

Measurement

PAUL LOCKHART

MEASUREMENT

MEASUREMENT

PAUL LOCKHART

THE BELKNAP PRESS OF HARVARD UNIVERSITY PRESS
CAMBRIDGE, MASSACHUSETTS • LONDON, ENGLAND

2012

Copyright © 2012 by the President and Fellows of Harvard College

All rights reserved

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Lockhart, Paul.

Measurement / Paul Lockhart.

p. cm.

Includes index.

ISBN 978-0-674-05755-5 (hardcover : alk. paper)

1. Geometry. I. Title.

QA447.L625 2012

516—dc23 2012007726

For Will, Ben, and Yarrow

CONTENTS

Reality and Imagination 1

On Problems 5

Part One: Size and Shape 21

In which we begin our investigation of abstract geometrical figures. Symmetrical tiling and angle measurement. Scaling and proportion. Length, area, and volume. The method of exhaustion and its consequences. Polygons and trigonometry. Conic sections and projective geometry. Mechanical curves.

Part Two: Time and Space 199

Containing some thoughts on mathematical motion. Coordinate systems and dimension. Motion as a numerical relationship. Vector representation and mechanical relativity. The measurement of velocity. The differential calculus and its myriad uses. Some final words of encouragement to the reader.

Acknowledgments 399

Index 401

REALITY AND IMAGINATION

There are many realities out there. There is, of course, the physical reality we find ourselves in. Then there are those imaginary universes that resemble physical reality very closely, such as the one where everything is exactly the same except I *didn't* pee in my pants in fifth grade, or the one where that beautiful dark-haired girl on the bus turned to me and we started talking and ended up falling in love. There are plenty of *those* kinds of imaginary realities, believe me. But that's neither here nor there.

I want to talk about a different sort of place. I'm going to call it "mathematical reality." In my mind's eye, there is a universe where beautiful shapes and patterns float by and do curious and surprising things that keep me amused and entertained. It's an amazing place, and I really love it.

The thing is, physical reality is a disaster. It's way too complicated, and nothing is at all what it appears to be. Objects expand and contract with temperature, atoms fly on and off. In particular, nothing can truly be measured. A blade of grass has no actual length. Any measurement made in this universe is necessarily a rough approximation. It's not bad; it's just the nature of the place. The smallest speck is not a point, and the thinnest wire is not a line.

Mathematical reality, on the other hand, is imaginary. It can be as simple and pretty as I want it to be. I get to have all those perfect things I can't have in real life. I will never hold a circle in my hand, but I can hold one in my mind. And I can measure it. Mathematical reality is a beautiful wonderland of my own

creation, and I can explore it and think about it and talk about it with my friends.

Now, there are lots of reasons people get interested in physical reality. Astronomers, biologists, chemists, and all the rest are trying to figure out how it works, to describe it.

I want to describe mathematical reality. To make patterns. To figure out how they work. That's what mathematicians like me try to do.

The point is I get to have them both—physical reality and mathematical reality. Both are beautiful and interesting (and somewhat frightening). The former is important to me because I am in it, the latter because it is in me. I get to have both these wonderful things in my life and so do you.

My idea with this book is that we will design patterns. We'll make patterns of shape and motion, and then we will try to understand our patterns and measure them. And we will see beautiful things!

But I won't lie to you: this is going to be very hard work. Mathematical reality is an infinite jungle full of enchanting mysteries, but the jungle does not give up its secrets easily. Be prepared to struggle, both intellectually and creatively. The truth is, I don't know of any human activity as demanding of one's imagination, intuition, and ingenuity. But I do it anyway. I do it because I love it and because I can't help it. Once you've been to the jungle, you can never really leave. It haunts your waking dreams.

So I invite you to go on an amazing adventure! And of course, I want you to love the jungle and to fall under its spell. What I've tried to do in this book is to express how math feels to me and to show you a few of our most beautiful and excit-

ing discoveries. Don't expect any footnotes or references or anything scholarly like that. This is *personal*. I just hope I can manage to convey these deep and fascinating ideas in a way that is comprehensible and fun.

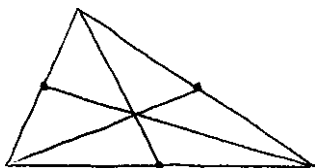
Still, expect it to be slow going. I have no desire to baby you or to protect you from the truth, and I'm not going to apologize for how hard it is. Let it take hours or even days for a new idea to sink in—it may have originally taken centuries!

I'm going to assume that you love beautiful things and are curious to learn about them. The only things you will need on this journey are common sense and simple human curiosity. So relax. Art is to be enjoyed, and this is an art book. Math is not a race or a contest; it's just you playing with your own imagination. Have a wonderful time!

ON PROBLEMS

What is a math problem? To a mathematician, a problem is a *probe*—a test of mathematical reality to see how it behaves. It is our way of “poking it with a stick” and seeing what happens. We have a piece of mathematical reality, which may be a configuration of shapes, a number pattern, or what have you, and we want to understand what makes it tick: What does it do and why does it do it? So we poke it—only not with our hands and not with a stick. We have to poke it with our minds.

As an example, let’s say you’ve been playing around with triangles, chopping them up into other triangles and so forth, and you happen to make a discovery:

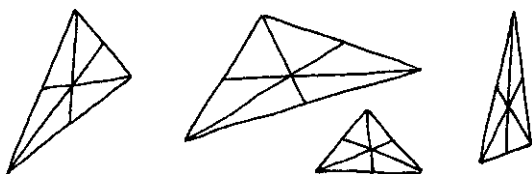


When you connect each corner of a triangle to the middle of the opposite side, the three lines seem to all meet at a point. You try this for a wide variety of triangles, and it always seems to happen. Now you have a mystery! But let’s be very clear about exactly what the mystery is. It’s not about your drawings or what looks like is happening on paper. The question of what pencil-and-paper triangles may or may not do is a scientific one about physical reality. If your drawing is sloppy, for example, then the lines won’t meet. I suppose you could make an extremely careful drawing and put it under a microscope,

but you would learn a lot more about graphite and paper fibers than you would about triangles.

The real mystery is about imaginary, too-perfect-to-exist triangles, and the question is whether these three perfect lines meet in a perfect point in mathematical reality. No pencils or microscopes will help you now. (This is a distinction I will be stressing throughout the book, probably to the point of annoyance.) So how are we to address such a question? Can anything ever really be known about such imaginary objects? What form could such knowledge take?

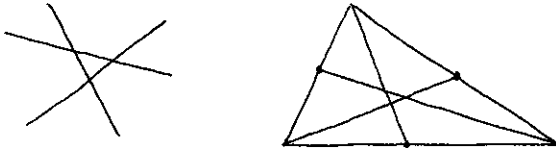
Before examining these issues, let's take a moment to simply delight in the question itself and to appreciate what is being said here about the nature of mathematical reality.



What we've stumbled onto is a conspiracy. Apparently, there is some underlying (and as yet unknown) structural interplay going on that is making this happen. I think that is marvelous and also a little scary. What do triangles know that we don't? Sometimes it makes me a little queasy to think about all the beautiful and profound truths out there waiting to be discovered and connected together.

So what exactly is the mystery here? The mystery is *why*. Why would a triangle want to do such a thing? After all, if you drop three sticks at random they usually don't meet at a point; they cross each other in three different places to form a

little triangle in the middle. Isn't that what we would expect to happen?



What we are looking for is an explanation. Of course, one reason why an explanation may not be forthcoming is that it simply isn't true. Maybe we fooled ourselves by wishful thinking or clumsy drawing. There's a lot of "fudging" in physical reality, so maybe we just couldn't see the little triangle where the lines cross. Perhaps it was so small that it got lost among all the smears and pencil crumbs. On the other hand, it's certainly the kind of thing that *could* be true. It has a lot of elements that mathematicians look for: naturalness, elegance, simplicity, and a certain inevitable quality. So it's probably true. But again, the question is why.

Now here's where the art comes in. In order to explain we have to create something. Namely, we need to somehow construct an argument—a piece of reasoning that will satisfy our curiosity as to why this behavior is happening. This is a very tall order. For one thing, it's not enough to draw or build a bunch of physical triangles and see that it more or less works for them. That is not an explanation; it's more of an "approximate verification." Ours is a much more serious philosophical issue.

Without knowing why the lines meet at a common point, how can we know that they actually do? In contrast with physical reality, there's nothing to observe. How will we ever

know anything about a purely imaginary realm? The point is, it doesn't matter so much *what* is true. It matters *why* it's true. The *why* is the *what*.

Not that I am trying to minimize the value of our ordinary senses—far from it. We desperately need any and all aids to our intuition and imagination: drawings, models, movies, whatever we can get. We just have to understand that ultimately these things are not really the subject of the conversation and cannot really tell us the truth about mathematical reality.

So now we really are in a predicament. We have discovered what we think may be a beautiful truth, and now we need to prove it. This is what mathematicians do, and this is what I hope you will enjoy doing yourself.

Is this such an extraordinarily difficult thing to do? Yes, it is. Is there some recipe or method to follow? No, there isn't. This is abstract art, pure and simple. And art is always a struggle. There is no systematic way to create beautiful and meaningful paintings or sculptures, and there is also no method for producing beautiful and meaningful mathematical arguments. Sorry. Math is the hardest thing there is, and that's one of the reasons I love it.

So no, I can't tell you how to do it, and I'm not going to hold your hand or give you a bunch of hints or solutions in the back of the book. If you want to paint a picture from your heart, there is no "answer painting" on the back of the canvas. If you are working on a problem and you are stuck and in pain, then welcome to the club. We mathematicians don't know how to solve our problems either. If we did, they would no longer be problems! We're always working at the edge of the unknown, and we're always stuck. Until we have a breakthrough. And I

hope you have many—it's an incredible feeling. But there is no special procedure for doing mathematics. You just have to think a lot and hope that inspiration comes to you.

But I won't just drop you into the jungle and leave you there. Your intelligence and your curiosity you will have to supply yourself—these are your machete and your canteen. But maybe I can provide you with a compass in the form of a few general words of advice.

The first is that *the best problems are your own*. You are the intrepid mental explorer; it's your mind and your adventure. Mathematical reality is *yours*—it's in your head for you to explore any time you feel like it. What are your questions? Where do you want to go? I've enjoyed coming up with some problems for you to think about, but these are merely seeds I've planted to help you start growing your own jungle. Don't be afraid that you can't answer your own questions—that's the natural state of the mathematician. Also, try to always have five or six problems you are working on. It is very frustrating to keep banging your head against the same wall over and over. (It's much better to have five or six walls to bang your head against!) Seriously, taking a break from a problem always seems to help.

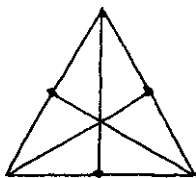
Another important piece of advice: *collaborate*. If you have a friend who also wants to do math, you can work together and share the joys and frustrations. It's a lot like playing music together. Sometimes I will spend six or eight hours working on a problem with a friend, and even if we accomplish next to nothing, we still had fun feeling dumb together.

So let it be hard. Try not to get discouraged or to take your failures too personally. It's not only you that is having trouble

understanding mathematical reality; it's all of us. Don't worry that you have no experience or that you're not "qualified." What makes a mathematician is not technical skill or encyclopedic knowledge but insatiable curiosity and a desire for simple beauty. Just be yourself and go where you want to go. Instead of being tentative and fearing failure or confusion, try to embrace the awe and mystery of it all and joyfully make a mess. Yes, your ideas won't work. Yes, your intuition will be flawed. Again, welcome to the club! I have a dozen bad ideas a day and so does every other mathematician.

Now, I know what you're thinking: a bunch of fuzzy, romantic talk about beauty and art and the exquisite pain of creativity is all very well and good, but how on earth am I supposed to do this? I've never created a mathematical argument in my life. Can't you give me a little more to go on?

Let's go back to our triangle and the three lines. How can we begin to cobble together some sort of an argument? One place we could start is by looking at a symmetrical triangle.

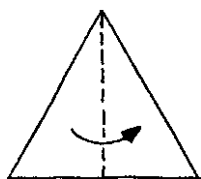


This kind of triangle is also called **equilateral** (Latin for "same sides"). Now, I know this is an absurdly atypical situation, but the idea is that if we can somehow explain why the lines meet in this special case, it might give us a clue about how to proceed with a more general triangle. Or it might not. You

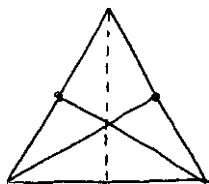
never know, you just have to mess around—what we mathematicians like to call “doing research.”

In any event, we have to start somewhere, and it should at least be easier to figure something out in this case. What we have going for us in this special situation is tons of symmetry. *Do not ignore symmetry!* In many ways, it is our most powerful mathematical tool. (Put it in your backpack with your machete and canteen.)

Here symmetry allows us to conclude that anything that happens on one side of the triangle must also happen on the other. Another way to say this is that if we flipped the triangle across its line of symmetry, it would look exactly the same.

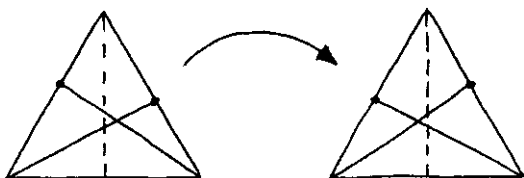


In particular, the midpoints of the two sides would switch places, as would the lines connecting them to their opposite corners.



But this means that the crossing point of these two lines can't be on one side of the line of symmetry, else when we flip the

triangle it would move to the other side, and we could tell that it got flipped!



So the crossing point must actually be *on* the line of symmetry. Clearly our third line (the one connecting the top corner to the middle of the bottom side) is simply the line of symmetry itself, and so that is why all three lines meet at a point. Isn't that a nice explanation?

This is an example of a mathematical argument, otherwise known as a *proof*. A proof is simply a story. The characters are the elements of the problem, and the plot is up to you. The goal, as in any literary fiction, is to write a story that is compelling as a narrative. In the case of mathematics, this means that the plot not only has to make logical sense but also be simple and elegant. No one likes a meandering, complicated quagmire of a proof. We want to follow along rationally to be sure, but we also want to be charmed and swept off our feet aesthetically. A proof should be lovely as well as logical.

Which brings me to another piece of advice: *improve your proofs*. Just because you have an explanation doesn't mean it's the best explanation. Can you eliminate any unnecessary clutter or complexity? Can you find an entirely different approach that gives you deeper insight? Prove, prove, and prove again. Painters, sculptors, and poets do the same thing.

Our proof just now, for instance, despite its logical clarity

and simplicity, has a slightly arbitrary feature. Even though we made an essential use of symmetry, there's something annoyingly asymmetrical about the proof (at least to me). Specifically, the argument favors one corner. Not that it's so very bad to pick one corner and use its line as our line of symmetry, it's just that the triangle is so symmetrical; we shouldn't have to make such an arbitrary choice.

We could, for instance, use the fact that in addition to having flip-symmetry, our triangle is also *rotationally* symmetric. That is, if we turn it one-third of a full turn around, it looks exactly the same. This means that our triangle must have a *center*.



Now, if we flip the triangle across any of its three lines of symmetry (favoring none of them), the triangle doesn't change, so its center must stay put. This means that the center point lies on all three lines of symmetry. So that's why the lines all meet!

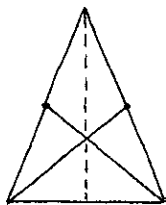
Now, I'm not trying to say that this argument is so much better or even all that different. (And in fact, there are lots of other ways to prove it.) All I'm saying is that deeper insight and understanding can be gained by coming at a problem in more than one way. In particular, the second proof not only tells me that the lines meet, it tells me where—namely, at the center of rotation. Which makes me wonder, where exactly is that? Specifically, how far up an equilateral triangle is its center?

Throughout the book, questions like this will come up. Part

of becoming a mathematician is learning to ask such questions, to poke your stick around looking for new and exciting truths to uncover. Problems and questions that occur to me I will put in boldface type. Then you can think about them and work on them as you please and hopefully also come up with problems of your own. So here's your first one:

Where is the center of an equilateral triangle?

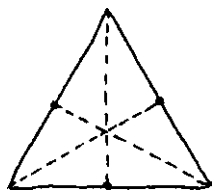
Now going back to the original problem, we see that we have barely made a dent. We have an explanation for why the lines meet in an equilateral triangle, but our arguments are so dependent on symmetry, it's hard to see how this will help in the more general situation. Actually, I suppose our first argument still works if our triangle has two equal sides:



The reason is that this kind of triangle, known as **isosceles** (Greek for “same legs”), still possesses a line of symmetry. This is a nice example of generalization—getting a problem or an argument to make sense in a wider context. But still, for the average asymmetrical triangle, our arguments clearly won't work.

This puts us in a place that is all too familiar to mathematicians. It's called stuck. We need a new idea, preferably one that

doesn't hinge so much on symmetry. So let's go back to the drawing board.



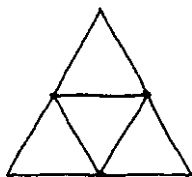
Is there something else we can do with these characters? We have a triangle, the midpoints of the sides, and the lines drawn to them from the corners.

Here's a thought. What if we connect the midpoints? Does anything interesting happen? This is the kind of thing you have to do as a mathematician: try things. Will they work? Will they yield useful information? Usually not. But you can't just sit there staring at some shapes or numbers. Try anything and everything. As you do more math, your intuition and your instincts will sharpen, and your ideas will get better. How do you know which ideas to try? You don't. You just have to guess. Experienced mathematicians have a great deal of sensitivity to structure, and so our guesses are more likely to be right, but we still have to guess. So guess.

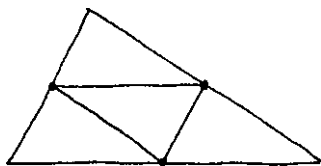
The important thing is not to be afraid. So you try some crazy idea, and it doesn't work. That puts you in some pretty good company! Archimedes, Gauss, you and I—we're all groping our way through mathematical reality, trying to understand what is going on, making guesses, trying out ideas, and mostly failing. And then every once in a while, you succeed. (Perhaps more frequently if you are Archimedes or Gauss.) And that feeling of unlocking an eternal mystery is

what keeps you going back to the jungle to get scratched up all over again.

So imagine you've tried this idea and that idea, and one day it occurs to you to connect the midpoints.



What do we notice? Well, we've divided the original triangle into four smaller ones. In the symmetrical case, they are clearly identical. What happens in general?



Are the triangles all the same? Actually, it looks like three of them might just be smaller (half-scale) versions of the original triangle. Could that be true? What about the middle one? Could it also be the same, only rotated upside down? What exactly have we stumbled onto here?

We've stumbled onto a glimmer of truth, pattern, and beauty, that's what. And maybe this will lead to something wholly unexpected, possibly having nothing to do with our original problem. So be it. There's nothing sacred about our three lines problem; it's a question like any other. If your thoughts on one problem lead you to another, then good for

you! Now you have *two* problems to work on. My advice: be open-minded and flexible. *Let a problem take you where it takes you.* If you come across a river in the jungle, follow it!

Are these four triangles identical?

Let's suppose this is true. And that, by the way, is a perfectly fine thing to do. Mathematicians are always supposing things and seeing what would happen (the Greeks even had a word for it—they called it *analysis*). There are thousands of apparent mathematical truths out there that we humans have discovered and believe to be true but have so far been unable to prove. They are called *conjectures*. A conjecture is simply a statement about mathematical reality that you believe to be true (usually you also have some examples to back it up, so it is a reasonably educated guess). I hope that you will find yourself conjecturing all over the place as you read this book and do mathematics. Maybe you will even prove some of your conjectures. Then you get to call them *theorems*.

Supposing that our conjecture about the four triangles is true (and, of course, we still want a nice proof of this), the next question would be whether this helps us solve our original problem. Maybe it will, maybe it won't. You just have to see if anything comes to you.

Essentially, engaging in the practice of mathematics means that you are playing around, making observations and discoveries, constructing examples (as well as counterexamples), formulating conjectures, and then—the hard part—proving them. I hope you will find this work fascinating and entertaining, challenging, and ultimately deeply rewarding.

So I will leave the problem of the triangle and its intersecting lines in your capable hands.

Which brings me to my next bit of advice: *critique your work*. Subject your arguments to scathing criticism by yourself and by others. That's what all artists do, especially mathematicians. As I've said, for a piece of mathematics to fully qualify as such, it has to stand up to two very different kinds of criticism: it must be logically sound and convincing as a rational argument, and it must also be elegant, revelatory, and emotionally satisfying. I'm sorry that these criteria are so painfully steep, but that is the nature of the art.

Now, aesthetic judgments are obviously quite personal, and they can change with time and place. Certainly that has happened with mathematics no less than with other human endeavors. An argument that was considered beautiful a thousand or even a hundred years ago might now be looked upon as clumsy and inelegant. (A lot of classical Greek mathematics, for example, appears quite dreadful to my modern sensibilities.)

My advice is not to worry about trying to hold yourself to some impossibly high standard of aesthetic excellence. If you like your proof (and most of us are fairly proud of our hard-won creations), then it is good. If you are dissatisfied in some way (and most of us are), then you have more work to do. As you gain experience, your taste will grow and develop, and you may find later that you are unhappy with some of your earlier work. That is as it should be.

I think the same could be said for logical validity as well. As you do more mathematics, you will literally get smarter. Your logical reasoning will become tighter, and you will begin to develop a mathematical "nose." You will learn to be suspicious,

to sense that some important details have been glossed over. So let that happen.

Now, there is a certain obnoxious type of mathematician who simply cannot allow false statements to be made at any time. I am not one of them. I believe in making a mess—that's how great art happens. So your first essays in this craft are likely to be logical disasters. You will believe things to be true, and they won't be. Your reasoning will be flawed. You will jump to conclusions. Well, go ahead and jump. The only person you have to satisfy is yourself. Believe me, you will discover plenty of errors in your own deduction. You will declare yourself a genius at breakfast and an idiot at lunch. We've all done it.

Part of the problem is that we are so concerned with our ideas being simple and beautiful that when we do have a pretty idea, we want so much to believe it. We want it to be true so badly that we don't always give it the careful scrutiny that we should. It's the mathematical version of "rapture of the deep." Divers see such beautiful sights that they forget to come up for air. Well, logic is our air, and careful reasoning is how we breathe. So don't forget to breathe!

The real difference between you and more experienced mathematicians is that we've seen a lot more ways that we can fool ourselves. So we have more nagging doubts and therefore insist on a much higher standard of logical rigor. We learn to play the devil's advocate.

Whenever I am working on a conjecture, I always entertain the possibility that it is false. Sometimes I work to prove it, other times I try to refute it—to prove myself wrong. Occasionally, I discover a counterexample showing that I was indeed misled

and that I need to refine or possibly scrap my conjecture. Still other times, my attempts to construct a counterexample keep running into the same barrier, and this barrier then becomes the key to my eventual proof. The point is to keep an open mind and not to let your hopes and wishes interfere with your pursuit of truth.

Of course, as much as we mathematicians may ultimately insist on the most persnickety level of logical clarity, we also know from experience when a proof “smells right,” and it is clear that we could supply the necessary details if we wished. The truth of the matter is that math is a human activity, and we humans make mistakes. Great mathematicians have “proved” utter nonsense, and so will you. (It’s another good reason to collaborate with other people—they can raise objections to your arguments that you might overlook.)

The point is to get out there in mathematical reality, make some discoveries, and have fun. Your desire for logical rigor will grow with experience; don’t worry.

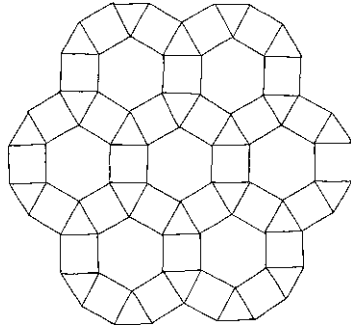
So go ahead and do your mathematical art. Subject it to your own standards of rationality and beauty. Does it please you? Then great! Are you a tormented struggling artist? Even better. Welcome to the jungle!

PART ONE

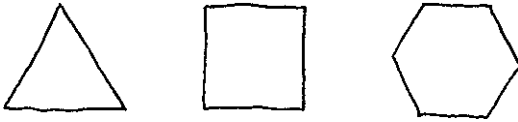
SIZE AND SHAPE

1

Here is a nice pattern.



Let me tell you why I find this kind of thing so attractive. First of all, it involves some of my favorite shapes.



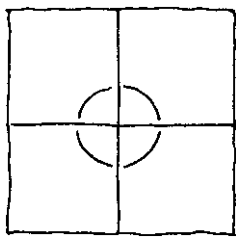
I like these shapes because they are simple and symmetrical. Shapes like these that are made of straight lines are called **polygons** (Greek for “many corners”). A polygon with all its sides the same length and all its angles equal is called **regular**. So I guess what I’m saying is, I like regular polygons.

Another reason why the design is appealing is that the pieces fit together so nicely. There are no gaps between the tiles (I like to think of them as ceramic tiles, like in a mosaic), and the tiles don’t overlap. At least, that’s how it appears. Remember, the objects that we’re really talking about are perfect, imaginary shapes. Just because the picture looks good doesn’t mean

that's what is really going on. Pictures, no matter how carefully made, are part of physical reality; they can't possibly tell us the truth about imaginary, mathematical objects. Shapes do what they do, not what we want them to do.

So how can we be sure that the polygons really do fit perfectly? For that matter, how can we know *anything* about these objects? The point is, we need to measure them—and not with any clumsy real-world implements like rulers or protractors, but with our minds. We need to find a way to measure these shapes using philosophical argument alone.

Do you see that in this case what we need to measure are the angles? In order to check that a mosaic pattern like this will work, we need to make sure that at every corner (where the tiles meet) the angles of the polygons add up to a full turn. For instance, the ordinary square tiling works because the angles of a square are quarter turns and it takes four of them to make a full turn.



By the way, I like to measure angles as portions of a full turn instead of using degrees. It seems simpler to me and more natural than using an arbitrary division of a turn into 360 parts (of course *you* may do as you please). So I'm going to say that a square has a corner angle of $\frac{1}{4}$.

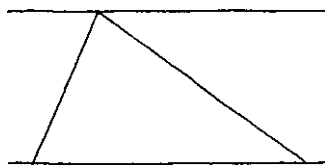
One of the first things people discovered about angles is the

surprising fact that for any triangle (no matter what shape) the sum of the angles is always the same, namely a half turn (or 180 degrees if you must be vulgar).

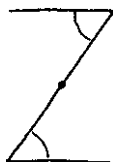


To get a feel for this, you might want to make some paper triangles and cut off their corners. When you join them together, they will always form a straight line. What a beautiful discovery! But how can we really know that it is true?

One way to see it is to view the triangle as being sandwiched between two parallel lines.

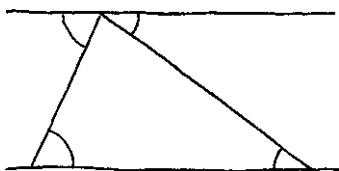


Notice how these lines form Z shapes with the sides of the triangle. (I suppose you might call the one on the right side a backward Z, but it doesn't really matter.) Now, the thing about Z shapes is that their angles are always equal.



This is because a Z shape is symmetrical: it looks exactly the same if you rotate it a half turn around its center point. That means the angle at the top must be the same as the angle at the bottom. Does that make sense? This is a typical example of a symmetry argument. The invariance of a shape under a certain set of motions allows us to deduce that two or more measurements must be the same.

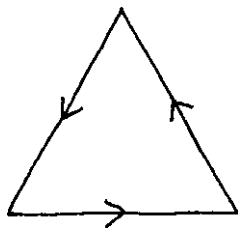
Going back to our triangle sandwich, we see that each angle at the bottom corresponds to an equal angle at the top.



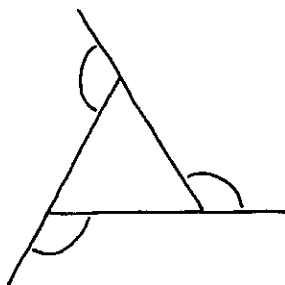
This means that the three angles of the triangle join together at the top to form a straight line. So the three turns add up to a half turn. What a delightful piece of mathematical reasoning!

This is what it means to do mathematics. To make a discovery (by whatever means, including playing around with physical models like paper, string, and rubber bands), and then to explain it in the simplest and most elegant way possible. This is the art of it, and this is why it is so challenging and fun.

One consequence of this discovery is that if our triangle happens to be equilateral (that is, regular) then its angles are all equal, so they must each be $\frac{1}{6}$. Another way to see this is to imagine driving around the perimeter of the triangle.



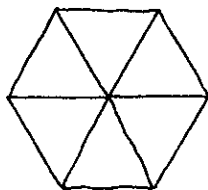
We make three equal turns to get back to where we started. Since we end up making one complete turn, each of these must be exactly $\frac{1}{3}$. Notice that the turns we've made are actually the *outside* angles of the triangle.



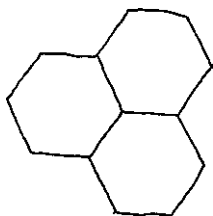
Since the inside and outside angles combine to make a half turn, the inside angles must be

$$\frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

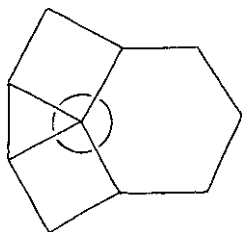
In particular, six of these triangles will fit together at a corner.



Hey, this makes a regular hexagon! So as a bonus, we get that the angles of a regular hexagon must be twice those of the triangle, in other words $\frac{1}{3}$. This means that three hexagons fit together perfectly.



So it is possible to have knowledge about these shapes after all. In particular, now we can see why the original mosaic design works.



At each corner of the mosaic, we have a hexagon, two squares, and a triangle. Adding up the angles we get

$$\frac{1}{3} + \frac{1}{4} + \frac{1}{4} + \frac{1}{6} = 1.$$

So it works!

(By the way, if you don't like doing arithmetic with fractions, you can always avoid it by changing your measuring units. For example, if you'd prefer, we could decide to measure angles

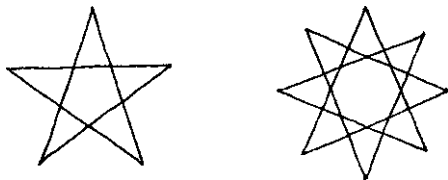
in twelfths of a turn, so that the angle of a regular hexagon would simply be 4, a square would have an angle of 3, and a triangle an angle of 2. Then the angles of our shapes would add up to $4 + 3 + 3 + 2 = 12$; that is, a full turn.)

I especially love how symmetrical this mosaic pattern is. Each corner has the same exact sequence of shapes around it: hexagon, square, triangle, square. This means that once we've checked that the angles fit at one corner, we automatically know that they work at all the other corners. Notice that the pattern can be continued indefinitely so that it covers an entire infinite plane. It makes me wonder what other beautiful mosaic patterns might be out there in mathematical reality?

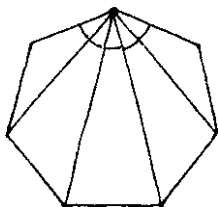
What are all the different ways to make symmetrical mosaic designs using regular polygons?

Naturally, we're going to need to know the angles of the various regular polygons. Can you figure out how to measure them?

What are the angles of a regular n -sided polygon?

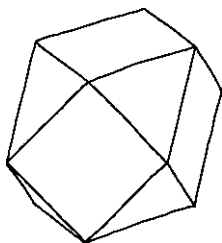


Can you measure the angles of a regular n -pointed star?

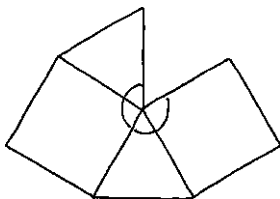


Do the diagonals drawn from one corner of a regular polygon always make equal angles?

While we're on the subject of pretty patterns made of polygons, I want to show you another of my favorites.



This time we have squares and triangles, but instead of lying flat, they are arranged to form a sort of ball shape. This kind of object is called a **polyhedron** (Greek for “many sides”). People have been playing around with them for thousands of years. One approach to thinking about them is to imagine unfolding them flat onto a plane. For example, one corner of my shape would unfold to look like this:



Here, we have two squares and two triangles around a point, but they leave a gap so that the shape can be folded up into a ball. So in the case of polyhedra, we need the angles to add up to less than a full turn.

**What happens if the angle sum is
more than a full turn?**

Another difference between polyhedra and flat mosaics is that the design involves only a finite number of tiles. The pattern will still go on forever (in a sense), but it will not extend indefinitely into space. Naturally, I'm curious about these patterns, too.

What are all the symmetrical polyhedra?

In other words, what are all the different ways to make polyhedra out of regular polygons so that at each corner we see the same pattern? Archimedes figured out all of the possibilities. Can you?

Of course, the most symmetrical kind of polyhedron would be one where all the faces are identical, like a cube. These are called **regular polyhedra**. It is an ancient discovery that there are exactly five of these (the so-called Platonic solids). Can you find all five?

What are the five regular polyhedra?

2

What is measuring? What exactly are we doing when we measure something? I think it is this: we are making a comparison. We are comparing the thing we are measuring to the thing we are measuring it with. In other words, *measuring is relative*. Any measurement that we make, whether real or imaginary, will necessarily depend on our choice of measuring unit. In the real world, we deal with these choices every day—a cup of sugar, a ton of coal, a thing of fries, whatever.

The question is, what sort of units do we want for our imaginary mathematical universe? For instance, how are we going to measure the lengths of these two sticks?

Let's suppose (for the sake of argument) that the first stick is exactly twice as long as the second. Does it really matter how many inches or centimeters they come out to be? I certainly don't want to subject my beautiful mathematical universe to something mundane and arbitrary like that. For me, it's the proportion (that 2:1 ratio) that's the important thing. In other words, I'm going to measure these sticks relative to each other.

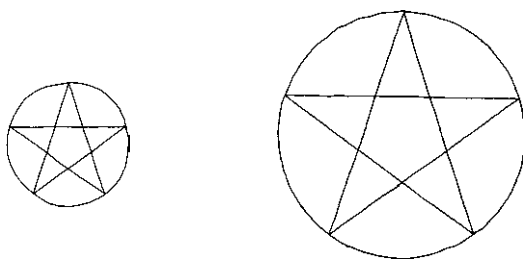
One way to think of it is that we simply aren't going to have any units at all, just proportions. Since there isn't a natural choice of unit for measuring length, we won't have one. So there. The sticks are just exactly as long as they are. But the first one is twice as long as the second.

The other way to go is to say that since the units don't matter, we'll choose whatever unit is convenient. For example, I could

choose the second stick to be my unit, or ruler, so that the lengths come out nice. The first stick has length 2, the second stick has length 1. I could just as easily say the lengths are 4 and 2, 6 and 3, or 1 and $\frac{1}{2}$. It just doesn't matter. When we make shapes or patterns and measure them, we can choose any unit that we want to, keeping in mind that what we are really measuring is a *proportion*.

I guess a simple example would be the perimeter of a square. If we choose our unit to be the side of the square (and why not?), then the perimeter would obviously be 4. What that really means is that for any square, the perimeter is four times as long as the side.

This business of units is related to the idea of scale. If we take some shape and blow it up by a certain factor, say 2, then all of our length measurements on the big shape will come out just as if we were measuring the original shape with a half-size ruler.

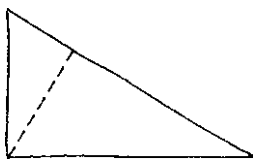


Let's call the process of blowing up (or shrinking down) **scaling**. So the second shape is obtained from the first by scaling by a factor of 2. Or, if we like, we could say that the first shape is the second one scaled by a factor of $\frac{1}{2}$.

Two figures related by a scaling are called **similar**. All I'm really trying to say here is that if two shapes are similar,

related by a certain scaling factor, then all corresponding length measurements are related by that same factor. People say that such things are “in proportion.” Notice that scaling doesn’t affect angles at all. The shape stays the same, only the size changes.

If two triangles have the same angles, are they necessarily similar? How about four-sided shapes?

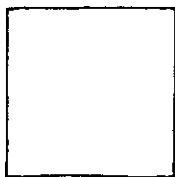


Show that if a right triangle is chopped into two smaller ones, they must both be similar to the original triangle.

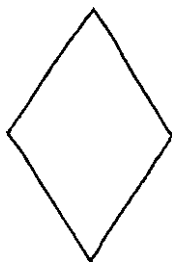
The nice thing about not having arbitrary units and always choosing to measure relative proportions is that it makes all our questions *scale independent*. To me, this is the simplest and most aesthetically pleasing approach. And given the fact that your shapes are in your head and mine are in mine, I really don’t see any other alternative. Is your imaginary circle bigger or smaller than mine? Does that question even have any meaning?

But before we can begin to go about measuring something, we need to know precisely what object it is that we are talking about.

Let’s suppose I have a square.



Now, there are some things I know about this shape right off the bat, such as the fact that it has four equal sides. The thing about information like this is that it is not really a discovery, nor does it require any explanation or proof. It's simply part of what I mean by the word *square*. Whenever you create or define a mathematical object, it always carries with it the blueprint of its own construction—the defining features that make it what it is and not some other thing. The questions we are asking as mathematicians then take this form: If I ask for such and such, what else do I get as a consequence? For example, if I ask for four equal sides, does that force my shape to be a square? Clearly, it doesn't.

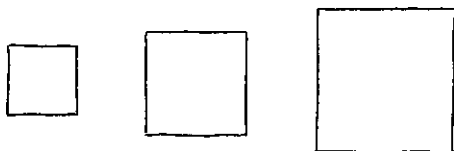


It could be a diamond shape with equal sides, a so-called **rhombus** (Greek for “spinning top”). In other words, the prescription of having four equal sides contains a certain amount of wiggle room. So one thing to always be aware of is whether you've pinned down your objects enough to get any information out of them. We can't precisely measure the

angles of an arbitrary rhombus, because that description still allows the shape the freedom to squirm around and change its angles. We need to be clear about the extent to which we have specified our objects so that we can ask well-posed, meaningful questions.

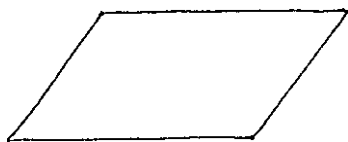
Are the opposite sides of a rhombus always parallel? Are the diagonals perpendicular?

Suppose we ask that the angles of our rhombus all be right angles. That certainly forces our shape to be a square, because that's what the word *square* means! Now is there any room left for it to wiggle around? There is in fact one more degree of freedom remaining, which is that it could change its size. (This would be relative, of course, to some other object we are considering. If all we had were a square, then size would have no meaning.)



All right, so let's suppose we select a particular length and imagine a square with sides that long. *Now* is it locked down? Yes, it is. And this has important consequences. This means that any further demands that we make on it might not be achievable. For example, if we want our square to have diagonals equal to its sides, well tough luck. We don't get to have that. Once a shape (or any structure in mathematics) is specified enough, the "forces of mathematical nature" then dictate all of its behavior. We can certainly try to find out what is true, but we no longer have any say in the matter.

In some sense, the real question about mathematical reality is: How much control do we have? How much can we ask of it? How many simultaneous demands can we make before it shatters like a glass sculpture in our hands? How pliable is mathematical reality? How forgiving and yielding? Where can we push, and where does it push back?

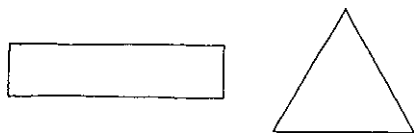


A parallelogram is a four-sided polygon with opposite sides parallel (i.e., a slanted box). Must the opposite angles of a parallelogram be equal?

Prove that a parallelogram with equal diagonals must be a rectangle.

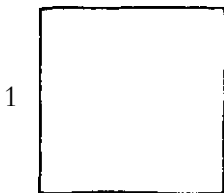
3

Two objects with the same shape (that is, similar objects) are easy to compare—the bigger one is bigger and the smaller one is smaller. It's when we compare different shapes that things get interesting. For instance, which one of these is bigger, and what does that even mean?



One idea is to compare the amount of space the two shapes take up. This measurement is usually called **area**. As with any measurement, there is no such thing as absolute area—only area relative to other areas. Our choice of unit is arbitrary; we could choose any shape and call the amount of space it occupies “one unit of area,” and all other areas could be measured against it.

On the other hand, once we make a choice of length unit, there is a natural (and traditional) choice of area unit, namely the amount of space occupied by a square of unit sides.



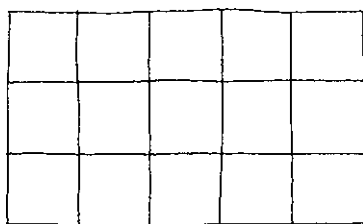
So the measurement of area really boils down to the question, how much room is my shape taking up compared to a unit square?



Suppose we cut a triangle from one corner to the middle of the opposite side.

Does the area get cut in half?

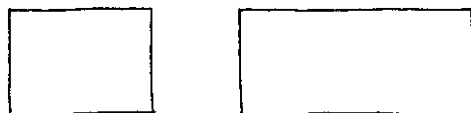
Some areas are relatively easy to measure. For example, suppose we have a 3 by 5 rectangle.



It is easy to see that we can chop this rectangle into fifteen identical pieces, each of which is a unit square. So the area of the rectangle is 15. That is, it takes up exactly fifteen times as much space as a unit square does. In general, if the sides of a rectangle are nice whole numbers, say m and n , then the area is simply their product, mn . We can just count the m rows of n squares each.

But what if the sides don't come out even? How can we measure the area of a rectangle if we can't chop it up nicely into unit squares?

Here are two rectangles of the same height.

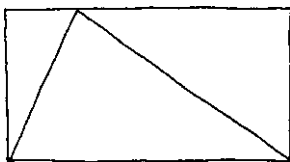


I like to think of the second one as a “stretched” version of the first. Is it clear that their areas are in the same proportion as their lengths? Stretching in one direction is called **dilation**. What we're saying is that dilation of a rectangle by a certain factor multiplies its area by that factor.

In particular, we can think of a rectangle with sides a and b as a unit square that has been dilated twice: by a factor of a in one direction and by b in another. This means that the area of the unit square will get multiplied, first by a and then by b . In

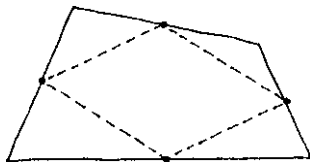
other words, it gets multiplied by ab . So the area of a rectangle is just the product of its sides. It doesn't matter whether the sides come out even or not.

What about the area of a triangle? My favorite way to think about it is to imagine a rectangular box built around the triangle. It turns out that the area of the triangle is always half that of the rectangle. Do you see why?



Why does a triangle take up exactly half of its box?

What happens to the area of the triangle as we slide the tip horizontally? What if it goes past the sides of the box?

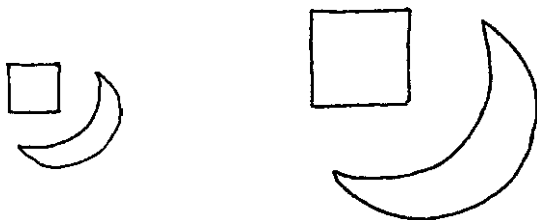


Show that when we connect the midpoints of the sides of any four-sided shape, it forms a parallelogram. What is its area?

Can a polygon always be chopped into pieces and reassembled to form a square?

One interesting feature of area is the way it behaves with respect to scaling. We can think of a scaling as being the result of two dilations by the same factor. If we have a square, and we scale it by a factor of r , then its area will get multiplied by r^2 . For example, if you blow up a square by a factor of 2, its perimeter will double, but its area will quadruple.

As a matter of fact, this will be true for any shape. The effect of scaling on area is to multiply by the square of the scaling factor, no matter what shape you're dealing with. A nice way to see this is to imagine a square with the same area as your shape.



After scaling by a factor of r , their areas will still be equal—the two shapes enclose the same amount of space whether or not I change my ruler. Since the area of the square gets multiplied by r^2 , so must the area of the other shape.

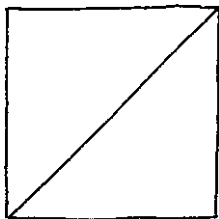
There is also the question of three-dimensional size. This is usually called **volume**. Naturally, we can take as our unit of volume that of a cube with unit sides. The first question is how to measure a simple three-dimensional box.

How does the volume of a box depend
on the lengths of its sides?

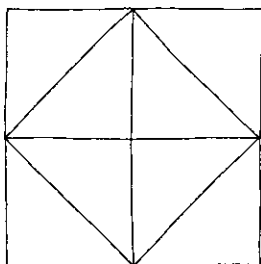
What is the effect of scaling on volume?

4

The study of size and shape is called **geometry**. One of the oldest and most influential problems in the history of geometry is this one: How long is the diagonal of a square?



Naturally, what we are really asking about is the proportion of diagonal to side. For convenience, let's take the side of the square to have length 1, and write d for the length of the diagonal. Now look at this design.



We have four unit squares coming together to make a 2 by 2 square. Notice that their diagonals also form a square. This square has sides of length d , so we can think of it as a unit square scaled by a factor of d . In particular, its diagonal must be d times as long as that of a unit square, so it must have length d^2 . On the other hand, just looking at the design we can see

that its diagonal has length 2. This means that whatever d is, d^2 must be equal to 2. Another way to see this is to notice that the d by d square takes up exactly half the area of the big square. Since the area of the big square is 4, this again says that $d^2 = 2$.

So what is d ? A good guess might be $1\frac{1}{2}$. But no, $\frac{3}{2} \times \frac{3}{2} = \frac{9}{4}$, which is greater than 2. This means d must actually be a little smaller. We can try other numbers: $\frac{7}{5}$ is too small, $\frac{10}{7}$ is too big, $\frac{17}{12}$ is very close but still not quite right.

So what are we going to do, keep trying numbers till the cows come home? What we are looking for is a proportion $\frac{a}{b}$ such that

$$\frac{a}{b} \times \frac{a}{b} = 2.$$

The only way this can happen is if the top number a when multiplied by itself is exactly twice as big as the bottom number b multiplied by itself. In other words, we need to find two whole numbers a and b so that

$$a^2 = 2b^2.$$

Since we're only interested in the ratio $\frac{a}{b}$, there is no point in looking at numbers a and b that are both even (we could just cancel any common factors of 2). We can also rule out the possibility that a is odd: if a were an odd number, then a^2 would also be odd, and there would be no way for it to be double the size of b^2 .

**Why is the product of two odd
numbers always odd?**

So the only numbers $\frac{a}{b}$ we need to consider are those where a is even and b is odd. But then a^2 is not only even but *twice* an even (that is, divisible by 4). Do you see why?

**Why is the product of two even
numbers always divisible by 4?**

Now, since b is odd, b^2 must also be odd, and so $2b^2$ is twice an odd. But we need a^2 to be *equal* to $2b^2$. How can twice an even be twice an odd? It can't.

What does this mean? It means that there simply aren't any whole numbers a and b with $a^2 = 2b^2$. In other words, *there is no fraction whose square is 2*. Our diagonal to side proportion d cannot be expressed as a fraction in any way—no matter how many pieces we divide our unit into, the diagonal will never come out evenly.

This discovery tends to have a rather unsettling effect on people. Usually, when we think about measuring something, we imagine it requiring only a finite number of applications of our ruler (including possibly dividing it into smaller, equal-size pieces). But this is not the case in mathematical reality. Instead, we find that there are geometric measurements (e.g., the diagonal and side of a square) that are *incommensurable*—that is, not simultaneously measurable as multiples of a common unit. This forces us to abandon the naïve idea that all measurements are describable as whole number proportions.

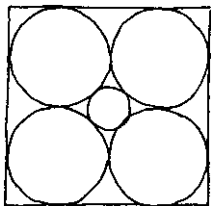
This number d we've discovered is called the square root of 2 and is written $\sqrt{2}$. Of course, this is really just a convenient shorthand way of saying, "the number that when multiplied by itself is 2." In other words, the only thing we really know

about $\sqrt{2}$ is that its square is 2. We have no hope of saying what this number *is* (at least as a whole number fraction), though of course we can approximate it. For example, $\sqrt{2} \approx 1.41432$. Whatever. That's hardly the point. We want to understand the *truth*.

