x-y-z-space, with addition and multiplication corresponding to translation and rotation in 3-space.

This proved impossible. Hamilton came as near as anyone could.

His quaternions include three independent "imaginaries," i, j, and k. Each of them squared yields −1!

A general quaternion has the form

$$a + bi + cj + dk$$

where a, b, c, d are real numbers.

To multiply quaternions, you have to multiply i, j, k by each other. This was the hard part. Hamilton discovered the system worked if

$$ij = k \qquad jk = i \qquad ki = j$$
$$ji = -k \qquad kj = -i \qquad ik = -j$$

The commutative law has vanished! Instead, an "anticommutative law." This was the first time anyone imagined an algebraic structure without commutativity. Hamilton was so delighted that he carved

$$ij = -ji$$

on a bridge he crossed going to church.

Hamilton and his disciples tried hard to make quaternions useful in mathematical physics. But Gibbs's vector analysis accomplished similar things more conveniently.

Quaternions are hyper-complex numbers. They add, subtract, multiply, and divide. Gibbs's vectors, on the contrary, have two different multiplications, but no division.

Quaternions are four-dimensional—a 3-vector linked to a number. They don't fit in higher dimensions. Gibbs vectors generalize to any dimensions.

Crowe reports the competition between quaternions and Gibbs-Heaviside vectors for modeling electromagnetism. Both formalisms can describe electromagnetic fields. But physicists preferred the one they found more convenient for calculation—Gibbs vectors.

Do and did quaternions exist? They existed as mathematical concepts from the day Hamilton discovered them. But they weren't sitting on that Irish bridge from the beginning of time, patiently waiting to be discovered. And they didn't start to exist on the day Hamilton started trying to fit them to physics.

They're a permanent piece of algebra, and they continue to be proposed for use in physics and engineering. But from the viewpoint of Platonist set theory, the quaternions were always ready and waiting in the grand abstract universal set structure, their anticommutative multiplication merely a certain subset of the set of all sets of sets of sets of sets of empty sets.

Extension of Structures and Equivalence Classes

The extensions of number systems we have just presented are in a sense optional, but in a stronger sense not optional. Nothing in the natural numbers *logically*

forces us to introduce negatives. The enlargements to integers, rational numbers, real numbers, and complex numbers were all compelled, slowly and reluctantly.

For another example of how these optional enlargements are in a deep sense compulsory, look at the Fibonacci numbers. These are the sequence

$$1, 1, 2, 3, 5, 8, 13, 21 \ldots$$

Each number after the first two is the sum of the two previous ones.

It's obvious that all the Fibonacci numbers are positive integers.

A little analysis shows that they're all combinations of the solutions of this quadratic equation:

$$x^2 = x + 1.$$

You can solve this equation with your high-school quadratic formula. You find two roots, both involving the square root of 5 (in a combination known to fame as the "golden ratio"). Both roots are irrational. But if you combine them, with the right irrational coefficients, you get the Fibonacci numbers! This is a sequence of natural numbers, yet to write a formula for them, we're forced to use an irrational number, $\sqrt{5}$!

It's enough to make you think $\sqrt{5}$ was already there when we learned to count—or even before, since, as Martin Gardner tells us,

$$2 + 2 = 4$$

was already true with the dinosaurs.

No wonder the mathematician in the street thinks $\sqrt{5}$ existed even before Fibonacci.

Another example. The infinite series

$$1 + x^2 + x^4 + x^6 + \ldots$$

converges if the absolute value of x is less than 1. It diverges if absolute x is greater than 1. To see why, notice that if absolute x is greater than 1, then the terms farther and farther out in the series get bigger and bigger. But for convergence they must get smaller and smaller.

For absolute value of x less than 1, this series sums to

$$\frac{1}{(1 - x^2)}$$

This fraction blows up when $x = 1$, because it becomes $1/0$. This is a good reason why the series can't converge for $x = 1$.

On the other hand, there's the series

$$1 - x^2 + x^4 - x^6 + \ldots$$

Like the previous one, this converges if absolute x is less than 1, and diverges for absolute x greater or equal 1. It sums to

$$\frac{1}{(1 + x^2)}$$

This denominator is always greater than or equal to 1, so this fraction doesn't blow up for any real x. Then why should the series diverge, if it's equal to a fractional algebraic expression that is well behaved for all real x?

Try replacing x by $z = x + iy$. That means, let the independent variable run around the complex x-y plane, not just the real x-axis. If you choose $z = i$ (the square root of -1) then the denominator *is* zero—the fraction blows up. There's a singularity on the *imaginary* axis at $z = i$, one unit away from the real x-axis. The singularity on the *imaginary axis* is responsible for divergence on the *real* axis! A phenomenon in real analysis, which, in a reasonable sense, can't be understood in terms of real numbers only. The complex numbers, whether or not we recognize them, are already controlling some of our real-number computations!

Finally, consider the trigonometric functions sin x and cos x and the exponential function e^{ax}, where a is some positive real number that we can choose at will. If the variable x is real, the behavior of this exponential function is completely different from that of the trigonometric functions. The exponential grows steadily from zero at $x = -\infty$, and its rate of growth is a. If a is any positive number, the exponential function grows faster and faster as x increases. The trigonometric functions, in contrast, remain bounded for all x, however large. They oscillate periodically between a minimum of -1 and a maximum of 1. By use of complex numbers, Euler made the astounding and brilliant discovery that these functions are "essentially" the same! If you do the unorthodox thing—choose a to be, not real, but imaginary—then you find that

$$e^{ix} = \cos x + i \sin x.$$

Making the domain of the exponential function imaginary turns it into a combination of sine and cosine!

A gap is yawning. A unification and deeper understanding beckon, which demand going out of the given mathematical structure—allowing the existence of $\sqrt{-1}$ —*changing the axioms.*

How to change the axioms? How to change or enlarge our mathematical structure? These questions go beyond axioms and theorems. As well as working within given axiomatic structures, mathematicians tear structures down, to replace them with others more powerful.

Calculus

Newton, Leibniz, Berkeley

Berkeley's famous *Analyst* (famous in the history of mathematics, forgotten in the history of philosophy) is an attack on the differential calculus of Newton and

Leibniz. The fallacy Berkeley exposed is simple. To compute the speed of a moving body, you divide the distance traveled by the time elapsed. If the speed is variable, this fraction depends on how much time elapses. But we want the speed at one instant—a time interval of length *zero*. For a falling stone, for example, we want its speed when it hits the ground—its final or "ultimate" velocity. But that seems to require dividing by zero—which is impossible.

Newton explained: "By the ultimate velocity is meant that with which the body is moved, neither before it arrives at its last place, when the motion ceases, nor after, but at the very instant when it arrives. . . . And in like manner, by the ultimate ratio of evanescent quantities is to be understood the ratio of the quantities, not before they vanish, nor after, but that with which they vanish."

This gives us a physical intuition of ultimate velocity. But when Newton calculated he used a mathematical algorithm, not physical intuition. Starting with a time interval of positive duration (call it h), he got an average speed depending on h. He simplified the answer algebraically, and finally set $h = 0$. The resulting expression was the instantaneous speed. Newton called it the "fluxion," and the associated distance function the "fluent."

"But," wrote Berkeley, "It should seem that this reasoning is not fair or conclusive. . . . For when it is said, let the increments vanish, let the increments be nothing, or let there be no increments, the former supposition that the increments were something, or that there were increments, is destroyed, and yet a consequence of that supposition, i.e., an expression got by virtue thereof, is retained. Which is a false way of reasoning. . . . Nor will it avail to say that [the term neglected] is a quantity exceedingly small; since we are told that *in rebus mathematicis errores quan minimi non sunt contemnendi.*" ("In mathematics not even the smallest errors are ignored.")

Berkeley admitted that Newton got the right answer, and that his use of it in physics was correct. He merely showed that Newton's reasoning was obscure.