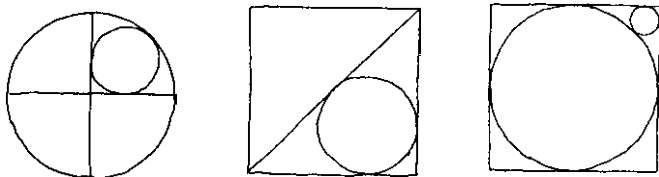
Well, the truth seems to be that we can't really measure the diagonal of a square. This is not to say that the diagonal doesn't exist or that it doesn't have a length. It does. The number is out there; we just can't talk about it in the way we want to. The problem is not with the diagonal; it's with our *language*.

Maybe it's the price we pay for mathematical beauty. We've created this imaginary universe (the only place where measurement is truly possible), and now we have to face the consequences. Numbers like this that cannot be expressed as fractions are called **irrational** (meaning "not a ratio"). They arise naturally in geometry, and we just have to somehow get comfortable with that. The diagonal of a square is precisely $\sqrt{2}$ times as long as the side, and that's all we can really say about it.

Is $\sqrt{3}$ irrational? What about $\sqrt{2} + \sqrt{3}$?



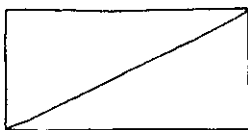
The big circles are clearly half as wide as the square. How about the small circle?



How big are these circles?

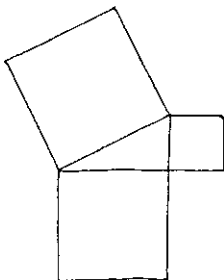
5

What about the diagonal of a rectangle?



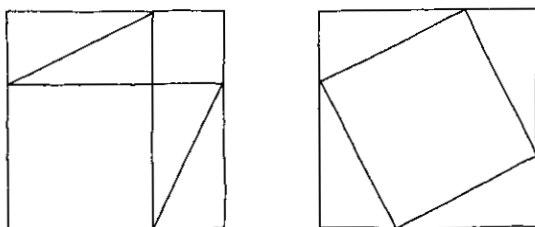
Of course, it depends on how long the sides are, but in what way? The relationship between the diagonal and the sides was discovered about four thousand years ago, and it's just as surprising now as it was then.

Notice how the diagonal cuts the rectangle into two identical triangles. Let's take one of these triangles and put a square on each of its sides.

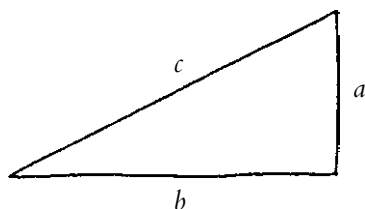


The amazing discovery is this: the big square takes up exactly as much area as the two smaller squares put together. No matter what shape the rectangle has, its sides and diagonal will always conspire to make these squares add up this way.

But why on earth should *that* be true? Here is a pretty way to see it using mosaic designs.



The first one uses the two smaller squares, together with four copies of the triangle, to make one big square. The second design uses the larger square (the one built on the diagonal) and those same four triangles to make another big square. The point is that these two big squares are identical; they both have sides equal to the two sides of the rectangle added together. In particular, this means that the two mosaics have the same total area. Now, if we remove the four triangles from each, the remaining areas must also match, so the two smaller squares really do take up exactly as much space as the larger one.



Let's call the sides of the rectangle a and b and the diagonal c .

Then the square of side a together with the square of side b has the same total area as the square of side c . In other words,

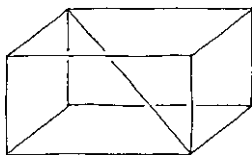
$$a^2 + b^2 = c^2.$$

This is the famous **Pythagorean theorem** relating the diagonal and sides of a rectangle. It's named after the Greek philosopher Pythagoras (circa 500 BC), although the discovery is actually far older, dating back to the ancient Babylonian and Egyptian civilizations.

For example, we find that a 1 by 2 rectangle has a diagonal of length $\sqrt{5}$. As usual, this number is hopelessly irrational. Generally speaking, a rectangle whose sides are nice whole numbers will almost always have an irrational diagonal. This is because the Pythagorean relation involves the square of the diagonal rather than the diagonal itself. On the other hand, a 3 by 4 rectangle has a diagonal of length 5, since $3^2 + 4^2 = 5^2$. Can you find any other nice rectangles like that?

**Which rectangles have whole number
sides and diagonals?**

How about the three-dimensional version? Instead of a rectangle, we can ask about a rectangular box.



**How does the diagonal of a box
depend on its three sides?**

Show that the height of an equilateral triangle is $\frac{1}{2}\sqrt{3}$ times as long as its side.

6

I think we're now in a position to do some serious measuring, but before we do, I want to address a serious question. Why are we doing this? What is the point of making up these imaginary shapes and then trying to measure them?

It's certainly not for any practical purpose. In fact, these imaginary shapes are actually harder to measure than real ones. Measuring the diagonal of a rectangle requires insight and ingenuity; measuring the diagonal of a piece of paper is easy—just get out a ruler. There are no truths, no surprises, no philosophical problems at all. No, the issues we're going to be dealing with have nothing to do with the real world in any way. For one thing, the patterns we will choose to measure will be chosen because they are beautiful and curious not because they are useful. People don't do mathematics because it's useful. They do it because it's *interesting*.

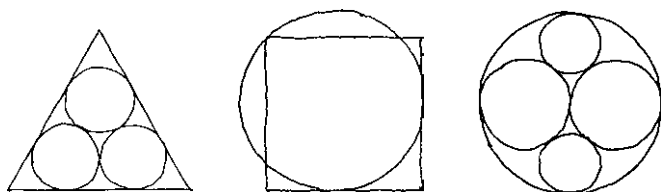
But what's so interesting about a bunch of measurements? Who cares what the length of some diagonal happens to be, or how much space some imaginary shape takes up? Those numbers are what they are. Does it really matter what?

Actually, I don't think it does. The point of a measurement problem is not what the measurement is; it's how to figure out what it is. The answer to the question about the diagonal of a square is not $\sqrt{2}$; it's the mosaic design. (At least that's one possible answer!)

The solution to a math problem is not a number; it's an argument, a proof. We're trying to create these little poems of pure reason. Of course, like any other form of poetry, we want our work to be beautiful as well as meaningful. Mathematics is the art of explanation, and consequently, it is difficult, frustrating, and deeply satisfying.

It's also a great philosophical exercise. We are capable of creating in our minds perfect imaginary objects, which then have perfect imaginary measurements. But can we get at them? There are truths out there. Do we have access to them? It's really a question about the limits of the human mind. *What can we know?* This is the real question at the heart of every mathematics problem.

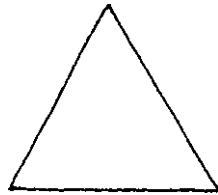
So the point of making these measurements is to see if we can. We do it because it's a challenge and an adventure and because it's fun. We do it because we're curious, and we want to understand mathematical reality and the minds that can conceive it.



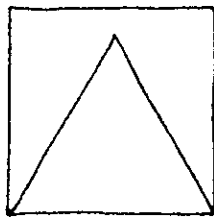
Some geometry problems speak for themselves.

7

Let's start by trying to measure the regular polygons. The simplest one is the equilateral triangle.



Having no diagonals to speak of, the interesting measurement is its area. But area compared to what? Since all measurement is relative, it makes no sense to ask how much space something takes up without something else to compare it with. I think the natural choice here would be a square with the same side length as the triangle. My favorite way to think of it is to imagine the triangle inside a square packing box.

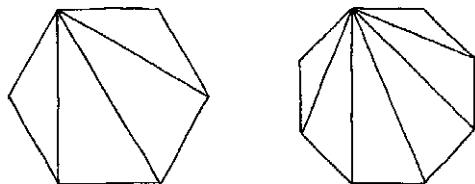


The question is, how much of the box does the triangle occupy? (Notice that this makes the question independent of any choice of units.) There's a certain number out there, intrinsic to the nature of triangles and squares, which is beyond

our control. What is it? More important, how can we figure out what it is?

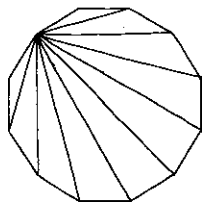
What is the area of an equilateral triangle?

It turns out that some regular polygons are easier to measure than others. Depending on the number of sides, these measurements can be more or less difficult to obtain. For instance, the regular hexagon (six sides) and octagon (eight sides) are relatively easy to measure, whereas the heptagon (seven sides) is quite spectacularly difficult.



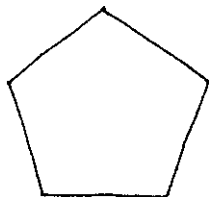
Can you measure the diagonals and areas of the regular hexagon and octagon?

Another one you might enjoy measuring is the regular dodecagon (twelve sides).

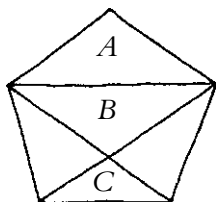


Can you measure the diagonals and area of the regular dodecagon?

One of the most beautiful (and challenging) problems in geometry is the measurement of the regular pentagon.



I want to show you a very pretty (and ingenious) way to measure the diagonal. As usual, we'll take the side of the pentagon to be our unit, and write d for the diagonal length. The idea is to chop the pentagon into triangles, like so:

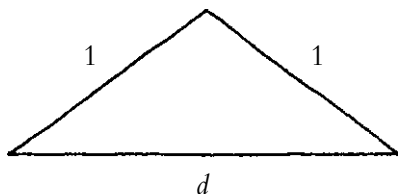


From the picture, it sure looks like triangles A and B are the same. Are they? It also looks as though triangle C has the same shape as the other two, only smaller. Does it? We're asking if the three triangles are similar. It turns out that they are. The question is why.

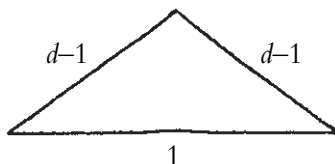
Why are the three triangles similar?

Why are the larger ones identical?

Let's start measuring the sides of these triangles. Triangle A has two short sides of length 1 and a long side of length d . The same goes for triangle B . So these two triangles look like this:



For triangle C , on the other hand, it's the long side that has length 1. What about the short sides? This is where the cleverness comes in: a short side of C and a short side of B join together to make a complete diagonal. This means that the sides of C must have length $d - 1$. Here's a picture of triangle C :



Now, the point is that these two triangles are similar. This means that the big one is a blowup of the little one by a certain factor. Comparing the long sides of the two triangles, we can see that the scaling factor must be d itself. In particular, the short sides of the little triangle, when scaled by this factor, must become the short sides of the big triangle. This means that our number d must satisfy the relation

$$d(d - 1) = 1.$$

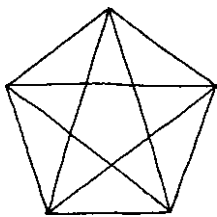
So that's it. That's our measurement. We wanted to know the proportion of diagonal to side of a regular pentagon, and now we know. It's the number that when multiplied by one less than itself equals one.

But what number is that? This is a lot like the situation we

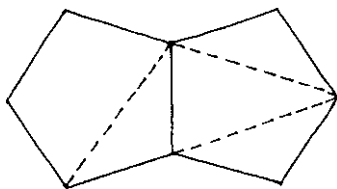
were in with the diagonal of a square; we had a number that behaved in a certain way (that time it was $d^2 = 2$) and we naturally wanted to know what that number was. What we discovered was a language problem. The irrationality of the square root of 2 meant that the language of whole number arithmetic (i.e., fractions) is not expressive enough for our needs. This forced us to fundamentally alter the way we think about measurement. Is the diagonal of the pentagon going to cause even more trouble?

We've already been obliged to enlarge our language to include not only addition, subtraction, multiplication, and division but square roots as well. This gives us a strong enough language to express the measurements of squares, and even rectangles. Is it enough for pentagons, too? Do we need to enlarge it even further?

The question is not what d is—we know what d is. It's the number that satisfies $d(d - 1) = 1$. The question is whether this number can be expressed in terms of square roots. Notice that we are no longer doing geometry. The problem has changed from one involving shapes and measurements to one about language and representation. Is our language powerful enough to allow us to untangle the relationship $d(d - 1) = 1$ to get at d itself? It turns out that it is.



How big is the small pentagon?



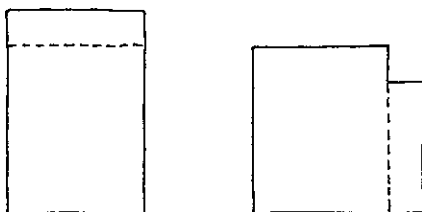
Use this configuration of two pentagons
to give an alternate proof that the
diagonal satisfies $d^2 = d + 1$.

8

The tangling and untangling of numerical relationships is called **algebra**. This kind of mathematics has a very long and fascinating history, dating back to the ancient Babylonians. In fact, the technique I want to show you is over four thousand years old.

The reason it's hard to untangle something like $d(d - 1)$ is that instead of being a square (which could then be square-rooted), it's a product of two different numbers. What the Babylonians discovered is that a product of two numbers can always be expressed as the *difference of two squares*. This makes it possible to rewrite relationships in terms of squares, so that square roots can be used to untangle them.

One way I like to think of it is to imagine the two numbers as sides of a rectangle, so their product is the area. Then the idea is to even out the sides of this rectangle by chopping some area off the top and reattaching it to the side.



This forms a square shape with a smaller square notch in it; in other words, a difference of two squares. In doing this, we are taking off exactly as much from the long side of the rectangle as we are adding to the short side. This means that the side of the square will be the *average* of the two sides of the rectangle.

As for the little square notch, its side length is just the amount by which the two sides of the rectangle differ from their average. Let's call that amount the *spread*. Then what we're saying is this: the product of two numbers is equal to the square of their average minus the square of their spread. For example, $11 \times 15 = 13^2 - 2^2$.

If a is the average of two numbers and s is their spread, then the numbers themselves must be $a + s$ and $a - s$. Our result can then be written

$$(a + s)(a - s) = a^2 - s^2.$$

This is sometimes called the *difference of squares* formula. What a beautiful piece of ancient Babylonian art! Now, here's some art for you to do.

Construct a mosaic design that demonstrates the algebraic relation $(x + y)^2 = x^2 + 2xy + y^2$.

Suppose you are given both the sum and difference of two numbers. How can you determine the numbers themselves? What if it's the sum and product that are given?

Let's use the Babylonian method to get a new description of our number d . The average of d and $d - 1$ is $d - \frac{1}{2}$, and the spread is $\frac{1}{2}$, so we have

$$d(d - 1) = (d - \frac{1}{2})^2 - (\frac{1}{2})^2.$$

Now we can rewrite our relationship $d(d - 1) = 1$ as

$$(d - \frac{1}{2})^2 - (\frac{1}{2})^2 = 1,$$

which means

$$(d - \frac{1}{2})^2 = \frac{5}{4}.$$

The point being that now we can unscramble this using square roots, to get

$$d = \frac{1}{2} + \sqrt{\frac{5}{4}},$$

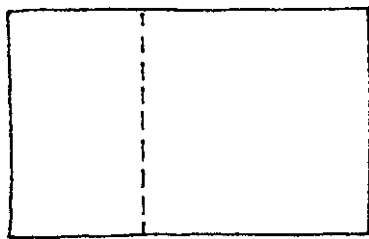
or if you prefer,

$$d = \frac{1 + \sqrt{5}}{2}.$$

Our number is now explicitly expressed in the language of whole number arithmetic and square roots.

Show that among all rectangles of a fixed perimeter, the square has the largest area.

Find a rectangle with the same area and perimeter as a given equilateral triangle.



A “golden rectangle” has the property that when a square is removed, the remaining rectangle is similar to the original. What are the proportions of a golden rectangle?

9

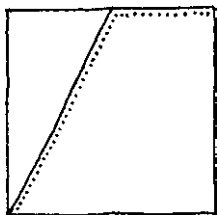
So it's true, the diagonal of a pentagon *can* be expressed in terms of square roots. In particular, it's easy to see from the expression $d = \frac{1}{2}(1 + \sqrt{5})$ that d is an irrational number, approximately 1.618. Of course, we could also have obtained that information directly from $d(d - 1) = 1$. In fact, the two expressions are equivalent in every way and tell us exactly the same things about d . There is not the slightest difference in mathematical content between the two.

I suppose the cynical view of the situation would be that we have expended a great deal of effort to go precisely nowhere. We began with a description of d as “the number that when multiplied by one less than itself equals 1,” and we ended with d described as “half of one more than the number whose square is 5.” That’s progress? If all the information about d is contained in the original equation, why did we bother solving it?

On the other hand, why bother baking bread? We could just eat the raw ingredients.

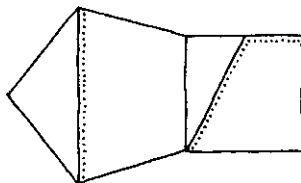
The point of doing algebra is not to solve equations; it’s to allow us to move back and forth between several equivalent representations, depending on the situation at hand and depending on our taste. In this sense, all algebraic manipulation is psychological. The numbers are making themselves known to us in various ways, and each different representation has its own feel to it and can give us ideas that might not occur to us otherwise.

For example, the representation $d = \frac{1}{2} + \sqrt{\frac{5}{4}}$ makes me think of this picture:



This path, between opposite corners of a unit square, is made up of two pieces. The horizontal part has length $\frac{1}{2}$, and the slanted part is just the diagonal of a $1 \times \frac{1}{2}$ rectangle. Pythagoras’s theorem says that the length of this diagonal is

$\sqrt{\frac{5}{4}}$. This means that the diagonal of a pentagon (which was very difficult to measure) is exactly as long as this simple path on a square.



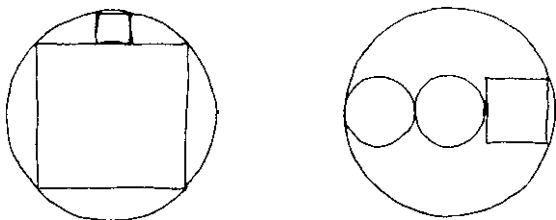
This is pretty. And totally unexpected. There's no way I would have guessed at this without that particular expression for d . In fact, even if I had guessed it (by playing with paper and rulers for instance), there would be no way for me to know if it really was a truth about mathematical reality, much less understand why, without in some way obtaining these two expressions for d and showing their equivalence. Of course, if I had somehow guessed that $d = \frac{1}{2}(1 + \sqrt{5})$ it would be easy enough to check that $d(d - 1) = 1$. The point of something like the Babylonian technique is that it allows us to express d explicitly in a certain language, without having to be such good guessers.

In general, the main task of the geometer is to translate geometric information into algebraic information, and vice versa. This is not so much a technical problem as it is a creative one. The *real* idea was the dissection of the pentagon into similar triangles. Where does such an idea come from? How can you invent something like that? I don't know. Mathematics is an art, and creative genius a mystery. Of course, technique helps—good painters understand light and shadow, good musicians have a thorough knowledge of functional harmony, and

good mathematicians can untangle algebraic information—but a beautiful piece of mathematics is just as hard to make as a beautiful portrait or sonata.

What is the area of a regular pentagon?

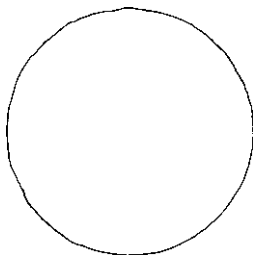
Again, I can't really help you here; you're on your own. There is a blank canvas in front of you, and you need an idea. Maybe you will have one, maybe not. That's art.



Two of my favorites.

10

How about a circle? You certainly can't ask for a prettier shape.

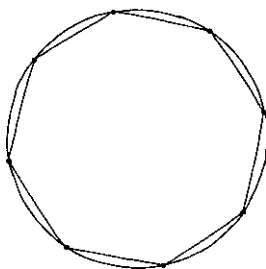


Circles are simple, symmetrical, and elegant. But how on earth are we going to measure them? For that matter, how will we measure *any* curved shapes?

The first thing to notice about a circle is that all the points on it are the same distance away from the center. That is, after all, what makes it a circle. That distance is called the **radius** of the circle. Since all circles have the same shape, it's really the radius that makes one circle different from another.

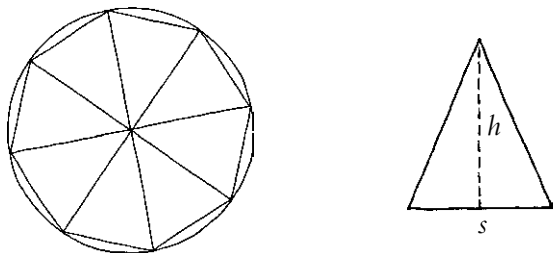
The perimeter of a circle is called its **circumference** (Latin for “carrying around”). I think the natural measurements to make of a circle are its area and circumference.

Let's start by making some approximations. If we place a certain number of equally spaced points around the circle, and then connect the dots, we get a nice regular polygon.



The area and perimeter of this polygon are smaller than the corresponding measurements for the circle, but they're pretty close. If we used more points, we could do even better. Suppose we use some large number of points, say n . Then we get a regular n -gon whose area and perimeter are really close to the true area and circumference of the circle. The important thing is that as the number of sides of the polygon is increased, the approximations get better and better.

What is the area of this polygon? Let's chop it into n identical triangles.



Each triangle has width equal to a side of the polygon, say s . The height of each triangle is the distance from the center of the circle to the side of the polygon. Let's call that distance h . Then each triangle will have area $\frac{1}{2}hs$. This means that the area of the polygon is $\frac{1}{2}hsn$. Notice that sn is just the perimeter of the polygon. So we can say

$$\text{area of polygon} = \frac{1}{2}hp,$$

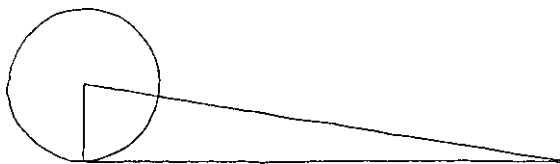
where p is the perimeter. We have a precise description of the area of a regular polygon in terms of its perimeter and the distance from center to side.

Now what happens as the number of sides n is increased indefinitely? The perimeter p will get closer and closer to the circumference C of the circle, and the distance h will approach the radius r . This means the area of the polygon must be approaching $\frac{1}{2}rC$. But the area of the polygon is also approaching the true area A of the circle! The only possible conclusion is that these numbers must be the same,

$$A = \frac{1}{2}rC.$$

The area of a circle is exactly half the product of its radius and circumference.

A nice way to think of this is to imagine unrolling the circumference onto a line, so that it forms a right triangle with the radius.

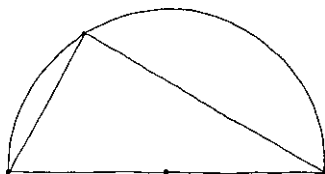


What our formula is saying is that the circle takes up exactly the same amount of space as this triangle.

Something really serious has just happened here. We have somehow obtained an exact description of the area of a circle using nothing but approximations. The point is that we didn't just make a few good approximations, we made *infinitely* many. We constructed an infinite sequence of increasingly better approximations, and there was enough of a pattern in those approximations that we could tell where they were heading. In other words, an infinite sequence of lies *with a pattern* can tell us the truth. It is arguable that this is the single greatest idea the human race has ever had.

This amazing technique, known as the **method of exhaustion**, was invented by the Greek mathematician Eudoxus (a student of Plato) around 370 BC. It allows us to measure curved shapes by constructing an infinite sequence of straight-line approximations.

The trick is to do this in such a way that the approximations have a pattern to them—an infinite list of random numbers doesn't tell us anything. It's not enough to have an infinite sequence; we have to be able to *read* it.



When a point on a circle is connected to both ends of a diameter it always makes a right angle. Why?



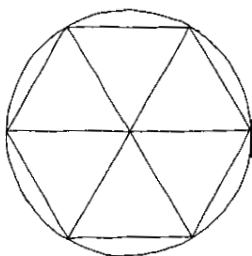
Show that if two points are connected to the same arc, the resulting angles must be the same.

11

We have expressed the area of a circle in terms of its circumference. But can we measure the circumference? With a square, it's natural to measure the perimeter in proportion to the side

length—the ratio of the length around to the length across. We can do the same thing with a circle. The distance across a circle is called its **diameter** (of course it's just twice the radius). So the analogous measurement for a circle would be the ratio of the circumference to the diameter. Since all circles are similar, this ratio is the same for every circle and is denoted by the Greek letter pi, or π . This number is to circles what 4 is to squares.

It's not hard to approximate pi. For example, suppose we put a regular hexagon inside a circle.



The perimeter of the hexagon is exactly three times the diameter. Since the circumference of the circle is a bit longer, we see that π is just a little greater than 3. If we use polygons with more sides we can get better estimates. Archimedes (circa 250 BC) used a 96-gon to get $\pi \approx \frac{22}{7}$. Many people are under the misconception that this is an exact equality, but it isn't. The actual value is a bit smaller, a decent approximation being $\pi \approx 3.1416$. Still better is the fifth-century Chinese estimate $\pi \approx \frac{355}{113}$.

But what is pi exactly? Well, the news is pretty bad. Pi is irrational (this was proved by Lambert in 1768), so there's no hope of expressing it as a ratio of whole numbers. In particular,

there is no way to measure both the diameter and circumference evenly.

The situation is actually even worse than for the diagonal of a square. Although $\sqrt{2}$ is irrational, it is at least describable as “the number whose square is 2.” In other words, this number $\sqrt{2}$ satisfies a relation that can be expressed in the language of whole number arithmetic; namely, it is the number x such that $x^2 = 2$. We may not be able to say what $\sqrt{2}$ is, but we can say what it does.

It turns out pi is different. Not only is it incapable of being expressed as a fraction, but in fact pi fails to satisfy any algebraic relationship whatsoever. What does pi do? It doesn’t do anything. It is what it is. Numbers like this are called **transcendental** (Latin for “climbing beyond”). Transcendental numbers—and there are lots of them—are simply beyond the power of algebra to describe. Lindemann proved that pi is transcendental in 1882. It is an amazing thing that we are able to know something like that.

On the other hand, mathematicians have found alternative descriptions of pi. For instance, in 1674 Leibniz discovered the formula

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \dots$$

The idea is that the more terms you add together on the right, the closer the sum gets to the number on the left. So pi can be expressed as an *infinite sum*. This at least provides us with a purely numerical description of pi, and it is also philosophically quite interesting. More important, such representations are all we’ve got.

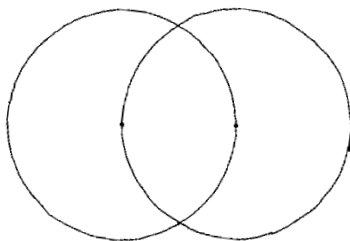
So that's the story. The ratio of circumference to diameter is pi, and there's nothing we can do about it. We'll simply have to expand our language to include it.

In particular, a circle of radius 1 has diameter 2, and so its circumference is 2π . The area of this circle is half the product of the radius and circumference, which is just π . Blowing up by a factor of r , we find that for a circle of radius r , the circumference and area are given by

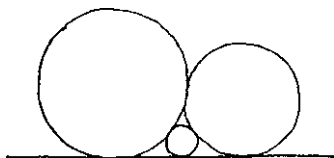
$$C = 2\pi r,$$

$$A = \pi r^2.$$

Notice that the first equation is practically content free; it is merely a restatement of the definition of pi. The second equation is the one with real depth and is equivalent to our discovery that the area of a circle is half the product of the radius and the circumference.



If two circles are arranged so that each passes through the center of the other, what are the area and perimeter of the overlap? What about for three overlapping circles?

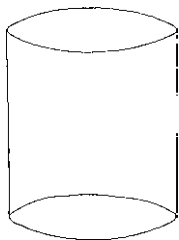


Two circles lie on a line, touching each other at a point. A small circle is inscribed in the space between. How does its radius depend on the radii of the two larger circles?

12

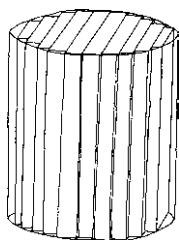
I want to say more about this method of exhaustion. The idea is to somehow get exact measurements using an infinite sequence of approximations, like how we measured circles using an infinite sequence of polygons. This is by far the most powerful and flexible measuring technique ever devised. For one thing, it reduces the measurement of curved shapes to the measurement of straight ones. It's amazing that we can say *anything* precise about curved shapes, let alone that we can do it in such a deep and beautiful way.

As another example of this technique, let me show you a nice way to measure the volume of a cylinder.



A cylinder is an interesting object. It's kind of round and it's kind of straight. It feels like it's halfway between a cube and a sphere. Anyway, it has these two ends that are circles (of the same size), one at a certain height above the other.

One way to approximate the volume of a cylinder is to imagine slicing it vertically into a large number of thin slabs and approximating the slabs by rectangular boxes.



Notice that the base rectangles of these boxes do a good job of approximating the area of the circular base of the cylinder. As the number of slices increases, the total volume of the boxes approaches the true volume of the cylinder, and the area of the rectangles approaches the true area of the circle.

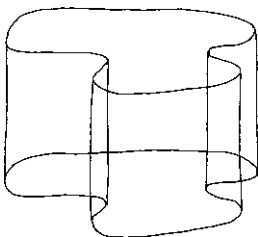
Now the volume of a rectangular box is just the product of its height and the area of its base, so the total volume of all the boxes will be their height times the total area of the rectangles. Here we have benefited from the fact that all of the boxes have exactly the same height. This means that the approximate volume of the cylinder is the height multiplied by the approximate area of the base.

This is enough of a pattern for us to read the true measurements of the cylinder. As the number of slices gets larger, and the approximations get better, the product of the height and the rectangular base area approaches both the volume of the

cylinder as well as the product of the height and the circular area. So they must be the same. In other words, the method of exhaustion has succeeded. The volume of a cylinder is simply the product of its height and the area of its base.

Two things occur to me here. One is that maybe this result is obvious to you. Is it intuitively clear that the amount of space taken up by a cylinder should be proportional to both the height and base area? I hate to be in the position of explaining something obvious. On the other hand, it's good to combine intuition with reasoning—that's what mathematics is.

The other thing is that perhaps slicing the cylinder into rectangular boxes in this way seems ugly and unnatural. After all, when we measured circles, we chopped them into a pleasing symmetrical arrangement of triangles. Why don't we chop the cylinder vertically through the center of the circle into triangular wedges? This is a perfectly valid criticism, actually. Let me answer it (as if I'm not also the one making it!) with another example.

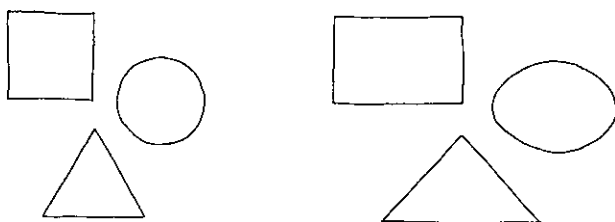


This object is made in the same way as the cylinder, except the top and bottom faces are no longer necessarily circles but perhaps some other figure. Let's call this sort of thing a **generalized cylinder**. In this case there isn't a nice symmetrical way to chop it up, so the rectangular slicing idea is as good a way as any. We still get the volume of the generalized cylinder as the product of

its height and base area. My point is that slicing this way works whether there is symmetry or not. It's a good example of the flexibility of the exhaustion technique.

**How can we measure the surface
area of a (generalized) cylinder?**

Now I want to show you just how powerful our method is. A while back, we were talking about dilation. This is the idea of stretching by a certain factor in one direction only. Sometimes I like to think of it as a transformation of the entire plane surface, like pulling opposite sides of a sheet of rubber. Any shapes or figures drawn on the plane will get dilated accordingly. Suppose we have some shapes and we subject them to a (horizontal) dilation by some factor.

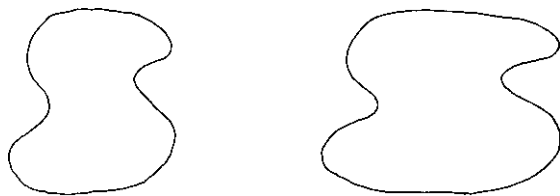


Notice how dramatically the shapes are affected. For instance, the square becomes a rectangle (so its sides aren't all the same length anymore). The equilateral triangle is transformed into a mere isosceles triangle, and the circle becomes an entirely new shape known as an **ellipse**.

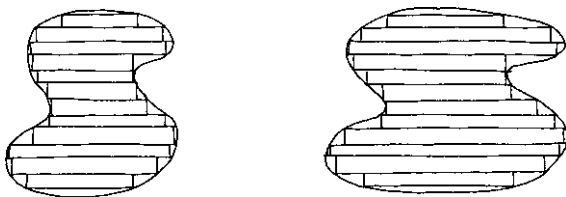
In general, dilation is a pretty destructive process. Lengths and angles often get severely distorted. In particular, there is usually no relationship whatever between the perimeter of a

shape before dilation and its perimeter afterward. The perimeter of an ellipse, for instance, is a very difficult classical measurement problem, mainly because it has no connection with the circumference of a circle.

On the other hand, dilation turns out to be very compatible with area. We already know how dilation affects the area of rectangles: if a rectangle is dilated by a certain factor (in a direction parallel to one of its sides), then its area gets multiplied by that same factor. Using the method of exhaustion, we can see that this remains true for any shape whatsoever. To be precise, let's suppose we have some shape, and we dilate by the factor r in some direction. We want to see that the area of the shape will get multiplied by r .



The idea is to slice our shape into thin rectangular strips, parallel to the direction of dilation, so that the area of our shape is closely approximated by the total area of the rectangles.



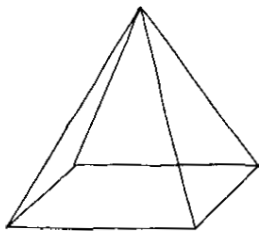
After dilation, the strips become dilated as well, so their areas get multiplied by r . This means that the approximate

area of the dilated shape is exactly r times the approximate area of the original. Letting the number of strips increase indefinitely (so that their thickness approaches zero), we see that the true areas must be related by a factor of r as well. I think it's quite surprising and wonderful that we can keep track of the area of a shape even after such an extreme distortion.

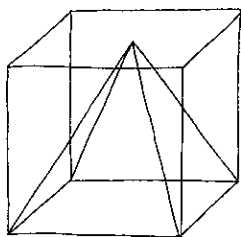
What is the area of an ellipse?

Similarly, a dilation of space by a certain factor in a given direction will magnify volumes by that factor. Do you see why? Since boxes behave well under dilation so will anything. Of course, we have to be somewhat careful. For instance, if a solid object is dilated by a factor of 2, its volume will indeed double, but its surface area generally will go nuts. Try it with a cube!

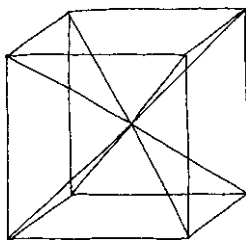
Now I want to show you a truly beautiful measurement (I hope those are the only kind I ever show you). We're going to measure the volume of a pyramid.



My favorite way to do this is to place the pyramid in a box with the same base and height. I like to think of this box as the carrying case for the pyramid.



The natural question is, how much of the volume of the box does the pyramid take up? This is a pretty hard problem. It is also very old, dating back to ancient Egypt (naturally). One way to begin is to notice—very cleverly—that a cube can be divided into pyramids by joining its center to each of its eight corners.

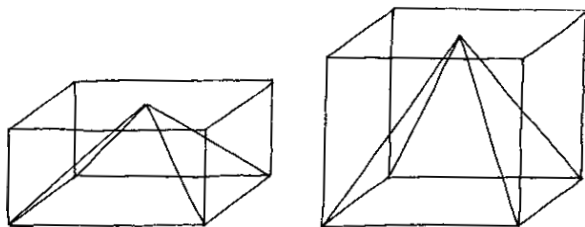


There are six of these pyramids because there is one for each face of the cube. These pyramids are all identical, so each must have a volume equal to one-sixth of the cube. The carrying case for one of these pyramids would be half the cube. So these pyramids take up exactly one-third of the volume of their carrying cases. I think that's a really pretty argument.

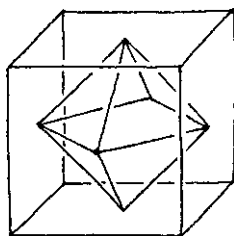
The trouble is this only works for that particular shape of pyramid (where the height is exactly half the length of a side of the base). Most pyramids won't fit together to make anything nice at all. They're either too steep or too shallow.

Does this mean that we can only measure one particular shape

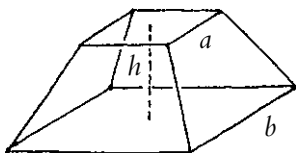
of pyramid? Not at all! The point is that *any* pyramid can be obtained from one of these special ones by appropriate dilations. If we want a steeper one, we can dilate vertically by whatever factor we need to until the height matches the one we want.



Now, here's my favorite part: dilation affects the volume of the pyramid and the volume of the carrying case in exactly the same way. They both get multiplied by the dilation factor. This means that the ratio of the two volumes is unaffected. Since the proportion was one-third for the special pyramids, it must be the same for any pyramid. So the volume of a pyramid is always exactly one-third the volume of the box it sits in. I just love that sequence of ideas. Notice how subtly and powerfully the method of exhaustion plays its role.

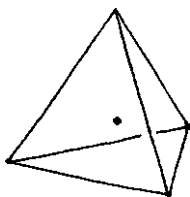


The centers of the faces of a cube can be joined to form a regular octahedron. How much of the volume of the cube does it take up?



A square of side a is placed at a height h above a square of side b , forming an incomplete pyramid.

How does its volume depend on a , b , and h ?



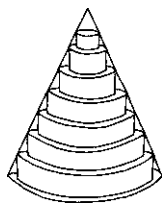
Where is the center of a regular tetrahedron?

13

Let's try to measure the volume of a cone.



I hope you agree that this is a beautiful and interesting shape. Naturally, some sort of exhaustion technique is called for, and the first idea that comes to mind is to approximate the cone by a stack of thin cylinders.



As the number of slices increases, and the cylinders get thinner, their total volume will approach the true volume of the cone. All we need to do is to figure out what the pattern is to these approximations and see where they are heading.

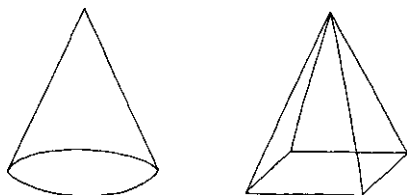
Unfortunately, this turns out not to be so easy. The volume of each cylinder depends on its radius, and these radii are varying as we move up and down the cone. It's a bit delicate. In fact, it would require a fairly expert algebraist to read this pattern and understand its behavior.

**Can you figure out the pattern
to these approximations?**

The truth is, exhaustion can be an almost impossibly difficult technique to put into practice. Even when a shape is relatively simple and we have a highly organized way to approximate it, the resulting sequence of approximations may have a pattern that is simply too subtle for us to predict. It's one thing for us to say that we will figure out where the sequence is heading, it's quite another to actually do it. If our shape is at all complicated, this is next to impossible. So how are we going to solve our problem?

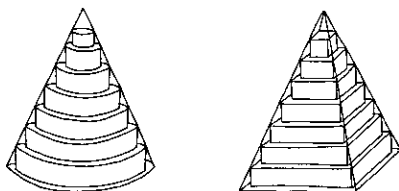
It is an aesthetic principle almost universally acknowledged among mathematicians that *the best way to solve a problem is to find an ingenious way not to have to solve it at all.*

So we're not going to measure the volume of the cone directly. Instead, we're going to compare it with another object whose volume we already understand—a pyramid.

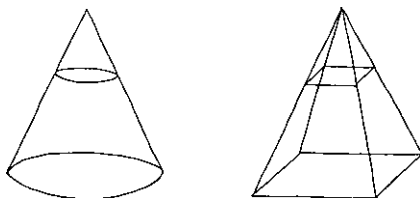


Imagine a pyramid of the same height as the cone, with a square base equal in area to the circular base of the cone. The idea is to show that these two objects have the same volume.

To do this, let's go back to our slicing idea. This time, we'll slice both the cone and the pyramid.



The cone is approximated by a stack of cylinders, and the pyramid by a stack of rectangular boxes. If we do this in the right way, then each cylinder will correspond to a box of the same thickness. The bases of these pieces will be **cross-sections** of the cone and pyramid.



Notice that when we slice through a cone like this we create a small cone on top. This little cone has exactly the same shape as the original, only smaller. In other words, it is a scaled-down version of the large cone. Same for the pyramid. In fact, since the cone and the pyramid have the same height, and the little ones also have the same height, the scaling factor must be the same for both shapes. Since the original cone and pyramid have equal base areas, so must their scaled versions.

The point I'm trying to make is that no matter what height we slice at, the cone and the pyramid will have equal cross-sectional areas. While I'm at it, I'd like to be able to say that they have "equal cross-sections" and have you understand that I mean the areas match, not that the cross-sections are necessarily the same identical shape. So the cone and pyramid have equal cross-sections.

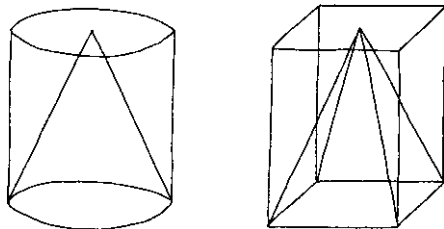
This means that back in our volume approximation, corresponding cylinders and boxes have equal bases. Since they also have the same thickness, their volumes must agree. So each little cylinder has the same exact volume as the corresponding box. In particular, the total volume of all the cylinders must equal the total volume of all the boxes. This means that no matter how many slices we make, the approximations for the cone and pyramid will always match.

As the slices get thinner, these approximations are then heading simultaneously toward the volume of the cone and the volume of the pyramid. It must be that these volumes are exactly the same. In other words, the volume of a cone is equal to the volume of a pyramid with the same height and base area.

What I love about this is that we don't have to figure out

what the patterns to these approximations are, only that they're the same as each other. We manage to avoid a difficult algebraic computation by making a well-chosen comparison.

To make this even nicer, let's put our cone inside a cylinder. This is analogous to the pyramid sitting inside its box.



Since the cylinder and the box have the same size base and height, they must have the same volume as well. In particular, the cone must take up exactly one-third of its carrying case just as the pyramid does.

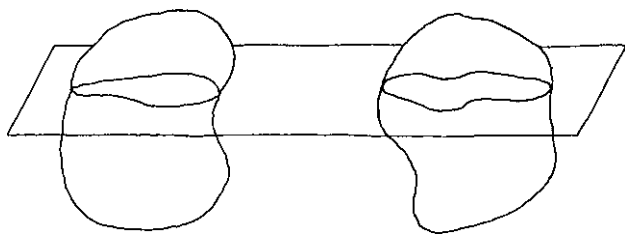
How can we measure the surface area of a cone?

Can you find a cross-section of a cube
that is a regular hexagon?

Now, it turns out that the cone is just the tip of the iceberg. This comparison idea is really quite general. Any solid object can be approximated by a stack of thin (generalized) cylinders, and if two such objects can be arranged so that they have equal cross-sections, then the method of exhaustion will guarantee that they have equal volumes.

This is a very old and beautiful result, known as the **Cavalieri principle**. (Although it was originally developed by Archimedes,

the technique was rediscovered in the 1630s by Galileo's student Bonaventura Cavalieri.) The idea is not to calculate volumes but to compare them; the trick is in choosing the right objects to compare.



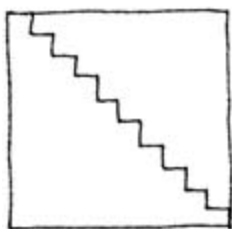
In order to use the Cavalieri principle, it is necessary to position the two objects in space so that for *every* horizontal slicing plane, the corresponding cross-sections always have equal areas. (In particular, the objects must have the same height.) This ensures that no matter how fine the cylindrical approximations become, they will continue to agree in volume.

It is also important to understand that this is not the only way that two objects can have the same volume. It is easy to find solids of equal volume that do not have equal cross-sections.

Unfortunately, the Cavalieri principle doesn't work for surface area. The cylindrical approximations that are good for volume do not do a good job of approximating the surface area. (This is a rather subtle point, actually.) In any case, there are plenty of objects with equal cross-sections and different surface areas. Can you find some?

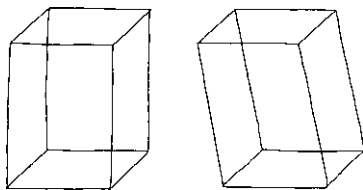
Can you find two objects with equal cross-sections and different surface areas?

Can you devise a Cavalieri principle
for areas in the plane?



Why can't the method of exhaustion be used in
this way to measure the diagonal of a square?

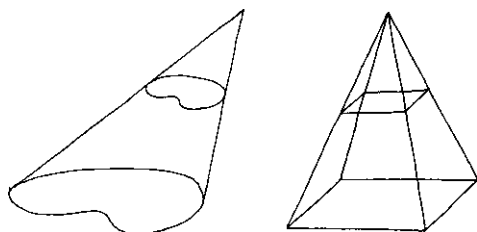
A really simple way to use the Cavalieri principle is when the two objects have identical cross-sections. That is, not only do the cross-sections have the same area, they are actually the exact same shape.



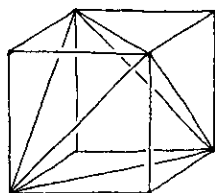
These two boxes have the same square base and the same height, but one is straight up and the other is slanted. If we look at the horizontal cross-sections, we can see that they are all the same—the same square as the base. It's as if the different cross-sections have merely slid into new positions with their shape unchanged. The Cavalieri principle tells us that these two boxes have the same volume.

The point is, as long as we simply move the various cross-sections around (we could even rotate them), their areas won't change, and the two solids will have the same volume. Of course, there is nothing special about squares; this would work with any shape.

What about a slanted pyramid? Or a slanted cone? We could even imagine a sort of **generalized cone**, which has any shape whatever for its base and comes to a point at some height above it.

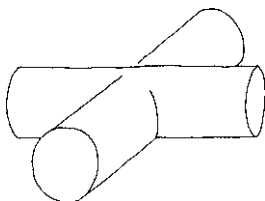


If we build a pyramid again, with the same base area and height as the generalized cone, the scaling argument from before tells us that the corresponding cross-sections are equal. What this means is that the volume of any generalized cone is simply one-third that of the corresponding generalized cylinder.



The diagonals of a cube form a regular tetrahedron. How much of the cube does it take up?

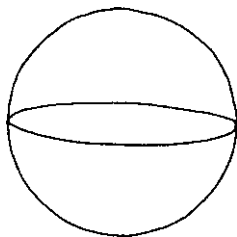
What are the volumes of the Platonic solids? How about the other symmetrical polyhedra?



Suppose two identical cylinders meet at right angles. What does their intersection look like, and what is its volume? What about three mutually perpendicular cylinders?

14

The most spectacular use of the Cavalieri principle was made almost two thousand years before Cavalieri was born. This was Archimedes's measurement of the volume of a sphere.

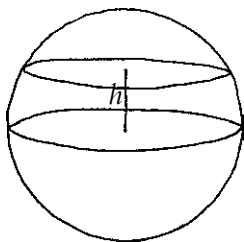


To my mind, this is the simplest and most elegant object imaginable. It is completely symmetrical—every point on the

surface is the same distance away from the center. We'll call that distance the *radius* of the sphere, just as we did for circles.