Leibniz was co-inventor, with Isaac Newton, of the infinitesimal calculus. Unlike Newton, Leibniz used "actual infinitesimals," though he couldn't explain coherently what they were. Cavalieri and others had calculated areas by dividing regions into infinitely many strips, each having infinitesimal positive area. Unfortunately, as Huygens showed, this method could give wrong answers. Problems of rates and velocities also led to infinitesimals. Think of a stone that in 2 seconds falls a distance of 4 feet, in 3 seconds a distance of 9 feet, and in general in t seconds a distance of t^2 feet. Leibniz got the stone's instantaneous speed by calculating its average speed over a time interval of infinitesimal duration. The calculation is so easy we do it right now.

Let dt be the duration of an infinitesimal time interval. At the beginning of the interval your watch reads t seconds, where t is some positive number. The distance fallen up to that time is t^2 feet. An infinitesimal time interval of duration dt elapses. Your watch then reads $(t + dt)$ seconds, and the stone has fallen $(t + dt)^2$ feet. So in the infinitesimal time dt, from instant t to instant $t + dt$, the distance

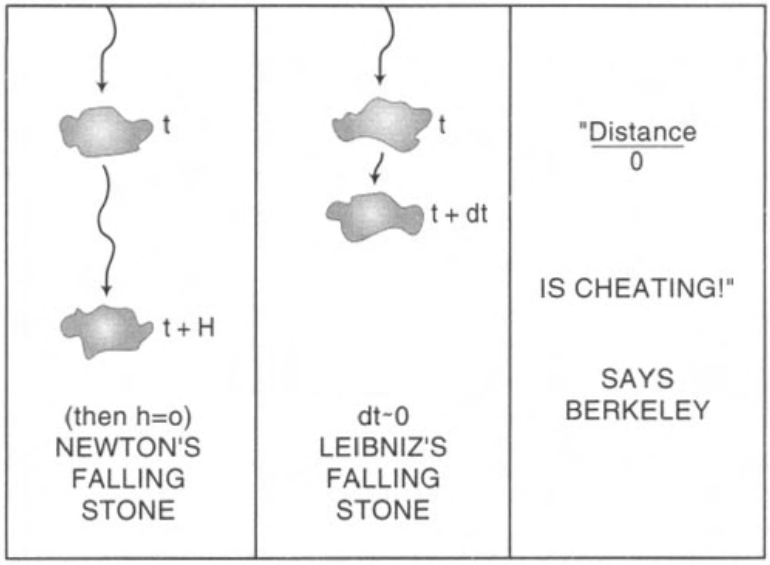

Figure 5. Falling body.

traveled is the distance from its starting point, which was t^2 feet below the stone's initial height, to its ending point, which is $(t+ dt)^2$ feet below the stone's initial height. The distance between the two points is the difference, $(t + dt)^2 - t^2$. "Average speed" is defined in general as a ratio: distance traveled divided by time elapsed. In the present case, that ratio is $[(t + dt)^2 - t^2]/(dt)$. A little algebra simplifies this expression to $(2t + dt)$. dt is infinitesimal, so we "neglect" it—throw it away—and find the instantaneous speed after t seconds: exactly 2t feet per second.

Leibniz's algebra was just like Newton's. He got the same answer as Newton after algebraic simplification, except that his formula had the infinitesimal dt where Newton had the small finite h. Then, instead of setting h equal to zero, as Newton did, Leibniz simply threw away the terms involving dt, *because they were infinitesimal*—negligible compared to the finite part of the answer.

This reasoning was also torn to shreds by Bishop Berkeley. He admitted that the answer, 2t, is right. Berkeley rightly objected to "throwing away" anything not equal to zero, no matter how small. He pointed out that, infinitesimal or not, dt has to be either zero or not zero. If it's not zero, then $2t + dt$ isn't the same as 2t. If it's zero, Leibniz had no right to divide by it. Either way, a fallacy.

Today Berkeley's objections don't disturb us. We show that the average speed *converges to a limit* as the time interval gets shorter. That limit is then *defined* as the instantaneous speed. This limit-and-continuity approach was developed by Cauchy and Weierstrass in the nineteenth century. It is adequate to demystify calculus.