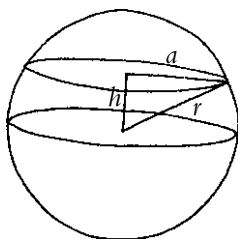
Now, the idea is to construct a different object that has the same cross-sectional areas as the sphere. Then the Cavalieri principle will tell us that their volumes are the same. Of course, this will be pointless unless the new shape is somehow easier to measure.

Suppose our sphere has radius r . Let's see if we can figure out the areas of the various cross-sections. Imagine that we make a horizontal slice at a certain height h above the equator.



The cross-section will be a circle. Let's say it has radius a . Of course, how big a is will depend on the height of the slice. When we slice through the middle of the ball (so that the height h is zero), the cross-section is the full equator. The radius a will be equal to r , the radius of the sphere. As we move up and h increases, the cross-sections will get smaller, so a will decrease. Finally, at the North Pole, a will be zero. The cross-section will be a single point.

We're going to need to know precisely how this cross-sectional area depends on the height. Luckily, this is not too difficult.



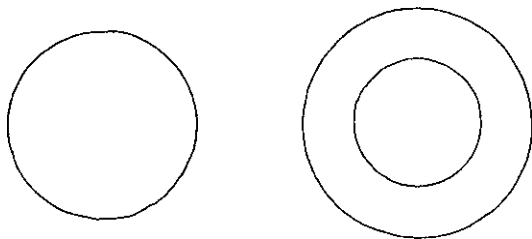
Any point on the circumference of the cross-section will be at distance r from the center of the sphere, so there is a right triangle with short sides h and a and long side r . Pythagoras's theorem tells us that

$$a^2 + h^2 = r^2.$$

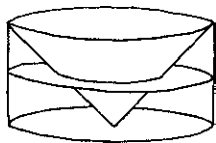
This means that the area of the cross-sectional circle is

$$\begin{aligned}\pi a^2 &= \pi(r^2 - h^2) \\ &= \pi r^2 - \pi h^2.\end{aligned}$$

This has a nice geometric interpretation. It says that the cross-sectional area is the same as the difference between the area of a circle of radius r and a circle of radius h . In other words, the area of a *ring*.

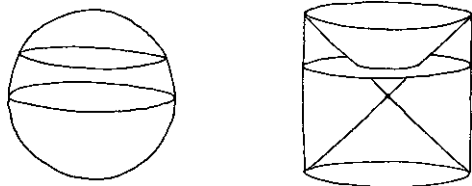


As the height h of the cross-section increases, this “pineapple ring” gets thinner. The outer radius stays the same, while the inner radius grows. Archimedes realized that these rings are precisely the cross-sections of a cylinder *with a cone removed*.



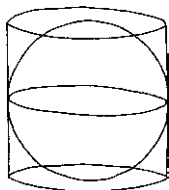
The cylinder has radius r and height r , and so does the cone. This makes it so that any slice of the cone will have the same radius as its height, which is precisely what the inner radius of the pineapple ring is supposed to do. Since the outer circle always has radius r , we can see that Archimedes was right.

We can now construct a solid with the same cross-sectional areas as the sphere. We simply glue together two copies of the cylinder with the cone removed, one for the top half of the sphere and one for the bottom. In other words, we get a cylinder with a *double cone* removed.



Since these two objects have the same cross-sections at every level, the Cavalieri principle says that they must have the same volume. Is that great, or what!

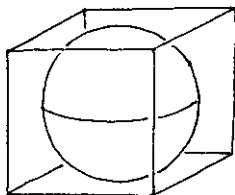
The thing is, we know how to deal with cones and cylinders. The two cones together must take up one-third of the cylinder since each one is taking up one-third of its own half of the cylinder. So the volume of Archimedes's solid is two-thirds of the volume of the cylinder. Notice that this cylinder has the same radius and height as the sphere itself. So we can say, as Archimedes himself did so long ago, that a sphere takes up exactly two-thirds of the cylinder it sits in.



This is measurement at its finest. Of course, if you prefer, we can express the volume of the sphere solely in terms of its radius. The volume of the cylinder would then be

$$\pi r^2 \times 2r = 2\pi r^3,$$

and so the volume of a sphere of radius r is $\frac{4}{3}\pi r^3$.



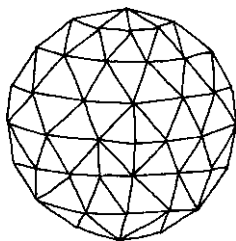
How much of a cube does a sphere occupy?

Is it more than half?



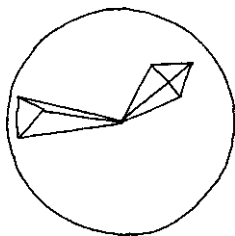
Show that a cone in a hemisphere occupies exactly half the volume.

While we've got the sphere sitting here, let's measure its surface area. I want to mimic our treatment of the circle, where we used polygons to approximate both its area and circumference. Now the idea will be to approximate the sphere using polyhedra with many faces.



It doesn't particularly matter how we do this, as long as the faces get smaller and smaller as we go. This will ensure that the volume and surface area of the polyhedron will approach those of the sphere. To keep it simple, let's suppose that all the faces are triangles.

To measure the volume of this polyhedron, we'll break it into pieces. If we connect the center of the sphere to the corners of each face, we form a bunch of thin triangular pyramids.



The volume of the polyhedron is the sum of the volumes of all of these little tetrahedrons. This is analogous to how we chopped up the polygonal approximation to the circle into lots of triangles.

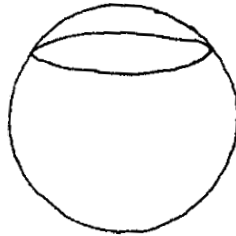
Now, here's the idea. The heights of these little pyramids are all very close to the radius r of the sphere. So the volume of each pyramid is roughly one-third the radius times its base area. Putting these together, we get that the total volume of the polyhedron is about one-third the radius times its surface area. This is only approximate, because those heights weren't quite equal to the radius, but it gets closer and closer to the truth.

What this means is that for the sphere, the volume V and the surface area S satisfy $V = \frac{1}{3}rS$ exactly. If we want to, we can combine this with $V = \frac{4}{3}\pi r^3$ to get

$$S = 4\pi r^2.$$

This is a beautiful measurement. It says that the surface area of a sphere is exactly four times the area of an equatorial circle.

Show that the surface area of a sphere is exactly two-thirds that of its (closed) cylinder.



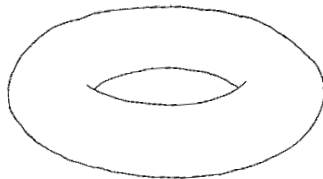
What are the volume and surface
area of a spherical cap?

15

Now I want to tell you about a really pretty discovery that was made in the early fourth century, as the classical period of geometry was coming to an end. The idea first appears in a collection of mathematical writings by the Greek geometer Pappus of Alexandria (circa 320 AD).

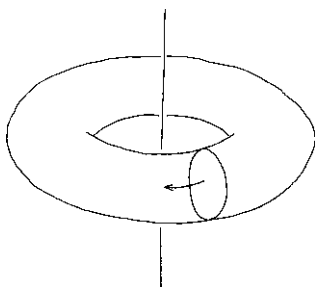
I have to say right from the start that I am a bit apprehensive about getting into this subject. Certain aspects of it are quite delicate, and it's not clear to me how I'm going to explain them. (There may be points where I simply have to throw up my hands.)

Let's start with a doughnut—I mean the shape, not the snack.



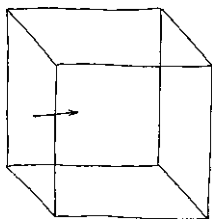
Up to now, we haven't really needed to be very precise about describing our shapes. They consist of points, in a plane or in space, arranged in some simple and pleasing way. We're familiar enough with the idea of a sphere, or a cone, or a rectangle. But what exactly *is* a doughnut?

The way I like to think of it is to imagine a circle being rotated around a line in space.



This abstract geometric kind of doughnut is called a **torus** (Latin for “cushion”). A torus is the object traced out by a circle that is moving through space along a circular path.

I think this is a very significant idea, this way of describing one shape via the motion of another. Not only does it provide us with new and exotic shapes like the torus, but it also allows us to look at some familiar objects in a new way. For instance, a cube can be thought of as being traced out by a square moving along a straight path.



Sometimes I like to pretend that the square is a prehistoric animal that crawled along this path millions of years ago. The cube is the “fossil record” of its struggle. Another image I have is of tracks in the snow. A rectangle is the “track” made by a sideways moving stick.



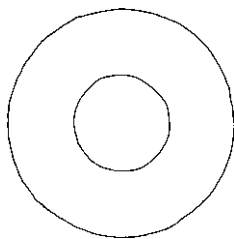
In any case, the point is that many beautiful shapes can be viewed as being the result of a motion of some kind.

Can you think of two different ways that a cylinder can be regarded as the result of a motion?

The question is whether thinking of a shape in this way helps at all in the measuring department. This is part of a recurring theme in geometry, the relationship between *description* and *measurement*. How does the measurement of an object depend on the way in which it is described?

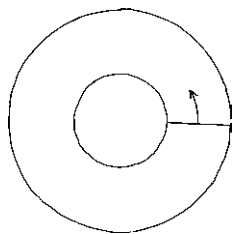
In particular, if an object is obtained from the motion of some simpler shape, precisely how are its measurements related to that shape and the way it moves? This is the question that Pappus of Alexandria was asking sixteen centuries ago, and it is his great discovery that I want to try to explain to you.

I'd like to start with the pineapple rings that we looked at before when we were measuring the sphere.



What we're talking about is the space between two concentric circles. This kind of region is called an **annulus** (Latin for "ring"). Naturally, we can think of it as a large circular region with a smaller one removed.

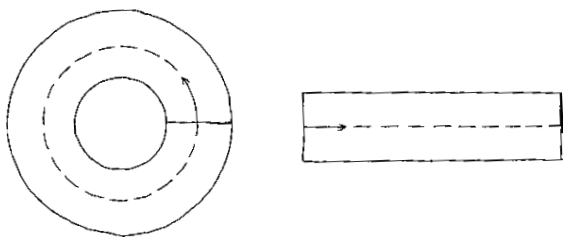
On the other hand, an annulus shape can also be viewed as being swept out by a stick moving along a circular path, like a snowplow going around a tree.



Of course, if the stick (or snowplow) were to travel along a straight path, it would sweep out a rectangle. We can now see the annulus and rectangle as being related aspects of the same idea—shapes formed by a moving stick. This is interesting because geometrically an annulus is very different from a rectangle. If you tried to bend a rectangle into a ring, for instance, it wouldn't work out very well; the inner edge would buckle and the outer edge would rip. Not a pretty picture.

The interesting question about rings and rectangles is how to compare their areas. Suppose we take a stick and drive it

around a circular path to form an annulus. How long should a straight path be in order to sweep out the same area? This is just the kind of thing Pappus was wondering about.



I think it's reasonable to expect the right length to be somewhere in between the inner and outer circumferences of the ring. A natural guess would be the middle circumference. Let's suppose we arrange it so that the rectangle is exactly as long as this "average" circle. Do the areas necessarily match?

It turns out they do. In fact, there is a very nice way to see this, which is connected to the Babylonian difference of squares relation, $x^2 - y^2 = (x + y)(x - y)$.

Here's the idea. Our annulus is completely determined by the radii of its inner and outer circles. Let's call the outer radius R and the inner radius r . Thinking of the annulus as the difference between two circles, we get that its area is simply $\pi R^2 - \pi r^2$.

For the rectangle, we'll need to know the length of the stick and the length of the path. The stick is easy; it's just $R - r$. Do you see why? The circle through the middle of the annulus really is an average, in the sense that its radius is the average of the inner and outer radii. In other words, the radius of the middle circle is $\frac{1}{2}(R + r)$.

Since the circumference of a circle is always 2π times as long

as its radius, the length of the path (and hence the length of the rectangle) must be

$$2\pi \times \frac{1}{2}(R + r) = \pi(R + r).$$

Finally, the area of the rectangle is the product of its length and width, or

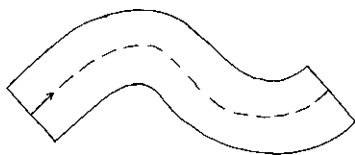
$$\begin{aligned}\pi(R + r)(R - r) &= \pi(R^2 - r^2) \\ &= \pi R^2 - \pi r^2,\end{aligned}$$

which is precisely the area of the annulus. I love how the algebra and geometry connect here. The difference of squares relation is reflected geometrically by the equivalence of ring and rectangle.

A nice way to think about it is to observe that the middle circle is simply the path traced out by the midpoint of the stick. In other words, it's the distance that the *center* travels that is important. Specifically, we have found that if the center of a stick travels along a circular path of a certain length, it sweeps out the same area as it would if the path were straight. In either case, the area is simply the product of the length of the stick and the length of the path.

This is a nice example of the way description (the annulus is described by the motion of the stick) affects measurement (the area depends in an elegant way on the stick and the path). Like I said, the connection between description and measurement is what geometry is all about.

We can take this example a bit further. Suppose we were to push the stick (by its midpoint) along some arbitrary path.

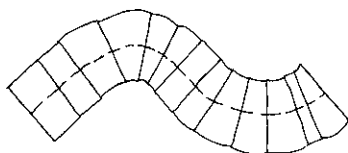


Does our result remain valid? Can we say that the area of the swept out region is the same as if it were straight? Is it simply the product of stick length and path length, or are we pushing our luck?

In fact, it's true regardless of the shape of the path. Let me see if I can explain why. First of all, notice that it works for paths that are partial circles, or *arcs*.



This is because both the arc length and the swept out area are in the same proportion to those of a complete annulus. In particular, the result holds for tiny annular “slivers” as well as for very thin rectangles. The idea is to piece these together to form more complicated shapes.



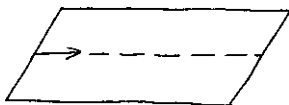
The various paths taken by the center of the stick fit together to form one big path made up of tiny circular and straight sections. By arranging these properly, we can make a path that approximates any desired path as closely as we wish.

In particular, we can (by creating an infinite sequence of such

approximations) make the total length of our path approach the length of the desired path, and the area of our conglomeration of slivers will approach the true area of the desired region. Since the approximate area is the product of the length of the stick and the length of the path, and this remains true as the approximations improve, it must be true for the actual region under consideration. Again, the method of exhaustion comes to the rescue.

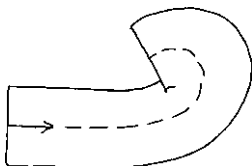
This is our first example of the amazing generality of Pappus's result: the area of a region swept out by a moving stick is the product of the length of the stick and the distance traveled by the center of the stick.

There are a couple of subtleties here. The first is that for this to work, the stick must remain perpendicular to the direction of motion at all times. Pushing the stick at an angle messes it up.



For example, the Pappus theorem fails miserably for a slanted rectangle. Since we assembled our shapes, at least approximately, from slivers of rings and rectangles, where the stick and the path are always at right angles, this is the only kind of motion our method can handle. Perpendicular motion is an essential ingredient of the Pappus philosophy.

The other issue is self-intersection.

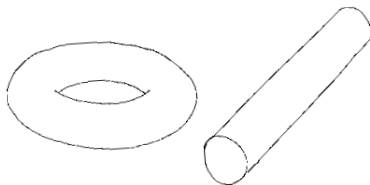


If the path curves too sharply, we'll end up tracing out parts of the region twice, and those areas of overlap will get counted double. As long as we stay perpendicular, and avoid sharp turns, we're fine.

**What is the perimeter of a region
formed by a moving stick?**

16

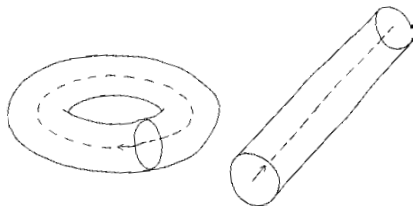
Now how about that doughnut? Since a torus is described by a circle traveling along a circular path, it makes sense to look at the object traced out by the same circle moving along a straight path. In other words, a cylinder.



This time, the snowplow is a circle. Actually, to be more precise, it is an entire **disk**, a solid filled-in circle. (It's customary to use the word *circle* for the curve itself and *disk* for the region it surrounds.) So we're pushing a disk through a solid mass of snow, creating both a torus and a cylinder.

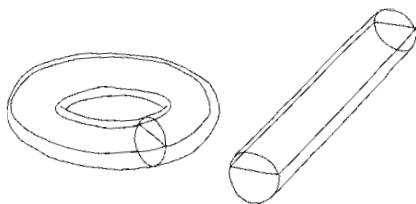
The question is, how long should the cylinder be to enclose the same volume as the torus? Notice that as the disk moves around the torus, its points trace out circular paths in space. These paths have various lengths. Which is the right one for the cylinder?

Our experience with the ring would suggest the average one. In other words, the path described by the center of the disk.



Suppose we arrange it so the length of the cylinder matches the length of this middle circle through the torus. Then we have the same disk moving through the same distance, but in two different ways, straight and circular. Are the volumes necessarily equal?

As a matter of fact, they are. Let me show you a pretty way to see this using the Cavalieri principle. We'll need to imagine taking horizontal slices of both objects.



The cross-sections of the cylinder aren't hard to visualize; they're rectangles. They all have the same length, namely the length of the cylinder, whereas their widths vary depending on the height of the slice. In fact, we can read the width of a cross-section on the circular base of the cylinder. It's the length of a horizontal slice through the disk at that height.

The cross-sections of the torus are a little bit more complicated; they are rings. As the height of the slice changes, both

the inner and outer edges vary in size. On the other hand, the middle circles of these rings are all the same. This is because horizontal slices through a disk are centered, due to the symmetry of the circle. So all these rings have the same length through the middle.

What this means is that the torus and the cylinder have cross-sections that are easy to compare. If we slice at the same height, we get an annulus and a rectangle that have the same width and also the same length. In particular their areas must agree. Since this is true no matter where we slice, the Cavalieri principle tells us that the volumes are equal.

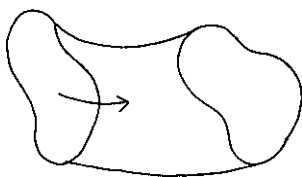
What we've shown is that the volume of a torus described by a disk moving along a circular path is simply the product of the area of the disk and the length of the path. So if we take a circle of radius a and push its center along a circle of radius b , we get a torus whose volume is given by

$$V = \pi a^2 \times 2\pi b.$$

This is a beautiful example of the Pappus philosophy. Again the notion of center plays an important role. Notice also that the moving disk remains at all times perpendicular to its direction of motion.

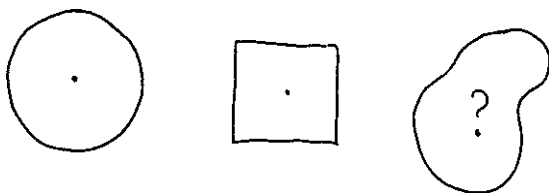
Just as before, we can generalize this result to arbitrary paths in space, using approximations made from slivers of tori and cylinders. So the volume of any solid formed by moving a disk perpendicularly along a path through its center can be obtained by multiplying the area of the disk by the length of the path.

The remarkable step that Pappus took was to generalize this further, replacing the disk by *any* plane figure.



Let's imagine a flat region of some fixed shape, dragged through space in such a way that it remains perpendicular to its direction of motion. This traces out some ridiculous solid. Pappus discovered that even the volume of this crazy blob obeys the same pattern as before: it is the product of the area of the original region and the length of a certain path. Naturally, this is the path taken by the average point in the region. But what does that mean exactly?

For symmetrical shapes like a circle or a square, there is a clear candidate, a center. Where is the center of an asymmetrical region?

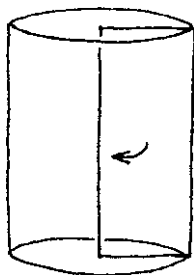


It turns out there is a way to define the center of an object, regardless of its shape. It even has a nice physical description: it's the place where if you put your finger there, the object would balance. This point is unique to each shape and is called its **centroid** (the corresponding physical notion is *center of mass*.) The problem for geometers is to make sense of this idea in a purely abstract way, since geometric objects are imaginary and don't have any actual mass or ability to

balance. That this can be done is wonderful, but not so easy to explain. I think I'll leave it for you as a nice, open-ended research project:

**How should we define the centroid of
a shape? Can we do it in such a way
that Pappus's theorem holds?**

In any case, the point is that every object has a centroid, and Pappus's great discovery is this: the volume of a solid described by a moving plane figure is equal to the product of the area of the figure and the length of the path traced out by its centroid. (Provided, of course, that the figure remains perpendicular to the direction of travel, and there is no self-intersection caused by sharp turns.) Notice that our discovery about moving sticks fits right in with this general philosophy. The centroid of a stick is just its midpoint.



**Show that Pappus's theorem works for a cylinder
formed by rotating a rectangle.**

Finally, there is the issue of surface area. How can we measure the surface area of a torus? This time we're only interested in

the crust of the doughnut. It is the circle itself that traces out the surface, not the disk. In other words, it's the points on the circle as it goes around that describe the surface we want to measure.

Again, it turns out to be the same as if we moved straight. The surface area of a torus is the product of the circumference of the moving circle and the length of the central path. So the surface area of the torus we looked at before would be

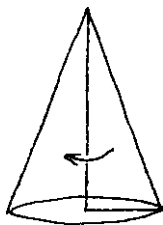
$$S = 2\pi a \times 2\pi b.$$

In general, the surface area of an object traced out by a moving plane figure is the product of the perimeter of the figure and the length of the path traced out by a certain point. (That is, assuming the shape doesn't rotate as it travels along the path.) Now, however, it's not the centroid of the region that matters, it's the centroid of the perimeter.

For a circle, or some such symmetrical object, these two notions of center coincide, but in general they don't. To get a rough idea, imagine two physical models made in the same flat shape. One is solid metal. The other has just a rim of metal and an interior made of a much lighter material. The balance points of the two models will not necessarily be the same. This sounds like another good research project:

How should we define the centroid of perimeter?

I hope this hasn't been too frustrating. These ideas are very deep and hard to explain. I just wanted to give you a taste of them now because I think they are so beautiful.



If we rotate a right triangle it forms a cone.

Assuming Pappus is right, where must
the centroid of the triangle be?

Can you find the centroid of a semicircle?

How about its centroid of perimeter?

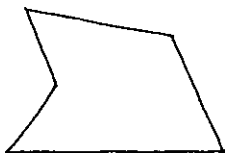
17

The shapes we've been dealing with so far—squares, circles, cylinders, and so on—are actually quite special. They are simple, symmetrical, and easy to describe. In other words, they're pretty. In fact, I would go so far as to say that they are pretty *because* they are easy to describe. The shapes that are the most pleasing to the eye are those that need the fewest words to specify. In geometry, as in the rest of mathematics, *simple is beautiful*.

But what about more complicated, irregular shapes? I think we need to look at them, too. After all, most shapes are not so simple and pretty. We'll surely miss the big picture if we restrict our attention to only the most elegant objects.

Take polygons, for instance. Up to now we've dealt almost exclusively with regular polygons (the ones with all their sides

the same length and all their angles equal). Certainly these are the prettiest. But there are lots of other polygons out there. Here is a not-so-regular one.



Of course, polygons like this are more complicated, and we're going to have to pay a price. The price is greater technicality—awkward shapes are going to be awkward to describe. Nevertheless, we need some way to indicate precisely which polygon we're talking about. We're not going to be able to make measurements or communicate ideas about a shape that is described only as “that thing that looks sort of like a hat.”

The most natural way to specify a particular polygon is to simply list all its angles and side lengths (in their proper order, of course). This information is like a blueprint; it pins down precisely which polygon we mean.

If you prefer, we can also think of a polygon as a sequence of distances and turns, as if we were traveling along its perimeter.



These outside turns will then add up to one complete turn. Of course, we have to be careful to count left and right turns oppositely. If we were traveling counterclockwise, for example, it would make sense to count left turns as positive and

right turns as negative. Then the grand total would be one full (counterclockwise) turn.

What do the inside angles of a polygon add up to?

Whatever way we choose to describe an irregular polygon, we still have to measure it. How, for instance, are we going to determine the area of such a polygon from a list of its angles and side lengths?

Even worse, it turns out that a shape described in this way might not even *be* a polygon. For example, it could intersect itself in the middle, or fail to close up at the end.



What should we do about these shapes? Should we call them polygons, too? What do we want the word *polygon* to mean? Of course, this is merely a question of terminology; the issue is not what is true but what is convenient.

Let's say we expand the meaning of the word *polygon* to include these new shapes. This at least has the advantage that *every* sequence of lengths and angles makes a polygon. Let's call a polygon **closed** if it closes up at the end, and **simple** if it never crosses over itself. Shapes that we used to call polygons would then be called *simple closed polygons*.

In any case, we have an interesting problem: How can we tell if a polygon, given by a sequence of lengths and angles, is simple or closed?

The point here is that angles and lengths are not independent of each other—there is a subtle connection between them. If we want a polygon to be closed, for instance, restrictions will then be placed on which lengths and angles we can use.

If all the angles of a simple closed four-sided
polygon are right angles, what condition
must the side lengths satisfy?

In general, the best strategy for dealing with polygons is to chop them into pieces. This is called a **dissection** of the polygon. In particular, we can always dissect a polygon into triangles.



This has the effect of reducing any problem about polygons down to a (possibly large) collection of triangle problems. For example, the area of a simple closed polygon would be the sum of the areas of the triangular pieces. To understand polygons, we need only understand the simplest ones: triangles. This is good! I'd much rather be thinking about triangles anyway; triangles are simpler, and simpler is better.

Make a short list of lengths and turns. What
triangle problems do you need to solve in order
to determine if your polygon is closed?

Are the midpoints of the sides of a triangle
enough information to reconstruct the triangle?

How about for four-sided polygons?

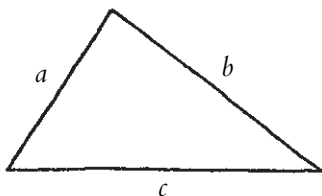
18

The study of triangles is called **trigonometry** (Greek for “triangle measurement”). The problem is to figure out how the various measurements of a triangle—angles, side lengths, and area—relate to each other. How, for example, does the area of a triangle depend on its sides? What is the relationship between the sides and the angles?

The first thing to notice about triangles is that they’re completely determined by their sides. If you tell me the three side lengths, I’ll know precisely which triangle you’re talking about. Unlike other polygons, triangles can’t wiggle.

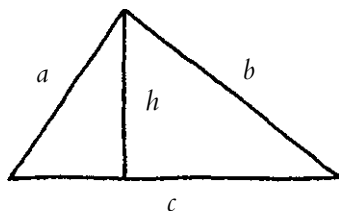
Do any three lengths form a triangle?

Suppose we have a triangle with certain side lengths a , b , and c (measured, of course, with respect to some suitably chosen unit).



What is the area of this triangle? Whatever it is, it must depend only on a , b , and c , since they determine the triangle, and hence its area, uniquely. The perimeter, for instance, is simply the sum of the three sides, $a + b + c$. Does the area have a similar algebraic description? If so, what is it? More important, how can we figure out what it is?

A natural way to begin is to drop a line from the top of the triangle down to its base.



Let's call this height h . Then the area A of the triangle can be expressed as

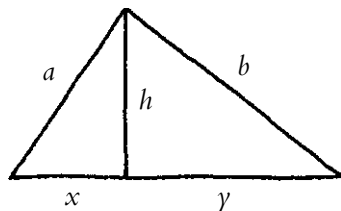
$$A = \frac{1}{2}ch.$$

The problem now becomes how to determine the height h in terms of the sides a , b , and c .

Before we get started, I want to say a few things about what we should expect. Our problem is to measure the area of a triangle given its sides. This question is completely symmetrical, in the sense that it treats the three sides equally; there are no "special" sides. In particular, there is no base involved in the question itself. What this means algebraically is that whatever our expression for the area turns out to be, it must treat the symbols a , b , and c symmetrically. If we were to switch all the a 's and b 's, for example, the formula should remain unchanged.

Another thing to notice is that because of the way that area is affected by scaling, our formula will have to be *homogeneous of degree 2*, meaning that if we replace the symbols a , b , and c by the scaled versions ra , rb , and rc , the effect must be to multiply the whole expression by r^2 . So we expect the area to be given by an algebraic combination of a , b , and c that is symmetrical and homogeneous. For example, it could look something like $A = a^2 + b^2 + c^2$. Unfortunately, it's not going to be quite that simple. Let's see what happens.

Notice how the height breaks the base c into two pieces. Let's call the pieces x and y . Our original triangle has been split into two right triangles.



Now we can use the Pythagorean relation to get information about x , y , and h . Hopefully, this will be enough information for us to actually figure out what they are. We have

$$\begin{aligned}x + y &= c, \\x^2 + h^2 &= a^2, \\y^2 + h^2 &= b^2.\end{aligned}$$

This looks a bit like alphabet soup. With so many letters and symbols flying around, it's important for us to keep the meaning and status of each one clear in our minds. Here a , b , and c refer to the sides of the original triangle. These are numbers that we

supposedly know from the start. The symbols x , y , and h , on the other hand, are unknowns. Their values are currently a mystery. We need to solve this mystery by somehow unscrambling the above equations to get x , y , and h expressed explicitly in terms of a , b , and c .

Generally speaking, this kind of problem can almost always be solved, provided there are enough equations. A good rule of thumb is that you need at least as many equations as you have unknowns (although that is no guarantee). In our case, since we have three of each, it should be possible to unscramble our equations. Of course, no rule of thumb can tell us *how* to unscramble them; that's where sheer algebraic skill comes in.

The first thing to do is to figure out what x and y are. See if you can rearrange our equations to get

$$x = \frac{c}{2} + \frac{a^2 - b^2}{2c},$$
$$y = \frac{c}{2} - \frac{a^2 - b^2}{2c}.$$

This tells us how the base of the triangle breaks up—the point where the height hits the base is precisely $(a^2 - b^2)/2c$ units away from the midpoint. This shift will be either to the left or right depending on which of a and b is larger.

The next step is to find the height h . Because of the way in which h appears in our equations, it's actually going to be a little easier to deal with h^2 instead. In fact, to make things prettier, let's rewrite x as $(c^2 + a^2 - b^2)/2c$ and use the equation $x^2 + h^2 = a^2$, so that

$$\begin{aligned}
 h^2 &= a^2 - x^2 \\
 &= a^2 - \left(\frac{c^2 + a^2 - b^2}{2c} \right)^2.
 \end{aligned}$$

Notice the asymmetry of this expression. This is partly due to the fact that we chose c as a base and h as the height to that base, so that c is being treated differently from a and b (we also used only the relation between x and h , and not the one involving y).

Now we can get at the area A . Again, it's a bit nicer to deal with A^2 instead. Since the area is given by $A = \frac{1}{2}ch$, we can write

$$\begin{aligned}
 A^2 &= \frac{1}{4}c^2h^2 \\
 &= \frac{1}{4}c^2a^2 - \frac{1}{4}c^2\left(\frac{c^2 + a^2 - b^2}{2c}\right)^2.
 \end{aligned}$$

This is not good. Although we've succeeded in measuring the area of the triangle, the algebraic form of this measurement is aesthetically unacceptable. First of all, it is not symmetrical; second, it's hideous. I simply refuse to believe that something as natural as the area of a triangle should depend on the sides in such an absurd way. It must be possible to rewrite this ridiculous expression in a more attractive form.

We can start by noticing that the whole thing can be written as the difference of two squares. Namely,

$$A^2 = \left(\frac{ac}{2} \right)^2 - \left(\frac{c^2 + a^2 - b^2}{4} \right)^2.$$

To simplify matters, let's multiply both sides of our equation by 16 to get rid of all the unpleasant denominators. We get

$$16A^2 = (2ac)^2 - (c^2 + a^2 - b^2)^2.$$

This is a definite improvement. Now, using the difference of squares relation, we can (cleverly) rewrite this as

$$\begin{aligned} 16A^2 &= (2ac + (c^2 + a^2 - b^2)) (2ac - (c^2 + a^2 - b^2)) \\ &= ((a^2 + 2ac + c^2) - b^2) (b^2 - (a^2 - 2ac + c^2)) \\ &= ((a + c)^2 - b^2) (b^2 - (a - c)^2). \end{aligned}$$

Again, we have differences of squares. This means we can break it down even further to get

$$16A^2 = (a + c + b)(a + c - b)(b + a - c)(b - a + c).$$

Now, that's more like it! The symmetry is finally revealed, and the pattern is actually quite beautiful.

Of course, we haven't really changed anything mathematically. These equations have all been saying the same exact thing about how the area depends on the sides—all of this clever algebraic manipulation hasn't changed that relationship. What has changed is its relationship to *us*. We're the ones who wanted to rearrange the information into a form that was more meaningful aesthetically. Triangles don't care. They do what they do regardless of how we choose to describe it. Algebra is really about psychology; it doesn't affect the truth, only how we relate to it. On the other hand, mathematics is not merely about truths; it's about beautiful truths. It's not enough to have

a formula for the area of a triangle; we want a pretty one. And now that's what we have.

Finally, to get the area A itself, we just need to divide our expression by 16 and take the square root. Notice that since there are four terms in the product, dividing by 16 is tantamount to cutting each one in half. Our formula for the area of a triangle becomes

$$A = \sqrt{\frac{a+c+b}{2} \cdot \frac{a+c-b}{2} \cdot \frac{b+a-c}{2} \cdot \frac{b-a+c}{2}}.$$

I have to admit that this looks rather complicated. Before we rush to judgment, however, let us remember that this formula is giving us the area of any triangle whatsoever, no matter what size or shape. This is no small feat. We should be grateful that there is any algebraic relationship at all, let alone one that involves the sides in such a simple way. This formula is actually quite elegant given what it has to do.

And in fact, we can make it even prettier by introducing a convenient abbreviation. Let $s = \frac{1}{2}(a + b + c)$. In other words, s stands for half the perimeter of the triangle (also known as the **semiperimeter**). The area can then be written very simply as

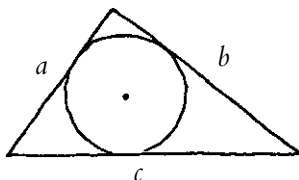
$$A = \sqrt{s(s-a)(s-b)(s-c)}.$$

This beautiful formula first appears in the writings of the Greek mathematician Heron of Alexandria (circa 60 AD), and for that reason is usually called **Heron's formula** (it's actually much older than Heron, and was probably known to Archimedes). Of course, none of the classical geometers would have approached the problem in quite the way we did; the style

at the time was much less algebraic. I wanted to do it this way because it is relatively straightforward and provides another nice illustration of the way algebra and geometry interact.

In any case, we now have a way to measure the area of any triangle. We simply take the three side lengths and stick them into Heron's formula, and the area pops right out. For example, the area of a triangle with sides 3, 5, and 6 would be $\sqrt{56}$.

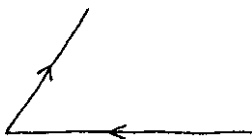
Can you find two different triangles
with the same area and perimeter?



If a triangle has sides a , b , and c , what
is the radius of the inscribed circle?

19

The most fundamental problems in geometry concern the relationship between length and angle. For example, suppose we travel a certain distance, turn a certain amount, and then go another distance. How far are we from where we started?



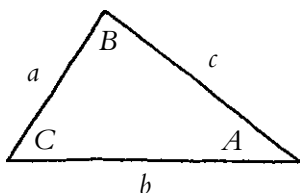
Another way to think of this question is to imagine two sticks that are held together at one end.



If we move the sticks apart, increasing the angle, the ends of the sticks get farther away; pushing the sticks together brings the ends closer. What exactly is the relationship between the angle of the sticks and the distance between their endpoints? This is perhaps the most basic question in all of geometry.

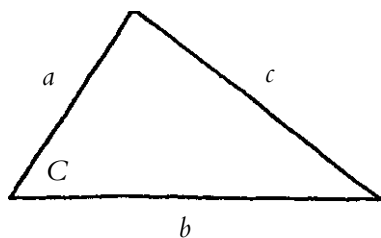
We can, of course, view this as being a problem about triangles. Essentially, we're asking how the side of a triangle depends on the opposite angle.

Perhaps it's time to introduce a convenient labeling scheme for triangles. The idea is to use small letters a , b , and c for the sides, and capital letters A , B , and C for the corresponding opposite angles.



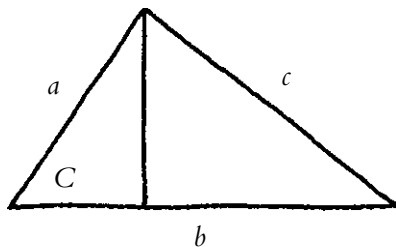
The point of this is to make it easy to remember which angle is opposite which side. (Of course, our ideas do not depend on labels, but the ease of communicating them often does.)

So the question is, given two sides a and b of a triangle, what is the relationship between the side c and the angle C ?



When C is a right angle, we know from Pythagoras's theorem that $c^2 = a^2 + b^2$. But what if C is not a right angle? What happens then?

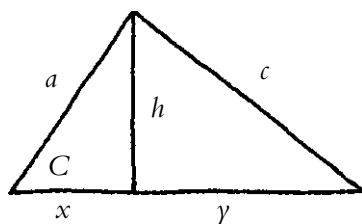
Let's first suppose that C is less than a right angle. The usual way to get at something like the length of c is to drop a perpendicular line so that c becomes the long side (or **hypotenuse**) of a right triangle.



As a matter of fact, the only method we've *ever* had for measuring lengths is to somehow get them involved in right triangles. This is why the Pythagorean relation is so important.

There is actually another technique for measuring lengths, which we used for the diagonal of a regular pentagon. What is it?

As before, let's call this height h , and the base pieces x and y .



Pythagoras then tells us that

$$c^2 = y^2 + h^2.$$

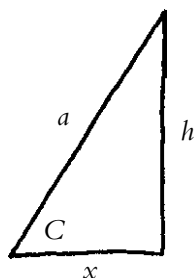
Of course, what we're after is a relation that tells us how c depends on a , b , and the angle C . Since $x^2 + h^2 = a^2$ and $x + y = b$, we can replace h^2 by $a^2 - x^2$ and y by $b - x$ to get

$$\begin{aligned} c^2 &= (b - x)^2 + a^2 - x^2 \\ &= a^2 + b^2 - 2bx. \end{aligned}$$

Notice the similarity between this equation and the Pythagorean relation—the $2bx$ piece must be some sort of correction term that measures the departure of C from being a right angle. We should consider this formula a generalized Pythagorean theorem that is valid for any angle, not just right angles.

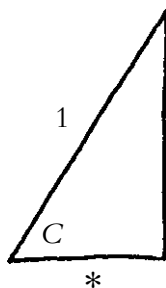
Of course, the present form of this expression is rather unsatisfactory, the two most glaring reasons being that it is not symmetrical in a and b (as it should be) and that the angle C itself does not make an appearance. Essentially, the problem comes down to determining this length x .

Let's take a closer look at the right triangle involving x , a , and the angle C .



Notice that this triangle is completely determined by the angle C and the hypotenuse a . In fact, C alone is enough to pin down the shape of this triangle. This is because the angles of a triangle always add up to a half turn; if we know one of the angles of a right triangle, we automatically know the other.

In particular, this means that our triangle is just a scaled version (by a factor of a) of the right triangle with angle C and hypotenuse 1.



So to find x , we just need to multiply the length of the side marked $*$ by the scaling factor a . Thus $x = a*$ and our formula for the third side of a triangle becomes

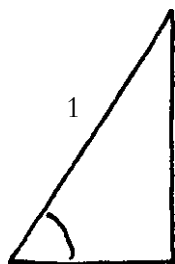
$$c^2 = a^2 + b^2 - 2ab*.$$

The point is that the length $*$ depends only on the angle C and not on the sides a and b . Our equation is now symmetrical

and reveals completely the dependence of c on the other two sides. The only thing remaining is to figure out exactly how $*$ depends on C . Notice that this question involves only this right triangle and not the original triangle we started with.

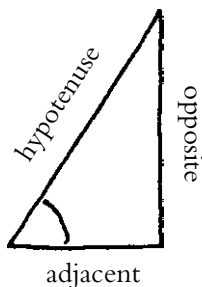
Something interesting has happened here. Our problem about triangles in general has been reduced to a problem about right triangles in particular. This is part of a general pattern: polygons are reduced to triangles; triangles are reduced to right triangles. A complete understanding of right triangles would tell us everything about polygons.

Our basic problem is this. We have a right triangle with a certain angle, and a hypotenuse of length 1. How long are its sides?



The sides of a right triangle are sometimes called its legs. In our case the two legs depend only on the angle. The vertical one, opposite the angle, is usually called the **sine** of the angle (it's where your sinuses would be if the triangle were your nose). The leg adjacent to the angle is called the **cosine** of the angle. I suppose what I actually mean is that the sine and cosine are the *lengths* of the legs, not the legs themselves. (Of course, we've been glossing over that sort of distinction this whole time, so why start worrying about it now!)

We can also think of the sine and cosine as being *proportions*.



The sine of an angle will be the ratio of the opposite side to the hypotenuse; the cosine is the proportion of adjacent side to hypotenuse. This is true regardless of whether the hypotenuse has unit length or not; the angle determines the shape of the right triangle, and these ratios are independent of scaling.

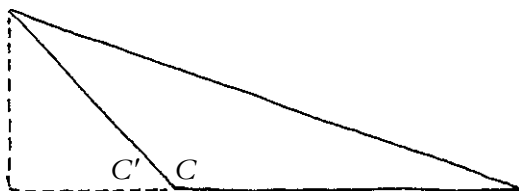
**How are the sines and cosines of the two angles
of a right triangle related to each other?**

In any case, the upshot is that to each angle there corresponds a pair of numbers, its sine and cosine, which depend only on that angle. If C is an angle, it is customary to write $\sin C$ and $\cos C$ for its sine and cosine. With this terminology, our formula now reads:

$$c^2 = a^2 + b^2 - 2ab \cos C.$$

This is our generalized Pythagorean theorem relating the third side of a triangle to the other two sides and the angle between them. Of course, all this really does is to transfer the problem to the right triangle situation. We still have to figure

out the cosine of the angle. Also, the assumption here was that the angle C was less than a right angle. What happens if it isn't?



Of course, we can still drop the same perpendicular, only this time it lies *outside* our triangle and forms a new angle C' , which sits next to our original angle C .

Show that in this case we get

$$c^2 = a^2 + b^2 + 2ab \cos C'.$$

So the Pythagorean relation for large angles is pretty much the same as before, only instead of subtracting the correction term $2ab \cos C$, we are adding $2ab \cos C'$.

We seem to have three separate cases (with three separate formulas), depending on whether the angle C is less than, equal to, or greater than a right angle. This kind of thing is always a bit galling; after all, two sticks can smoothly open and close on their angle hinge, and the distance between their endpoints will vary continuously. Shouldn't there be one nice, simple pattern?

One way to proceed is simply to be clever with our definitions. Since $\cos C$ (at present) only has meaning when C is less than a right angle, we are free to give it any meaning we wish when C is larger. The idea is to do this in such a way

that our Pythagorean relation $c^2 = a^2 + b^2 - 2ab \cos C$ remains valid in all three cases. That is, *we let the pattern determine our choice of meaning*. This is a major theme throughout mathematics; it could even be said that this is the essence of the art—listening to patterns and adjusting our definitions and intuitions accordingly.

This leads us first to define the cosine of a right angle to be zero (so that we recover the usual Pythagorean theorem) and then, more strangely, to define the cosine of C when C is larger than a right angle to be the *negative* of the cosine of C' , the angle next to C .

What we have done here is to *expand the meaning of cosine*. Originally, we defined the cosine of an angle in terms of side lengths of a right triangle. Now we are choosing to give $\cos C$ meaning even when C is too big to fit in a right triangle. We are doing this so that we get one universal pattern instead of three separate ones. But more important, we are letting math do the talking. We are being sensitive to what angles and lengths want. They want cosine to generalize, and they are telling us what they need that generalization to be. Now it is up to us to reconcile that with our intuition.

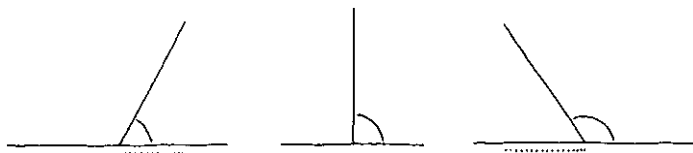
One way to do this is to imagine a stick (of unit length, say) at an angle with the ground.



Depending on this angle, the shadow of the stick will be longer or shorter (I'm assuming the metaphorical sun is directly

overhead). In fact, we can see that the length of this shadow is precisely what we have been calling the cosine of the angle.

Now, as the angle increases, the shadow gets shorter, until the stick is straight up (at a right angle with the ground) and the shadow has length zero. If we keep going, the shadow reappears, only on the *other side*. Its length is now the cosine of the angle next to ours.



So a nice way to think about our expanded definition of cosine is that we are redefining the cosine of an angle to be the shadow of a unit length stick, keeping track not only of the length of the shadow but also its *direction*. That is, shadows on the same side as the angle are counted positively, and shadows on the other side are measured as negative. With this choice of meaning for cosine, we get a single Pythagorean relation

$$c^2 = a^2 + b^2 - 2ab \cos C$$

valid for all angles.

One thing that this formula tells us is that angles and lengths don't directly relate to each other; the angle information must be delivered indirectly, via the cosine. It's as if angles need an attorney, in the form of their cosine, to represent them in their dealings with lengths. Angles and lengths live in different worlds and speak different languages. Sine and cosine serve as a dictionary, converting angle measurements into length measurements.

Show that if a triangle has sides a and b meeting at an angle C , then its area is $\frac{1}{2}ab \sin C$.

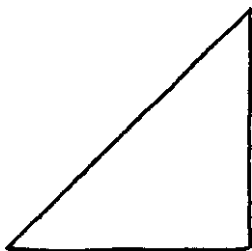
What is the angle between the faces of a regular tetrahedron? How about for the other regular polyhedra?

Show that you can fill space completely using regular octahedrons and tetrahedrons. Can you find any other ways to tile three-dimensional space with symmetrical polyhedra?

20

Given an angle (measured, say, as a portion of a full turn), how can we figure out its sine and cosine? Conversely, if we are told what its sine and cosine are, how can we determine the angle itself?

Some angles have sines and cosines that are fairly easy to measure. For instance, an angle of $\frac{1}{8}$ (or 45 degrees) makes a right triangle that is half a square.



This means its sine and cosine are both equal to the ratio of the side of a square to its diagonal, or $\frac{1}{\sqrt{2}}$.

**What are the sine and cosine
of one-sixth of a turn?**

By the way, it turns out that the sine and cosine of an angle are a bit redundant; if you know one of them, you can deduce the other. The connection between them comes from the Pythagorean relation. Can you figure out what that connection is?

**What is the relationship between
the sine and cosine of an angle?**

The natural thing to do at this point would be to start compiling a table of sines and cosines for various angles. This is exactly what astronomers and navigators were doing six hundred years ago; ships were sailing long distances, and reasonably accurate navigational measurements had to be made.

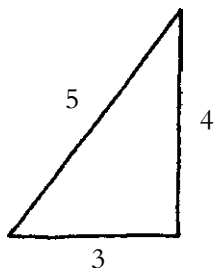
Of course, in that situation, all you need are approximations. The sine of 45 degrees is about .7071, and that's good enough for any practical purpose. For impractical purposes such as geometry, however, this simply will not do. If we want to measure perfect imaginary shapes—and we do—we'll need to figure out the exact values of sine and cosine.

Unfortunately, this is quite a difficult thing to do, even when the angle is a nice fraction of a full turn. For instance, the sine and cosine of $\frac{3}{13}$ are extremely unpleasant numbers. They are certainly irrational, and although they can be expressed in

terms of roots of various kinds, it is not a pretty picture. It is much better to just call them the sine and cosine of $\frac{3}{13}$ and leave it at that.

Even worse, if the angle itself is bad, say an irrational amount of a full turn, then the sine and cosine are generically transcendental numbers. This means we have no algebraic way whatsoever to refer to them. As with the number π , we simply have no choice but to accept numbers like the sine of $\frac{1}{\sqrt{17}}$ as having no simpler description. Once again we must enlarge our language and learn to make the best of it.

A similar thing happens when we try to determine an angle from knowledge of its sine and cosine. For example, the angle that occurs in the beautiful 3, 4, 5 right triangle has a sine of $\frac{4}{5}$ and a cosine of $\frac{3}{5}$.



But what is the angle? What portion of a full turn is it? This number turns out to be transcendental as well. What this means is that “the angle whose sine is $\frac{4}{5}$ ” is as good a description as we’re ever going to get. There’s simply no way to take the numbers 3, 4, and 5 and do some finite sequence of algebraic operations with them to arrive at the measurement of this angle.

All in all it’s a very depressing (and somewhat embarrassing) situation. We’ve managed to reduce *every* problem concern-

ing the measurement of polygons down to this one essential question of how the sine and cosine of an angle depend on the angle itself, and what I'm telling you is that this problem is (in general) intrinsically unsolvable. This is not to say that there aren't certain pretty angles—like $\frac{1}{8}$ or $\frac{1}{6}$ —whose sine and cosine are nice numbers that can be expressed algebraically, but they are a small minority indeed.

What I think is interesting about a situation like this is that we are able to ask perfectly natural geometric questions that we can't answer. Moreover, we can *prove* that they are unanswerable. In other words, we can know that something is unknowable. Maybe this is not so depressing after all—it's a pretty amazing human accomplishment!

Of course, I've done nothing to help explain how it is that we do know such things. It's all very well for me to say that such and such a number is transcendental; it's quite another for me to show you why.

I'm in a truly unfortunate predicament here. It's important to me that you understand the positive nature of statements like " π is transcendental" or " $\sqrt{2}$ is irrational." When a mathematician like me says that something is impossible, be it that π cannot be represented algebraically, or that there is no fraction whose square is 2, I'm not saying something negative about what we can't do or don't have. I'm talking about what we do have: an explanation! We know that $\sqrt{2}$ is irrational, and we understand why. We have a perfectly reasonable explanation—namely, Pythagoras's argument about even and odd numbers.

Over the centuries, mathematics, like any art form, has achieved a certain depth. Many works of art are extremely sophisticated and require years of study to properly understand

and appreciate. This is the case with the transcendence of π , unfortunately. Proofs exist, even very beautiful ones, but that doesn't mean that I can easily explain them to you here. For now, I think you're just going to have to take my word for it.

**Can you use a regular pentagon to find the
sine and cosine of one-fifth of a turn?**

21

What do we want out of trigonometry? In the best of all possible worlds, we would like to be able to determine *all* the measurements of any given triangle. Let's say that a triangle has been completely measured once we know its angles, side lengths, and area. Of course, we would have to know some of these measurements to begin with in order to specify which triangle we're even talking about.

How much information do we need? Which combinations of angle and side information are sufficient to pin down a triangle precisely? There are several possibilities:

Three sides. In this case the triangle is certainly determined uniquely. The generalized Pythagorean theorem can then be used to find the angles (or at any rate their cosines, which is morally the same, and all we can reasonably hope for). Heron's formula gives us the area directly from the three sides, so in this case we can always measure the triangle completely.

Two sides. This is generally not enough information to spec-

ify a particular triangle, unless we have some additional angle information. If we know the angle between the two sides, or at least its cosine, then the generalized Pythagorean theorem will give us the other side, and we're done. Otherwise, if all we have is one of the other angles, that won't be enough to determine the triangle. Do you see why?

**Why are two sides and an angle insufficient
in general to specify a triangle?**

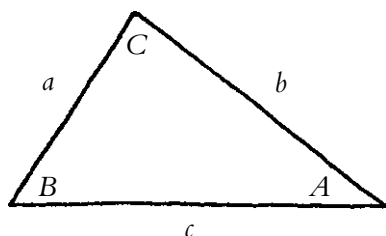
Of course, if we were given two of the angles it would be a different story. Since the angles of a triangle always add up to a half turn, knowledge of two of them amounts to knowing all three. In particular, if we had two sides and two angles, we could figure out the angle between the sides, and we would be in business.

One side. This is pretty scant information; we'll definitely need to know all the angles. One side and one angle is not going to cut the mustard. On the other hand, knowing all three angles would tell us the shape of the triangle, which determines the triangle up to scaling. Any one of the sides would then lock down the triangle completely (we would also need to specify which side opposite which angle). The problem would then be to figure out the lengths of the other two sides. Given the angles of a triangle, and one of its sides, how can we determine the other two?

A more elegant (and symmetrical) way to deal with this question is to think of it in terms of proportions. We really only need to know the ratios of the sides to each other; if we then

had any one of the sides, we could easily figure out the others. The nice thing about proportions is that they are independent of scaling; they depend only on the angles of the triangle. So let's rephrase our question. Given the angles of a triangle, how can we determine the relative proportions of the sides?

In terms of our labeling conventions, we're asking how the proportion $a : b : c$ depends on the angles A , B , and C .



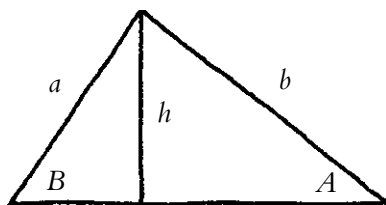
It's easy to see that longer sides are opposite larger angles; the question is whether we can say anything more precise than that.

Since we're dealing with angles and lengths, we naturally expect sines and cosines to make an appearance, and in fact they do. The relationship between the sides of a triangle and their opposite angles is one of the most beautiful patterns in geometry: the sides are in the same proportion as the *sines* of the angles. In other words,

$$a : b : c = \sin A : \sin B : \sin C.$$

This result is usually referred to as the **law of sines** (our generalized Pythagorean theorem is often called the law of cosines, but I think that's a silly name for it).

To see why this is true, let's drop the usual perpendicular.



Notice that this height h is opposite both angles A and B . This means that

$$\sin A = \frac{h}{b},$$

$$\sin B = \frac{h}{a}.$$

Dividing these equations, we get

$$\frac{\sin A}{\sin B} = \frac{h/b}{h/a} = \frac{a}{b}.$$

Thus $a : b = \sin A : \sin B$, and the sides are in the same proportion as the sines of the opposite angles. As with the generalized Pythagorean theorem, we're seeing how angles communicate length information via their sines and cosines. I like the present version of this sentiment because of its symmetry.

One thing I just realized about this argument is that it presupposes that the angles are all acute (that is, less than right angles). What happens if we have a triangle with a larger, obtuse angle? Do such triangles still obey the law of sines? For that matter, what do we even want the sine of such an angle to mean?

How should we define the sine of an obtuse angle? Can we do it so the law of sines still holds?

Using the law of sines, the generalized Pythagorean theorem, and Heron's formula, we can completely measure any triangle—at least in the sense that we can reduce the measurement of any triangle (and hence any polygon) to the determination of a bunch of sines and cosines. The buck usually stops here, unless there's some sort of amazing symmetry or coincidence, because of the transcendental nature of sine and cosine. The goal of trigonometry then becomes not to calculate these numbers but to find patterns and relationships among them.

How are the sine and cosine of an angle related to the sine and cosine of an angle twice as large?

I should point out that everything we've been saying about polygons works the same way in three dimensions for polyhedra. In particular, polyhedra can always be dissected into various pyramids, and these can be measured using triangles. In this way, all problems concerning polyhedra come down to sines and cosines as well.

Prove that if two angle bisectors of a triangle are equal, then the triangle must be isosceles.

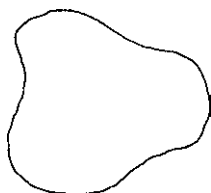
Show that if a four-sided shape with sides a , b , c , and d is inscribed in a circle, then its area is given by Brahmagupta's formula:

$$A = \sqrt{(s-a)(s-b)(s-c)(s-d)},$$

$$\text{where } s = \frac{1}{2}(a+b+c+d).$$

22

What shapes are left for us to measure? The answer is, most of them! In fact, we haven't even begun to deal with the vast majority of shapes out there. Everything we've looked at so far has had some sort of special property, like straight sides or symmetry, that sets it apart and makes it atypical. Most shapes have no such distinguishing features. Most shapes are asymmetrical and ugly, curved, and not in any particularly pleasing way.



But why would we want to work with something like that? Why should we (meaning you) expend time and energy trying to understand some ugly blob? And even if we did want to, how would we do it? How do we even describe, let alone measure, an irregularly curved shape like this one? For that matter, what do I even mean by “this one”—*what* one? Exactly which shape am I talking about here?

If I were doing something practical, I could simply say “the shape in the diagram” and be done with it. The picture itself would be the shape, and rough measurements could be made right from it.

Mathematically, however, the picture is no help at all. A diagram, being a part of the physical world we live in, is much too crude and imprecise to refer to a specific mathematical object. And it's not merely a question of accuracy. A circle etched

in gold by a laser to within a billionth of an inch is just as irrelevant (if not more so) than one made by a kindergartener out of construction paper. Neither one is anything like a true circle.

The important thing to understand is that diagrams and other such models are made of atoms, not idealized imaginary points. In particular, this means that a diagram cannot accurately describe anything. Not that diagrams are completely useless; we just need to understand that their role is not to specify or define but to stimulate creativity and imagination. A construction paper circle may not be a circle, but it still might give me ideas.

So how, then, are we going to describe a particular irregularly curved shape? Such a shape would contain infinitely many points, and unlike a polygon, no finite collection of them would be enough to pin the shape down—we would need an infinite list of points. But how can I think about a shape, or tell you about it, if I need to provide an infinite amount of information? The question is not what shapes do we *want* to talk about, but what shapes *can* we talk about.

The disturbing truth is that most shapes cannot be talked about. They're out there all right; we just have no way to refer to them. Being human, using finite languages over finite lifetimes, the only mathematical objects we can ever deal with are those that have finite descriptions. A random spatter of infinitely many points can never be described, and neither can a random curve.

What I'm saying is, the only shapes that we are ever going to be able to specify precisely are those that have enough of a pattern to them to allow an infinity of points to be described in a finite way. The reason we can talk about a circle is not because of the kindergarten cutout, but because of the phrase "all the points at a certain distance from a fixed center." Since the circle has such a simple pattern, I don't need to tell you

where each of its individual points are; I can just tell you the pattern they obey.

My point is that's all we can ever do. The only shapes we can talk about are those with a pattern, and it is the pattern itself—a finite set of words in a finite language—that defines the shape. Those shapes that do not have such a pattern (the vast majority, I'm afraid) can never be referred to, let alone measured, by any human beings, ever. The set of objects that we can think about and describe to others is limited from the start by our own humanity. This is actually something of a theme throughout mathematics. For instance, the only numbers we can talk about are those with a pattern; most numbers can never be referred to either.

Geometry, then, is not so much about shapes themselves as it is about the verbal patterns that define them. The central problem of geometry is to take these patterns and produce measurements—numbers which themselves must necessarily be given by verbal patterns. We have already talked about polygons, which can be specified easily by a finite list of sides and angles, and circles, which have their own very simple pattern. What are some other patterns we can think of? What sorts of descriptions are possible? What curves besides circles can we talk about?

23

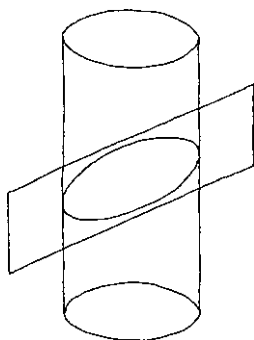
There is one curve besides the circle that we have already come across, and in fact it's one of the oldest and most beautiful objects in all of geometry: the ellipse.



An ellipse is a dilated circle—a circle that has been stretched by a certain factor in one direction. As such, it is a very precise and specific shape. I suppose I should say it is a specific class of shapes, since there are different ellipses depending on the stretch factor. If you like, you can even think of a circle itself as being a special type of ellipse—with a stretch factor of 1!

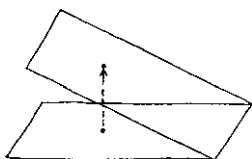
The point being that an ellipse is not just any old oval shape; it's a particular curve with a particular pattern, namely that of being a dilated circle. Actually, it turns out that there are several different ways to describe an ellipse, and the interplay among these various descriptions makes for some very fascinating and beautiful mathematics.

For example, one of the nicest ways to think of an ellipse is as a circle viewed at an angle. Another way to say this is that an ellipse is what you get when you slice a cylinder with a slanted plane.

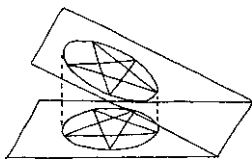


It's certainly clear that if you slice a cylinder in this way you get some sort of curved shape, but how can we be sure that it is an ellipse, as opposed to some other oval-shaped curve? What exactly is the connection between slanted cross-sections and dilations?

I think the simplest way to understand the situation is to imagine two planes in space that meet at an angle.

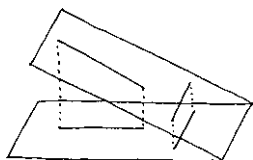


For convenience, let's think of the first plane as being horizontal. Any point on this plane could then be lifted straight up to a corresponding point on the slanted plane. In this way, any shape on the first plane can be transformed into a new shape on the second plane.



This kind of transformation is called a **projection**. So what we're saying is that an ellipse is a projection of a circle. Of course, this is just new language; we still have to figure out why it's true. Why, when a circle is projected, does it get dilated?

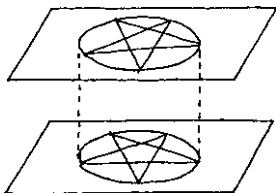
The reason is that projection *is* dilation. They are exactly the same process. Or rather, they are two different processes that have the exact same effect. To see this, consider the line where the two planes intersect.



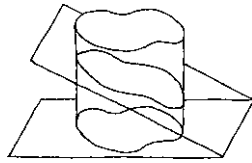
Imagine two sticks on the first plane, one parallel to the line and one perpendicular to it. After projection, the first stick is still parallel to the line, and its length is the same as before. The perpendicular stick remains perpendicular, but it gets longer—projection expands distances in one direction and not the other. In other words, projection produces a dilation in the direction perpendicular to the line of intersection of the two planes. Notice that as the angle between the planes increases, so does the stretch factor.

**How exactly does the dilation factor
depend on the angle between the planes?**

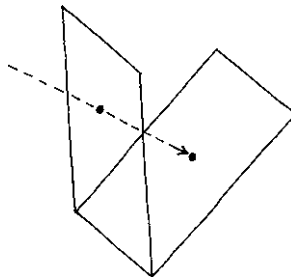
Notice also that if the two planes happen to be parallel, then projection doesn't do anything at all—it's dilation by a factor of 1!



At any rate, we now have a radically different way to think about dilation. Rather than a stretching of a single plane, we can view it as a projection through space of one plane onto another. In particular, the dilated form of *any* object (not just a circle) occurs as a suitably slanted cross-section of the generalized cylinder with that object as its base.



We can also imagine projections where neither plane is necessarily horizontal, and the direction in which the points are being projected is not necessarily vertical. In other words, we can choose any two planes in space, and any direction, and obtain a projection that transforms the shapes on one plane into shapes on the other.



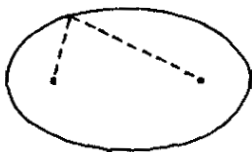
The question is, do these more general projections yield anything new, or do they still just dilate?

**Do projections in any direction
always produce dilations?**

24

An entirely different approach to ellipses is through their so-called focal properties. It turns out that inside every ellipse are two special points, called **focal points**, which have the

amazing feature that every point on the ellipse has the same *combined* distance to them.



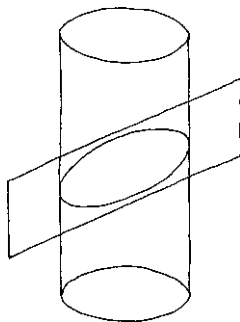
In other words, as a point travels along the perimeter of the ellipse, the distances to the two focal points will change, but the sum will remain constant. This makes it possible for us to describe an ellipse in a new way, as “the set of points whose distances to two fixed points have a fixed sum,” or some such phrase. Some people even choose to take this as their definition of an ellipse.

Of course, it doesn't really matter whether you think of an ellipse as a dilated circle that happens to have an interesting focal property, or if you think of the focal property as the defining characteristic of ellipses, which then happen to be dilated circles. Either way, we have some work to do. I mean, a dilated circle is one thing, a curve with focal points is another. Why should they be the same? More to the point, how can we *prove* they are the same?

This is the sort of thing I love about mathematics. Not only are there amazing discoveries to be made, but you have the additional challenge of understanding why such a thing should be true and of crafting a beautiful and logical explanation. You get all the pleasure of art and science all in one package, plus it's all in your head!

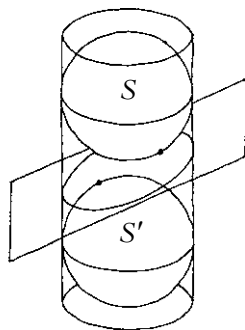
I want to show you an ingenious argument (discovered by Dandelin in 1822) that explains why dilated circles have this

focal property. Let's start by viewing our ellipse as a cross-section of a cylinder by a slanted plane.



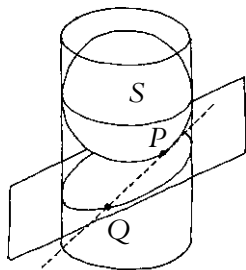
If we're going to be able to show that this curve satisfies the focal property, where on earth are the focal points going to be? The answer is shockingly beautiful.

Take a sphere S (of the same diameter as the cylinder) and drop it into the cylinder from above, so that it falls and hits the slicing plane at a point P . Do the same thing with another sphere S' from below, pushing up until it hits the plane at a point P' .



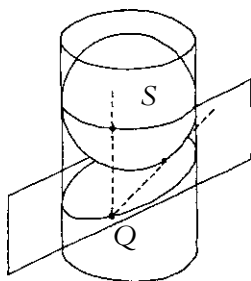
These two points P and P' (where the spheres hit) turn out to be the focal points. Is that gorgeous, or what!

Of course, to confirm this, we need to show that no matter what point we choose on the ellipse, the total distance to these two points will always be the same. Let's suppose Q is an arbitrary point somewhere on the ellipse. Now imagine the line through Q and P .

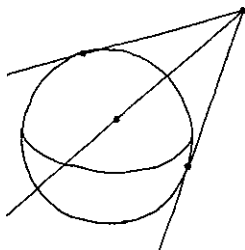


This line has a very interesting feature: it touches the sphere S exactly once. This is quite unusual. Most lines either miss a sphere entirely or pass through it, hitting it twice. A line that touches a sphere only once is called a **tangent** (Latin for “touching”). The line through Q and P is a tangent to the sphere S because it is contained in the plane, which hits the sphere only at P .

There is another way to make a tangent to S , and that is to take a vertical line through Q , intersecting the sphere S on its equator.



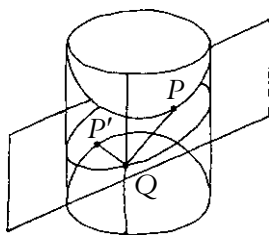
In general, there are many tangents to a sphere from a given point. The interesting thing is that they all have the same length.



That is, the distance from a point outside the sphere to a point on the sphere where the tangent hits is the same no matter which tangent you use.

**Why do the tangents from a given point
to a sphere all have the same length?**

In particular, the distance from our point Q to the alleged focal point P is the same as the vertical distance from Q to the equator of S . To make it simpler, let me chop off our original cylinder at the equators of the two spheres.



Then what we're saying is that the distance from Q to P is the same as the distance from Q to the top of this cylinder, and by similar reasoning, the distance from Q to P' must be the same as the distance from Q to the bottom of the cylinder.

This means that the total distance from Q to the two points P and P' must simply be the height of the cylinder. Since this height is independent of the position of the point Q , our ellipse really does satisfy the focal property, and this beautiful proof shows us why. What an inspired work of art!

How do people come up with such ingenious arguments? It's the same way people come up with *Madame Bovary* or *Mona Lisa*. I have no idea how it happens. I only know that when it happens to me, I feel very fortunate.

A circle is a special type of ellipse.

Where are its focal points?

25

Now I want to tell you about another remarkable property of ellipses, which is interesting not only mathematically, but also from a “real world” point of view. Probably the simplest way to describe it is to think of an ellipse as a sort of pool table with a cushion running around its perimeter. Imagine a hole at one of the focal points and a ball placed at the other. Then it turns out that no matter which direction you shoot the ball, it will always bounce off the cushion straight into the hole!



In other words, ellipses are bent in just the right way so that lines from one focal point get reflected into lines to the other. Geometrically, this is saying that the two lines meet the ellipse in *equal angles*.



What makes this a little confusing is that we're dealing with a curve; what does angle mean exactly?

The most elegant way out of this dilemma is to use a tangent: a line that touches the ellipse at exactly one point. Each point on the ellipse has a unique tangent line through it that indicates the direction in which the curve is bending there.



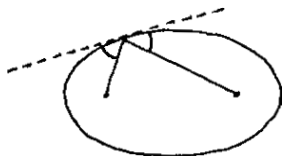
This gives us a way to talk about angles made by curves. The angle between two curves is just the angle made by their tangents.



The use of tangent lines to help understand curves is an ancient and traditional technique. Tangents convey a lot of

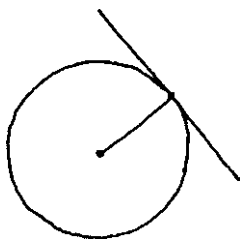
information about how a curve behaves, and being straight lines, they are much easier to deal with than the curve itself.

Now the “pool table property” can be stated precisely. It says that for any point on an ellipse, the lines to the focal points make equal angles with the tangent.



I suppose we should really call this the *tangent property* of an ellipse instead (it’s a bit more dignified). Whatever we call it, it’s a truly beautiful and surprising fact about ellipses, and it’s crying out for explanation.

By the way, a nice special case of this property occurs with circles: a ball shot from the center bounces straight back to the center. The tangent property says that for a circle the tangent must be perpendicular to the radius.



Why is the tangent to a circle
perpendicular to the radius?

As I’ve said before, the task of the mathematician is not only to discover fascinating truths but also to explain them. It’s one thing to draw some ellipses and lines and say that such and such

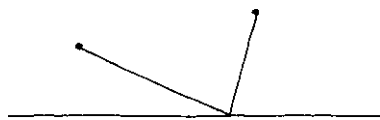
is happening—it's quite another to prove it. So I want to show you a proof of the tangent property. The explanation I have in mind is not only simple and pretty but is general enough to apply to many other situations besides ellipses.

In fact, let's start by looking at a different (but related) problem. Suppose we have two points situated on the same side of an infinite line (it's nicer to deal with an infinite line because its length and position don't become an issue).

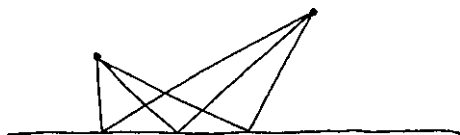


The question is, what is the shortest path from one point to the other that touches the line? (Naturally, the part about touching the line is the interesting part. If we dropped that requirement then the answer would simply be the straight line connecting the two points.)

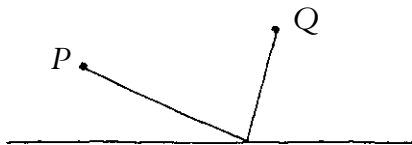
Clearly the shortest path must look something like this:



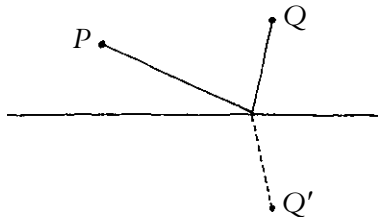
Since our path has to hit the line somewhere, we can't do better than to go straight there. The question is, where is *there*? Among all the possible points on the line, which one gives us the shortest path? Or does it even matter? Maybe they all have the same length!



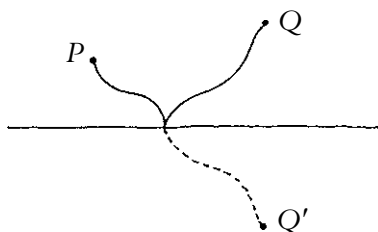
The truth is, it does matter. There is only one shortest path, and I'll tell you how to find it. Let's first give the points names, say P and Q . Suppose we have a path from P to Q that touches the line.



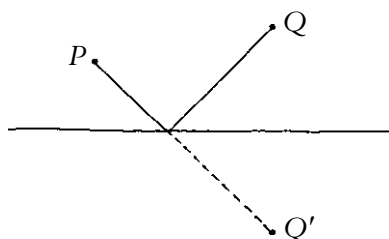
There's a very simple way to tell if such a path is as short as possible. The idea, which is one of the most startling and unexpected in all of geometry, is to look at the *reflection* of the path across the line. To be specific, let's take one part of the path, say from where it hits the line to where it hits the point Q , and reflect that part over the line.



We now have a new path that starts at P , crosses the line, and ends up at the point Q' , the reflection of the original point Q . In this way, *any* path from P to Q that touches the line can be transformed into a new path from P to Q' .



Now, here's the thing: the new path has exactly the same length as the original. This means that the problem of finding the shortest path from P to Q that hits the line is really the same as finding the shortest path from P to Q' . But that's easy—it's just a straight line. In other words, the path we're looking for, the shortest path between the points that touches the line, is simply the path that *when reflected* becomes straight.



Apart from its sheer loveliness, this argument is also an excellent example of the modern mathematical viewpoint that considers problems as occurring within a framework of structures and structure-preserving transformations. In this case, the relevant structures are paths and their lengths, and the key to the problem is recognizing reflection as the appropriate structure-preserving transformation. This is admittedly a rather professional point of view, but I think it's a valuable way for anyone to think about math problems.

Now that we know precisely what the shortest path looks like, we can think about alternative descriptions of it. One of the simplest is that it's the path that makes equal angles with the line. The shortest path is the one that “bounces off.”

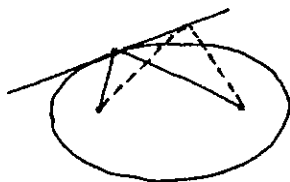


Why does the shortest path make equal angles with the line?

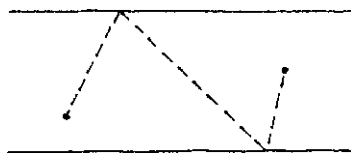
Of course, the reason I bring all this up is that it helps explain the tangent property of ellipses. In that situation, we have a path running from one focal point to the other, by way of a point on the perimeter. We want an explanation as to why the angles made with the ellipse (that is to say, with the tangent) must be equal.



Well, the reason is that this path happens to be the shortest path between the focal points that touches the tangent line. This is easy to see from the focal property of the ellipse: all the points on the ellipse have the same total distance to the focal points. Naturally, points inside the ellipse will have a smaller total distance, and points outside will have a larger total distance. In particular, since any point of the tangent line (other than the actual point of contact with the ellipse) is strictly outside the ellipse, the path through such a point must be longer than the one through the contact point itself.



Since our path is shortest, it must then make equal angles with the tangent. The tangent property, or “pool table” effect, comes directly from the focal property of ellipses and the fact that shortest paths always bounce.



Suppose two points lie between parallel lines.

What is the shortest path from one to the other that hits both lines?

26

I want to say a few more things about the relationship between geometry and reality. Of course, in some sense, they are entirely different, one being a completely imaginary construction of the human mind and the other (presumably) not. Physical reality was here before there were conscious human beings, and it will still be here when we're gone. Mathematical reality, on the other hand, depends on consciousness for its very existence. An ellipse is an idea. There are no actual ellipses out there in the real world. Anything real is necessarily a wriggling, jiggling mass of trillions of atoms and is therefore far too complicated to ever be described by human beings in any precise way.

There are two important differences between physical atoms (which real things are made of) and the mathematical points

that make up our imaginary geometric objects. First, atoms are constantly in motion, flying on and off and smashing into each other. Points do what we tell them to do; the center of a circle doesn't wiggle around. Secondly, atoms are discrete—they stay away from each other. Two atoms can be brought only so close together; the forces of nature (apparently) do not allow them to get closer. Of course, we place no such restriction on our imaginary points. Mathematical objects are governed by aesthetic choices, not physical laws. In particular, a line or curve of points is impossible to realize physically. Any "curve" made of real particles is necessarily going to be lumpy, with all kinds of gaps in it—more like a string of pearls than a strand of hair (and that includes, of course, an actual strand of hair).

On the other hand, it's not true that there is absolutely *no* connection between geometry and reality. There may not be any perfect cubes or spheres in the world, but there are some pretty good approximate ones. Any property that the mathematical cube and sphere enjoy must be roughly true for a wooden box and a bowling ball.

A good example is the tangent property of the ellipse. The pool table analogy is not merely a brilliant rhetorical device; we actually could build such a pool table, with green felt and everything. It might take a little trial and error to adjust the size of the hole and the springiness of the cushions, but we could definitely get it to work; we could shoot an actual ball in any direction and it would always go in. People have also built elliptical rooms that exhibit the tangent property in a different way. Two people stand at the focal points and whisper to each other. All of the sound waves emitted by one person bounce off the walls and end up in the other person's ear. The result is

that they can hear each other, and no one else in the room can hear a thing.

So how is it that these things actually work? If atoms and points are so different, why does a pool table made of atoms behave so much like an imaginary ellipse made of points? What is the connection between real objects and mathematical ones?

First of all, notice that something like the elliptical pool table wouldn't work if it were too small; for instance, if it were only a few hundred atoms wide. This object would behave nothing like an ellipse. An atom-sized ball would simply fly through the gaps in the wall or get involved in some complicated electromagnetic interaction with it. In order to behave at all geometrically, an object has to contain enough atoms to statistically cancel out these kinds of effects. It has to be big enough.

On the other hand, if the pool table were too big, say the size of a galaxy, it would also fail due to gravitational and relativistic effects. To be like a geometric thing, a real thing has to be the right size; namely, it has to be about *our* size. It has to be roughly at the scale at which we humans operate. Why? Because we're the ones who made up the mathematics!

We are creatures of a certain size, and we experience the world in a certain way. We're much too big to have any direct experience with atoms; our senses can't pick up anything that small. So we have no intuition at that scale. Our imaginations are informed by our experiences; it's only natural that the kind of imaginary objects we would create in our minds would be simplified and perfected versions of the things we've seen and felt. If we were a radically different size, we would have developed a very different type of geometry—at least initially. Over the centuries, people have invented lots of different geometries,

some of which work well as models of reality at very small or very large scales and some that have nothing to do with the real world whatsoever.

So the connection between geometry and reality is *us*. We are the bridge between the two. Mathematics takes place in our minds, our minds are a by-product of our brains, our brains are part of our bodies, and our bodies are *real*.

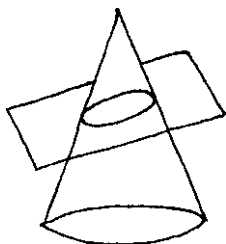
**Can you see how to make a rough
model of an ellipse using a pencil, two
thumbtacks, and a piece of string?**

27

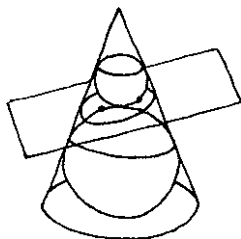
An ellipse is one of those rare shapes that we can actually talk about; it has a definite, precise pattern that can be put into words. Of course, all we've really done is to take a preexisting pattern (namely that of a circle) and modify it slightly; it's not like we built up the ellipse pattern from scratch. An ellipse is a *transformed* circle, and it is the transformation itself (namely dilation) that endows the ellipse with its various properties and allows us to speak of it at all. The classical focal definition is another way; it's a generalization of the idea of a circle and its center.

The point is that we created a new shape by modifying an old one. Any geometric transformation can be used in this way, provided we have a precise description of how it works ("a sphere with a dent in it" is a bit too vague). In particular, if a shape has a definite, describable pattern, then so will any dilation of it.

One simple way to get new shapes from old ones is by taking cross-sections. Ellipses occur as cross-sections of a cylinder. What happens when we slice other three-dimensional objects? A sphere is certainly an attractive candidate. Unfortunately, all of its cross-sections are circles, so we get nothing new. What about the cross-sections of a cone?



Surprisingly, these turn out to be ellipses. This might at first seem very strange, seeing as how cones are so different from cylinders. You would expect them to have more asymmetrical, egg-shaped cross-sections. On the other hand, it's not hard to modify our earlier argument with the Dandelin spheres to show that the cross-sections of a cone satisfy the exact same focal property.



Again we have two spheres, of different sizes this time, which each hit the plane at exactly one point. The main difference now is that the spheres no longer touch the cone along

their equators, but along parallel circles above their equators. Nevertheless, the same argument with the tangents shows that the cross-sectional curve has the right focal property, so it is in fact an ellipse.

Can you work out the details of this proof?

This is a rather depressing state of affairs—the switch from cylinders to cones doesn't seem to give us any new curves. But wait! There are *other* ways to slice a cone.

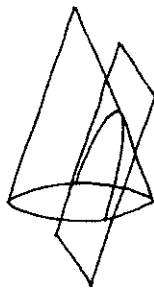


Because a cone is slanted, we can get different types of cross-section depending on whether the slicing plane is more or less slanted than the cone itself. When the plane was less slanted than the cone we got an ellipse. What curve are we getting now?

It's certainly not an ellipse, that's for sure. For one thing, it never closes up—it just keeps getting bigger and bigger as we extend the cone. Of course, we could chop it off at some point, but that seems rather arbitrary. A simpler and prettier idea (at least to me) is to imagine an *infinite* cone, so that the cross-sectional curve is infinite as well. The plane never comes out the other side!



Curves that occur as cross-sections of a cone are called **conic sections** or conics for short. There are actually three types of conic section, depending on the slant of the slicing plane. If the plane is less slanted than the cone, we get an ellipse. If the plane is more slanted than the cone, we get the kind of infinite curve we've just been talking about, called a **hyperbola** (Greek for "thrown beyond"). The remaining possibility is when the slicing plane is at exactly the same slant as the cone itself.

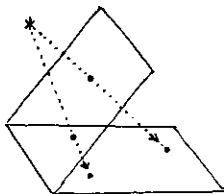


This kind of conic, called a **parabola** (Greek for "thrown beside"), is also infinite, but (as we will soon discover) is shaped quite differently than a hyperbola.

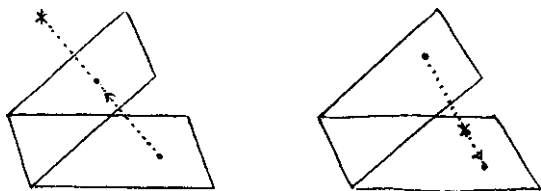
The point is, there are different ways to slice a cone, and depending on what kind of slice you make, you get different types of curves with very different properties. The conic sections were extensively studied by the classical geometers, most notably by Apollonius (circa 230 BC). One of the great

discoveries of this period was that hyperbolas and parabolas have their own focal and tangent properties, just as ellipses do. Of course, I want to tell you about them, but first I thought it would be nice to show you a somewhat different, more modern way to think about the conic sections.

The idea is to think of them *projectively*. Let's imagine two planes in space. Instead of choosing a particular direction to project in, let's fix a certain point in space (not on either plane) to project *from*.



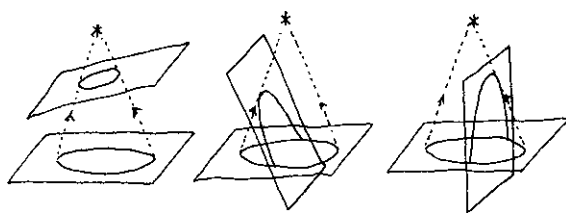
Points on the first plane are projected onto the second plane by straight lines through this projection point. (Sometimes I like to think of the projection point as the sun, and the projected images as shadows.) Of course, there's nothing saying that the second plane has to be behind the first; we might be projecting toward the point instead of away from it. We could even place the projection point *between* the two planes.



In any case, we have a new type of projection, usually called a **central projection**, as opposed to the kind of parallel projection we talked about before.

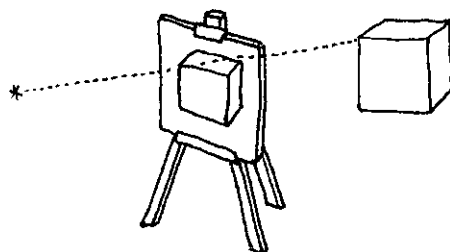
What is the effect of central projection when the planes are parallel? What if the projection point lies between the planes?

Both types of projection are transformations—ways to change one shape into another. This gives us a systematic way to create new shapes and relate them to old ones. In particular, we can view the conic sections—ellipse, hyperbola, and parabola—as being various central projections of a circle.



One interpretation of this is that these curves actually *are* circles, only viewed from different *perspectives*.

In fact, the whole issue of perspective comes down to central projection. The very act of seeing is a projection: the external world is projected through the pupil of the eye onto the retina. A perspective drawing is an attempt to mimic this process, using an imaginary observer as a projection point. Geometric projection is just the ultimate idealization of this process.



Of course, once you have a mathematical idea, whatever its origin, it quickly breaks free from all ties to reality. A completely new type of geometry, the so-called **projective geometry**, arose in the early seventeenth century out of the attempt to understand the mathematics of perspective.

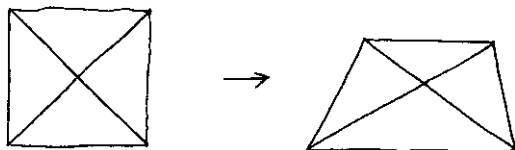
Can any three points on a line be
projected to any other three collinear
points? How about four points?

28

Since projection corresponds to a change in perspective, it's natural to think of two objects related by a projection as being the same; that is, two different views of the same object. The philosophy of projective geometry is that the only properties of a shape that matter are those that are unaffected by projections. What is intrinsic, what is "real" about a shape, should not depend on one's point of view; beauty should *not* be in the eye of the beholder. Any feature that changes under projection is not so much a property of the object itself but of the way in which it is being viewed. This is a rather modern way of thinking. We have a certain type of transformation (in this case projection), and we are interested in those structures that are *invariant*.

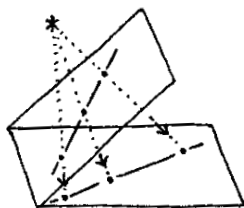
Are all triangles the same projectively?
How about all four-sided polygons?

The biggest difference between classical and projective geometry is that the traditional measurements—angle, length, area, and volume—no longer have any meaning. Projection warps a shape so much that all such measurements get radically changed. In this sense, projection is extremely destructive.



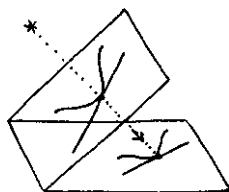
So if projective geometry is not concerned with measurements, what *is* it concerned with? What sorts of things are unaffected by projection?

A good example is *straightness*—any projection of a straight line is still a straight line. From a projective point of view, straightness is “real.” In particular, if a collection of points is collinear (meaning the points all lie on the same line), they remain collinear after any projection. If things line up, they line up; it doesn’t depend on your point of view.



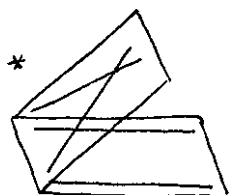
Is a projection of a polygon always a polygon?

Another projective invariant is tangency: if a line is tangent to a curve at a certain point prior to projection, it will still be tangent afterward, even though the shape of the curve and the position of the line may change.

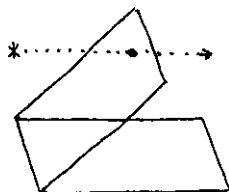


Generally speaking, anything having to do with intersections is usually a projective invariant. For two curves, both the question of whether they cross each other, as well as the number of times they cross each other, are projective invariants. But not, for instance, the *angle* at which they cross each other. That would get totally messed up.

Actually, it turns out that the intersection issue is a bit more complicated. Intersection is *not* in fact a projective invariant. It is even possible to have two intersecting straight lines that, when projected, become parallel.

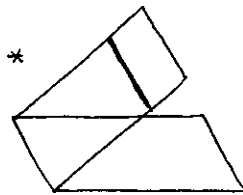


What's happening here is that the point where the two lines intersect doesn't appear on the target plane at all. In fact, in any central projection, there will be special points on one plane that don't make it onto the other.



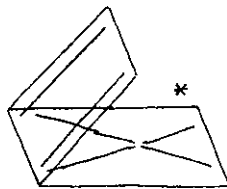
The trouble is that sometimes the line of sight from the projection point is *parallel* to the target plane. This is something of a disaster. It means that projection is bad—it loses information. In particular, it can lose the information of whether two lines intersect or not.

The extent of the damage is that there is an entire infinite *line* of points that will disappear under projection.



Specifically, it's the line parallel to the target plane that is at the same height as the projection point. All the points on this line will be lost in the projection process.

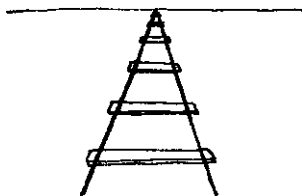
Something similarly disastrous happens in reverse. If we start with parallel lines on a plane and project them onto another plane, we obtain something truly monstrous—a pair of crossed lines with the crossing point *missing*.



This is, of course, a completely unacceptable state of affairs. We simply cannot allow this sort of ugliness!

What does a projection of three
parallel lines look like?

By the way, it is precisely this feature of central projection that is responsible for the vanishing point phenomenon—parallel lines, such as railroad tracks, appear to meet at a point on the horizon.



From a practical point of view, say to an artist or architect, this is all good news. It's very nice to be able to draw convincing pictures of railroad tracks, and nobody needs to lose any sleep about some missing points. Mathematically, however, it's quite disturbing. Disturbing enough, in fact, to lead geometers to take a very bold and imaginative step—to *redefine space itself*.

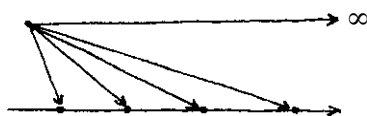
The idea is really quite ingenious. What goes wrong with projection is that not all lines through a point necessarily hit a given plane.



The trouble is, things can be parallel. Lines can be parallel to lines, planes can be parallel to other planes, and, as in this case, lines and planes can be parallel to each other. Since it is parallelism that causes the problem, the idea is to get rid of it—to make it so that lines or planes that lie in the same direction actually do meet.

The plan is this: for each direction in space, we imagine a new point somehow infinitely far away in that direction. The idea is that all lines that lie in that direction will now meet at the new imaginary point. It's that simple. We just throw in enough new points (one for each direction is enough) so that parallel lines and planes now intersect each other.

One nice way to think about it is to imagine a line and a point and see how various lines through that point intersect the line.



As the lines get closer to being parallel, the intersection points move farther and farther to the right. The philosophy is that when the line becomes exactly parallel, it still has an intersection point, one that is infinitely far away to the right. Interestingly, the same thing happens with lines slanted to the left. The new points we are adding lie both infinitely far to the right *and* to the left. It's as if our lines are somehow like circles that pass through infinity and come back around the other side.

Does this sound like the insane ravings of a lunatic? I admit it takes a bit of getting used to. Perhaps you object to these new points on the grounds that they are imaginary—they're not really there. But none of the things we've been talking about are real anyway. There's no "there" there in the first place. We made up imaginary points, lines, and other shapes so that things could be simple and beautiful—we did it for art's sake. Now we're doing it again, this time so that projections will be

simple and beautiful. It's really nice, once you get accustomed to it.

These points we're adding in are called *points at infinity*. The new enlarged space we've created, consisting of ordinary three-dimensional space plus all the points at infinity, is known as **projective space**. It is customary to add the appropriate points at infinity onto all the various lines and planes as well. A *projective line* is thus an ordinary line together with the point at infinity corresponding to its direction. A *projective plane* is a plane, along with all the points at infinity that you would expect—the ones corresponding to the various directions on that plane.

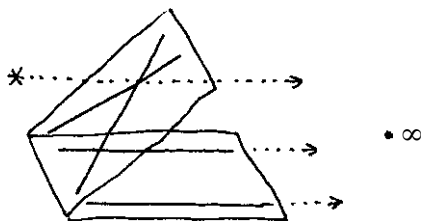
The upshot of this is that we have a new geometry, one where parallelism has been banished. Two lines on a plane intersect, period. If they intersected before, they still do. If they were parallel before, they now meet at infinity. This is a much prettier, more symmetrical situation than in classical geometry.

What about two planes? Normally, two planes intersect in a line. What happens when the planes are parallel? Notice that parallel planes have the exact same points at infinity, and that these points then constitute the intersection of the two planes. This makes it desirable to view the points at infinity of a plane as lying on a *line at infinity*. Now we can say with complete generality that two projective planes in projective space always intersect in a projective line.

Similarly, it is nice to think of the complete set of points at infinity in projective space as forming a projective plane at infinity. Then we can say, for instance, that a line and a plane always meet at exactly one point (unless, of course, the line happens to lie *in* the plane).

Do two lines in projective space necessarily intersect?

Now that we have a better environment for it to operate in, projection becomes a very nicely behaved transformation indeed. Instead of parallel lines being turned into a disgusting pair of crossed lines minus the crossing point, we can see now that the lines were crossed all along, and all that's happened is the crossing point has moved from infinity to an ordinary point.



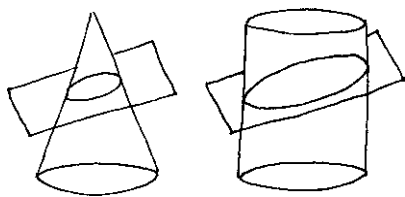
Of course, the right way to deal with projective space is to forget about the distinction between ordinary points and points at infinity. Projectively, there is no such distinction; what is ordinary from one perspective is infinite from another. Projective space is a completely symmetrical environment, and all of its points are created equal.

In particular, the distinction between parallel and central projection is rather spurious. Parallel projection is just central projection from a point at infinity. So we might as well drop the adjectives, which reflect a classical bias, and simply call them both *projection*.

We now have a completely reformed projection transformation, and we've identified a few of its invariants—straightness, tangency, and intersection. Can you find any others?

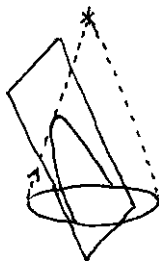
Can you discover a projective invariant?

Now I can be a little more precise about something I mentioned earlier—that the conic sections can be thought of as projections of a circle. For an ellipse there is not much more to say; we saw that certain slices of a cone or cylinder give us ellipses, and these are certainly projections of a circle.



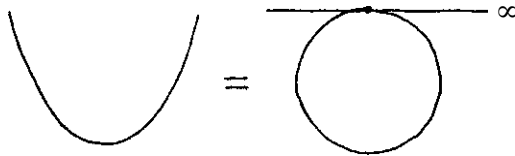
The cone corresponds to a central projection, while the cylinder is giving us a parallel projection of a circle. Since these are really the same projectively, it makes sense to think of a cylinder as a special type of cone—one whose tip is at infinity.

A parabola occurs when we slice a cone at the same slant as the cone itself.



In this case we are again projecting the circle from the horizontal plane onto the slanted plane using the tip of the cone as our projection point. So parabolas are certainly projections of a circle. Notice that there is exactly one point of the circle that

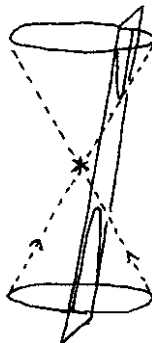
does not project onto the slicing plane proper; it ends up being projected to a point at infinity on the slicing plane. This means that a parabola is simply a circle with one of its points at infinity. The line at infinity is then tangent to the circle.