Newton and Leibniz didn't have an explicit definition of limit. The careful use of limits requires explication of the real number system. This subtle task even

now may not be quite finished. Still, we see today that Newton was essentially using limits.

Leibniz explained that his infinitesimal dt is "fictitious." This fiction is like an ordinary positive number, but smaller than any ordinary positive number. This is not easy to grasp. How do we decide which properties of ordinary positive numbers apply to dt because it's "just like an ordinary positive number," and which ones don't apply because "it's smaller than any ordinary positive number"? What's the square root of dt? It must be infinitesimal, yet bigger than dt. How many infinitesimals are we going to need? What about the cube root, the fourth root, the tenth root? These puzzles were solved by Abraham Robinson 200 years later, using the theory of formal languages—modern mathematical logic.

The infinitesimal has a fascinating history. At least as far back as Archimedes, it's been used by mathematicians who were perfectly aware that it didn't make sense. It surfaced from underground in the 1960s, when Robinson legitimized it with his "nonstandard analysis."

Nonstandard analysis is the fruit of a century's development of mathematical logic. The basis of it is to regard the language in which we talk about mathematics as itself a mathematical object, obeying explicit formal rules. This formal language then is subject to mathematical reasoning. (Which we carry on, as usual, in ordinary, everyday language, just as we use ordinary language to talk about Basic or C.) Then it makes mathematical sense to say that an infinitesimal is greater than zero and smaller than all the positive numbers *expressible in the formal language.* When Robinson rehabilitated infinitesimals with his nonstandard analysis, he borrowed the word "monad" from Leibniz's metaphysics. In his nonstandard analysis, a monad is an infinitesimal neighborhood (the set of points infinitely close to some given point.)

So today we have two distinct rigorous formulations of calculus. The creators of the calculus were using tools whose theories were centuries in the future.

A Calculus Refresher

Calculus is the heart of "modern mathematics"—mathematics since Newton. It's the part of mathematics most important in science and technology, the part engineers must know.

It's built around two main problems. The central discovery of calculus is that these problems are related—in fact, as we will see, they're opposites.

The first problem is speed. How fast is something changing? The second main problem is area. How big is some curved region?

First, speed. The speed is simple if it's constant:

$$\text{Speed} = \frac{\text{Distance}}{\text{Time}}$$

Divide distance traveled by time elapsed.

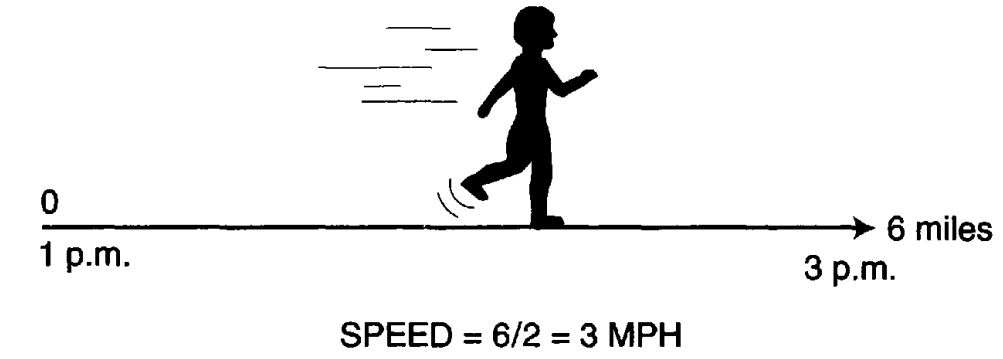

Figure 6. Motion at constant speed.

But speed isn't constant. When you drive you start at speed zero, gradually go to the speed limit, then finally slow down to zero. Your speed varies from instant to instant. What is your speed at some particular instant?

Figure 7.

Example: a body falling in vacuum near the surface of the earth travels $16\ t^2$ feet in t seconds. How fast is it falling after 2 seconds?