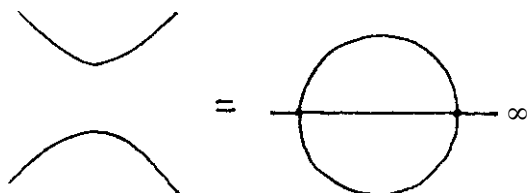
As for a hyperbola, something a little strange happens.



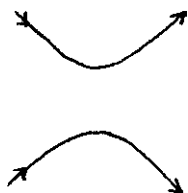
The part of the circle *behind* the slicing plane projects nicely, forming the characteristic bowl shape, but where is the rest of the circle? Surprisingly, it appears *above* the cone. In other words, central projection from the tip of the cone doesn't only shine down, it also shines up.



So the circle projects to *two* bowl-shaped curves, one pointing up and the other pointing down. A hyperbola, then, should be thought of as consisting of two pieces. Again it is a projection of a circle, only this time there are two points at infinity.



As we travel along the hyperbola, we go off to infinity in one direction, pass through the point at infinity corresponding to that direction, and return along the other piece on the opposite side.

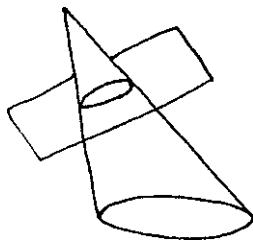


When a cone is sliced by a plane to form
a hyperbola, which two points on the
circle are projected to infinity?

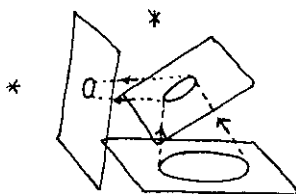
So the conic sections, properly understood, are just projected circles. This means that, projectively speaking, they *are* circles. The differences between them from a classical point of view depend on how the circle intersects the line at infinity—whether in zero, one, or two points.

Not only that, it turns out that *every* projection of a circle is

a conic section. No matter how you project a circle, you will always get either an ellipse, parabola, or hyperbola. There aren't any other curves out there that are projectively equivalent to a circle. In particular, this means that a slanted cone gives us the same cross-sectional curves as an upright cone.



Even if the base of the cone is itself a conic section, say an ellipse, we still don't get anything new. That is, a projection of a conic is still a conic.



In general, a projection of a projection is always a projection. Roughly speaking, a perspective view of somebody else's perspective view is still a perspective view. This is one of the nicest features of projective geometry—projective space and projective transformations form a closed system, which is in many ways simpler and more beautiful than classical geometry.

Shine a flashlight on the wall at various angles.

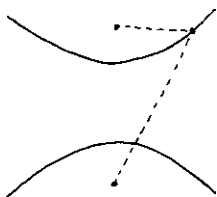
Can you see all three types of conic section?

29

As gratifying as it may be to view the conic sections projectively—to see them as different perspective views of the same circle—it doesn't actually tell us that much about the geometry of these curves. It's all very well to know that hyperbolas, parabolas, and ellipses are projectively equivalent, but they are still different shapes. What do they look like exactly? How, for instance, does a parabola differ from a hyperbola?

At this point, we know a lot more about ellipses than we do about the other conics. We know that ellipses are dilated circles, and we know they have particularly nice focal and tangent properties. Can we say anything similar about hyperbolas and parabolas? It turns out that we can.

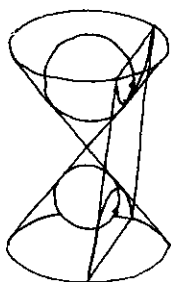
Hyperbolas have a very beautiful focal property, as a matter of fact. Like an ellipse, a hyperbola contains two special focal points, and as a point moves along the hyperbola, its distances to these focal points follow a simple pattern.



This time, however, it's not the sum of the distances that remains constant, it's the *difference*. That is, a hyperbola is the set of points whose distances to two fixed points differ by a fixed amount.

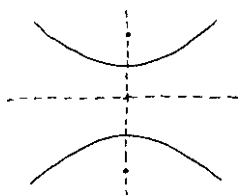
Naturally, such an outrageous claim requires some sort of proof. We need to show that if a cone is sliced steeply (so as

to make a hyperbola) then the points of the cross-section must obey this new focal property. As you might expect, this can be done with spheres and tangents in the usual way.



Can you work out the details of this proof?

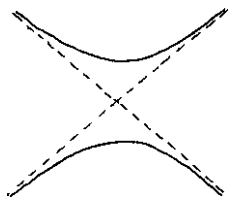
The focal property tells us quite a bit about hyperbolas. For one thing, it means they must be fairly symmetrical.



Not only are each of the two pieces of the hyperbola themselves symmetrical, but they are mirror images of each other. There is symmetry across the line connecting the focal points and also across the perpendicular line between them.

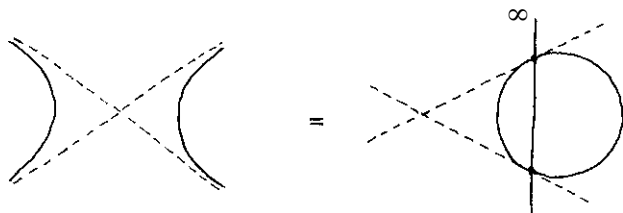
Why do hyperbolas have so much symmetry?

Another beautiful feature of hyperbolas is the way they nestle so nicely between a pair of crossed lines.

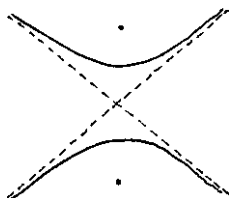


Neither of these lines actually touches the hyperbola, but as you travel out along the hyperbola you get closer and closer to them. In other words, these lines are simply the tangents to the hyperbola at infinity.

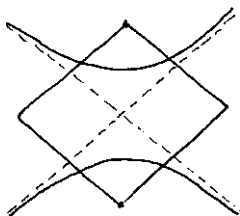
The simplest way to think of it is to view the hyperbola as a circle in projective space that meets the line at infinity at two points. (That is, after all, what a hyperbola is.) The circle then has two tangent lines at these points, and these are the lines that we're seeing.



Since hyperbolas are symmetrical, the crossing point of the tangents must be exactly halfway between the two focal points.



There is a very pretty connection between these tangent lines and the focal property of the hyperbola. If we draw lines through the focal points, parallel to the tangents, we get a diamond shape.



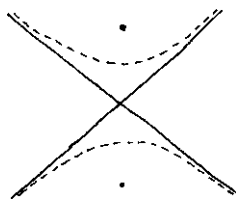
Because of the symmetry, the sides of this diamond must all be the same. The angles won't necessarily be right angles, so we can't say it's a square, but it's still a nice diamond. (You could also call it a rhombus if you like that word better.)

The focal property says that the distances from each point on the hyperbola to the focal points have a constant difference. It turns out that this constant difference, what we might call the **focal constant** of the hyperbola, is exactly equal to the side length of the diamond.

**Why is the focal constant of a hyperbola
equal to the side of the diamond?**

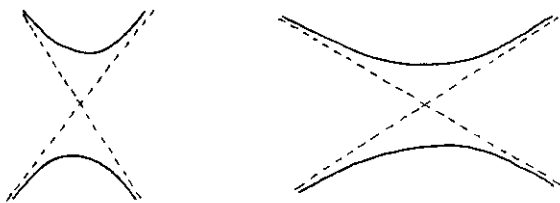
(Probably the easiest way to see it is to imagine a point moving along the hyperbola toward infinity, and think about what happens to the lines connecting it to the focal points.)

One consequence of this is that a hyperbola is completely determined by its tangents at infinity (the two crossed lines) and its focal points.

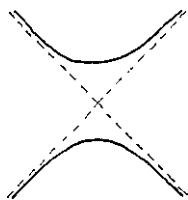


Any pair of crossed lines, together with a pair of symmetrically placed points, determines a unique hyperbola. The reason is that if we know the lines and the points, we can make the diamond and get the focal constant. This, along with the location of the focal points, determines every point on the hyperbola. So, in order to specify a particular hyperbola, it is enough to know the angle between the lines and the distance between the focal points.

In fact, since scaling does not affect angles, the shape of a hyperbola depends only on the angle between the tangents. Two hyperbolas with the same angle but different focal distances are simply scaled versions of each other; they are similar hyperbolas. So the different hyperbola shapes correspond to the different possible angles between the tangent lines.



In particular, there is the very special **right hyperbola** whose tangents meet at a right angle.



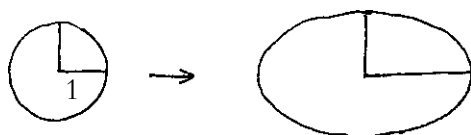
This shape is to hyperbolas what the circle is to ellipses; it is the standard object from which all others are obtained. That is, every hyperbola occurs as a dilation of a right hyperbola.

Why is every hyperbola a dilation of a right hyperbola?

(There is a subtlety here: How do we know that a dilation of a hyperbola even *is* a hyperbola?)

Hyperbolas and ellipses have a lot in common. They have very similar focal properties, involving the distances to two fixed points, the only difference being that it is the sum of these distances that provides the focal constant for an ellipse, whereas for a hyperbola it is their difference. Both are infinite classes of shapes that can be obtained as dilations of a single prototype—the circle and the right hyperbola.

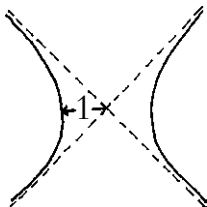
More specifically, once we have chosen a length unit we can speak of the **unit circle** whose radius is that unit. Then any ellipse can be obtained by dilating this circle twice—by a certain factor in one direction and by another factor in a perpendicular direction.



In this way we can think of an ellipse as having a **long radius** and a **short radius**. The ellipse is then completely determined by these two lengths.

If an ellipse has long radius a and short radius b , where are its focal points?

Similarly, we can talk about a **unit hyperbola**. This would be the right hyperbola whose distance from center to edge is exactly one unit.

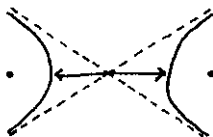
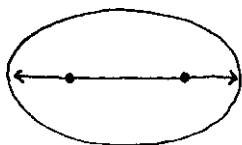


Every hyperbola then occurs as a dilation (by two factors, in two directions) of this one.

Where are the focal points of a unit hyperbola?

What if we dilate it by factors a and b ?

Another amusing similarity between ellipses and hyperbolas is the way the focal constant appears geometrically.

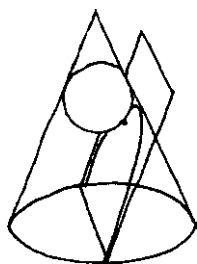


**Show that the focal constant of an ellipse
or hyperbola is equal to its diameter.**

It also happens that ellipses and hyperbolas have similar tangent properties. For the ellipse it's the "pool table" effect. What is it for the hyperbola?

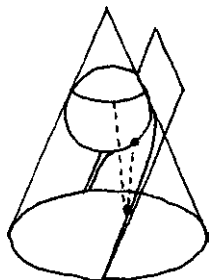
**Can you discover the tangent
property of a hyperbola?**

The parabola, it turns out, is another story altogether. It does have a focal property, but it is of a very different character from those of the ellipse and hyperbola. Instead of two focal points, a parabola has only one. When we slice a cone at exactly the same slant as the cone itself, we create only one compartment capable of housing a sphere in the right way.

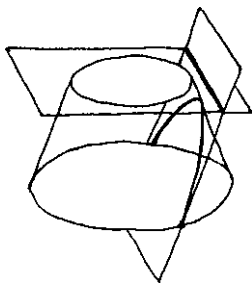


That is, there is only one sphere that is simultaneously tangent to the cone and the slicing plane. As usual, the focal point of the parabola is the point where this sphere hits the plane. The distance from a point on the parabola to the focal point is the same, then, as the distance from the point to the sphere, along

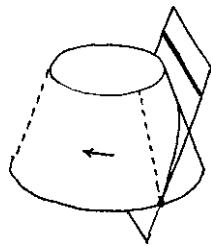
the cone. In other words, the distance to the circle where the sphere meets the cone.



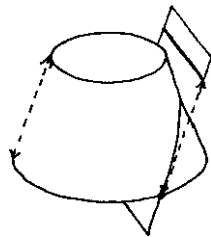
In the case of the ellipse and the hyperbola, we had another focal distance we could compare this with. Here we've got bupkes. How can we understand what this length means geometrically? I think the best way to see what's going on is to slice the cone twice horizontally, both through the circle and through our chosen point, to make a sort of lamp-shade thing.



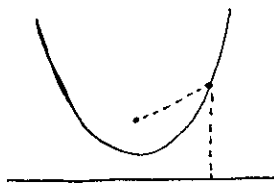
This removes any unnecessary cone baggage. Notice that the plane through the circle intersects the slicing plane in a certain line. This line turns out to be the key to the whole business. The important thing is that it depends only on the parabola itself, not on which point we happened to choose.



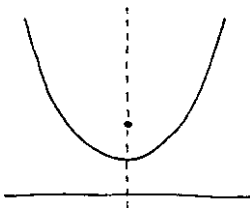
Now, here's the beautiful observation. The length that we're interested in (the distance from our point to the focal point) is just the distance along the lamp shade between the two horizontal planes. We can swing this length around the lamp shade without changing it. In particular, we can roll it around until it's directly opposite the slicing plane.



Now it's easy to see what this length is—it's just the distance from our point to the special line. How pretty! So the deal with parabolas is that not only is there a focal point, there is also a **focal line**, and the points of the parabola obey the beautiful pattern of being equidistant to both.



This focal property of a parabola has a number of interesting consequences. For one thing, it means parabolas must be symmetrical (not that that's any great surprise).



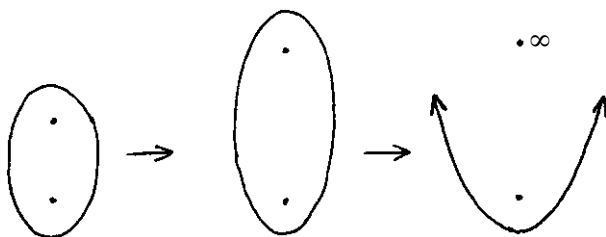
Since a parabola is completely determined by a point and a line, it's only the distance between them that makes one parabola different from another.



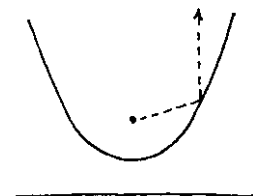
What this means is that any two parabolas are just scalings of each other. In other words, all parabolas are similar. There's really only one parabola shape. There are lots of different ellipses and hyperbolas depending on how you stretch them, but there's only one parabola. That makes it very special.

What about dilations of a parabola?

One nice way to think about parabolas is to view them as *infinite ellipses*: a parabola is what you get when you fix one focal point of an ellipse and send the other one off to infinity.

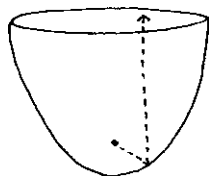


(We could also think of them as infinite hyperbolas in the same way.) Parabolas lie, in a sense, on the borderline between ellipses and hyperbolas. Right away, this tells us what the tangent property of a parabola must be: if we shoot out from the focal point, we will hit the wall of the parabola and bounce straight out to infinity.



Can you prove this tangent property directly,
without any “infinity” mumbo-jumbo?

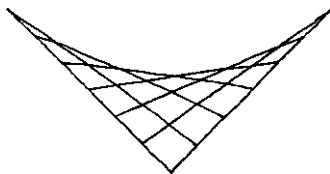
Perhaps even more beautifully, the surface formed by rotating a parabola, usually called a **paraboloid**, has the exact same tangent property in all directions.



This has a number of amusing practical applications. For one thing, it says that if we make a parabolic mirror (a mirror in the shape of a paraboloid) and place a lightbulb at the focal point, then all the radiation will be sent straight out; none of the energy will be wasted. This is exactly how flashlights and automobile headlights are designed. Running this in reverse, a parabolic mirror also makes a terrific solar oven. All the sunlight entering the mirror is focused at a single point. (That's why it's called a focal point.) Conic sections make good lenses; they bend light in an interesting and useful way.

If I have lingered on the subject of conic sections for so long, it is because they are so beautiful and have so many interesting properties, and I can't resist telling you about them. The other reason is that it is *possible* to tell you about them. It's not that easy to talk about curves, and conics are relatively simple as far as things go.

I want to stress something. These conic sections are very particular and specific curves—not every bowl-shaped object is a parabola or hyperbola. Most curves don't have anything like a focal or tangent property. These things are special, and we should cherish them!

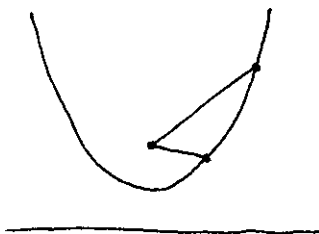


If you connect lines in this evenly spaced pattern, a parabola appears. Why?

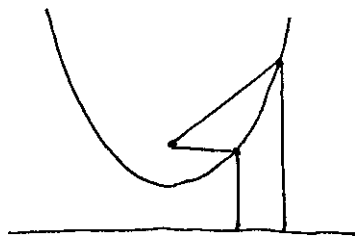
One final word about conics, and that concerns their measurement. We've already discussed the ellipse situation. Since an ellipse is simply a dilated circle, its area is easy to measure; for the same reason, its perimeter is not. To be precise, if we have an ellipse whose long and short radii are a and b , say, then the area is simply πab . Do you see why? The perimeter, on the other hand, depends on a and b in a transcendental way. There is no formula in the sense of a finite algebraic description.

Unfortunately, that's par for the course; the same is true for the parabola and hyperbola. Of course, those curves are infinite, so we can't really talk about their perimeters as such. But even if we chop them off at some point, their lengths are not algebraically describable. Not that they aren't very interesting. In fact, we will return to the subject of conic section lengths a little later on, when we will have some more powerful measurement techniques.

There is one measurement that we are in a position to make, and that is the area of a *parabolic sector*.

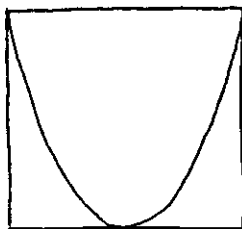


This is the kind of region formed by two straight lines drawn from the focal point out to the curve itself. The nicest way to measure this area is to compare it to the *parabolic rectangle* made by dropping lines straight down to the focal line.



Using the method of exhaustion, Archimedes was able to show that the area of the sector is exactly half that of the rectangle. Can you do the same?

Why is the area of a parabolic sector equal to half the area of the parabolic rectangle?

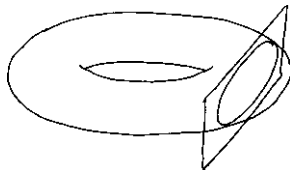


Show that a parabolic section takes up exactly two-thirds of its box.

30

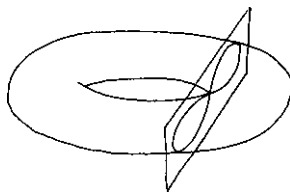
What an amazing can of worms we opened up just by slicing a cone! If such a simple shape as that has such interesting cross-sections, what will happen if we slice something more

complicated? What sort of curves do we get when we slice, say, a doughnut?



It turns out that this curve, despite the nice symmetrical oval shape, is definitely *not* an ellipse. It doesn't have the right focal or tangent properties, and it isn't a dilated anything; it's an entirely new kind of curve that we haven't seen before. I suppose we could call it a **toric section** if we wanted to.

If we move the slicing plane over a little, so that it just touches the inner rim of the torus, we get an even more exotic cross-section.



Of course, this is not just any old figure eight sort of shape but a very specific kind of curve with a very specific kind of pattern—the one that comes from being a slice of a doughnut. This is a fairly sophisticated geometric object. What sort of properties might this curve have? How on earth would we measure such a thing?

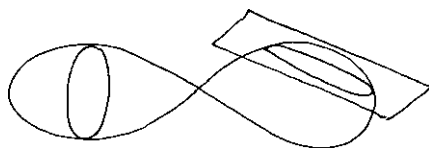
A while back I was talking about the description problem. The only shapes we can talk about are the ones that have a describable pattern. The geometer's job is to somehow turn this

pattern information into measurement information. Naturally, this is going to be a whole lot easier to do when the pattern is a simple one. The more elaborate our descriptions, the harder it is to say anything about the shapes they describe.

The sad truth is that measurement is almost always impossible. It is only the simplest objects that we have any hope of measuring. Even then it's no picnic. Remember how clever we had to be to measure a sphere? What chance do we have against a shape whose description is at all involved?

What I'm saying is that in addition to a description problem, we also have a complexity problem. Not only do our shapes have to have a pattern, they have to have a *simple* pattern. The problem is that the only way we have to measure curved shapes is the method of exhaustion, and if the patterns get too complicated, it quickly becomes unwieldy.

The situation is kind of ironic in a way. Before, we were worried about not being able to describe any new shapes at all. Now we have lots of ways to do that. For instance, we could start with one of these figure eight toric sections, rotate it in space to form a surface, and take a cross-section of that.

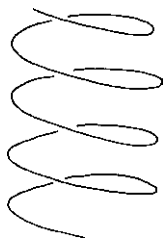


God only knows what sort of curve *this* is! No way is it an ellipse. No, our problem is not a shortage of new patterns. In fact, we've developed quite a little arsenal of description tools: we can dilate and project, take cross-sections, make Pappus-

type constructions, and perform any and all of these operations in succession. We are in a position to create some truly nightmarish mathematical objects, and there is absolutely *no hope* of being able to measure them. We may be out of the description frying pan, but we're definitely into the measurement fire.

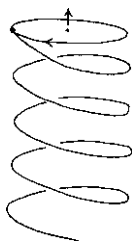
And you know what? I don't care. As our descriptions get more and more elaborate, not only does measurement become more and more difficult, but I get less and less interested. I really don't care about the cross-sections of a rotated toric section. For me, the point of doing mathematics is to see something beautiful, not to create a bunch of increasingly rococo patterns just because we can.

So are there any beautiful shapes left? As a matter of fact, there are. One particularly nice example is the **helix**.

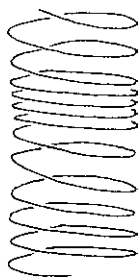


Now, that's the kind of simple, elegant shape I'm talking about! I would love to think about something pretty like that. Of course, before we do, we'll need some sort of precise description. What exactly is a helix?

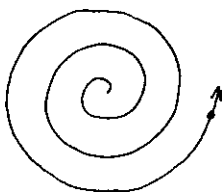
My favorite way to think of it is to imagine a circular disk in space, lying horizontally let's say, with a specially marked point on its rim. If we rotate the disk, and at the same time raise it up vertically, the special point should trace out a perfect helix.



Actually, there is an interesting subtlety here. To make a really nice helix, the rotating and the raising need to be done at *constant speed*. If we speed up and slow down, the helix will get stretched and squished in a most unpleasant (and sickeningly familiar) way:



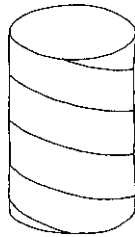
This means that our description of a helix (assuming we want a nice one) depends not only on the fact that the circle is moving but on the *way* that it's moving.



How can we view a spiral
as the result of a motion?

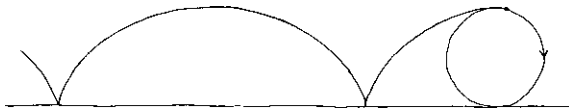
There are many different styles of helix, of course, depending on how fast the circle rises relative to how quickly it turns. An easy way to determine a particular helix shape is to specify both the radius of the rotating circle and the height increase that the point makes after one full rotation.

Sometimes it's nice to imagine a helix living on the surface of a cylinder, like a barber pole. The helix can then be described in terms of the size and shape of the cylinder and the number of full turns made by the helix.



How can we measure the length of a helix?

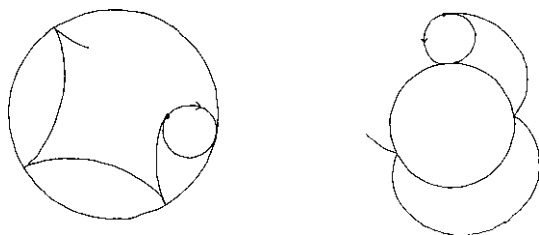
A helix is an example of what are called **mechanical curves**; that is, curves that are described as the path of a point on a moving object. Among the most fascinating and beautiful mechanical curves is the **cycloid**, the curve traced out by a point on a rolling circle.



This is a completely new shape, unlike anything we've seen before. It also turns out to have an amazing number of interesting properties; if there were a "Most Interesting Curve

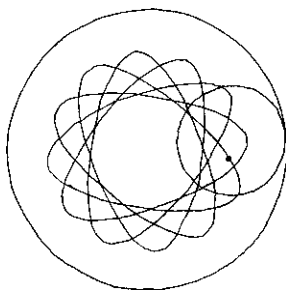
of the Seventeenth Century” award, the cycloid would win hands down.

There are several interesting variations on this cycloid idea. One is to have the circle rolling around inside another circle. This traces out a so-called **hypocycloid**. Of course, it could also roll on the outside, making an **epicycloid**.



How does the number of cusps of a hypocycloid depend on the radii of the two circles? What about for an epicycloid?

Another idea is to allow the tracing point to be in the interior of the rolling disk. In the hypocycloid case, this produces the very beautiful **spirograph** curves.



What happens if the tracing point is at the center?

The thing about something like a cycloid or a spirograph is that it is a natural, engaging pattern that is simple, satisfying, and attractive. This is not some contrived cross-section of a rotated projection of a slice of a whatever. For both aesthetic as well as practical reasons, these curves are interesting and crying out to be measured and understood. Of course, the only way to understand a mechanical curve like a cycloid is to understand the motion that creates it.

Which puts us in an entirely new situation. Up to now, the shapes and patterns we've been interested in have been static; they've just been sitting there. Now we're talking about things that move. We'll need to shift our emphasis away from shapes, and start thinking about *motions*.

Can you think of a way to
describe a helix on a torus?

A ladder slips down the wall until it hits the
floor. What curve does its midpoint describe?

PART TWO

TIME AND SPACE

1

What is motion? What exactly do we mean when we say that something is moving? We mean that as time goes by its position changes. When something moves, *where* it is depends on *when* it is, and the precise way that *where* depends on *when* is what makes the motion what it is. In other words, motion is a relationship between time and space.

In order to describe and measure motions, then, we'll need to be able to tell where something is—to record the position of an object and to know at what time that position occurred.

Needless to say, we're not talking about the way real objects (whatever they are) move around in the real universe (whatever that is). This, of course, turns out to be ridiculously complicated and unpleasant. Instead, we're talking about purely imaginary mathematical motions taking place in a purely imaginary mathematical reality.

So, our first problem is how to describe a particular location in space. This is not something we had to worry about before when we were measuring size and shape; the volume of a cone doesn't depend on where it is or what time it happens to be. But once things start moving, we definitely need a way to distinguish one place from another.

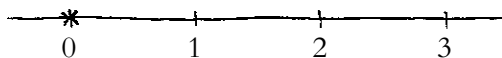
The simplest scenario I can think of is a single point moving along a straight line.



To describe the motion of this point, we need to be able to specify its position at any time. What we need is a map—some sort of reference system, some way of keeping track of where things are.

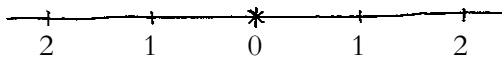
The easiest way to do this is to select an arbitrary point on the line as a reference point. Then we can specify any particular location on the line by saying how far away it is from this special point. This necessarily means choosing some sort of length unit, which, like the reference point itself, is completely arbitrary. A line, unlike the earth, has no natural landmarks; we have to put them there ourselves.

The upshot of this is that each point on the line gets assigned a numerical label, and we can refer to any position by giving its number (once we've selected a reference point and a unit).



Notice that the reference point itself receives the number 0. This point is usually called the **origin** of the system.

Actually, there is a slight problem with this plan: two different locations can receive the same label. For instance, the point that is one unit to the left of the reference point will get the label 1, as will the point one unit to the right.



We'll need a way to distinguish the two directions; otherwise, if we say that something is at position 1 we won't know which one we mean.

So our reference system not only needs an origin and a unit, but also an **orientation**. That is, we need to decide which direction is forward and which is backward. Of course, it doesn't matter which we choose. A line in the abstract has no left or right, and it is completely up to us to decide what those words mean.

In any case, once we've made our choices of origin, unit, and orientation, we can then refer to any position on the line unambiguously. We could say, for example, that a point was at position 3 in the backward direction, and that would pin it down completely.

An even nicer way to proceed is to use positive numbers for one direction and negative numbers for the other. Then we could simply say that our point was at position -3 .



There are several advantages to this scheme. For one thing, it means that all locations can be described by a single number, instead of a number and a direction. More important, it allows us to connect geometry and arithmetic in a very pleasing and beautiful way.

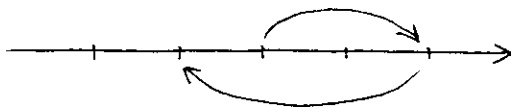
First of all, notice that moving one unit in the positive direction simply increases the position number by one.



I like to think of such a move as a shift. I imagine the entire line shifting over (to the right in this case) so that what was at position 0 is now at position 1, and so on. There is a shift for every number; we can shift by 2 or by the square root of 2 or by pi. We can also shift the other way; a shift (backward) by 2 units would correspond to the number -2 .

Geometrically, shifting is very nice because it preserves distances. If two points are at a certain distance from each other before shifting, they will be at the same distance afterward. A

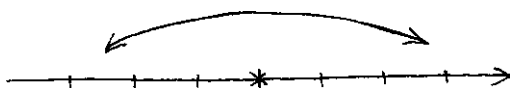
geometric transformation like this that preserves distances is called an **isometry** (Greek for “same measure”). A nice feature of isometries is that if you perform one isometry and then another, the result is also an isometry. In particular, if you shift by a certain amount, say 2, and then shift by another amount, say -3 , the result is also a shift, in fact a shift of -1 . So not only do two shifts make a shift, but the corresponding positive or negative numbers get *added*.



This means that the geometry of shifts has the same structure as the addition of numbers. In mathematical parlance, the two systems are *isomorphic*. This is what mathematicians are always on the lookout for—*isomorphisms* between apparently different structures.

So the main benefit of using positive and negative numbers to indicate direction is that we get this nice isomorphism between the group of shift isometries and the group of numbers under addition.

Actually, we get a lot more. There are other natural geometric transformations besides shifts; for instance, there are reflections. Reflections are nice because they are isometries. What happens if we reflect a point from one side of the origin to the other? Of course, its position number gets negated. Position 3 reflects to position -3 and vice versa. So we can say that the arithmetic operation of negation corresponds to the geometric idea of reflection.

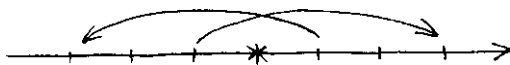


What is the arithmetic version of reflecting across a point other than the origin?

Are there any other isometries of a line?

Another nice example is scaling. If we blow up the line by a factor of 2, all distances will double, so a point that was at a certain distance from the origin will now be twice as far away. In other words, its position number will get multiplied by 2. This means that scaling corresponds to multiplication.

What happens if we multiply by a negative number? In this case, not only will there be a scaling (by a factor corresponding to the size of the number) but also a reflection. Multiplying by -3 is the same as stretching by a factor of 3 and flipping.

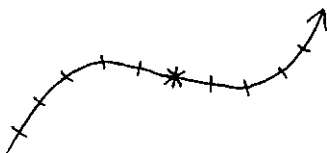


This means that the entire system of arithmetic, including positive and negative numbers, addition, subtraction, multiplication, and division, is mirrored completely by the natural geometry of the line. I especially like this correspondence because it explains such choices as having $(-2) \times (-3) = 6$: scaling by 2 and reflecting, then scaling by 3 and reflecting, is the same as scaling by 6.

In any event, we now have a very convenient way to locate points on a line—we use a numerical reference system. Once

again, the ingredients of such a system are an origin (reference point), a unit (for measuring distance), and an orientation (a choice of direction, to be called positive). The important thing to understand is that these are arbitrary choices having less to do with the space at hand and more to do with ourselves. Space has no orientation, no natural unit, and no special place. There is no such thing as left or right, up or down, big or small, here or there, until we make these *choices*.

As a matter of fact, there's really nothing about what we've set up that has anything to do with our space being a straight line; we could do exactly the same thing with any curve.



If we're interested in a point moving along a curve, we can set up a reference system just as before—choosing an arbitrary origin, unit, and orientation. Then every location on the curve would again have a unique numerical label.

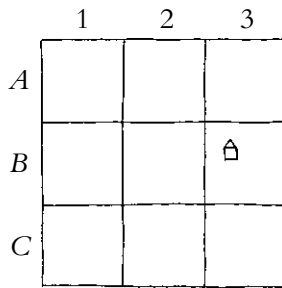
**What happens if the curve is
closed or intersects itself?**

**What is the distance between two points at
positions a and b ? Where is their midpoint?**

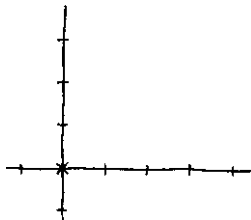
**What about the point one-third
of the way from a to b ?**

2

Having devised a way to locate points on a line, we can now try to do the same thing for a plane. How do we make a map of the plane? One idea is to simply mimic the system used for street maps:



Of course, this kind of grid system (where such and such a street might be found in square B-3) is much too crude for our needs. If we want to describe the motion of a point in a plane, we'll need to know precisely where it is at all times. We need the finest grid possible—one with no space at all between the gridlines. In other words, *every* horizontal and vertical position needs to receive a label. The customary way to do this is to use two number lines, one horizontal and one vertical. (It is also traditional to use the same unit for both lines and to have them intersect at their origins.)



Then any point in the plane can be referred to by its horizontal and vertical position numbers. It is just the same as a street map except that instead of blocks, we have a whole continuum of possible positions in each direction.

Another major difference is that the plane has no intrinsic landmarks. There is no “center of town,” no “north,” and no customary unit of distance like a mile. A grid, or **coordinate system**, on the plane is a completely arbitrary construct that we impose on it. There is no such thing as horizontal or vertical on an imaginary plane. These are choices that we make for our own convenience. When we coordinatize the plane, we are choosing two arbitrary (usually perpendicular) directions and deciding to call one of them horizontal and the other vertical. Obviously, there’s no one best way to do this.

I think it’s important to understand the choices that we’re making in more detail. First of all, there is the choice of reference point or origin. Of course, it can be anywhere; you get to decide where you want to put it. Then there’s the choice of unit, which is again entirely up to you. I usually like to choose a reference point and a unit that have something to do with the objects and motions under consideration—to tailor them to the situation at hand.

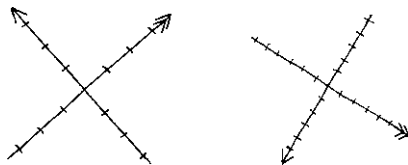
The really interesting choices involve the two lines. Since each line will have to be oriented, meaning that a choice of forward and backward along each line will have to be made, it is nicest to think that instead of two lines we’re really choosing two directions. These will be the positive directions along the horizontal and vertical lines of our grid.

But there’s one more choice to make after we’ve chosen the two directions; namely, which is which. With street maps, it is customary to use letters for one direction and numbers

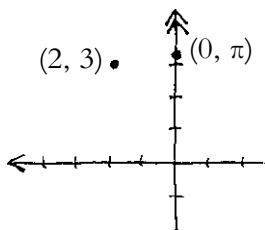
for the other. This avoids confusion. In our case, we can't get away with that because we don't have an infinite continuum of letters. So we distinguish them using order. We'll pick one direction to be first and the other to be second. If you want to call them horizontal and vertical, or the other way around, fine, just be aware that those words are meaningless. Words like up and down, clockwise and counterclockwise, left and right, horizontal and vertical refer to the way things are oriented with respect to your *body*. When Australians and Canadians point up, they are both pointing in the direction from their feet to their heads, but they point in (roughly) opposite directions in space.

The point is that we need to choose two directions and designate one of them as the first direction and the other as the second. This set of choices is what constitutes an orientation of the plane. In particular, we could designate a certain rotational direction, say from the first direction toward the second, as clockwise. So just as for the line, a reference system for the plane consists of an origin, a unit, and an orientation.

Here are two perfectly good coordinate systems (I've marked the first direction with an arrow and the second with a double arrow).

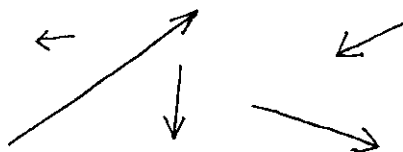


Once we've set up a system like this, each point in the plane will get a unique label consisting of two numbers. It is customary to write such a label as a number pair, such as $(2, 3)$ or $(0, \pi)$.

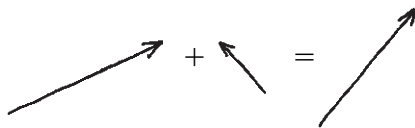


How does the distance between two points in a plane depend on their coordinates?

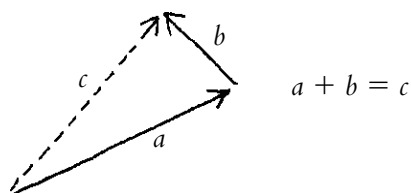
Just as we did with the line, we can relate the geometry of position in the plane to the algebra of shifts. A shift of the plane moves every point a certain distance in a certain direction. Notice again that a shift of a shift is still a shift. A nice way to represent such a shift is by an arrow of the appropriate length in the appropriate direction.



An arrow like this is called a **vector** (Latin for “carrier”). Since every shift corresponds to a vector and vice versa, we can talk about adding two vectors to get another vector. This would correspond to two shifts resulting in a total shift.



The easiest way to think of it is that a vector is an instruction, and adding is simply following the shifting instructions.

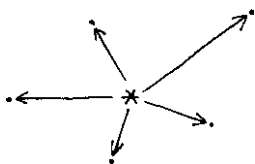


So the geometry of shifting is the same as the algebra of vectors. The reason we didn't see this before is that there are only two directions on a line. Of course, we can think of positive and negative numbers as vectors if we want to.

The effect of scaling on vectors is easy to see: it simply stretches (or shrinks) them without changing their direction. So we can talk about doubling a vector or dividing a vector by pi. We can also talk about the negative of a vector, namely the vector of the same length that points in the exact opposite direction. Naturally, multiplying a vector by a negative number would both dilate it and reverse its direction.

How do you subtract two vectors?

Once again we have a nice isomorphism between the shift isometries and an algebra of some kind. The point of vectors is that they encode geometric information algebraically. In particular, we can imagine a very simple vector-based reference system for locations in the plane. If we choose a fixed reference point, then every location in the plane can be thought of as being the tip of an arrow emanating from this origin.

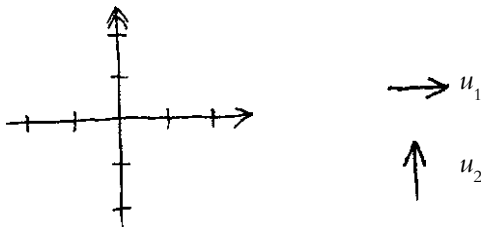


In other words, every point in the plane corresponds to a vector (with respect to a certain choice of origin). This sort of reference system is more like a radar screen than a street map.

If two points are described by the vectors a and b , what vector refers to their midpoint?

Can you use vector algebra to show that the lines drawn from the corners of a triangle to the midpoints of the opposite sides all meet in a single point?

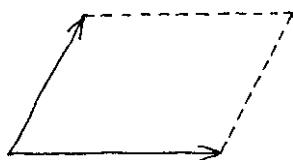
There is a very simple and natural relationship between vectors and coordinates. When we set up a coordinate map we first choose an origin, and then two directions. We can think of these directions as vectors. The nicest way to do this is to represent the two directions by so-called **unit vectors**; that is, vectors of unit length. So we have a first unit vector and a second unit vector.



Then instead of saying that a point has coordinates $(2, 3)$ we can say that the corresponding position vector is the sum of two vectors: the first unit vector scaled by 2, plus the second unit vector scaled by 3. That is, we can write algebraic descrip-

tions like $p = 2u_1 + 3u_2$ to describe where we are. Not that there's any real difference between the two schemes, just a slight change in viewpoint and notation. By the way, it's not at all necessary for our system to be rectangular; that is, the two directions or unit vectors need not be perpendicular. We still get a perfectly usable (albeit crooked) map of the plane.

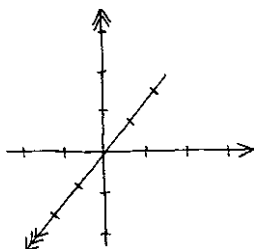
How does the angle between two vectors
depend on their coordinates?



How does the area of the parallelogram formed
by two vectors depend on their coordinates?

3

What about three-dimensional space? Can we do something similar? Yes, we can! Only we'll need three directions:



An orientation in this case will consist of three (usually mutually perpendicular) directions, given in order: first, second, and third. Then every location in space can be referred to by a number triple (a, b, c) or, what is the same, a weighted sum of unit vectors $au_1 + bu_2 + cu_3$. Everything goes through as before, including the isomorphism between shift isometries and vector algebra. Again, the thing to keep in mind is that a reference system is something we humans impose on space; it is not an intrinsic feature of space itself. We do it to keep track of moving things. If we were smart enough to do that without such a system, we would. It's not a pleasant thing to do, this plopping down of a coordinate system onto a space. It's ugly, and should be avoided whenever possible.

At any rate, we now have a technique for mapping out space: points on a line can be represented by numbers, points on a plane by number pairs, and points in space by number triples. Each time the dimension of the space goes up, we need an additional number slot in our representation. Each new slot corresponds to a new independent direction.

This is really what dimension means—the number of coordinates required to specify the different locations in space. So a line or curve is one-dimensional, a plane or the surface of a sphere is two-dimensional (longitude and latitude do the trick), and the space inside a cube is three-dimensional. The dimension of a space is a number that describes in a rough, qualitative way what life is like in that space—how much freedom you have to move around.

What about four-dimensional space? Is there such a thing? If we're asking whether four-dimensional space is real we might as well ask about three-dimensional space: Is there such

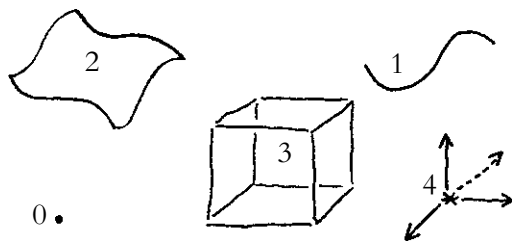
a thing? I suppose it *appears* that there is. We're walking around (apparently), and things certainly look and feel as though they are part of a three-dimensional universe, but when you come right down to it, three-dimensional space is really an abstract mathematical object—inspired by our perception of reality, to be sure, but imaginary nonetheless. So I don't think we should put four-dimensional space in any special mystical category. Spaces come in all sorts of dimensions, and none are any more real than any other. There are no one-dimensional or two-dimensional spaces in real life, and the only thing that gives the number 3 any special status is that our senses appear to offer us that particular illusion.

What I'm saying is that we are free to speak of locations in four-dimensional space in the same way as before, namely as quadruples of numbers. Alternatively, we could imagine four mutually perpendicular unit vectors, and then the various combinations of them would describe the positions in four-dimensional space. The major difference is logistical—we have no visual or tactile experience in four dimensions, and seeing as how we like to draw things on paper, which is itself (roughly) two dimensional, pictures of four-dimensional space are a bit problematic. As annoying as that may be, we can take comfort in the fact that pictures don't mean all that much anyway, and anything we wanted to understand or prove about objects in four-dimensional space would have to be handled ultimately by reason alone, just like anything else in mathematics.

**How many corners are on a
four-dimensional cube?**

So yes, there is such a thing as four-dimensional space: it is simply the collection of all quadruples of numbers. And the same goes for any number of dimensions. There is no reason why we couldn't work in eight- or thirteen-dimensional space if we wanted to. I suppose to be really precise, I should say that four-dimensional (Euclidean) space is actually the set of all possible four-tuples of *points on a line*—that is, each point in four-dimensional space is a quadruple of points in one-dimensional space. The set of four-tuples of numbers is in fact a *map* of that space, and not the space itself.

One thing I find particularly annoying is when people (especially in science-fiction movies!) talk about “the fourth dimension.” There is no fourth dimension—just as there is no first, second, or third dimension. (Which dimension is the third one, width?) Dimensions do not come in an order; they're not out there as preexisting entities. So there is no fourth dimension; there are *spaces* (lots of them) some of which happen to be four-dimensional. In other words, dimension is a number attached to a space, and every space has a dimension; namely, the number of coordinates a map of the space requires.



The modern way to think about dimension is as an invariant—one of the sturdiest invariants there is, in fact. A space can be subjected to the most drastic deformations and distortions,

and its dimension will not change. The surface of a sphere is two dimensional, and no matter how you stretch it, dent it, or twist it, it will remain two dimensional.

The main use of dimension in mathematics is as a classification tool. Most people enjoy sorting things into groups, and geometers are no different. Just as biologists like to divide living things into various categories (plants, animals, fungi, etc.), geometers are also faced with a rich and varied multitude of shapes and the desire to classify them into groups is irresistible.

The most important feature of living things is that they are alive. They convert energy, and distinctions can be made based on the manner in which this is done (e.g., photosynthesis, respiration, fermentation). This is what makes animals different from plants, for instance.

To me, the most important thing about a geometric object is that it can be *measured*. It makes sense to divide shapes into categories depending on the manner in which this is done. Curves are different from surfaces because length is different from area.

In a way, this is a subtle point. When we talk about measuring the circumference and area of a circle, we are actually talking about two completely different objects. The circumference is a length measurement of the *curve* known as a circle. The area measurement is referring to the *surface* consisting of the interior of the circle, usually called a *disk*. Similarly, geometers use the word *sphere* to mean the two-dimensional surface, whereas the solid object is referred to as a *ball*. So we measure the length of a circle, the area of a disk or sphere, and the volume of a ball.

Thus, one-dimensional spaces (also known as curves) correspond to one-dimensional measurement, namely length.

Two-dimensional spaces (surfaces) are measured in terms of area; and solids occupy volume.

What is the volume of a four-dimensional box? How long is its diagonal?

Show that a four-dimensional pyramid occupies one-quarter of its four-dimensional box.

What is the four-dimensional analog of a cone? Can you measure its volume? How about a sphere?

Dimension plays the same role in geometry as kingdom does in biology: it is the top-level hierarchical subdivision. As different as the surface of a cone is from that of a cube, they are closer to each other than either is to a straight line or the space inside a ball.

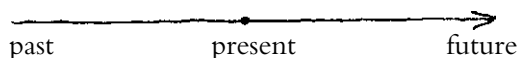
Here is another nice way to think about dimension. Suppose we have some shapes of various dimensions, and we scale them by a certain factor r . The natural size measurements are then affected in different ways depending on the dimension of the shape. The lengths of curves get multiplied by the scaling factor r , the areas of surfaces by r^2 , and the volumes of solids by r^3 . So the dimension appears as the power of the scaling factor.

What is the dimension of an angle?

4

In order to describe motion, we not only need a way to locate position but also the ability to tell time. Of course, we're not talking about real time, the kind of time that goes by in the physical world (and God only knows what the deal is with that!), but rather a purely abstract mathematical version of time, which, as with everything mathematical, we get to invent. What do we want mathematical time to mean?

The most elegant answer is this: time is a line. The points on this time line represent moments, and moving around on the line corresponds to going forward or backward in time. The choice of a line to represent time is interesting, because it gives us a geometrical way to think about something that (at least to me) is not in itself particularly visual.



So how do we tell what time it is? Naturally, we need some sort of clock. But what is a clock, exactly? A clock is a reference system! It is a way of assigning numbers to moments in time. To set up our clock, we simply do the same thing for time that we did for space: choose an origin (a reference time), a time unit, and an orientation (clockwise?). Having done this, every moment in time can then be represented by a single (positive or negative) number.