In the time interval between 2 seconds and 2.1 seconds—time lapse of .1 second—it falls

$$16\ (2.1)^2 - 16\ (2^2)\ \text{feet} = 6.56\ \text{feet}.$$

Dividing distance by time (.1 second), its average speed was 65.6 ft/sec.

Exercise. Repeat the calculation with a time lapse of .01 second. (You'll get an average speed of 64.16 ft/sec, between time 2 seconds and time 2.01 seconds.) Do it again, with a ***very small*** time lapse, .001 seconds. (Its average velocity over this time period is 64.016 ft/sec.) ### (### means "end of exercise.")

But I don't want an ***average*** speed. I want the ***exact*** speed after 2 seconds! That means a time lapse of *zero*. Division by zero is impossible. The formula

$$\text{Speed} = \frac{\text{Distance}}{\text{Time}}$$

becomes meaningless.

However, without setting time lapse to 0, you've crept closer and closer to 0. You used lapses of .1, .01, .001. and found speeds of 65.6, 64.16, and 64.016.

NOW! A giant conceptual leap! If the average speeds approach a limit as the time lapse approaches zero, we declare, *as a definition*, the instantaneous speed *is* that limit! In this example, the limit is 64 ft/sec when t = 2. It makes sense! We agree, that's what we'll mean by instantaneous velocity.

The notion of speed as a limit took centuries to formulate. Medieval and Renaissance mathematicians calculated rates of change without defining mathematically what they wanted. The founders of the calculus, Isaac Newton and Gottlob Leibniz, fought bitterly about who had priority in the fundamental theorem of calculus (explained below.)

Exercise. Make a graph of this falling body function: distance = time squared, or $d = t^2$.

(I dropped the 16 to simplify your graphing and my calculating.) This is a quadratic function. Its graph is a parabola. Mark the points (2, 4) and (2.1, 4.41) on the parabola. The second is above and right of the first. Draw a straight line (*secant*) between the two. What's the slope of this line? ("Rise over run.")

$$\text{Rise} = 2.1^2 - 2.0^2 = .41$$
$$\text{Run} = 2.1 - 2.0 = .1$$

Slope = .41/.1 = 4.1, which we just found is the average velocity (allowing for the factor of 16 which we took out). *The average rate of change of a function*

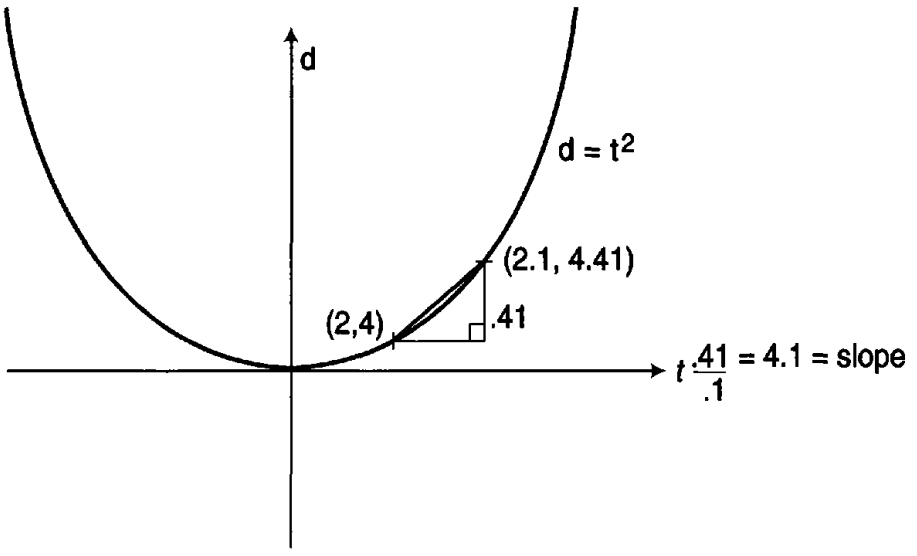

Figure 8. Differentiating x^2 = finding its slope (this graph is called a parabola).