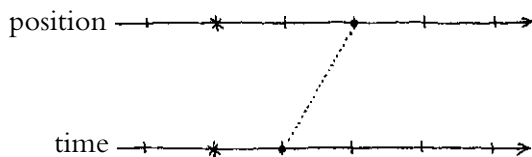
Sometimes I like to think of motions as experiments, and the time line as my stopwatch. The origin is then the moment I start my experiment. If we call our time unit a second (and we're free to call anything anything) then the number 2 would

refer to the precise instant two seconds after the start of the experiment, whereas the number $-\pi$ would correspond to the time exactly pi seconds before my experiment began.

Naturally, as with spatial reference systems, any clock is completely arbitrary, and we are free to design clocks to suit our present purposes, whatever they may be.

Let's imagine a point moving along a line in a certain way, perhaps speeding up and slowing down, maybe changing direction from time to time—whatever. Suppose we've chosen convenient reference systems for both time and space, so that the position of the point and the time of day are each represented by a single number.

Then the motion of the point can be completely described by knowing exactly which time numbers go with exactly which position numbers. If, for instance, the point is at position number 2 when the clock reads 1, then that information (position is 2 when time is 1) constitutes an **event** in the history of the motion, and complete knowledge of all such events is tantamount to the motion itself. Geometrically, we can represent an event like this as a pair of points, one in space and one in time, linked together by the motion of the point itself:



Of course, knowledge of one or even a million such correspondences between space and time is not enough. We need to know *all* of them. Just as we cannot measure a shape unless we know exactly where every one of its points is, we can't measure a motion unless we can say precisely where the object

is at every instant. This brings us right back to our description problem: we can't talk about a motion unless it has a pattern, and a pattern we humans can describe.

This means that with respect to a given reference system, the position numbers and the time numbers must satisfy some sort of numerical relationship that we can state in a finite amount of time.

For example, suppose that a point is moving at a steady rate and we've chosen a time unit (call it seconds) and a space unit (let's say inches) and that the constant speed of the point is, say, two inches per second. If we calibrate our clock so that the point is at position number 0 at the start of the experiment, then we know that when the time number is 0, the position number is 0 also.

Abbreviating the time number by the letter t and the position number by the letter p , we can say that when $t = 0$, $p = 0$. Also, when $t = 1$, $p = 2$; and when $t = 2$, $p = 4$. We could even make a little chart:

t	p
0	0
1	2
2	4

Since the point is moving at a steady rate, we know that the position number will always be exactly twice as big as the time number. So when $t = \frac{1}{2}$, $p = 1$; when $t = \sqrt{2}$, $p = 2\sqrt{2}$; and when $t = -\pi$, $p = -2\pi$ (assuming the point was moving before we started the stopwatch).

This means that we know *every* event in the history of this motion. This is because the pattern is describable. Either the

phrase “a point moving at a constant rate of two inches per second” or the more succinct $p = 2t$ serves to describe the pattern completely. Notice that both these descriptions depend on the choice of units: if we chose a different unit of time or distance or both, we would get a different description of the same motion.

In fact, if a point is moving along a line at a constant speed in a certain direction, we can always choose our orientations, units, and origins so that the pattern of position and time is simply $p = t$. Of course, if we had two points moving on the same line in different ways, we wouldn’t be able to choose a reference system in which both motions could be described so simply.

Design a pair of moving points on a line.

When and where do they collide?

Another (perhaps silly) example of a motion is a point standing still. If we choose this point as our origin, then the motion (or lack thereof) could be encoded as $p = 0$ with no mention of t at all.

The point is, any relationship of the form “position number equals blah blah blah” (where the blahs may or may not depend on the time number) describes a particular motion (with respect to that reference system). All that is necessary is that for each value of the time number the pattern produces a single, definite value for the position number. That’s what a motion is—a relationship between time and space that tells us the exact position at every moment.

Just as for shapes, where we wanted to somehow get measurement information (e.g., lengths, areas, angles) from pattern descriptions, we’ll want to figure out how to take patterns

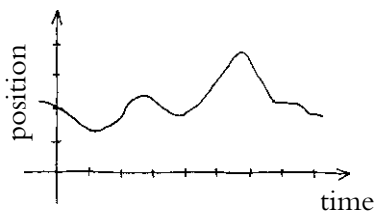
of motion and turn them into measurements. How fast is it moving? In what direction is it heading? How far did it travel, and how long did it take to get there? That sort of thing.

Which relationships between position number and time number correspond to constant speed motions on a line?

5

One thing that makes thinking about motion somewhat more difficult, or at least feel different from thinking about size and shape, is that we have no picture. A shape has a shape, but a motion is a *relationship*. How can we “see” a relationship?

One idea, of course, is to make a graph. In the case of a point moving on a line, we could imagine a chart with, say, time as the horizontal and position as the vertical.



At each moment in time, given by a number on the time line, there would correspond a position number, and we can simply plot these numbers on the chart to give us a visual representation of the pattern of motion.

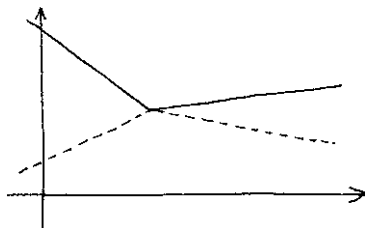
Notice that what we end up making is a curve. It is absolutely crucial to understand the status of this curve. This is *not* the

path that the point traces out—after all, the point is traveling along a straight line—rather, it is the record of its motion. The point itself is traveling within a one-dimensional space, whereas this curve here, this graph of the motion, is sitting in a two-dimensional space.

This two-dimensional space is very interesting. It is not entirely spatial, since one of its dimensions corresponds to time, and not entirely temporal either, since the other dimension refers to position. This environment is called **space-time**. The points of space-time can be thought of as events, and a motion is then a curve of events.

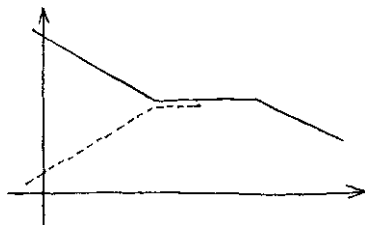
What is the space-time representation of a constant speed motion along a line?

In this way, all motions can be thought of statically—a motion in space is the same thing as a curve in space-time, and this curve does not move. As an example, imagine two points speeding toward each other on a line, colliding like billiard balls, and then bouncing away. The space-time picture might look something like this:



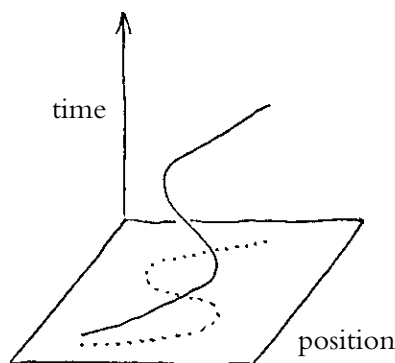
The important thing is to be able to read such a diagram: the points are not moving in a plane; they are moving along a straight line. It is the introduction of time as an extra dimension that gives us a planar diagram. Motion in one-dimensional

space corresponds to a curve in two-dimensional space-time. Rather than asking how things move in our universe, a physicist trying to understand billiard balls and other moving things can rephrase the question as, what curves in space-time are possible?



If two bugs crawl along the edge of a table in
such a way that their motion has this space-time
diagram, what must have happened?

Of course, the same thing is true in higher dimensions. Motion in a plane corresponds to a curve in three-dimensional space-time. I like to imagine the vertical direction coming up out of the plane as representing time, so that as a point roams around the plane, its space-time curve rises up through space directly above where the point is at that moment.



Again, we get a fixed, nonmoving geometric representation of the original motion. The deal we're making is to trade motion for dimension. We get to have a static picture instead of a more complicated moving one, but the price we pay is that we have to bump up the dimension. It's not necessarily a trade we'll always want to make, but at least it's an option.

Incidentally, this means that a motion in three-dimensional space corresponds to a curve in four dimensions, which means that you—your whole life in this world—is just a single curve in four-dimensional space-time. Actually, since you are composed of trillions of particles all wiggling and jiggling in space, it's more like your life is a braid of trillions of infinitesimally thin threads twisting around each other, some of them flying off, and so on. All the events—past, present, and future—of our whole ridiculous universe are writ on this one four-dimensional canvas, and we are but the tiniest brush strokes.

Ahem. The point is that at a cost of one dimension, we can replace motions (relationships between space and time) with curves—single, motionless, geometric objects.

This means that **kinematics** (the study of motion) is really the same as geometry. How something moves—the style of its motion—is completely reflected in the shape of its space-time curve. Understanding motion in a certain space is the same as understanding shape in a space one dimension higher.

So not only can we use geometry to study motion, we can also go the other way as well. Sometimes, a particular curve, which we are interested in for purely geometric reasons, can be thought of as the space-time image of a motion. This whole connection between motion and shape comes from our choice of a line to represent time. In particular, any geometric line can be thought of as a time line if we so desire.

Which curves in the plane can occur as space-time images of a one-dimensional motion?

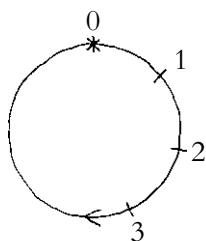
I think the most important thing to understand here is that motion is a relationship, and it is the relationship itself that is the central object of our inquiry. Whether you want to think of it as a motion in space or as a curve in space-time is secondary. Ultimately, when we do geometry or mechanics, we are not investigating shapes or motions but relationships. A graph of a relationship may be a nice visual representation, and such a picture might give us ideas, just as thinking of it as a motion might give us different ideas, but precise measurement information can come only from the relationship pattern itself.

What if time were two dimensional or circular?

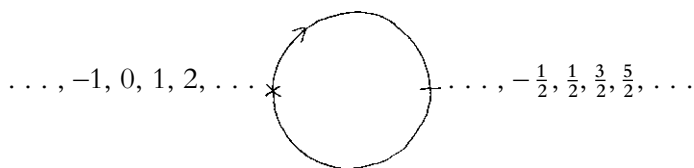
6

The thing that got me started talking about motion in the first place was that I wanted to understand mechanical curves like the helix and the spirograph curves. These shapes are described by points moving on rolling circles. Next to constant speed motion along a straight line, this is the simplest motion I can think of—the motion of a point, at constant speed, around a circle.

Of course, if the only thing that's going on is a point moving along a circle, then we're in essentially the same situation as with a line. We can choose any point on the circle as our reference point, pick one direction around the circle as positive, and obtain a number circle in the same way as a number line.



In this manner, we can record the position of a moving point at all times. The only new wrinkle is that, since the circle is closed, the numbers will wrap around and each point will receive infinitely many labels. The circle will have a certain length (depending on our choice of unit), and the various numbers corresponding to a particular position will differ from each other by multiples of that amount. For example, if we choose our unit so that the circle has length 1 (and why not?), then the origin of our coordinate system will receive the labels 0 and 1, as well as 2, 3, -1 , and all the other positive and negative whole numbers.



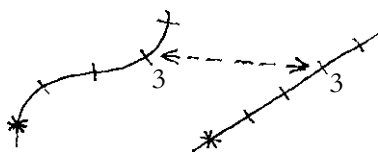
Except for this slight twist of having multiple coordinates, the circle behaves the same as the line. A motion on the circle can be described in the usual way, as a numerical pattern relating the position number to the time on the clock. The simple relationship $p = t$, which describes constant speed motion on a line, also describes constant speed circular motion.

In fact, the same goes for any curve whatsoever; all curves can be coordinatized in this way, so describing motion on one

curve is the same as for another. In other words, all curves are intrinsically the same. Well, actually, that's not quite true; there is the difference between open and closed. But that is the only difference. Structurally, any two open curves are identical, and any two closed curves are also. That is, if your universe were one dimensional and so were mine, we couldn't tell the difference between them. If we both chose reference points and units and set up coordinate systems, then every location in my world would have a corresponding place in yours, and no experiment we could perform could detect the difference—except, of course, for the experiment of going off in one direction and seeing if we ever come back or not. From a classification point of view, there are exactly two one-dimensional geometries. (I'm leaving out the unpleasant possibility of boundary points where space suddenly ends, such as in a line segment with endpoints.)

A geometry, in the modern sense, is a space of some sort endowed with a **metric** (that is, a notion of distance), and two geometries are considered the same if there is a correspondence between them that preserves the distance between points. In other words, the structure-preserving transformations are the isometries.

So if we have two curves, let's say a wiggly one and a straight one, we can coordinatize each in whatever way we please, and this sets up an isometry between them. Namely, we just correspond points with the same numerical label.



But if a curved line and a straight one are geometrically identical, then what does *curved* mean? What are we detecting about these two shapes when we look at one and call it straight and the other not?

Intrinsically—that is, from the inside—the experiences of people living in these two spaces are absolutely identical. What is different about them is extrinsic: the view from the outside. The two curves are the same in and of themselves; the difference is the way in which they have been embedded in the plane. Differences between the two curves *can* be detected by two-dimensional creatures living in the plane. For instance, the distance between two points can be measured on one of the curves (and I mean the distance between them in the plane) and compared with the corresponding measurement on the other curve.



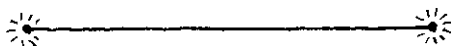
Now these measurements do *not* come out the same. The point is that one-dimensional creatures use rulers that exist inside their universe, so they can't measure how their world might be bending with respect to some larger ambient universe.

So what *curved* means is that one space has been stuck inside another in such a way that its intrinsic metric disagrees with that of the larger outside space. A straight line in the plane is straight because whether you measure it from the inside or the outside, you get the same distances (assuming, of course, that the one-dimensional creatures are using the same measuring unit as their two-dimensional brethren).

So curved and straight are *relative* notions. A one-dimensional space is neither straight nor curved until you inject it into a higher-dimensional space. Then the two metrics can be compared. It's not the curve itself that is curved so much as it is the manner in which it is embedded.

In general, whenever one space sits inside another—whether it is a curve in a plane, an arc lying in a sphere, or a torus floating in space—the larger, “parent” space induces a metric on the smaller space. Any other metric that this subspace may have intrinsically can then be compared with the one it inherited. If they agree, it means that the smaller space was injected into the larger isometrically—straight, or flat, or whatever you want to call it. Otherwise, it got bent.

This is, of course, the modern viewpoint. Under this interpretation, the circle, as well as every other curve, is intrinsically flat. A nice way to think of a flat circle is to imagine a stick with “magical” endpoints:



The idea is that when you sail off one endpoint, you immediately reappear on the other. In other words, the two endpoints represent the same exact place. The point is, there is no intrinsic difference between this magical space and the customary idea of a circle. What makes a circle circular is the way it is situated in the plane.

Can you design a “magic surface”
representation for a flat cylinder? How
about a flat cone or doughnut?

All of this is a very long-winded way of saying that circular motion is only really circular when there's something else going on to compare it with. The cycloid is a good example. A point is not simply moving along a circle; the circle is rolling along a line. To understand this motion, we need to know what circular motion looks like not from the inside point of view of the circle but from the outside view of the plane.

Every curve can be straightened without metric distortion. Is the same true for surfaces?

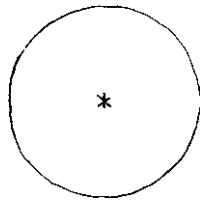
What are the straight lines on a sphere? How about for a cylinder, cone, or torus?

7

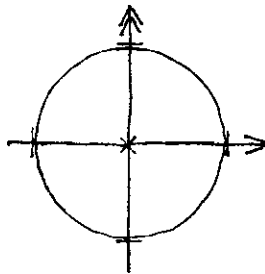
So the right questions to ask are about *how* a circle sits in the plane. In essence, we have two competing coordinate systems: the intrinsic circular one and the one coming from the ambient space of the plane. The question is how these two systems compare.

Of course, there's no such thing as *the* coordinate system for either the circle or the plane. Coordinate systems depend on choices. If we make ugly, unpleasant choices, the systems will relate to each other in an ugly, unpleasant way.

So what would be the nicest choices? We have a circle sitting in a plane. The first thing to do is to choose a reference point in the plane. I can't imagine a nicer, more symmetrical location than the center of the circle.



As for the two directions, we might as well make them perpendicular, and then, of course, the symmetry of the circle makes it pretty irrelevant which two directions we choose. So, let's pick some random direction and call it horizontal and call the other vertical. It is customary to orient these on the page as left to right and down to up, respectively, but that is, of course, entirely up to you. Let's say we do it the usual way. It is also customary to choose the horizontal as the first coordinate. Having oriented our system (or ourselves, whichever way you want to think of it), we need to metrize it by choosing a unit. Since the circle is the only interesting thing in sight, we might as well choose its radius as our unit.

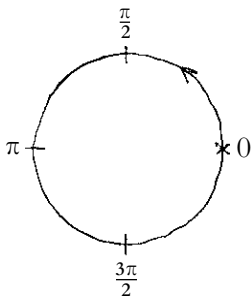


Now our rectangular coordinate system in the plane is all set up. Every point in the plane (including, especially, those points on the circle) can now be given a coordinate label consisting of two numbers. The top of the circle, for instance, would be assigned the pair $(0, 1)$.

The other coordinate system we are interested in is the one

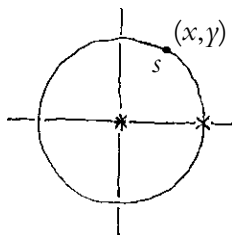
coming from the circle itself. This is a one-dimensional system. Any time a curve sits in a surface, the geometry of the situation will come down to a comparison of a one-dimensional system with a two-dimensional system.

To set up the circular system, we will need to choose a reference point on the circle. I can't say that there's any particularly strong candidate. I suppose we might as well choose one of the four points where the perpendicular axes cross the circle, the classic choice being the right-most point $(1, 0)$. Not that it matters in the least. Then, there is the clockwise versus counterclockwise issue. Which direction will be positive? Again it doesn't matter. Custom dictates making it counterclockwise (i.e., from the first direction toward the second). So we'll start at the right-most point of the circle and lay off units counterclockwise around the circumference. Naturally, we will use the same units as we did for the rectangular system, so we won't have any unnecessary conversions to do. In other words, we're measuring the circle using its own radius as our ruler. Under this system, the total length of the circle is 2π , so for instance, the top of the circle will receive the label $\frac{\pi}{2}$, being one-quarter of the way around. (Of course, it will also receive the labels $\frac{5\pi}{2}$, $-\frac{3\pi}{2}$, and infinitely many others, circles being closed and all).



Now, here's the point. Every location on the circle receives both a circular and a rectangular coordinate label. The top of the circle is at position $\frac{\pi}{2}$ along the circle, meaning the distance along the circle from the starting point is $\frac{\pi}{2}$, whereas its rectangular label is $(0, 1)$. The origin of our circular system, of course, gets the label 0, and (by our choice) it has rectangular coordinates $(1, 0)$. The fundamental question about circles in the plane is how to convert between the two systems. There is absolutely no way to understand something like a rolling ball without being able to go back and forth between rectangular and circular reference systems.

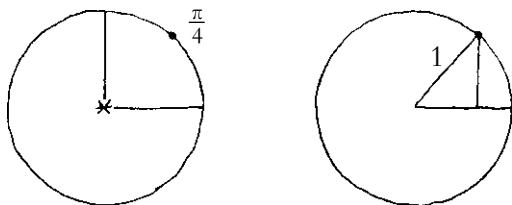
Suppose we have a point somewhere on the circle. Let's call its circular coordinate s . Then the question is how exactly its rectangular coordinates, say x and y , depend on s .



We know, from the way we set it up, that when $s = 0$, then $x = 1$ and $y = 0$. We can even make a little chart of the four corners:

s	x	y
0	1	0
$\frac{\pi}{2}$	0	1
π	-1	0
$\frac{3\pi}{2}$	0	-1

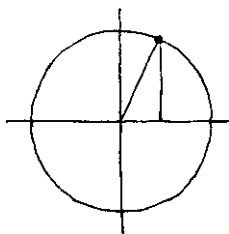
For other points, the correspondence is subtler. Consider, for instance, the point halfway from the origin to the top of the circle. Its circular coordinate is, of course, just $\frac{\pi}{4}$, one-eighth of the way around the full circle. But where is that point in the up-and-down, side-to-side sense?



One way to see it is to make a little triangle. Since the angle of this right triangle is one-eighth of a full turn (or 45 degrees), we know that the triangle is half of a square. The long side of the triangle has length 1, since our unit was chosen as the radius of the circle. So the two short sides must both be $\frac{1}{\sqrt{2}}$ (the diagonal of a square being $\sqrt{2}$ times its side). Thus when $s = \frac{\pi}{4}$, we get $x = \frac{1}{\sqrt{2}}$ and $y = \frac{1}{\sqrt{2}}$.

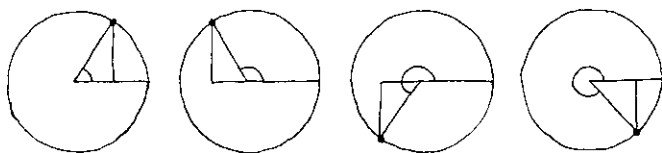
Alternatively, we could reason that since this is a right triangle of hypotenuse 1, its legs are precisely what we called the sine and cosine of the angle, which in this case is one-eighth of a turn.

Generally speaking, this is the best we can do. For a random point on the circle, the only way to talk about its rectangular coordinates is via the sine and cosine of the angle formed by this little right triangle.



Essentially, what's going on is that circular coordinates correspond to length along the circle, which is related directly to the angle made at the center of the circle. The rectangular coordinates refer to the perpendicular lengths formed by this angle, and that is precisely what we have been calling the sine and cosine of the angle. I suppose it's not too surprising that such simple objects as circles and right triangles would have some sort of connection.

There are a number of subtleties and details to work out here. The first is that the angles can get too big to fit in a right triangle.



As our point swings around, the angle that it makes with the horizontal increases from nothing to a full turn. When the angle, let's call it A , is small, then the horizontal and vertical coordinates of the point are simply $\cos A$ and $\sin A$, respectively. When the point passes the top of the circle (the $\frac{\pi}{2}$ mark), the corresponding right triangle is now on the other side of the circle, and its angle is not A anymore but the angle next to A . This is exactly the same thing that happened to us when we were measuring triangles. We ended up deciding that it would be most convenient to *define* the cosine of an angle A in this range to be the exact negative of the cosine of the angle next to it. This is lucky for us, because that is precisely what the horizontal coordinate of a point in this range should be. It's no coincidence that the two problems—measuring the distance

between two sticks at an angle and determining the location of a point on a circle—should require the same choice of extension of sine and cosine. We build mathematical objects to be beautiful, and beautiful things, like crystals, have tremendous consistency: they follow patterns, and they don't like to have those patterns disrupted.

**What are the rectangular coordinates of
the point with circular coordinate $\frac{3\pi}{4}$?**

Similarly, the nicest extension of sine to angles in this range (between one-quarter and one-half of a turn) is to have $\sin A$ be the *same* as the sine of the angle next to A , not the negative of it. This choice allows the law of sines to remain valid for large angles, as well as giving us the right vertical coordinate for points in this quarter of the circle.

Really, what's going on here is this: we have two problems, triangle measurement and the comparison of circular and rectangular coordinate systems. Well, they turn out to be the same. More precisely, the sine and cosine of an angle are a special case of the circle problem—the case of smallish angles. So we have an old definition of sine and cosine in terms of right triangle proportions. Now we're forging a new definition, and lucky for us, it's not conflicting with the old one. This is a recurring theme throughout mathematics—the extension of a naïve concept to a wider and more general context.

So the idea is to give a meaning to the sine and cosine of any angle whatsoever. If the angle is small (between zero and one-quarter of a turn) then we know what sine and cosine mean, namely the sides of the corresponding right triangle of hypot-

enuse 1. Between one-quarter and one-half of a turn, we look at the outside turn and its sine and cosine. Then the sine of our angle is the same as the sine of the outside angle, and the cosine is the negative. In both cases, the sine and cosine of our angle are just the rectangular coordinates of the corresponding point on the circle. Naturally, the plan is to define the sine and cosine of any angle in this way. So here we go: the cosine of an angle is the horizontal coordinate of the point on the circle described by that angle, and the sine is the vertical coordinate.

**Make a sine and cosine table for all multiples
of one-twelfth of a turn (30 degrees).**

**If two angles add up to a full turn, what is the
relationship between their sines and cosines?
What if they add up to a half turn?**

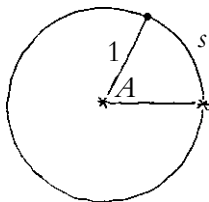
8

The situation is now this: for a point on the circle, with circular coordinate s and rectangular coordinates x and y , we have

$$\begin{aligned}x &= \cos A, \\ y &= \sin A,\end{aligned}$$

where A is the angle formed at the center of the circle by the point, counterclockwise from the horizontal. (Of course, in some sense this is totally content free; it's really just a restatement of our

failure to measure triangles algebraically.) In any case, the whole issue now comes down to how this angle A depends on s .



The number s represents a length—the length around the circle to our point—and A is the corresponding angle. Traditionally, the relationship between lengths and angles is somewhat strained. There is a lot of mistrust and resentment, and also sines and cosines. But that’s really about angles and *straight* lengths. The relationship between angles and *circular* lengths is a whole different story. In fact, it’s about as simple as can be: they’re proportional. A full turn corresponds to a complete circumference length, a half turn to half a circumference, and so on. So depending on your choice of length and angle units, the two will just be off by some factor. In particular, if we measure length using the radius and angle using full turns (as we have been), then the relationship is simply $s = 2\pi A$.

Right away we could end this discussion by saying that the conversion between circular and rectangular systems is simply this:

$$\begin{aligned}x &= \cos(s/2\pi), \\ y &= \sin(s/2\pi).\end{aligned}$$

And that’s that. If you have the circular coordinate s , all you have to do is scale it down by 2π to convert it to an angle

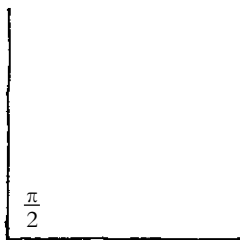
measured in full turns, then convert the angle back to a pair of lengths x and y using sine and cosine. The positive and negative signs are taken care of by our clever new definition of sine and cosine. And so we get the rectangular coordinates.

The only thing that is a little obnoxious about this is that we have to convert our arc length to an angle and then convert the angle back into a pair of lengths. This is happening for two reasons. One is our choice of units—we're measuring angles in full turns. Of course, if we measured them in degrees, it would be even worse; the conversion from arc length to angle measurement would be $A = \frac{360}{2\pi}s$. The question is, what are the best units for angle measurement? Should a full turn be thought of as 360 degrees, or one full turn, or what? Of course, it doesn't really matter; it's just a question of convenience. But convenience is a nice thing anyway. My feeling is that for polygon measurement (e.g., when we were looking for possible tiling patterns), measuring angles as portions of a full turn is simple and natural. Now that we're comparing circular and rectangular coordinate systems, though, it seems a bit clunky. I don't really like that 2π conversion factor.

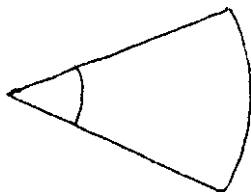
The other thing that's getting in our way is our interpretation of what it is that sine and cosine do for a living. We've been thinking all along, and naturally enough, that they convert angles into lengths—or more precisely, ratios of lengths. This necessarily means that we have to go through angles any time we want to measure circles or circular motions, and that just doesn't seem right.

So here's my proposal. It's rather modern, and it may seem strange and arbitrary, but bear with me. First of all, we're going to choose a new way of measuring angles. A full turn will not

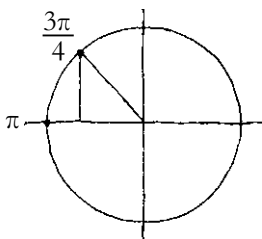
be 360 something-or-others, nor will it be our unit. A full turn will be 2π . That is, we're going to use the circular coordinate system itself to measure angles. So a right angle receives a measurement of $\frac{\pi}{2}$.



The advantage of this is that there is now no conversion between angles and arc lengths. Arc length *is* angle. More precisely, we're measuring angles by the ratio of arc length to radius.



So angles are really just length proportions. Further, let's stop thinking of sine and cosine as operating on angles and instead think of them more abstractly as converting numbers into numbers. We can define them via our understanding of the circular and rectangular coordinate systems: the sine and cosine of a number are just the rectangular coordinates of the point on the circle with that number as its circular coordinate. For example, the cosine of π is -1 , and the sine of $\frac{3\pi}{4}$ is $\frac{1}{\sqrt{2}}$.



Not that anything is really any different from before, just the units and the attitude. The nice thing is that we can eliminate angles from the situation and simply say that if s is the circular coordinate of a point on our circle, then its rectangular coordinates are

$$\begin{aligned}x &= \cos s, \\y &= \sin s.\end{aligned}$$

And this makes complete sense for any number s whatsoever. Of course, this is really just a restatement of our new definition of sine and cosine. I suppose the real content of this is that there is no disagreement with any of our prior interpretations. What sine and cosine do is convert circular measurements into rectangular ones. They are the abstract mathematical version of “putting a round peg into a square hole.”

Show that $\cos(-x) = \cos(x)$ and $\sin(-x) = -\sin(x)$.

How do the sine and cosine of $a + b$ depend on the sine and cosine of a and b themselves?

Make a graph showing how the sine and cosine of a number varies depending on the number. What do you notice?

9

The simplest nonlinear motion I can think of is a point moving in a circular path at a constant speed, usually referred to as **uniform circular motion**. To describe such a motion, we would need, as always, to choose coordinate systems for time and space—to build a clock and a map suitable for the situation.

Naturally, the simplest choices would be a length unit equal to the radius of the circle and a time unit chosen so that the speed of the point was equal to 1 (in other words, to choose our unit of time to be the amount of time it takes the point to travel an arc length equal to 1 length unit). With these choices, the description of the motion is as simple as can be: if s is the circular coordinate and t is the time, then the motion is given simply by the pattern $s = t$.

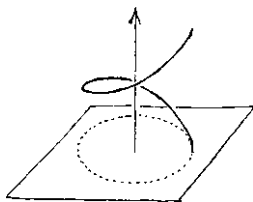
Of course, if we are concerned with the relationship between our point and some external object, say another point or line in the plane, we would prefer a description of the motion from the plane's point of view. We could do this by choosing a rectangular system for the plane, with coordinates x and y say, and describe the motion of the point in those terms. The simplest setup would be what we had before, with the center of the circle as our origin, and so on. If we orient our system so that the motion of the point is counterclockwise and its initial position (at time $t = 0$) is the customary starting point $x = 1$, $y = 0$, then the motion can be described by the set of relations $s = t$, $x = \cos s$, $y = \sin s$. More simply, we could just write:

$$\begin{aligned}x &= \cos t, \\y &= \sin t.\end{aligned}$$

This provides a complete description of the pattern of motion. At each time t , we get an explicit (albeit transcendental) specification of the precise location of our moving point.

What if our point were moving clockwise?

The space-time view of this motion is particularly interesting. Since the motion is taking place in a plane, the corresponding space-time is three-dimensional, with coordinates x , y , and t . The curve in space-time corresponding to uniform circular motion is a helix.



The idea is that as our point travels around the circle, it is lifted up in the time direction. So a motion around a circle in the plane is the same as a static helix in three-dimensional space-time.

The amazing thing about this is that since a helix appears as the space-time representation of circular motion, we can use our motion to describe the ordinary three-dimensional (non-space-time) sort of helix. That is, if we have a helix in space, we can describe it as the set of points (x, y, z) , where

$$\begin{aligned}x &= \cos z, \\y &= \sin z.\end{aligned}$$

The point being that z doesn't care whether we think of it as a space coordinate or a time coordinate: it's just a number! The modern philosophy is to make everything—shapes, motions, angles, speeds, what have you—numbers, so that we have maximum flexibility. In particular, any space-time picture can be viewed as a space-only picture if we so desire.

**How could we describe a constant
speed motion along a helix?**

So not only can we set up coordinate systems and use relationships between the various coordinates to describe motions, we can do the same thing to describe static objects. For instance, a sphere can be thought of as the set of points (x, y, z) in a three-dimensional rectangular coordinate system satisfying the relationship $x^2 + y^2 + z^2 = 1$. One can then make measurements and deduce properties of the sphere from this numerical description.

**Why does the relationship
 $x^2 + y^2 + z^2 = 1$ describe a sphere?**

**Can you construct a coordinate
representation of a cone?**

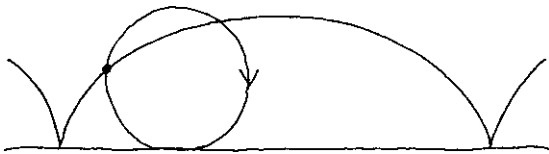
This idea of representing geometric objects via coordinate descriptions provides us with a very rich and flexible language for describing shapes, and the connections between algebra and geometry that are revealed by this point of view are among the most fascinating and beautiful results in all of mathematics.

Can you design equations for uniform circular motion with radius r and speed v ?

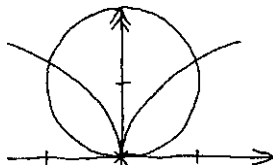
Show that every line in the plane has a coordinate description of the form $Ax + By = C$.

10

Now let's try to describe the cycloid. This curve is traced out by a point on a rolling circle, so what we need is a precise description of this motion. Where exactly is the moving point at any given moment? Of course, the first job is to design an appropriate coordinate system.

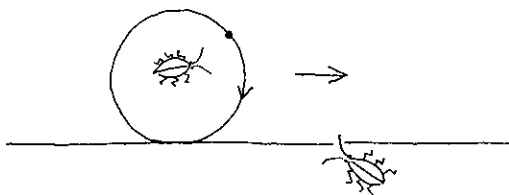


I like to choose the radius of the circle as my spatial unit, the line it's rolling on as my first direction (oriented in the direction that the disk is rolling), and my origin (in both space and time) to be a moment when the point is touching the line; that is, when the point has rolled completely underneath the circle.



The only thing left is to choose the time unit. This is tantamount to choosing the *speed* of the rolling. Of course, it makes no real difference; the same curve will be traced out whether it rolls quickly or slowly. So we may as well choose our units so that the speed is pretty. Let's say that the speed is 1. By that I mean that if you look at the disk in isolation, independent of the line it's rolling on, it rotates so that the moving point has constant unit speed along the circle.

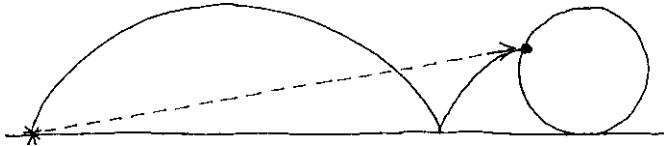
Actually, this idea of looking at the motion from different points of view is extremely valuable. It usually goes by the name of *relativity*. A bug sitting somewhere in the plane would see this motion as a point on a disk rolling on a fixed line, whereas another bug who was riding on the disk (sitting at the center, let's say) would simply see the point rotating around it, with the line speeding by.



The point is, neither is right or wrong; they're both right from their own point of view. The important thing is for them to be able to communicate with each other. That is one thing that makes the vector approach to motion representation very convenient: since positions are already described in terms of shifts, it's very easy to adjust to someone else's perspective—we just add on another shift!

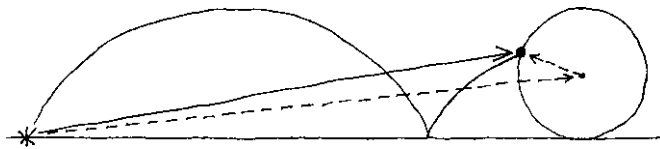
Let me be as clear about this as I can (as if up until now I've been purposely vague). Let's look at the vector representation

of our moving point; that is, the shift that takes us from the origin to the point itself.



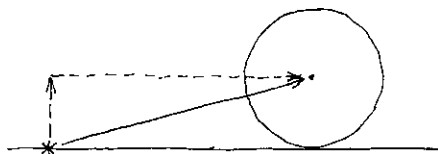
Of course, this vector is changing all the time and in a complicated way. That's the whole point; the cycloid motion is not so simple, and this vector is getting longer and rotating up and down in a subtle way that we are trying to describe precisely.

The idea of relativity is to try to find another perspective from which the motion is simpler, for instance from the center of the circle.



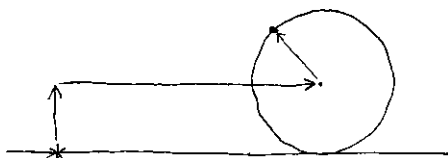
Now we can view our vector (which describes the cycloid motion) as being a sum of two simpler vectors, namely the one from the origin to the center of the circle and the radial vector from there to the point itself. The motion of the center is simple because there is no rotation, and the radial vector is simple because it is purely rotation.

This is an extremely useful technique: a clever change of perspective breaks down a complex motion into a sum of simpler motions. We can go even further with this. Instead of watching the center from the point of view of the origin, it's a little nicer to watch it from a position one unit higher.



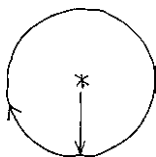
This means we're breaking the vector to the center into a sum of two pieces—a vector up one unit and a horizontal vector from that position to the center of the circle. These are both simpler motions, since their directions don't change. In fact, the first vector doesn't change at all.

So we have broken the relatively complicated motion of a point on a rolling disk into three much simpler motions: a constant vector to get us up to the level of the center of the disk, a purely horizontal vector from there to the center itself, and then finally the vector from the center to the rotating point.



Of course, we still need to give precise descriptions of how each of these vectors depend on time. Let's have t denote the time on our clock, u_1 be the unit vector in the (positive) horizontal direction, and u_2 the vertical unit vector.

Let's start with the vector from the center of the circle to the moving point. This is simply uniform circular motion, where we start at the bottom of the circle (that was our choice of where the point is at time $t = 0$) and rotate clockwise (that was our choice of which direction the disk travels).



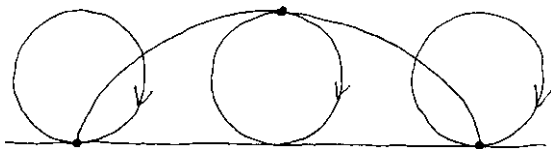
If we were starting from the right-hand side of the circle and traveling counterclockwise, we would have exactly the situation we looked at before with circular motion, namely the coordinates of the point would be $\cos t$ and $\sin t$. That is, the vector from the center of the circle to the point would simply be $(\cos t)u_1 + (\sin t)u_2$. Since we're starting at the bottom and moving clockwise, this needs to be modified to

$$(-\sin t)u_1 + (-\cos t)u_2.$$

Maybe the best way to see this is to think about the coordinates separately. The horizontal position needs to begin at 0, decrease to -1 , and then move back to 0, up to 1, and back to 0 again. That's exactly what $\sin t$ does—only negated. So the horizontal coordinate is moving in the $-\sin t$ pattern; similarly for the vertical, only with cosine. Alternatively, our point is moving in the customary way from the perspective of someone standing on the other side of the plane, looking sideways. This has the effect of reversing the coordinates and negating them—more relativity, I suppose. In any case, we have a precise description of the motion of the point from the point of view of the center.

**What are the various coordinate descriptions
for a point moving uniformly on a circle,
starting at any of the four corners
and moving in either direction?**

Next we need to describe the motion of the center itself. The vertical part of it is easy; it's just u_2 . It's the horizontal part that's going to be a bit tricky. Probably the simplest way to measure the horizontal motion is to let the circle make one complete rotation.



Now because the disk is rolling (that is, it's not slipping or skidding), the full circumference is laid down horizontally. In other words, the distance along the road that the disk travels is one circumference. This means that in the amount of time it takes the point to make one full rotation (and thus travel a distance of one circumference), the center moves the exact same distance horizontally.

This means the horizontal speed of the center is the same as the speed of the rotating point along the circle. But we chose our time unit so that this speed is 1. Thus the horizontal speed of the center is also 1. Is that at all understandable? This is by far the hardest part of the problem—interpreting what “rolling” means exactly.

So that's pretty. The center travels horizontally at exactly the same speed as the point rotates. Algebraically, what this means is that the horizontal vector is simply tu_1 . That is, it points in the positive horizontal direction, and its length is always equal to t , since it starts off at zero and grows at a constant unit rate.

Putting everything together, we get that the position vector p of our point is given by

$$p = (-\sin t)u_1 + (-\cos t)u_2 + u_2 + tu_1.$$

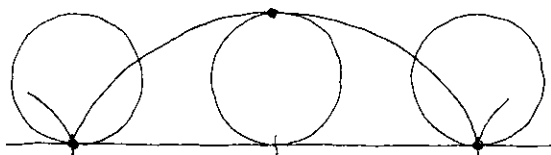
This is as nice an example of mechanical relativity as you will ever see. Our complicated motion is broken into a sum of simpler motions corresponding to different points of view.

If you prefer, we can rewrite this description in terms of coordinates. The amount of u_1 that p contains is $t - \sin t$, and the amount of u_2 is $1 - \cos t$, so we could write our motion as

$$\begin{aligned}x &= t - \sin t, \\y &= 1 - \cos t,\end{aligned}$$

where as usual, x denotes the horizontal and y the vertical coordinate of the point at any time t . This is a fairly nice description, given how complex the motion appears to be.

Let's test this out a bit. When $t = 0$, this says that $x = 0$ and $y = 0$. So that's good. It means that the point starts out at the origin, as planned. When $t = 2\pi$, we get $x = 2\pi$, $y = 0$, which also agrees with what we decided before—that the disk travels a distance of 2π after one full rotation.



We can also see that halfway along, when $t = \pi$, we get $x = \pi - \sin \pi = \pi$, and $y = 1 - \cos \pi = 2$, again in agreement

with what we expect. So that's nice; it means we probably didn't make any gross blunders.

We now have a precise description of the motion of a point on a rolling disk.

If two points in the plane move at constant speeds on a collision course, what will the motion of each one look like from the point of view of the other?

Two points move on a line with constant velocities. What is the view from their (moving) midpoint?

Can you construct descriptions for the motions that determine a hypocycloid or an epicycloid? How about a spirograph?

11

So we have solved the description problem for the cycloid. We can now say exactly where the moving point is at all times; namely, we have the precise description

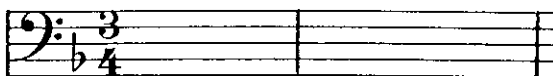
$$\begin{aligned}x &= t - \sin t, \\ y &= 1 - \cos t.\end{aligned}$$

What we have here is an *encoding* of the cycloid. That is, a symbolic representation of the shape information. We have

the idea. Has the romance of music been lost? No, but a cultural barrier has been erected. Those who cannot read music are to some extent excluded. Whenever a symbolic encoding system is invented, there is a literacy issue. It's great to be able to communicate your ideas precisely in a compact form; the problem is you then need to learn how to *read*.

We are now in this exact situation with respect to shape and motion. Descartes' idea of using coordinate systems to describe locations numerically allows us to represent even very complex motions in a precise and succinct form (namely, as a set of equations). The advantages are that we can move information around easily on paper, and we don't have to make any difficult drawings. The disadvantage is that we have to be careful. Notice how fragile musical and algebraic notations are; one small change in a symbol could ruin the whole thing! But the biggest problem is that we have to learn to become fluent in the symbolic language. The point is not to be mere scribes who can operate the symbols and make translations but to be composers who can use the language to create and investigate beautiful things. The invention of a symbolic language creates a culture, be it musical, literary, or mathematical, and there is romance in that as well.

As a matter of fact, we can carry the musical analogy even further. After all, what is a piece of music? Isn't it a motion of sounds? Let's take the notes of the piano keyboard as our pitch space; that is, the space in which our point/note is going to wander. The musical staff is nothing more than a map of this space; the lines and spaces are the possible locations, and they have coordinates such as middle C, high B \flat , and so on.



Like any map, the musical staff has an orientation (high notes toward the top, low notes toward the bottom) and a unit (one step). The various clefs and key signatures determine the origin of the system. A piece of music can then be graphed in pitch-time. The horizontal direction measures time (the unit is the beat, the origin being the start of the piece). The little black dots denote the musical “events.” (I suppose we could consider loudness as another dimension in our space of note points, so a piece of sheet music is really a graph of a two-dimensional motion.)

So both composers and mathematicians construct coordinate systems appropriate to their description problem and use a symbolic language to encode the patterns. And just as a good violinist can glance at a line of sheet music and hear the tune in her head, an experienced geometer can see and feel the shapes and motions described by a system of equations (at least if they are reasonably simple).

Can you construct a coordinate
description of a spiral motion?

12

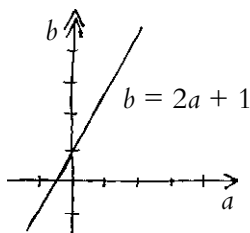
What we have been talking about is *representation*. Whenever one thing is used to represent another, there are always interesting philosophical consequences. For one thing, there is the

question of exactly who is representing whom. Is the sheet music a transcription of the sound, or is the performance an enactment of the sheet music? Or is it that both the writing and the playing are representations of the same abstract musical idea?

In our case, we have a geometric object (a shape or motion) and its representation by a set of equations. But is it the shape that is the real thing and the equations only a convenient algebraic encoding, or could we just as easily view the equations (i.e., the number pattern) as the true object of interest and the shape or motion as a mere visual or mechanical representation of it?

Of course, we have known all along that pictures themselves are not very useful description tools (their value is mostly psychological) and that when we speak of a circle we are not really talking about a picture but rather a linguistic pattern: the collection of points at a certain distance from a fixed center. What Descartes realized is that any such verbal description that is precise enough to specify a shape or motion exactly can be replaced by a numerical pattern and represented as a set of equations. For instance, the circle can be encoded (with our usual choice of coordinates) as $x^2 + y^2 = 1$.

On the other hand, any equation or set of equations involving any number of varying numerical quantities (usually called **variables**) can be interpreted as describing a shape or motion. That is, every numerical relationship has a certain “look” to it. The relation $b = 2a + 1$ (which encodes the purely abstract numerical information that the variable b is always one more than twice the value of a) can, if we wish, be thought of as a line in two-dimensional space:



Or, if we prefer, we could view this as a space-time picture and imagine that a is time and b is position. In this way, the relationship $b = 2a + 1$ becomes the record of a constant speed motion beginning at position 1 and moving forward at a rate of 2.

Ultimately, of course, it is neither shapes, motions, nor equations that are the real object of study but *patterns*. If you choose to represent your pattern geometrically or algebraically, that's fine. Either way, it is the abstract pattern relationship that you are really talking about.

What happens when we view shapes as mere visual representations of number patterns? Well, for one thing we get a lot of new shapes! This so-called **coordinate geometry** (initiated by the publication of Descartes' *La Géométrie* in 1637) not only provides a convenient solution to the description problem—providing us with a uniform linguistic framework in which to describe geometric patterns—but at the same time gives geometers an entirely new way to construct shapes and motions. We now have almost unlimited descriptive ability.

The question then becomes which equations correspond to which shapes? (Of course, I'm including motions here, since we can always think of a motion as a curve in space-time.) What we need is a “dictionary” to help us translate between geometric and algebraic descriptions. We could start with:

dimension of space \leftrightarrow number of variables

shape or motion \leftrightarrow relations among variables

So that if, for example, we were interested in the equation $x^2 + y^2 = z^2$, we could “see” it as the set of points (x, y, z) in three-dimensional space whose coordinates satisfy that relationship. We would then be able to tell that the point $(3, 4, 5)$ was included as part of this shape and that the point $(1, 2, 3)$ was not. Whatever shape this is, it has been completely and precisely determined.

What shape is it?

I like to think of a set of variables as creating an ambient space and the equations as carving out the shape. In the case of two variables, the ambient space would be two-dimensional and the relation between the variables would carve out a curve determined by that relationship. In particular, we can add to our dictionary:

line in the plane $\leftrightarrow Ax + By = C$

circle $\leftrightarrow x^2 + y^2 = 1$

There is a subtlety here, actually. Whenever we go back and forth between shapes and equations, there is always something lurking in the background—namely, the coordinate system. For instance, if you (for whatever reason) were to take the origin of your coordinate system to be a point other than the center of the circle and your unit to be something other than its radius, then $x^2 + y^2 = 1$ would no longer be the correct description.

**What is the equation of a circle of radius r
centered at the point (a, b) ?**

One of the most beautiful discoveries of this period (the early 1600s) was that simple equations correspond to simple shapes. The simplest numerical relationships are those that involve no multiplications among the variables, only addition and scaling by constants. In two dimensions, these have the form $Ax + By = C$ and correspond to lines. (For this reason such equations are often called *linear*.) In three dimensions we would have another variable, $Ax + By + Cz = D$. The picture (or graph) is now a plane in space.

**Why does $Ax + By + Cz = D$
describe a plane?**

More complicated equations would involve products among the variables, the simplest examples being equations with only pair-wise products such as x^2 or xy . The circle $x^2 + y^2 = 1$ would be one example of such a *degree 2* equation. Another would be the squaring relation $y = x^2$. The graph of this equation turns out to be a parabola.

**Why does $y = x^2$ carve out a parabola?
Where are its focal point and focal line?**

In fact, it turns out that the graph of *any* degree 2 equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey = F$$

is always a conic section. That is, the class of curves we have been calling conic sections corresponds exactly to the set of degree 2 equations in two variables. In other words, the simplest nonlinear curves correspond to the simplest nonlinear equations. So we have another entry in our dictionary:

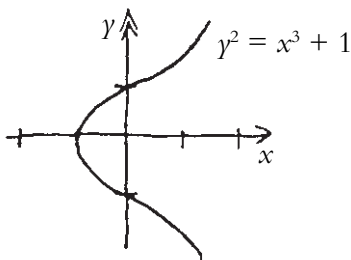
conic sections \leftrightarrow degree 2 equations

There is a small technicality here, actually. Some degree 2 equations (e.g., $x^2 - 4y^2 = 0$) have graphs that are not in fact conic sections. They instead form so-called *degenerate* conics (a pair of crossed lines in this case). So there is a bit of a subtlety to the correspondence.

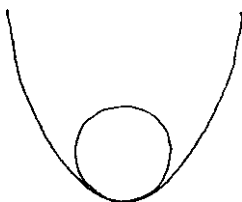
How do the coefficients A, B, \dots, F
determine which type of conic section
(ellipse, parabola, or hyperbola) is described?

When is a degree 2 equation degenerate, and
how many types of degeneracy are there?

What about equations of degree 3, such as $y^2 = x^3 + 1$? What shape does this make?



It turns out that this one is new. It's not a circle or a conic or a cycloid or a spiral, or anything else we have a name for. It is "the graph of $y^2 = x^3 + 1$," and that's the simplest description we're ever going to have. This is what I meant by *a lot* of new shapes. Any numerical relationship you want to write down will carve out some sort of shape, and all but the simplest few will be absolutely brand-new. This is the expressive power of algebra—the moment we put numbers on a line to make a map, we get this amazing wealth of new shapes.



**What is the largest circle that can
sit at the bottom of a parabola?**

13

Having solved the description problem for motions (at least in the sense of having a universal language in which to describe them), it is clearly time to start measuring.

What is there to measure about a motion? A coordinate description tells us where it is when. Questions like, "where was it at such and such a time?" or "what time was it when it was here?" can be answered directly from the equations describing the motion. This would come down to some scrambling and

unscrambling of our numerical relationships—in other words, doing some algebra. This could conceivably be quite unpleasant in practice, but it presents no particularly deep philosophical problems.

Far more interesting are the questions: How fast was it going? How far did it travel? These are clearly related, since how far you go depends very much on how fast you go. So our first really interesting problem about motion is the measurement of *speed*.

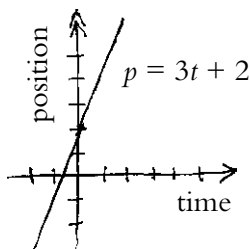
Suppose we have a moving point described by a set of coordinate equations. How can we determine its speed? Since the entire motion is completely and precisely specified by the numerical relationships (i.e., the way the coordinates depend on time), the equations must somehow hold the speed information within them. How do we get that information out?

Let's start with the simplest possible situation: uniform linear motion in one dimension. (I suppose the *simplest* situation would be no motion at all, but that is not terribly exciting.) Here our motion can be described by a simple equation like $p = 3t + 2$ (where as usual p is the position number and t is the time coordinate). In this case, it is particularly easy to read off the speed information: the point is traveling (forward) at a speed of 3 (space units per time unit). In other words, the speed is simply the coefficient of time—the factor by which t is being multiplied. Thus for any uniform linear motion $p = At + B$, the initial position is B and the speed is A .

**What if the coefficient of
time is negative or zero?**

Of course, all of this is very dependent on our coordinate choices. If we reverse the orientation, rescale our units, or shift the origin, we can always rewrite any uniform linear motion as $p = t$ so that the speed will be 1. (Alternatively, we can recalibrate our clock so that the motion has unit speed.) It's the same problem we've always had with measurement: it's all relative. There really is no such thing as absolute speed, only speeds relative to other speeds. When we say that the motion $p = 3t + 2$ has a speed of 3, both the description of the motion and the value of the speed depend on the choice of units for the coordinate system. (A more abstract—and therefore simpler—point of view would be to forget about time and space altogether and simply view $p = 3t + 2$ as a relationship between two numerical variables p and t . There are no units, only numbers. Then we can say that p is moving three times as fast as t is.) In any case, no matter how you want to think of it, the point is that for uniform linear motion, the speed is easy to get out of the equation—it's just sitting there!

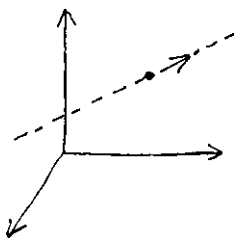
Speaking of alternate viewpoints, what is the space-time picture?



The space-time view of a uniform linear motion is a straight line. The speed appears as the *slantedness* of this line. That is, if the speed is 3, it means that the line is slanted at a 3:1 ratio

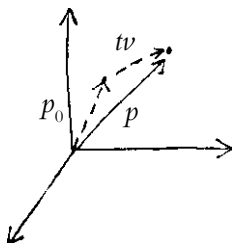
(assuming we represent time and space units equally). So at least for uniform motion in one dimension, the situation is pretty simple. The speed is obtained easily from the equation and has a natural geometric representation (as slantedness) in space-time. What happens for more general motions?

One complication is that the point might be moving around in a higher-dimensional space. Suppose we have a point moving in three dimensions, again at constant speed in some fixed direction.



Of course, if this were the only object of interest, we could simply take the path of the point as our universe and view the motion as being one dimensional. But in general (which is the nicest way to work), we might require a three-dimensional ambient space. There may be other moving points in the picture, for instance.

What are the equations of such a motion? The simplest way to think about it is to use a vector description.



As usual, let's have p denote the position vector of our moving point. So p depends on the time t in some way. As before, we can break this vector (which grows and shrinks and turns in a subtle way as time goes by) into a sum of simpler pieces. The first piece is the initial position vector; that is, the vector pointing to the location of our point at time $t = 0$. This is often written as p_0 to denote that it is the value of p at time 0. The other piece is the vector from the initial position to the current position. Notice that this vector points in the direction that our point is heading, and since the motion has constant speed, its length grows at a constant rate. This means that it must have the form tv for some fixed vector v . Putting the pieces together, we get that every uniform linear motion in space must have the form

$$p = p_0 + tv.$$

At time $t = 0$, this says that $p = p_0$, the initial position. As time goes by, the position changes, so that every second (or whatever you want to call your time unit) the position shifts by the vector v . This means that the vector v not only holds the heading information but also the speed. In fact, the speed is simply the *length* of v , since that is how much distance is traveled every second. So we see that in higher dimensions, the speed and direction are most nicely considered together as a vector. This vector is called the **velocity** of the motion. (In the one-dimensional case, the velocity is just a single number; its *sign* then carries the heading information.) Notice, as in the one-dimensional setting, the velocity is easily read off from the equation as the (vector) coefficient of time.

If we wish, we can always rewrite any vector description as a set of equations in the coordinates, for example:

$$x = 3t + 2,$$

$$y = 2t - 1,$$

$$z = -t.$$

This would correspond to the vector description $p = p_0 + tv$ with $p_0 = 2u_1 - u_2$, or $(2, -1, 0)$, and $v = 3u_1 + 2u_2 - u_3$, or $(3, 2, -1)$. More succinctly, we could write

$$p = (2, -1, 0) + t(3, 2, -1).$$

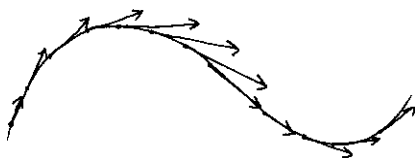
Either way, we get a description of a linear motion in three-dimensional space with velocity vector $(3, 2, -1)$. The speed of this motion is simply the length of this vector, which (using the Pythagorean relation) we find to be $\sqrt{14}$.

So the dimension issue really isn't a problem; we simply move from one equation to several (or what is the same, from numbers to vectors) and treat speed and heading simultaneously as a velocity vector. Of course, there is nothing special about three dimensions; the same idea works for motion in any dimension whatsoever.

Suppose two motions in space are given
by equations $p = p_0 + tv$ and $q = q_0 + tw$.
What conditions on the vectors p_0 , q_0 , v ,
and w will ensure collision?

The real problem is that most motions aren't uniform. In general, a moving point does not keep a steady velocity; it speeds up and slows down and changes its direction constantly. In other words, the velocity vector itself depends on time.

The usual way to picture it is to imagine the velocity vector as an arrow situated at each point along the path:



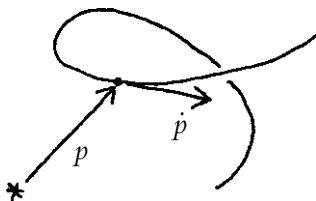
Here we have a motion in the plane, and the velocity arrows show the point speeding up and then slowing down again. Notice that since the velocity vector always points in the direction of motion, these arrows will always be tangent to the path. I like to think of our point as a tiny moving car equipped with a speedometer and a compass. At every moment they together indicate the velocity (e.g., northwest at 40 mph).

So our fundamental problem is this: given a motion (that is, a description of how the position vector varies with time), to determine its velocity (also a vector varying with time). This is what the measurement of motion comes down to; transforming one vector equation (position) into another (velocity).

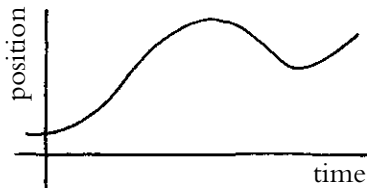
**What is the velocity vector for
uniform circular motion?**

14

We now know what we want to measure, but how do we go about measuring it? We have a moving position vector p , and we want to determine the corresponding velocity vector (usually denoted by \dot{p}). For example, if $p = 2t - 1$ is a one-dimensional motion, then $\dot{p} = 2$ is its (constant) velocity. In general, of course, things are not so simple. The position vector moves around in a complicated way, and it's not at all obvious how we are going to use its description to obtain the velocity information.



Let's start with the one-dimensional situation and imagine a point moving along a line in some complicated way. The space-time picture might look something like this:



We saw before that for constant speed motion the velocity could be viewed as the slantedness of the space-time curve (which was, of course, a straight line). It was Isaac Newton's insightful observation that this remains true for *any* motion. More precisely, Newton recognized the steepness of the tangent line at a point

of the space-time curve to be a geometric representation of the velocity at that precise instant. What a beautiful connection between shape and motion! The seventeenth-century problem of velocity is *the same* as the classical Greek problem of finding the tangent to a plane curve.

If you like, you can imagine that each point on the space-time curve carries with it its tangent line. As the moving point speeds up, the tangent line gets steeper, and as it slows down, the tangent line flattens out. If the point starts backing up, the tangent line slants down. Notice that at the precise moment that the point reverses direction, its velocity is exactly zero. The tangent line is horizontal!



What this means is at that *precise instant* the point is neither traveling forward nor backward. When you throw a ball in the air, it goes up and then comes down (so they say), but there is a split second there where it “hangs.” (Of course, we are concerned with imaginary idealized motions. What *really* happens with a ball is anyone’s guess!)

**Can the velocity be zero without the
point stopping or changing direction?**

Reaction to Newton’s idea varies. Some feel it to be patently obvious that the slantedness of the tangent line is the same as

the velocity, while to others it makes absolutely no sense at all. In fact, some question the validity of instantaneous velocity in the first place. How can a moving object have a speed at a precise instant? If you stop time, doesn't speed have no meaning? Of course, we're not really stopping time, we're *selecting* a time. (I'm sure you can well imagine the ensuing philosophical and religious quarrels, the most famous being Bishop Berkeley's treatise, *The Analyst*, addressed to Newton as "an infidel mathematician.")

Probably the simplest approach is to take the idea of instantaneous velocity (the speed and heading at a precise instant) as an intuitively clear notion (much like the length of a curve) and use the tangent line interpretation as a means of measuring it. We could even answer any philosophical qualms by simply *defining* instantaneous velocity as the slantedness of the tangent line.

Newton's idea allows us to reinterpret our problem geometrically: How can we measure the slantedness of a given curve at a given point?



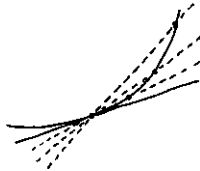
More precisely, suppose we have an equation of the form $p =$ something involving t . This determines a space-time curve, and if we select a certain moment t , we can ask how slanted the tangent line is at that point. How do we get this information out of the equation? In what way does the geometric idea of "tangent line to a curve" get translated into the language

of variables and equations? This is the problem that Newton solved.

The idea is to use the method of exhaustion. We will get at the true tangent line using an infinite sequence of approximations. Specifically, we can approximate the tangent line at a point on the curve by choosing a nearby point and connecting the dots.



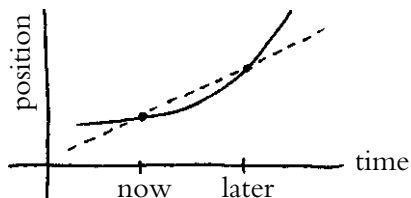
Of course, this line does not have the right slantedness (that's what being an approximation means), but as we move the nearby point closer and closer to our point, the approximation gets better and better.



So we can get the true slantedness of the tangent line from these approximating lines—provided there is some sort of pattern to their slantedness. Naturally, the slantedness pattern will have to come from the curve itself; that is, from its equation.

Of course, the Greek geometers knew that the tangent problem could be approached in this manner; the new idea was to combine it with Descartes' method of coordinates. In the case of a one-dimensional motion and its associated space-time

curve, what we are doing is applying the method of exhaustion to *time*.



If we select a particular moment, which we might as well call “now,” we can approximate the slantedness at that point (that is, the velocity) by selecting slightly nearby moments—let’s say slightly later ones—and figuring out the slant of the line connecting them (the *approximate velocity*). If we’re lucky (and we often are), there will be a pattern to this slantedness as the “later” point gets closer and closer to “now”—that is, as the elapsed time shrinks to zero. If we’re clever (and we often are), we can read this pattern and see where it’s heading. And that’s how we’ll get the exact velocity.

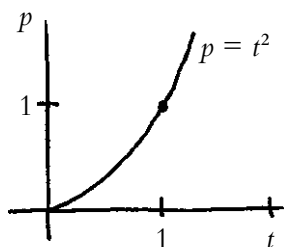
Does this all sound a bit farfetched? There are certainly a number of ways this plan could go awry. What if we can’t figure out the approximate velocities? What if they don’t have a pattern? What if they have a pattern, but it is too hard for us to read?

It turns out that the first of these is no problem at all. The approximate velocity (or slantedness, if you prefer) is simply the ratio of position change to time change:

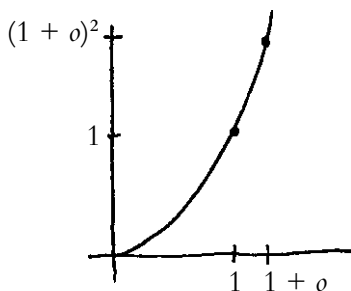
$$\text{approximate velocity} = \frac{p(\text{later}) - p(\text{now})}{t(\text{later}) - t(\text{now})}.$$

So given two points whose time and position coordinates are known, it is a relatively simple matter to calculate the slantedness of the line connecting them. The tricky part is going to be figuring out where these approximations are heading.

As an example, suppose we have the motion $p = t^2$ (the simplest nonuniform motion). Let's try to determine the velocity at the moment $t = 1$, $p = 1$.



In other words, we want to figure out the slantedness of the space-time curve (which happens to be a parabola) at the point $(1, 1)$. In this case our “now” is the moment when $t = 1$. A slightly later moment would be $t = 1 + o$, where o is a very small positive number. (This choice of notation was a little joke on Newton’s part—he chose the letter o to represent a variable that is heading toward zero. His detractors were apparently not amused. Berkeley derided o as “the ghost of a departed quantity.”)



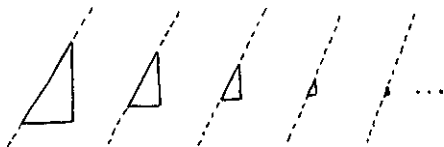
If we write $p(t)$ (as has become customary) for the value of the position p at the time t , then the change in position is simply

$$p(1 + o) - p(1) = (1 + o)^2 - 1,$$

and the elapsed time is simply o itself. Thus, the approximate velocity would be

$$\frac{(1 + o)^2 - 1}{o}.$$

Now the question is, where is this number heading as o approaches zero? Notice that as o gets smaller and smaller, both the top and bottom of this fraction approach zero. Essentially what's happening is that we are trying to calculate a certain slantedness using a sequence of ever-shrinking little triangles:



Even though the triangles themselves are shrinking away to nothing, their slantedness is not: it's heading toward the truth, namely the velocity we are after. The problem is how to tease that information out of our approximation pattern. We can't simply watch idly by as a fraction becomes $0/0$. We need to understand *how* it's getting there. Is the numerator approaching zero twice as fast as the denominator? Half as fast? Where is the proportion heading? To paraphrase Newton, we want the ratio of the quantities not before they vanish, nor afterward, but *with which* they vanish.

This is our first potential disaster. The fraction

$$\frac{(1 + o)^2 - 1}{o}$$

is heading toward the true velocity/slantedness, and it is definitely following a pattern. (I just wrote it down, didn't I?) The question is whether we are clever enough to read this pattern. We have a psychological problem—the expression of this pattern is not in a form that makes it easy for us to see what is happening. The solution is to rearrange it algebraically; not to change it, but to change its *form* so we can better understand it. In this case it is not particularly hard to do:

$$\begin{aligned}\frac{(1 + o)^2 - 1}{o} &= \frac{2o + o^2}{o} \\ &= 2 + o.\end{aligned}$$

Now, that's more like it! Not only is $2 + o$ much simpler looking, but it's also quite easy to see where it is heading, namely 2. In other words, the instantaneous velocity at the precise moment when $t = 1$ is exactly 2. More succinctly, we could write $\dot{p}(1) = 2$. So if a point is moving in the pattern $p = t^2$ (with respect to a certain map and clock), then at time $t = 1$, it is moving forward at a rate of two space units per time unit. Or, if you prefer, we can say that the tangent line to the parabola $p = t^2$ at the point $(1, 1)$ has a slant of 2. At least in this very simple case, our plan has been completely successful.

In fact, we can calculate in the same way the velocity of this motion $p = t^2$ at any moment whatsoever. At time t , the approximate velocity is

$$\begin{aligned}\frac{(t + o)^2 - t^2}{o} &= \frac{2to + o^2}{o} \\ &= 2t + o,\end{aligned}$$

and this clearly approaches $2t$ as o approaches zero. Thus we get $\dot{p}(t) = 2t$. The velocity at any moment is simply twice the time (in agreement with our intuition that the point should be speeding up). So here is our first nonobvious fact about velocity:

$$p = t^2 \Rightarrow \dot{p} = 2t.$$

On the face of it, it would seem that we got very lucky: we were able to rearrange the approximations in a way that allowed us to see what they were up to. Does the ability to calculate velocities necessarily come down to a question of algebraic skill?

In general, for any one-dimensional motion $p(t)$ (regardless of how complex the dependence on time), we can say that as o approaches zero,

$$\frac{p(t+o) - p(t)}{o} \text{ approaches } \dot{p}(t).$$

This gives us a systematic way to calculate the velocity pattern $\dot{p}(t)$ from the motion pattern $p(t)$ itself. The only question is whether we are clever enough to tell where the approximations are heading.

Check that for $p(t) = At + B$ we get
the expected velocity $\dot{p}(t) = A$.

If $p(t) = At^2 + Bt + C$, what is $\dot{p}(t)$?

How about if $p(t) = t^3$?

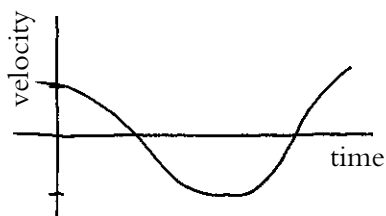
15

Let's step back from the details for a minute and think about exactly what we are doing. As usual, we have three equivalent ways to view the situation. The geometric view is that the objects we are interested in are *curves*, and we want to measure their slantedness and how it changes. The kinetic view is that we have a *motion* and we want to calculate its velocity at all times. More abstractly, we can regard our problem as taking one number pattern (that which describes the curve or motion in some coordinate system) and from it deriving another pattern (that of the slantedness or speed). For this reason, the second pattern is usually called the **derivative** of the first.

Suppose we graph our motion pattern in space-time:



We can then make a new graph by plotting the velocity at each time:



Notice that the vertical scales in these two pictures are completely different. The first is a coordinate map of the one-dimensional space in which the point moves, whereas the second graph is plotted on a scale of possible *rates*, a very different thing entirely. We could say that the first picture is a curve in space-time and the second is a curve in rate-time (so in particular the units are quite different).

Our velocity project then comes down to transforming the first picture into the second. Qualitatively, we can see that since the derivative picture records the slantedness, it will have large values where the original curve is steep, small values where it flattens out, and negative values where it slants down. To say anything more precise, we need a way to take a number pattern p and produce its derivative number pattern \dot{p} . In the abstract, we have just solved this problem, namely $\dot{p}(t)$ is precisely the number that is approached by

$$\frac{p(t + o) - p(t)}{o}$$

as o gets closer to zero. The only question is whether we can always get an *explicit* description of how \dot{p} depends on time. For instance, we saw that when $p = t^2$, we could actually calculate $\dot{p} = 2t$.

The abstract viewpoint allows us to shed any geometric or mechanical prejudices and simply view our problem as the study of the transformation $p \rightarrow \dot{p}$. What does this transformation look like algebraically? How does it behave? As p gets more complicated (that is, the way it depends on t gets more algebraically involved), presumably \dot{p} does also. But precisely how?

Here are a few things we do know:

If p is constant, then $\dot{p} = 0$.

If $p = ct$ for some constant c , then $\dot{p} = c$.

If $p = t^2$, then $\dot{p} = 2t$.

The first two of these are obvious—after all, \dot{p} is supposed to be the velocity. The third we calculated using the method of exhaustion. What happens if we have something more complicated, like $p = t^2 + 3t - 4$? In this case, the approximate velocity is

$$\frac{(t + o)^2 + 3(t + o) - 4 - (t^2 + 3t - 4)}{o} = 2t + 3 + o,$$

which approaches the true velocity $2t + 3$. So

$$p = t^2 + 3t - 4 \implies \dot{p} = 2t + 3.$$

Notice that this is exactly what we would have gotten if we had simply “dotted” each piece of p separately. That is, if we had thought of p as a sum of three pieces: t^2 , $3t$, and -4 . Then dotting each piece gives us the correct total. This means that dotting is a very well behaved operation. An algebraist would say that it “respects addition,” meaning that if you have a motion of the form $p = a + b$, where a and b are themselves motion patterns (variables which depend on t in some way), then the simple and beautiful truth is that

$$p = a + b \implies \dot{p} = \dot{a} + \dot{b}.$$

In other words, *the velocity of a sum is the sum of the velocities*. Of course, we can't assume that this is always true just because it happened to work for the one special case we just looked at. But it is in fact universally valid, and it's not hard to see why. The reason is that if $p = a + b$, then for any time t ,

$$p(t) = a(t) + b(t).$$

In particular,

$$\begin{aligned} p(t + o) - p(t) &= (a(t + o) + b(t + o)) - (a(t) + b(t)) \\ &= (a(t + o) - a(t)) + (b(t + o) - b(t)). \end{aligned}$$

In other words, the amount p moves in a short time interval is the sum of how much a moves and how much b moves. Dividing by the elapsed time, we get the approximate velocity relationship

$$\frac{p(t + o) - p(t)}{o} = \frac{a(t + o) - a(t)}{o} + \frac{b(t + o) - b(t)}{o}.$$

Letting o approach zero, we see that the left-hand side approaches \dot{p} and the right-hand side approaches $\dot{a} + \dot{b}$, so they must be equal. And, of course, the same goes for any number of pieces, so we have

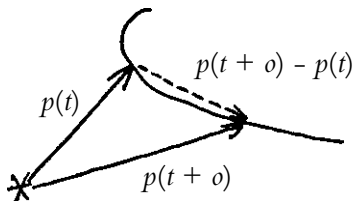
$$p = a + b + c + \cdots \implies \dot{p} = \dot{a} + \dot{b} + \dot{c} + \cdots.$$

Suppose $p = ca$ where c is constant. Show that $\dot{p} = c\dot{a}$. Does this make sense intuitively?

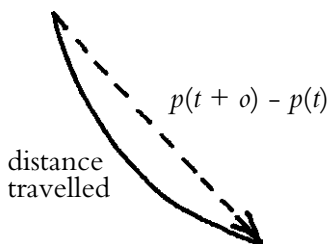
So far we have only considered one-dimensional motions. How do these ideas play out in higher-dimensional settings? Suppose we have a point moving in three-dimensional space in some way, described by a time-dependent position vector p .



Let's try the same idea as before, letting a small amount of time o elapse. The position vector then changes from its current value $p(t)$ to the nearby vector $p(t + o)$.



The difference $p(t + o) - p(t)$ is then also a vector, namely the shift from the current position to the slightly later position. This vector is tiny, but points very nearly in the direction of motion. In other words, its heading is very close to that of the true velocity at time t . As for its length, it is, of course, approaching zero, but when divided by o it should give a good approximation to the speed, since (at least for small values of o) the length of $p(t + o) - p(t)$ is pretty much the same as the distance traveled by the point during that small interval of time. (Note that division by o doesn't change the direction, only the length.)



Thus as o approaches zero, the approximate velocity vector

$$\frac{p(t + o) - p(t)}{o}$$

not only points more and more in the direction of the true velocity, but its length gets closer and closer to the true speed as well. So these approximations (which are now vectors) do in fact approach the precise velocity vector $\dot{p}(t)$.

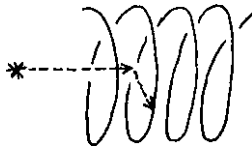
This means we don't have to make any major changes when we go to higher dimensions. We can still use the method of exhaustion in the same way. Of course, there may be computational (not to mention conceptual) differences when we move to a higher-dimensional setting, but the algebraic form of the approximations is exactly the same, and that is very good news indeed.

In fact, this allows us to immediately generalize our results on velocity addition: if $p = a + b$ is a *vector* sum of motion patterns (that is, a and b are time-dependent vectors and p is their sum) then we can still say that $\dot{p} = \dot{a} + \dot{b}$, and we don't need any fancy new explanation; the exact same argument works as before, since it required only purely algebraic rearrangements that work just as well for vectors as for numbers. Since this may well be the most important discovery ever made

about motion, I'll say it again: *the velocity of a sum is the sum of the velocities.*

**Does $p = ca \Rightarrow \dot{p} = c\dot{a}$ still work
in higher dimensions?**

Intuitively, one can imagine that each of the parts of the sum are compelling the point to travel in a certain direction at a certain speed, tugging on it, as it were, in its own way, and the resulting motion is the effect of these separate tugs acting simultaneously. For example, we can view a helical motion as a sum of a rotational (uniform circular) motion together with a linear motion.



The linear motion is pulling the point forward at a certain speed, and the circular motion is pushing it around the circle.



The combined effect of these two (their vector sum) is then the actual velocity of the helical motion.

It is this addition law of velocities that makes mechanical relativity—the breaking down of motions into sums of simpler

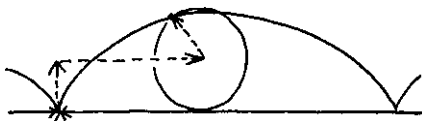
motions—such a useful idea. If the velocity of a compound motion could not be easily recovered from the velocities of its separate pieces, there would not be so much value to the breakdown in the first place.

What we have now is a *reduction strategy*. If we want to understand a complex motion, we can look for ways to break it up into simple parts and then study the parts separately. The good news is that (at least in the case of velocity) we can easily reassemble the information piece by piece.

**How does the speed of a point in the plane
depend on its horizontal and vertical speeds?**

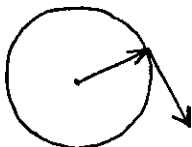
16

Now let's see if we can use these ideas to find the velocity of the cycloid motion. We have already broken this motion down into a sum of three pieces: a constant vector, a uniform linear motion, and a uniform circular motion.

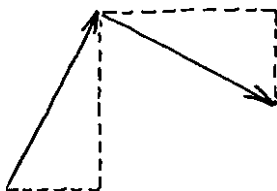


So all we have to do is to calculate the separate velocity vectors and then add them up. The velocity of a constant is zero (shifting our reference point doesn't affect velocities) and the uniform linear motion has a constant velocity, which due to our coordinate choices is simply u_1 , the unit vector in the

first direction. The velocity of uniform circular motion is also not hard to determine.



The velocity vector always points directly along the circle, so it must be perpendicular to the radial position vector. Since we chose our units so that the radius and the speed are both equal to 1, these two vectors are both of unit length. The radial vector starts its journey at the bottom of the circle (that is, it is equal to $-u_2$ at time $t = 0$) and rotates clockwise, so we found it to have the coordinate description $(-\sin t)u_1 + (-\cos t)u_2$, which for simplicity we could write as $(-\sin t, -\cos t)$. What are the coordinates of the velocity vector?



Notice that when two vectors in the plane are perpendicular, they both form the same little right triangle—only one's up is the other one's across (and the orientation gets flipped). More precisely, if a vector has coordinates (x, y) and we rotate it one-quarter of a turn clockwise (that is, from the second direction toward the first), the new coordinates will be $(y, -x)$.

What if we rotate it counterclockwise?

This means that the velocity vector of our (clockwise, starting at the bottom) uniform circular motion must have coordinates $(-\cos t, \sin t)$. We could also see this by observing that the velocity vector itself is undergoing uniform circular motion, beginning at $(-1, 0)$ and proceeding clockwise. In any case, we can now assemble the pieces:

$$\begin{aligned} p &= \text{constant} + \text{linear} + \text{circular} \\ &= u_2 + tu_1 + (-\sin t)u_1 + (-\cos t)u_2 \end{aligned}$$

and therefore

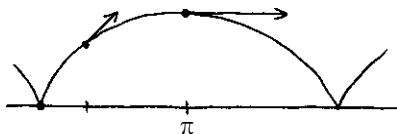
$$\dot{p} = 0 + u_1 + (-\cos t)u_1 + (\sin t)u_2,$$

which we could also write as $(1 - \cos t, \sin t)$ for short. So we now know the exact velocity of our moving point at all times. In particular, its speed at time t is given by

$$\sqrt{(1 - \cos t)^2 + \sin^2 t} = \sqrt{2 - 2\cos t}.$$

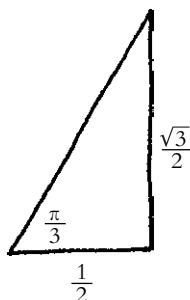
Here I'm making use of the customary abbreviation $\sin^2 t$ in place of the more cumbersome $(\sin t)^2$.

Let's take a look at some specific moments in the history of this motion.



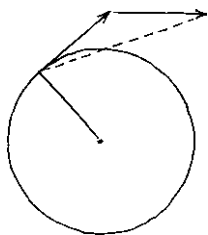
At time $t = \pi$, when the point has reached the top of the rolling disk, the position vector $p = (\pi, 2)$ and the velocity vector

(according to our formula) is $\dot{p} = (2, 0)$, meaning that our point is moving directly forward at a rate of 2, twice the speed at which the center of the disk is traveling. Notice also that at time $t = 0$ (and again at times $2\pi, 4\pi$, etc.), the velocity vector is 0. These are the moments when the motion of the point reverses direction and “hangs.” Finally, at time $t = \frac{\pi}{3}$ (one-sixth of the way through the first rotation), we have $\dot{p} = (\frac{1}{2}, \frac{\sqrt{3}}{2})$, which means our point is heading forward and up at an angle of $\frac{\pi}{3}$:



and thus its speed at this precise instant is exactly 1.

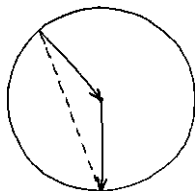
Another way to understand this velocity vector is to forget about coordinates and simply add the linear and circular velocities geometrically.



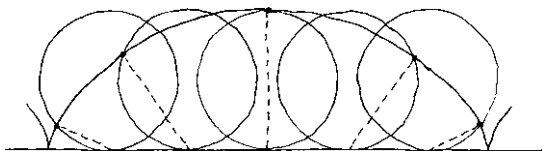
The circular velocity (as we have seen) is perpendicular to the radial vector, and the linear velocity adds to this a horizontal shift. Both of these vectors have length equal to the radius by

our choices. Our velocity vector is just the sum of these two vectors, so they make a little triangle.

Now here's the clever observation: if we rotate this triangle 90 degrees clockwise, then the circular velocity becomes a radius and the horizontal push turns into a downward vertical vector.



The velocity vector has become a so-called **chord** of the circle, connecting our moving point to the point where the circle touches the ground. In other words, we can see the velocity as simply being a rotated version of the chord.



Show that the heading changes at a uniform rate.

What happens to the speed and heading
as we pass through the time $t = 2\pi$?

Notice, by the way, that we have inadvertently (or in my case quite verterntly!) discovered an expression for the length of a circular chord: if two points on a unit circle are separated by an arc of length t , then the chord connecting them has length $\sqrt{2 - 2\cos t}$.

**Show that this length can also
be written as $2 \sin \frac{t}{2}$.**

The upshot of this calculation is that the speed of the cycloid motion is equal to the length of the chord. Which is not to say that such information could not be obtained by other means. The cycloid, for instance, is simple enough that one does not require vector or coordinate descriptions at all, as long as one is sufficiently clever. (In fact, the measurement of the area of a cycloid preceded Descartes' work by several years.)

The point is not that these techniques—vectors and coordinates, relativity, exhaustion—are always necessary (although they often are), but that they are so wonderfully general and require no particular inspiration or genius on the part of the user. That is, we have a *uniform* way to treat geometric and mechanical problems. Of course, there will be occasions when a simpler or more symmetrical approach is possible, but these tend to be rather ad hoc and special, though undeniably quite beautiful and imaginative.

**Can you measure the speed of helical motion
without the use of coordinates and relativity?**

**Can you measure the velocity of a spirograph
motion? (I suggest coordinates and relativity!)**

One very powerful consequence of the way velocity respects vector addition is that it allows us to view high-dimensional motions as a set of simultaneous one-dimensional motions. For example, any motion in two dimensions can be written

$$p = xu_1 + \gamma u_2,$$

where x and γ are the separate horizontal and vertical components, which, since they depend on time in some way, can be viewed as one-dimensional motions in their own right. Then our addition law tells us that

$$\dot{p} = \dot{x}u_1 + \dot{\gamma}u_2$$

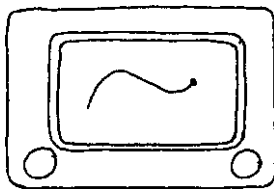
or, if you prefer a coordinate description,

$$p = (x, \gamma) \Rightarrow \dot{p} = (\dot{x}, \dot{\gamma}).$$

In particular, the speed of the moving point (being the length of this vector) is just the Pythagorean combination of the separate one-dimensional speeds, namely

$$\sqrt{\dot{x}^2 + \dot{\gamma}^2}.$$

Sometimes I like to imagine my point as being controlled by the knobs of an Etch A Sketch:



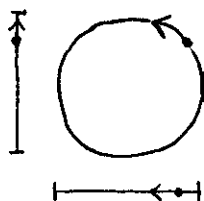
Then what we are saying is that not only is the position of the point simply the pair of positions of the dials, but the velocity of the point is also just the pair of velocities. So if at some instant

the horizontal knob is moving at a rate of 3 and the vertical knob at a rate of 4, then the point itself has a speed of 5 at that moment and is moving in the “over 3, up 4” direction. Of course, this also works in three dimensions or higher, the only difference being one of visualization (and having more knobs). So in general, for a motion in any dimension whatsoever,

$$p = (x, y, z, \dots) \Rightarrow \dot{p} = (\dot{x}, \dot{y}, \dot{z}, \dots).$$

For example, the (admittedly rather contrived) three-dimensional motion $p = (t^2, t + 1, 3t)$ would have the velocity vector $(2t, 1, 3)$. Thus the velocity problem in any dimension can always be reduced to the one-dimensional case. The velocity of a moving point is easily recovered from those of its various coordinate “shadows.”

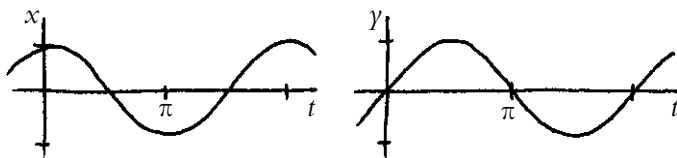
An interesting example of this phenomenon is uniform circular motion.



In this case (assuming the standard choices), the horizontal and vertical components of this motion are just $\cos t$ and $\sin t$ respectively. That is, we could think of uniform circular motion as the pair of one-dimensional motions

$$\begin{aligned} x &= \cos t, \\ y &= \sin t. \end{aligned}$$

In other words, to draw a circle on an Etch A Sketch, you would need to turn the dials in the pattern of sine and cosine waves:



Notice that these patterns are just shifted versions of each other; the cosine of a number is always equal to the sine of a number $\frac{\pi}{2}$ greater.

Why is $\cos t = \sin(t + \frac{\pi}{2})$?

This pair of space-time diagrams contains all the information of uniform circular motion. In particular, their separate velocities must be the components of the two-dimensional velocity. Since we already know the velocity of uniform circular motion (it's just the position vector rotated a quarter turn counterclockwise), we have

$$p = (\cos t, \sin t) \Rightarrow \dot{p} = (-\sin t, \cos t),$$

and the velocities of the components must match up:

$$\begin{aligned} x = \cos t &\Rightarrow \dot{x} = -\sin t, \\ y = \sin t &\Rightarrow \dot{y} = \cos t. \end{aligned}$$

So the slantedness of the sine wave at any time is just the height of the cosine wave at that moment and vice versa (with

the added twist of the negative sign). The sine and cosine patterns form a very incestuous pair—each one is (essentially) the derivative of the other. By the way, the annoying negative sign is unavoidable; if we changed our conventions regarding orientation, it would still be there, only in a different place.

What sort of curves do we get when the
horizontal and vertical wave motions have
different frequencies? For example,
what if $x = \cos(3t)$, $y = \sin(5t)$?

17

At the risk of being redundant (a risk I seem to be quite willing to take), I want to say a few more words about the philosophy we have adopted. The idea is to subsume the study of shape and motion into the larger, more abstract world of numerical variables and relationships. This viewpoint not only has the benefit of simplicity (there are no units to worry about and we don't need to be able to visualize anything) but also tremendous flexibility and generality.

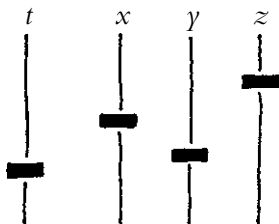
In fact, it would be hard to find any scientist, architect, or engineer who is not in some way engaged in the process of *modeling*—creating an abstract, simplified representation of their problem by a set of variables and equations (e.g., a biologist's model of mammalian territorial behavior, a cardiologist's model of vascular pressure, or an electrical engineer's model of energy capacity). Of course, in cases like these,

there is a considerable difference between the real object of interest (i.e., nature) and a mathematical model of it. This is pretty much what scientists spend their time worrying about—the aptness of their mathematical models of reality. For example, when new experiments are performed, or new data is collected, it often leads to the rejection of the current model and its replacement by an updated version.

The situation is quite different for mathematicians: for us, the mathematical model *is* the object of study! There is nothing empirical here; we are not awaiting any confirmation or test results. A mathematical structure is what it is, and anything we discover about it is the truth. In particular, if we choose to model an imaginary curve or motion by a set of equations, we are not making any guesses or losing any information through oversimplification: our objects are already (for aesthetic reasons) as simple as they can be. There is no possibility of conflating reality and imagination if everything is imaginary in the first place.

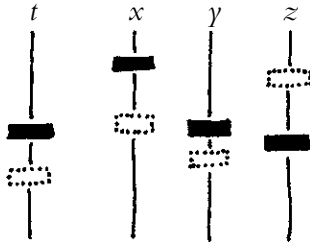
So henceforth, our objects of study will be systems of variables (which Newton called *fluents*, Latin for “that which flows”) and relations; that is, equations expressing the relationships among the variables.

Sometimes I like to think of my variables as being sliders on an imaginary multichannel mixing board:



For example, a motion in three-dimensional space could be represented by a system of four variables, together with a set of equations that tell us how the three spatial coordinates depend on time. These equations constitute the wiring or programming of the mixing board. As we move the t slider, the x , y , and z sliders respond automatically, moving according to the wiring pattern.

With this image in mind, what we have been doing to calculate velocities (what might be called the Newtonian methodology) is to “nudge” the sliders a little bit and see how far they move:



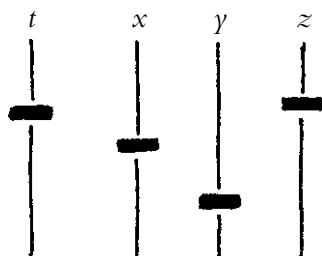
It has become customary to use the abbreviation Δx in place of the more cumbersome $x(t + o) - x(t)$, so that Δx simply measures the amount of *change* in the variable x . In particular, the elapsed time o could also be thought of as Δt .

When we vary the t slider, changing its value by the tiny amount Δt , the other sliders x , y , and z respond, and they increase by Δx , Δy , and Δz respectively. (Some of these increments may be negative if the corresponding variable decreases). We get the approximate velocities

$$\frac{\Delta x}{\Delta t}, \frac{\Delta y}{\Delta t}, \frac{\Delta z}{\Delta t}.$$

As Δt shrinks away to zero, these will approach the true instantaneous velocities \dot{x} , \dot{y} , \dot{z} . The velocity vector of our three-dimensional motion is then $(\dot{x}, \dot{y}, \dot{z})$ and its speed is the length of this vector, $\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}$.

As an example of our new abstract outlook, suppose we have this system of variables and relations:



$$\begin{aligned} x &= t + \cos t \\ y &= \sin t \\ z &= t \end{aligned}$$

From our previous results on the derivatives of sine and cosine, we immediately get

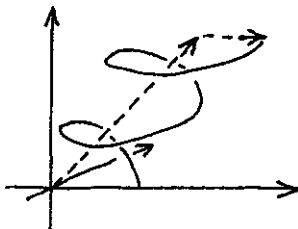
$$\begin{aligned} \dot{x} &= 1 - \sin t, \\ \dot{y} &= \cos t, \\ \dot{z} &= 1, \end{aligned}$$

and this computation requires no geometric or kinetic interpretation. Of course if, for psychological or romantic reasons, we wanted to think of it as a motion, we certainly could. In fact, it's not too hard to see these equations as describing a slanted helical motion. Maybe the easiest way to see this is to write the motion as a vector,

$$p = t(1, 0, 1) + (\cos t, \sin t, 0).$$

Here we have a sum of a uniform circular motion (with all the standard choices) and a uniform linear motion that, rather

than being perpendicular to the rotating disk (as in a conventional helix), is tilted at a 45-degree angle:



Our derivatives could then be viewed as components of the velocity vector

$$\dot{p} = (1 - \sin t, \cos t, 1)$$

whose length gives us the speed of our moving point:

$$\begin{aligned} \text{speed} &= \sqrt{(1 - \sin t)^2 + (\cos t)^2 + 1} \\ &= \sqrt{3 - 2\sin t} . \end{aligned}$$

This tells us that our point is speeding up and slowing down in a fairly subtle way, and since $\sin t$ varies between -1 and $+1$, the speed ranges from 1 to $\sqrt{5}$.

The point being that these measurements come directly from the abstract numerical relationships and not from any visual or kinetic image. The model doesn't know, or need to know, what it is a model of (if anything). Our project has subtly shifted (if my constant harping on this point can be called subtle) from being the study of velocities of motions to the study of the derivative—the abstract transformation by which a variable p produces a new variable \dot{p} , which Newton referred to as the *fluxion* of the fluent p .

This immediately suggests the possibility of double dotting—viewing \dot{p} itself as a variable, which could then be dotted to produce \ddot{p} and even three dots, and so on. If we interpret p as a motion (that is, as the position of a moving point), then \dot{p} would measure the rate at which the velocity \dot{p} changes; in other words, the **acceleration**. Geometrically, \ddot{p} could be seen as a way to measure the rate at which the slantedness of a curve changes—what a geometer would call its **curvature**. As mathematicians, of course, we are free to make either interpretation, or neither. We can simply speak of *higher derivatives* in the abstract and then study their interesting properties.

For example, as we take more and more derivatives, the squaring function ($p = t^2$) transforms into doubling ($\dot{p} = 2t$), which in turn becomes constant ($\ddot{p} = 2$) and finally zero for all higher derivatives.

What are the higher derivatives
of sine and cosine?

18

Before we develop these ideas any further, I want to show you an even more general and abstract approach that I vastly prefer. Maybe the best way to start is with an analogy. We've seen many times that measurements are always relative and that any well-posed measurement question always (at least implicitly) comes down to a comparison of some sort. For example, if we wanted to measure a certain length or area, we would be asking about the extent of a line or the space enclosed by a region,

measured in comparison to some other object of the same kind. We could, of course, choose some standard of comparison, such as a certain fixed square whose side length and area we could take to be our units of measurement. Any new object could then be compared with this standard and measured against it.

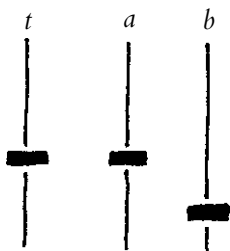
To me, this is a repugnant idea. I don't want any unnecessary and contrived units cluttering up my beautiful imaginary universe. If I want to measure the diagonal of a pentagon compared to its side, I don't need to measure both with respect to some preexisting standard length and then compare the two; I can compare them directly to each other. (I know I've talked about this a million times, but bear with me.)

The way I like to think of it is that lines have length or extent whether or not we measure it. A region encloses space whether or not I choose to compare it with anything else. So length and area aren't *numbers*, they are abstract geometric quantities. Only when we compare them and form ratios do we obtain numerical values. The diagonal of a square has a length, and so does the side. Neither of these is a number, but nevertheless one is exactly $\sqrt{2}$ times the other.

If this all sounds like I'm flogging a dead horse, the point is that we have been unconsciously putting this same kind of arbitrary and unnecessary obstacle in our way when we measure velocity. If you watch a cheetah running, it has a speed or a rate independent of any measuring, just as a circle encloses an amount of space. This rate is not a number and there are no units, and yet, if a horse were running alongside it (so now I'm flogging a *live* horse?), we could tell that the cheetah was going twice as fast. That is, we could wait a certain amount of time (no need to measure it in seconds or anything) and see how far the two animals traveled (also no need to measure in any

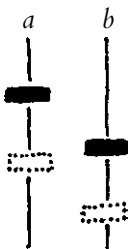
units) and compare the distances to each other. So the abstract, unmeasured rate of something is meaningful. What we have been doing up to now is choosing a standard unit of speed—namely, the speed of our clock! That is, the rate of *time itself* has been our (soon-to-be-discarded) unit of measurement.

Let's imagine that we have two time-dependent variables a and b , related to each other by some equation.



If we were interested in the relative speeds of a and b at a particular moment, we could, of course, calculate \dot{a} and \dot{b} and form their ratio \dot{a}/\dot{b} . But this is every bit as unnecessary (and aesthetically appalling) as the geometric examples I mentioned. We shouldn't need to involve time at all.

Suppose, for instance, that only the a and b sliders were accessible to us, and the t slider was hidden behind the scenes somewhere. We could still give the mixing board a little kick, and the sliders would each move slightly.



As usual, we would obtain the small variations Δa and Δb . Only now, instead of comparing them both to Δt , we simply compare Δa and Δb directly. Then as these small differences both approach zero, we get the true proportion of their instantaneous velocities.

Does this make sense? To make this easier to talk about, let me introduce some notation—this is, after all, the whole point of notation. Let's write dx for “the instantaneous rate of change of the variable x .” That is, dx is the abstract, nonnumerical velocity, analogous to “cheetah speed.” (This notation was first introduced by Leibniz in the 1670s.) Then what we are saying is that the proportion of small changes $\Delta a : \Delta b$ approaches the true velocity proportion $da : db$ as these tiny increments simultaneously vanish.

In terms of our new Leibnizian notation, the fluxion \dot{x} is simply the ratio dx/dt . This means that all of our results concerning fluxions can be rephrased easily in this new abstract language. For instance, our previous computation

$$p = t^2 \implies \dot{p} = 2t$$

can be rewritten simply as

$$d(t^2) = 2t \, dt.$$

This is not a special statement about time; this is true for any variable whatever. So (using w for “whatever”) we have

$$d(w^2) = 2w \, dw,$$

and this says that “the rate at which the variable w^2 changes is always exactly twice the current value of w times as fast as the rate of w itself.” Note the economy of the notation; we don’t need to give names to the patterns and then dot the names, we can just d the patterns directly. Thus we also have

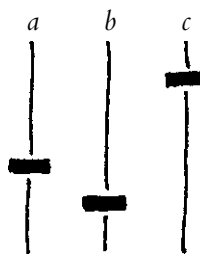
$$\begin{aligned}d(cw) &= c \, dw, \text{ for any constant } c, \\d(\sin w) &= \cos w \, dw, \\d(\cos w) &= -\sin w \, dw.\end{aligned}$$

I want to make two things perfectly clear. The first is that dx (the so-called **differential** of x) is not a number; it is an abstract rate. Cheetah speed is not a number and neither is horse speed, but we can still say one is twice the other (as with lengths, areas, and all other measurements). The other is that this d we are using (the **Leibniz d -operator**) is not a number either. When we write dx we are not multiplying d by x , we are *applying* the d -operator to the variable x to obtain the differential of x . The notational ambiguity is slightly annoying I admit, but as long as we are reasonably careful (which would include not choosing d as one of our variable names!), it’s not really much of a problem. On the contrary, Leibniz’s notation is extremely flexible and convenient once you get used to it.

Generally speaking, most of our measurement problems will come down to finding the relative velocities of a set of variables. Whether time is included among them is up to you and depends on the specific problem at hand. If you were interested in a particular motion, then perhaps it would make sense to think of time as one of your variables and all the others depen-

dent on it. A purely geometric question, on the other hand, has no need for any ticking clocks.

Suppose we have a set of variables a , b , and c connected to each other by a set of equations:



$$a^2 = b^2 + 3$$

$$c = 2a + b$$

Notice that in this example none of the variables is special. None of them plays the role of time—there is no “master” slider that controls the others. Instead we have an *interdependence* among the variables. At any given moment, the variables will have certain values, as well as differentials (i.e., their instantaneous rates of change at that moment). The question is, how exactly do the relationships among the variables control the relative proportions of their differentials? How can we take the information

$$a^2 = b^2 + 3,$$

$$c = 2a + b,$$

and determine the ratios $da : db : dc$?

The direct approach would be to give the mixing board a kick, check out the variation of the sliders, and figure out where their proportions are heading as the kick gets smaller. But here’s the point: we don’t actually need to go through this laborious process. Instead, we can simply apply the d -operator to both sides of the equations:

$$\begin{aligned}d(a^2) &= d(b^2 + 3), \\dc &= d(2a + b).\end{aligned}$$

After all, if two variables are always equal their rates must also be equal. Expanding these accordingly, we obtain the differential equations

$$\begin{aligned}2a\,da &= 2b\,db, \\dc &= 2\,da + db.\end{aligned}$$

So, for instance, at the moment when $a = 2$, $b = 1$, and $c = 5$ (which does in fact satisfy our original equations and so qualifies as an actual moment), we have

$$\begin{aligned}4\,da &= 2\,db, \\dc &= 2\,da + db.\end{aligned}$$

Thus at that precise instant, b is moving twice as fast as a , and c four times as fast. In other words, the ratio $da:db:dc$ is $1:2:4$. We now have a simple and direct method for solving any problem concerning relative rates of change—just d everything!

Incidentally, Leibniz's original interpretation was somewhat different. His view was that dx , rather than representing the instantaneous rate at which x changes, is instead an infinitesimal change in x itself. That is, as Δx shrinks away to nothing, it sort of “hovers” at the value dx , which, though not exactly zero, is nevertheless smaller than any positive quantity. (Imagine what his critics had to say about that!) Actually, there is no real problem with this view, as long as you are sufficiently careful. After all, in a small interval of time, the proportion of two

velocities is the same as the proportion of the distances traveled. The point is that the approximation $\Delta a : \Delta b$ approaches the true proportion $da : db$. We may interpret it however we wish.

19

The problem of velocity can now be reduced to the study of the Leibniz d -operator. Given any set of equations which describe a motion, we can simply d them to obtain the relative rate information. The only remaining problem is to determine exactly how the d -operator behaves. How exactly does an interdependence among variables get transformed into a relationship among differentials?

We saw before that if a and b are two variables and we form a new variable $c = 2a + b$, then the rate at which c changes can be easily determined from the rates of a and b :

$$\begin{aligned} dc &= d(2a + b) \\ &= 2 da + db. \end{aligned}$$

Here we have used the fact that d behaves *linearly*; that is, for any variables x and y , and any constant c , we have

$$\begin{aligned} d(x + y) &= dx + dy, \\ d(cx) &= c dx. \end{aligned}$$

But what if the relationships among the variables are more complicated? Suppose, for example, we wanted to compare

the rates of x and y , where $y = x^3 \sin x$? We can certainly say that $dy = d(x^3 \sin x)$, but in order to relate this to dx itself, we need to understand more about the d -ing process. In particular, we need to know how d acts on *products* of variables. How exactly does $d(ab)$ depend on da and db ? This is where our study of motion has taken us—we're now asking questions about the abstract behavior of a differential operator. How does d act on square roots? On division? Any operation that can be performed on a number or set of numbers could conceivably be used to describe an interdependence among variables, and to understand their relative rates of change, we would need to know how d behaves when confronted with such an operation. Of course, many operations (such as the $x^3 \sin x$ example above) can be viewed as having been built up from simpler ones (e.g., $x^3 \sin x$ is x^3 times $\sin x$), so that if we can figure out the behavior of d for a few simple operations (in particular, multiplication), we can hopefully deal with more complex combinations of them.

So let's try to determine $d(xy)$ in terms of dx and dy . We'll do this "by hand" as it were, imagining that x and y somehow depend on t (which we can think of as time if we want to), and then we'll see what happens as we vary t a little bit. Essentially, what this does is to choose dt as our unit of speed. This is analogous to the situation in geometry where we choose an arbitrary unit, use it to make measurements, and then discard it once we have discovered the correct relationships. (It's not unlike the scaffolding used during the construction of a building. It is temporarily quite helpful, but it is ultimately removed.)

So imagine that t changes a little, let's say by an amount Δt . Then x and y react, becoming $x + \Delta x$ and $y + \Delta y$ respectively.

The amount of change in xy is then

$$\begin{aligned}\Delta(xy) &= (x + \Delta x)(y + \Delta y) - xy \\ &= x \cdot \Delta y + y \cdot \Delta x + \Delta x \cdot \Delta y.\end{aligned}$$

Dividing both sides of this equation by Δt , we get

$$\frac{\Delta(xy)}{\Delta t} = x \frac{\Delta y}{\Delta t} + y \frac{\Delta x}{\Delta t} + \frac{\Delta x}{\Delta t} \cdot \frac{\Delta y}{\Delta t} \cdot \Delta t,$$

where the last term has been reorganized for the sake of symmetry. Letting Δt approach zero, we see that the last term shrinks away to nothing, and we get

$$\frac{d(xy)}{dt} = x \frac{dy}{dt} + y \frac{dx}{dt}.$$

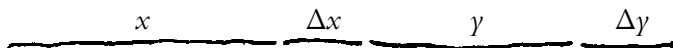
At this point, we no longer require the services of our background variable t , so multiplying both sides by dt , we get our sought-after relationship

$$d(xy) = x \, dy + y \, dx.$$

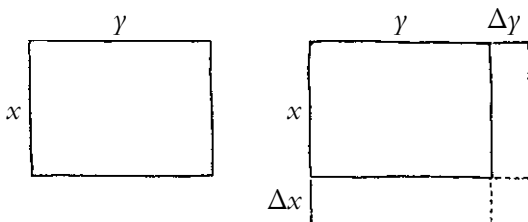
This is sometimes called **Leibniz's rule**. Before we investigate the consequences of this discovery, I want to say a couple of things about what it means and why it makes good sense. First of all, what is being said is that the speed of a product of two variables is just the sum of the speeds of each variable magnified by the value of the other. There is a nice way to see this geometrically. First, let's think about addition for a second. Imagine two sticks with lengths x and y (so the sticks are growing and shrinking). If we put the sticks together, we get a new stick of length $x + y$.



If at a given moment x and y have certain speeds, it seems clear that $x + y$ has a speed equal to their sum. We can even imagine a little time going by and observing the small changes:



So the change in $x + y$ is simply $\Delta x + \Delta y$. Since speed is proportional to the change in length, we get $d(x + y) = dx + dy$. (Leibniz would say that the infinitesimal change is $dx + dy$.) Now for multiplication. We can imagine the area of a rectangle of sides x and y :



We see that for small changes in x and y we get a change in area equal to that of an L-shaped sliver, so we again get

$$\Delta(xy) = x\Delta y + y\Delta x + \Delta x \cdot \Delta y.$$

For tiny increments Δx and Δy , the final term is of much smaller magnitude than the other terms. Both Newton and Leibniz recognized that this term is comparatively negligible and ultimately makes no contribution to the velocity. (Berkeley was less than convinced.) A more modern explanation (which, ironically, is practically the same one that Archimedes or

Eudoxus would have given) is that $\Delta x \cdot \Delta y$ is never equal to zero, but that its proportion to the other terms *approaches* zero (the first two terms are on the order of Δt , whereas the last term is more like $(\Delta t)^2$ in magnitude). So in fact, we are justified in replacing all terms of the form Δw with the corresponding differential dw , so long as we omit all “higher-order” terms involving products of Δ ’s. Thus we again obtain Leibniz’s beautiful formula,

$$d(xy) = x \, dy + y \, dx.$$

Now let’s examine some of its consequences. First of all, notice that our previous result $d(w^2) = 2w \, dw$ follows immediately from our product formula:

$$\begin{aligned} d(w^2) &= d(w \cdot w) \\ &= w \, dw + w \, dw = 2w \, dw. \end{aligned}$$

We can even compute $d(w^3)$ in the same way:

$$\begin{aligned} d(w^3) &= d(w^2 \cdot w) \\ &= w^2 \, dw + w \, d(w^2) \\ &= w^2 \, dw + w \cdot 2w \, dw = 3w^2 \, dw. \end{aligned}$$

Show that in general,

$$d(w^n) = n w^{n-1} \, dw, \text{ for } n = 2, 3, 4, \dots$$

Now we can calculate things like $d(x^3 \sin x)$. From our earlier work, we know that $d(\sin x) = \cos x \, dx$, so our product formula gives us

$$\begin{aligned}
 d(x^3 \sin x) &= x^3 d(\sin x) + \sin x d(x^3) \\
 &= x^3 \cos x dx + \sin x \cdot 3x^2 dx \\
 &= (x^3 \cos x + 3x^2 \sin x) dx.
 \end{aligned}$$

So, for example, when $x = \pi$, the variable $x^3 \sin x$ is traveling exactly $-\pi^3$ times as fast as x is (i.e., π^3 times as fast in the opposite direction). It's pretty amazing that we have access to information like that at all, let alone that we can get at it so easily (if you call the development of an entire theory of mathematical motion over a period of twenty centuries easy).

As a further consequence of Leibniz's rule, we can easily obtain a formula for the differential of the reciprocal of a variable, $d(1/w)$. The simplest way to proceed is to go back to the very definition of reciprocal, namely

$$w \cdot \frac{1}{w} = 1.$$

Applying d and using the product formula, we get

$$w d\left(\frac{1}{w}\right) + \frac{1}{w} dw = 0.$$

Rearranging this, we find

$$d\left(\frac{1}{w}\right) = -\frac{dw}{w^2}.$$

Now we know the precise rate at which $1/w$ changes, depending on how w itself is varying.

Show that for any variables a and b ,

$$d\left(\frac{a}{b}\right) = \frac{b \, da - a \, db}{b^2}.$$

Show that $d(\sqrt{w}) = \frac{dw}{2\sqrt{w}}$.

20

At this point we have compiled a fairly extensive library of facts about the Leibniz d -operator. Here is a summary of what we know so far:

Constants: $dc = 0$ for any constant c .

Sums: $d(a + b) = da + db$,

$$d(a - b) = da - db.$$

(You can derive the second formula from scratch or use the fact that $(a - b) + b = a$. Of course, it is rather obvious in the first place.)

Products: $d(ab) = a \, db + b \, da$.

(In particular, we have $d(cw) = c \, dw$ for any constant c .)

Quotients: $d\left(\frac{a}{b}\right) = \frac{b \, da - a \, db}{b^2}$.

Sine and cosine: $d(\sin w) = \cos w \, dw$,
 $d(\cos w) = -\sin w \, dw$.

Powers: $d(w^n) = n w^{n-1} \, dw$,
 $n = 2, 3, 4, \dots$

Square roots: $d(\sqrt{w}) = \frac{dw}{2\sqrt{w}}$.

Of course, we will be adding to this list, but not very much; this is already an extremely powerful collection of results and allows us to calculate the differential of almost any combination of variables you can imagine. Here are a few illustrative examples:

$$\begin{aligned} d(a^3 b^2) &= a^3 d(b^2) + b^2 d(a^3) \\ &= a^3 \cdot 2b db + b^2 \cdot 3a^2 da \\ &= 3a^2 b^2 da + 2a^3 b db. \end{aligned}$$

$$\begin{aligned} d(\sqrt{u^2 + v^2}) &= \frac{d(u^2 + v^2)}{2\sqrt{u^2 + v^2}} \\ &= \frac{2u du + 2v dv}{2\sqrt{u^2 + v^2}} \\ &= \frac{u}{\sqrt{u^2 + v^2}} du + \frac{v}{\sqrt{u^2 + v^2}} dv. \end{aligned}$$

$$\begin{aligned} d\left(\frac{\sin w}{\cos w}\right) &= \frac{\cos w d(\sin w) - \sin w d(\cos w)}{\cos^2 w} \\ &= \frac{\cos^2 w dw + \sin^2 w dw}{\cos^2 w} \\ &= \frac{dw}{\cos^2 w}. \end{aligned}$$

Sometimes I like to think of the d -operator as a sort of enzyme acting on long, complicated molecules (the atoms are the variables themselves). For instance, if x and y are my atoms, I can construct the complex molecule $(y \cos \sqrt{x})^3$. We can view this as being constructed hierarchically as follows: start with x , square-root it, take the cosine of that, multiply by y , and then cube the whole thing. So structurally, I can think of it as being

a cube. That is, I can blur my eyes (so to speak) and picture it as $(\text{blah})^3$, where for the moment I ignore the details of what “blah” stands for. Then my d -enzyme goes to work:

$$d((\text{blah})^3) = 3(\text{blah})^2 d(\text{blah}).$$

This is because $d(w^3) = 3w^2 dw$ for *any* variable w , no matter what it looks like. So d doesn't care what “blah” is; it just goes to work unraveling the molecule step by step. (This process is commonly referred to as *chaining*.)

We are now reduced to finding $d(\text{blah})$. Now “blah” itself is a product, namely $\gamma \cos \sqrt{x}$. So using the product pattern, we get

$$\begin{aligned} d(\text{blah}) &= d(\gamma \cos \sqrt{x}) \\ &= \gamma d(\cos \sqrt{x}) + \cos \sqrt{x} d\gamma. \end{aligned}$$

This reveals the next layer of structure in our molecule, so we then need to break down $d(\cos \sqrt{x})$:

$$d(\cos \sqrt{x}) = -\sin \sqrt{x} d(\sqrt{x}).$$

Finally, from our table we find

$$d(\sqrt{x}) = \frac{dx}{2\sqrt{x}}.$$

Putting everything together (and rewriting it in terms of our variables x and γ), we get

$$d((\gamma \cos \sqrt{x})^3) = 3(\gamma \cos \sqrt{x})^2 (\cos \sqrt{x} d\gamma - \frac{\gamma \sin \sqrt{x}}{2\sqrt{x}} dx).$$

In principle, there is nothing to stop you from computing the differentials of even the most complex combinations of variables, breaking everything down until it depends only on the differentials of the “atomic” variables themselves. And think of all the work we are saving! Imagine trying to calculate the relative velocities by hand, using small increments and trying to figure out where huge masses of ratios of vanishing quantities are heading. The d -operator essentially takes care of the bookkeeping for the method of exhaustion and saves us the gory details.

This is very much like the situation we find in ordinary arithmetic. We have an encoding scheme whereby a quantity of rocks (or whatever) can be represented, namely as a Hindu-Arabic digit sequence (e.g., 231 encodes two piles of one hundred rocks, three piles of ten, and one leftover). We can then ask how these codes behave when the piles are combined and rearranged in various ways.

For example, if we have two piles of sizes 231 and 186, what is the code for their sum; that is, the pile we get when we push the two piles together? As I am sure you are aware, there is a well-known system for determining this sort of thing: 6 plus 1 is 7, 8 plus 3 is 11, carry the 1, add that to 2, plus 1 is 4, so it’s 417.

The point is that we don’t need to have any actual rocks; the computation (in this case addition) can be performed with the symbols alone. We don’t have to push piles of rocks around and then count them by hand; the system takes care of it for us. (Of course, someone had to invent the system!)

A symbolic computation system like this is called a **calculus** (Latin for “counting stone”). A calculus typically consists of

three ingredients: a notation scheme for representing the relevant objects symbolically, a (hopefully small) set of manipulation procedures (e.g., carrying), and a (also hopefully small) table of basic facts—the single-digit sums, for example. The idea is to use the procedures to break a complex problem down into simpler pieces, and then these can be looked up in the table (or memorized, if you wish).

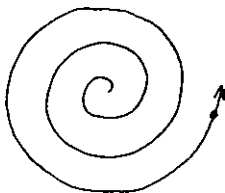
Thus we have the *multiplicative calculus* of elementary arithmetic, which includes a few procedures such as carrying and place shifting, as well as the infamous times tables. The amazing power of this system is that it allows us to quickly and easily perform calculations that would otherwise be very time consuming and laborious. To multiply 1876 by 316 we merely have to jiggle a few symbols around (or better yet, let a machine do it). Nobody has to do the backbreaking work of laying out 1876 rows of 316 rocks and then counting the whole mess by hand. A calculus is a fantastic thing to have.

It is also a very rare thing to have. Most problems in mathematics do not afford such systematic treatment. In fact, most of the really great problems in mathematics are still largely unsolved, and those on which progress has been made have required tremendous ingenuity and specific individual treatment. Every once in a great while, a class of problems comes along for which a calculus can be developed, and it is always a major achievement.

So it is a cause for great celebration that we have a differential calculus—a systematic mechanical procedure for calculating differentials that does not require “pushing rocks together”; that is, it spares us having to apply the method of exhaustion from scratch. The *calculus differentialis* is Leibniz’s great masterpiece,

and I intend to spend the rest of this book showing off its amazing power and versatility.

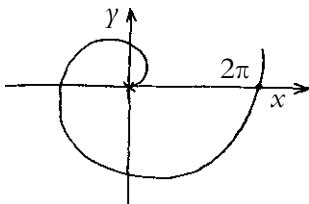
As an illustration of our new techniques, let's measure the velocity of a spiral motion.



The first question is, what exactly do we mean by spiral? I like to think of it as a point on the end of a rotating stick that gets longer as it turns. For simplicity, let's say that the rate of turning and the rate of lengthening are both uniform. In fact, let's take them both to be 1. If the stick were simply rotating, we could use the standard description for uniform circular motion: $x = \cos t$, $y = \sin t$. Because the stick is growing, this will have to be modified to

$$\begin{aligned}x &= t \cos t, \\y &= t \sin t.\end{aligned}$$

This is the hard part—choosing a model and setting up a coordinate system. Now our spiral motion begins at $(0, 0)$ when $t = 0$ and travels counterclockwise.



At time $t = 2\pi$, our point has position $(2\pi, 0)$. How fast is it traveling? Applying d to our equations gives

$$\begin{aligned} dx &= (-t \sin t + \cos t) dt, \\ dy &= (t \cos t + \sin t) dt, \end{aligned}$$

so we get a velocity vector (\dot{x}, \dot{y}) equal at all times to $(\cos t - t \sin t, \sin t + t \cos t)$. In particular, at the end of one rotation (when $t = 2\pi$), we have a velocity of $(1, 2\pi)$ and hence a speed of $\sqrt{1 + 4\pi^2}$.

Show that this spiral motion has the same speed at all times as the parabolic motion $x = t$, $y = \frac{1}{2}t^2$.

21

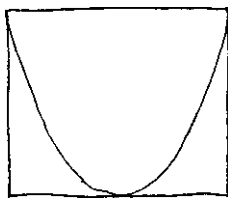
We now have a simple and reliable way to measure the relative rates of a set of interrelated numerical variables. In particular, we have completely solved the problem of velocity. Given any motion (that is, a set of time-dependent variables and equations expressing this dependence), we can simply apply the Leibniz d -operator to these equations and, using our differential calculus, obtain the velocity components \dot{x} as ratios dx/dt .

This alone would be more than enough justification for all the effort (both conceptual and technical) that went into the development of the differential calculus, but the fact is that not only velocity but virtually *all* measurement problems can be expressed in the language of variables and differentials,

and the differential calculus allows us to solve a great many of them quite easily. (Of course, I would argue that the real justification lies in the beauty and profundity of the ideas themselves.)

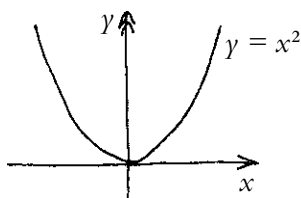
In particular, it was discovered quite early in the development of the differential calculus that these methods can even be applied to the problems of classical geometry—that is, the measurement of angle, length, area, and volume. In many ways, this is quite surprising. After all, differentials are instantaneous rates of variable quantities, whereas geometric measurements are fixed and static.

A while back, I was telling you about Archimedes's measurement of the parabola. The very beautiful discovery was that a parabolic section always takes up exactly two-thirds of its box.

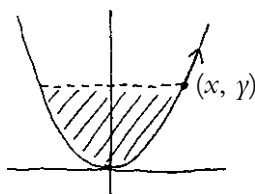


Archimedes proved this using the method of exhaustion, chopping the parabolic region up into approximating triangles and rearranging them cleverly. It is a masterpiece of classical technique, and the details are fairly intricate. Now I want to show you a different way.

Let's set up the usual coordinate description of the parabola. The axes are chosen with respect to the symmetry of the parabola, and the units are chosen to give the simplest possible algebraic description, $y = x^2$.

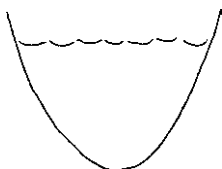


So the shape information is now in the hands of two abstract numerical variables x and y and the relationship between them. Now here is the key idea. Instead of choosing some particular chopping place to create our parabolic area, we imagine the point at which we slice the parabola to be moving, so that the area is *variable*.



As the point moves along the parabola, the area of the enclosed region changes. Our problem becomes not merely the determination of one particular parabolic area, but the measurement of *all* such areas. In other words, we are interested in the relationship between where we chop the parabola and how much area is enclosed.

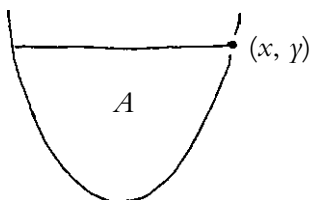
If you like, we can think of the parabola as a bowl slowly filling with liquid.



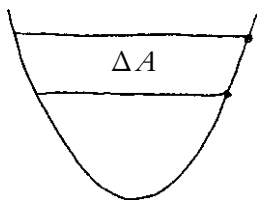
As the level of the liquid increases, so does its area (this is imaginary two-dimensional liquid), and our question becomes, how does the area depend on the height?

The important point is that now that the area is variable, it has a *rate*. In place of a cold, dead area just sitting there, we have an active, exciting area, which, as a consequence of its fluent nature, possesses a differential. Let's see if we can get our hands on it.

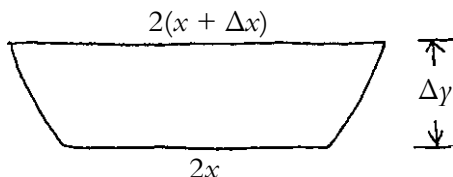
Going back to our coordinate description, we see that at any moment in the life of this motion we have three related variables: x , y , and the area A . The problem is to determine the precise relationship among them.



If we give our picture a little kick, the point moves, x and y change slightly, and so does A .



The small change ΔA appears as a thin sliver of area. How big is it? Intuitively, we would say that it is about the same size as a rectangle of the same height and width; that is, about $2x\Delta y$.



More precisely, this sliver (being curved) must be slightly larger than the inside rectangle, whose area is $2x\Delta y$, and slightly smaller than the outside rectangle with area $2(x + \Delta x)\Delta y$. This means that

$$2x\Delta y < \Delta A < 2(x + \Delta x)\Delta y.$$

Of course, all three of these quantities are approaching zero, but their relative proportions are not. In particular, we have $\Delta A/\Delta y$ sandwiched between two values,

$$2x < \frac{\Delta A}{\Delta y} < 2x + 2\Delta x.$$

Since both the upper and lower bounds approach the same thing, namely $2x$, it must be that $\Delta A/\Delta y$ does also. That is,

$$\frac{\Delta A}{\Delta y} \rightarrow 2x.$$

On the other hand, $\Delta A/\Delta y$, being the ratio of small changes in the variables A and y , must also approach the true proportion of their differentials. Thus $dA/dy = 2x$. Multiplying by dy , we obtain a differential equation for the area of a parabolic section:

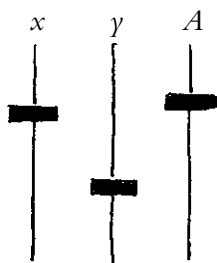
$$dA = 2x \, dy.$$

In essence, our rectangular approximation $\Delta A \approx 2x\Delta y$ becomes, at the moment of vanishing, an exact equality $dA = 2x \, dy$.

This tells us how the area A depends on x and y , but only indirectly. We are being told information about the variables via their differentials. This is the *opposite* of the velocity problem. Instead of a known motion whose velocity we seek, here we have the rate information, and we want to recover the relationship between the variables themselves. (The fact that area measurement is inversely related to velocity—which is in turn connected to slantedness—was discovered early in the seventeenth century by Fermat and others, but waited upon the development of the differential calculus to be put into clear focus.)

So we find that to measure the area of a parabolic section (and indeed any area), we are reduced to solving a differential equation. That is, we need to determine what precise relationship A must have to x and y so that when we d it we get the above relation among the differentials.

It turns out (at least in this case) that this is not particularly hard to do. Thinking about it abstractly, we have three variables and two relationships:



$$y = x^2$$

$$dA = 2x \, dy$$

The first equation indicates the shape we are measuring, and the second comes from our geometric reasoning. At this

point, we can forget about origins and motivations and view the problem as a purely abstract question about three variables. How do we express A in terms of x and y ?

The first step might be to eliminate y from the discussion. After all, we know what it is, namely x^2 . So we can rewrite our differential equation as

$$\begin{aligned} dA &= 2x \, dy \\ &= 2x \, d(x^2) \\ &= 4x^2 \, dx. \end{aligned}$$

So now we have a purely technical question about differentials. What must A be so that $dA = 4x^2 \, dx$? Instead of asking how to d a particular combination of variables, we're asking how to un- d something!

This is a major theme in mathematics. Anything interesting enough to do is almost always interesting enough to undo. Adding gives rise to the desire to un-add (i.e., subtract), squaring leads to square-rooting, and so forth. The reason for this is linguistic. Any operation or process that can be described invariably leads to an expansion of language. Questions can then be asked using this new expressive ability (what do I add to 7 to make 11?), and our curiosity inevitably leads us to invert the process. Whenever you tie a knot, you immediately create the desire (or at least the possibility of desire) to untie it. From this viewpoint, we can see the essentially inverse relationship between arithmetic and algebra, for instance.

In any case, we have an interesting practical and philosophical problem: How do we un- d something? It turns out, contrary to the case of d -ing itself, that there is no calculus for this. That

is, there does not (and cannot) exist a systematic, step-by-step procedure for determining the solution to a differential equation. Which is not to say that there aren't many cases (including our present one) where we *can* succeed, just that there is no universal formula for success.

So solving differential equations is something of an art. Imagination and intuition play at least as large a role as technical facility. That is both sad (in that we will not be able to solve many of our most interesting problems) and also quite fascinating. No matter how clever we are, and how tightly we grasp, mathematics always manages to squirt out between our fingers.

This is a lot like our experience with numbers. We have no problem squaring any fraction we choose, but square-rooting leads to new numbers not describable in that language. The same goes for the d -operator. Only in the most fortunate circumstances are we able to explicitly determine the solutions to differential equations. Most of the time we will be resigned to implicit descriptions (e.g., “the number whose square is 2”).

Luckily, in the case of the parabola, this is not one of those times. We can in fact solve the differential equation $dA = 4x^2 dx$, and so determine the area of a parabola. (Of course, we know this has to be possible since Archimedes did it!) Not only that, but I can even offer a general methodology for solving differential equations—not one that will always (or even often) be successful, to be sure—but here it is: *guessing*. Not random, out-of-the-blue guessing, but intelligent, conscientious guessing, informed by experience and sensitivity to pattern. Needless to say, the more experience one has with the differential calculus, the better one will be at making good guesses.

For instance, experience tells me that the un-*d*-ing of $4x^2 dx$ will probably involve x^3 , since I know that $d(x^3) = 3x^2 dx$. And in fact this tells me exactly what to do. Since what I want ($4x^2 dx$) is just a constant multiple of what I have ($3x^2 dx$), all I need to do is adjust my guess accordingly. That is, I should *modify* my guess to $\frac{4}{3}x^3$. Sure enough,

$$\begin{aligned} d\left(\frac{4}{3}x^3\right) &= \frac{4}{3} \cdot 3x^2 dx \\ &= 4x^2 dx. \end{aligned}$$

So we can rewrite our differential equation for parabolic area as

$$dA = d\left(\frac{4}{3}x^3\right).$$

Of course, I would love to be able to conclude that A itself must then be exactly equal to $\frac{4}{3}x^3$, and this will in fact turn out to be the case, but we have to be a little careful with our reasoning here. Just because two variables have the same differential does not make them equal—a car and its trailer have the exact same speed at all times, but not the same position. This is again just like the squaring and square-rooting situation: both 4 and -4 have the same square, but they are nevertheless unequal. The point is that *d*-ing, like the squaring operation, *loses information*. So whenever we invert these processes, there is always a certain amount of ambiguity.

In particular, since $d(c) = 0$ for any constant c , it will always be true that $d(w)$ and $d(w + c)$ are indistinguishable. In other words, two variables that are off by an additive constant will always have the exact same differential. Is that the *only* way that two variables can have the same differential? If $dw = 0$,

does that mean that w must be constant? Of course it does! The equation $dw = 0$ means that w is *not moving at all*. So if two variables a and b have the same differential ($da = db$), then

$$d(a - b) = da - db = 0,$$

and this means that $a - b$ must be constant. So there is some ambiguity in un- d -ing, but not too much. Just as a number has two square roots, a differential has an infinite number of un- d -ings, all of them differing from each other by additive constants.

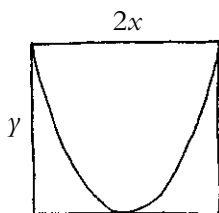
So we cannot *directly* conclude from our differential equation $dA = d(\frac{4}{3}x^3)$ that A itself equals $\frac{4}{3}x^3$, but we do know that at worst they differ by a constant. That is,

$$A = \frac{4}{3}x^3 + \text{constant}.$$

That is the most we can say from the differential equation alone. There is no way to rule out a possible constant on differential grounds, just as there is no way to tell from the speedometer alone whether you are in the car or the trailer. This ambiguity comes from the fact that our geometric argument only considered the *change* in the area, not where we started measuring it from.

But in fact we do have slightly more information—namely, we have a so-called *initial condition*. At the bottom of the parabola, we clearly have both x and A equal to zero. Since the above equation expresses a relationship between our variables which is valid at all times, it must hold at this particular moment as well. This implies that our (putative) constant must in fact be zero. So we can conclude that $A = \frac{4}{3}x^3$ after all.

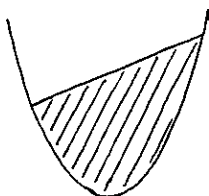
Generally speaking, there are two parts to solving a differential equation: intelligent guessing with modification to get a so-called generic solution and then using special values of the variables—typically initial conditions—to determine any ambiguous constants.



Going back to our coordinate picture, we see that the rectangle containing our parabolic region has area $2xy = 2x^3$. Thus the parabola-to-box proportion is

$$\frac{\frac{4}{3}x^3}{2x^3} = \frac{2}{3}.$$

Since this ratio is independent of any units, we can throw away all the scaffolding—coordinate systems, variables, equations, and all—and simply say (along with Archimedes) that a parabolic section always takes up two-thirds of its box.



What about the area of
a slanted parabolic section?

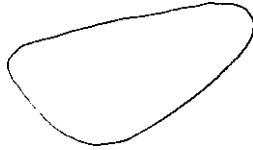
22

Let's step back a little bit and think about what just happened. I don't want the big ideas to get lost in the computational details. The point is that we can apply our differential methods even to something as seemingly static as a geometric measurement. The key idea is this: *get your measurements moving*. Every application of the differential calculus—to geometry, mathematical physics, electrical engineering, and anything else—comes down to this one idea. If you want to measure something, wiggle it. Once a measurement is in motion, it has a rate of motion, and if we are at all fortunate (and we usually are), we can derive some sort of differential equation describing the way our measurement behaves.

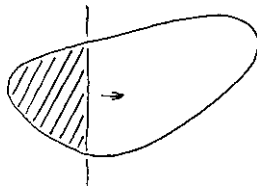
What this means is that the study of measurement ultimately reduces to the study of differential equations (a possible exception being the measurement of polygons, i.e., trigonometry, where simpler methods are available). Questions concerning the existence and uniqueness of solutions (as well as their ability to be explicitly described) dominated the mathematics of the eighteenth century and continue to be an active area of mathematical research.

The method we used to obtain a differential equation for the area of the parabola is actually quite general. First, we found a simple way to view the area we wanted as a variable quantity; that is, we got it moving. Then, we estimated the change in area in terms of the changes in the coordinate variables. Finally, we let the small changes approach zero so that our approximation became an exact statement about instantaneous rates; that is, a differential equation.

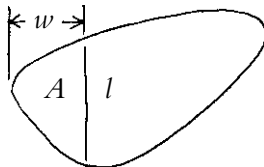
Suppose, for instance, that we had a closed curve whose area we wanted to measure.



A simple way to get the area moving is to choose a direction and “sweep out” the area in that direction, as though we were putting the curve through a scanner:

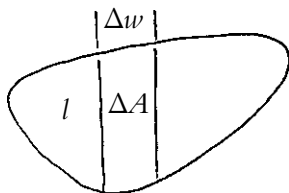


In this way, the variable area depends on the location of the scanning line. Let's denote the position of the line (i.e., the width of the area collected so far) by w and the swept-out area by A . At any given moment, we have a certain cross-sectional length, say l , and as the scanner moves along, we get variations in w , l , and A .



Of course, the way that w and l are related depends on the shape of the curve (in fact, it practically *defines* the shape of the curve). If we make a slight change in the scanner position, say

from w to $w + \Delta w$, we get corresponding changes in the length l and the area A .



When these changes are extremely small, the region of area ΔA is essentially a thin rectangular strip of width Δw and height l . So we have an approximation

$$\Delta A \approx l \Delta w.$$

Another way of saying this is that $\Delta A / \Delta w$, being in some sense the “average” cross-sectional length during this small scanning interval, must be roughly equal to l . Of course, as the small changes approach zero and the thin sliver of area ΔA gets thinner, this average length approaches l exactly. So we get a differential equation of the form

$$dA = l \, dw.$$

What this says is that the rate of change in the scanned area is just the product of the cross-sectional length and the rate of the scanning motion. Does this remind you of the Pappus philosophy? Leibniz’s own view was that areas are comprised of infinitely many infinitesimally thin rectangles, so that the above differential equation is essentially an infinitesimal version of the “length times width” formula for a rectangular area.

However you wish to interpret it, the above equation is quite general; we can say this for any curve and any scanning direction. Because of this, it is up to us to choose our orientations wisely so that we get the simplest differential equation possible. (In particular, our choices will determine the precise form of the relationship between w and l , which will seriously—and subtly—affect our ability to solve such an equation.)

A nice example of this method is the measurement of the area of a **sinusoidal arch**; that is, one of the humps in the graph of the relation $y = \sin x$.



Here it makes good sense to sweep horizontally so that the position of the scanning line is simply the coordinate x itself (running from 0 to π) and the cross-sectional length is just $\sin x$. Then our differential equation for area reads

$$dA = \sin x \, dx.$$

A reasonable guess for a solution is $A = \cos x$, but actually (according to our calculus), we have $d(\cos x) = -\sin x \, dx$, so we're off by a minus sign. So $A = -\cos x$ does the job. Of course, we could still be off by an additive constant, and in fact the initial condition $x = A = 0$ tells us that the constant must be 1. That is, we find the area to be

$$A = 1 - \cos x.$$

This tells us what the swept-out area is for *all* positions of the scanning line. In particular, when $x = \pi$, we get the nice result that the area of a complete arch is exactly

$$1 - \cos \pi = 1 - (-1) = 2.$$

How beautiful! I've always found this result surprising (and somewhat ironic, given the transcendental nature of the sine function).

I think it is important to understand the close connection between these techniques and the classical method of exhaustion.

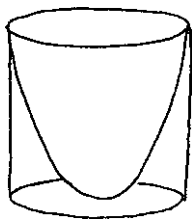


The classical idea would be to choose a direction and slice up our area into tiny approximating rectangles. If we got incredibly lucky, we might notice a pattern to the approximations and be able to figure out where they are heading. With the differential approach, we don't need to be clever at all; we simply write down the equations and let the d -operator do all the pattern bookkeeping for us. The difficulty is transferred from the polygons and the details of the approximation pattern to the solving of a differential equation. This is almost always a trade worth making. Even forgetting about the technical details, at least the differential method is completely uniform, whereas in

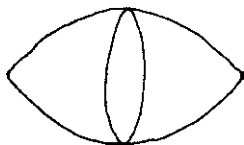
the classical case, each new shape has to be handled in its own ad hoc and idiosyncratic way.

So the “rocks and symbols” analogy is really quite apt. The classical method of exhaustion is like dealing with massive piles of rocks, and the differential calculus is like adding columns of digits—so much so that we can even build machines to do the computations for us. This fits into the larger historical trend in mathematics known as the *arithmetization of geometry*. Shapes become number patterns, and their measurements are governed by differential equations.

Can you find a differential equation for
swept-out volume? How does this
compare with the Cavalieri principle?

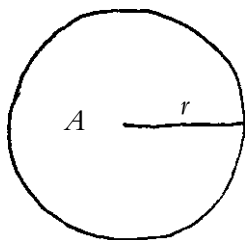


Show that a paraboloid (a rotated parabola)
takes up exactly half its cylinder.

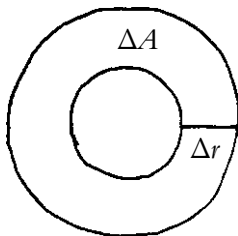


Can you measure the volume
of a rotated sinusoidal arch?

As a final example, let me show you an amusing way to determine the area of a circle (this is, after all, the prototypical example of classical exhaustion). I know we already understand the circle (as much as it can be understood, anyway), but my point is to show that our new methods can give us fresh perspectives on old problems. In this case, instead of sweeping out area, I'm going to grow area out from the center.



So both the radius r and the area A will be variables. In this case, small changes lead to a circular ring of area.



When Δr is very small, this ring can be unfolded (approximately) to form a rectangle of width Δr and length equal to the circumference $2\pi r$. Thus $\Delta A \approx 2\pi r \Delta r$, so we get the differential equation

$$dA = 2\pi r \, dr.$$

This, together with the initial condition $r = A = 0$, tells us that A must be πr^2 , as expected.

Can we do something like this to relate
the measurements of a sphere?

23

So all the subtle and beautiful variety of shapes and motions and all the fascinating questions concerning their measurements can be reduced to the problem of applying and inverting the Leibniz d -operator. Whereas d itself transforms variables into differentials, many of the most interesting measurements (such as area and volume) involve doing the opposite. So in many ways un- d -ing is the more interesting process, especially since we have a calculus for d itself.

Leibniz, of course, had a somewhat different interpretation. He considered dx not to be the instantaneous *rate* of x (although he certainly understood Newton's theory of fluxions perfectly), but rather (and somewhat more mystically) as the *infinitesimal change* in x . A useful analogy here might be to think of x as a list of numbers—a so-called *discrete* variable:

$$x: \quad 0, 1, 3, 2, 5, 6, 4, \dots$$

Then dx is analogous to the list of consecutive differences, or jumps:

$$dx: \quad 1, 2, -1, 3, 1, -2, \dots$$

The un-*d*-ing process could then be thought of as a way to go from these differences back to the original list. Clearly, the way to do this is with *running totals*. That is, if we add up the first however many differences, we get back the numbers we started with. Notice that the usual ambiguity is present: if all the x numbers are shifted by some amount (say we add 3 to all of them), the differences are unaffected. So the running totals won't necessarily give us back the original numbers exactly, but at worst, we'll be off by a constant shift.

So Leibniz thought of un-*d*-ing as a sort of summing operation, albeit a very strange one—the idea being that we are “smoothly” adding infinitely many infinitesimal differences dx to recover our original varying quantity x . For this reason, he was led to introduce the symbol \int for the un-*d*-ing operator (it's a fancy capital *S* from the Latin word *summa*). In any event, the notation—like that for square root—is pretty handy and certainly doesn't do any harm.

In fact, the analogy with square root is a good one—which is why I keep making it! Suppose we have a number x and we know that $x^2 = 16$. Rewriting this as $x = \sqrt{16}$ doesn't particularly change anything, but it does provide an easy-to-use abbreviation and gives us the option of referring to our number as “the square root of 16” as opposed to “that which when squared is 16.” Of course, in this case we could also say that our number must be either 4 or -4 .

Similarly, if we had variables x and y related by a differential equation, say

$$dy = x^2 dx,$$

we could, using Leibniz's notation, rewrite this in the possibly more psychologically satisfying form,

$$y = \int x^2 dx.$$

By the way, this is usually read as “the integral of $x^2 dx$ ” rather than using the older word *summa*. (The word *integral* comes from the Latin *integer*, meaning “whole.”) Leibniz's symbol is known as the *integral sign* and the un-*d*-ing process is usually referred to as *integration*.

In this particular case, we can guess and modify to obtain the result

$$\int x^2 dx = \frac{1}{3}x^3 + \text{constant}.$$

Now in practice, most people use both the square root sign $\sqrt{}$ and the integral sign \int in a somewhat cavalier manner. That is to say, when I write $\sqrt{16} = 4$, I am quite aware of the fact that -4 is also a square root of 16. Sometimes I might even remind myself of this possibility by writing $\sqrt{16} = \pm 4$. Similarly, I will often write things like

$$\int x^2 dx = \frac{1}{3}x^3,$$

knowing full well that there is a potential additive constant. The same goes for any operation that collapses information; if several different numbers all go to the same place, the inverse operation will carry a certain amount of ambiguity. How you

deal with that fact notationally is your business, but confusing things can happen if you aren't careful!

Anyway, most working mathematicians use the integral sign (at least in this context) to mean *any* variable with the prescribed differential, and the ambiguity is not usually explicitly written, though, of course, it is understood. When one speaks of “*the* square root of 16” or “*the* integral of $x^2 dx$ ” one needs to understand this somewhat professional meaning of the word *the*.

So the art of measurement pretty much comes down to understanding the behavior of the \int -operator. As I mentioned already, this turns out not to be so simple. Which is not to say that we know nothing. Over the last 350 years, people have discovered and compiled hundreds of patterns and formulae in the form of so-called integral tables, which, in effect, give us a sort of integral calculus (albeit a rather humiliating one, since many of the most interesting and naturally occurring differentials do not appear).

Continuing the square root analogy, it does, of course, happen that one gets lucky (e.g., $\sqrt{16}$) and can rewrite an expression in a more explicit form, but most of the time, as in the case of $\sqrt{2}$, it is not a question of finding a simpler form; the number itself is simply not expressible in the language you wish to use.

Similarly, most integrals are not expressible in terms of so-called elementary operations (e.g., addition and subtraction, multiplication and division, square roots, sine and cosine). For example, the integral

$$\int \sqrt{1 + \sin^2 x} \, dx$$

is certainly “out there” as a variable, depending on x in some definite way (at least up to the usual ambiguous constant), but that dependence is *provably not describable* in terms of algebraic and trigonometric patterns. This is a pretty spectacular example of the power of modern mathematics that we can even devise such arguments (and, of course, I can’t explain them to you here, which is admittedly rather frustrating).

So we are in a very amusing position philosophically. Just considering the area of a closed curve, for instance, we have first of all the rather humbling state of affairs that almost all curves are indescribable *in principle* (because they have no pattern that can be encoded in a finite language), and then on top of that, even the ones that can be talked about (i.e., the ones that can be described by a set of variables and relations) almost always lead to differential equations whose solutions are not explicitly describable. We have this amazingly beautiful and powerful theory of differentials (including a *calculus* for crying out loud), but the powers that be (the mathematical gods?) have decreed that we are only to have definite explicit knowledge in the tiniest fraction of cases.

On the bright side, at the very least we have a uniform language for measurement description, and through this means we are able to make connections and see relationships between measurements, even if we are forbidden from knowing them explicitly. In particular, if two seemingly unrelated problems lead to the same differential equation, then even if we cannot solve it, we still know there is a deep underlying connection between them. This is, ultimately, the only real value of linguistic constructs and the only thing that conscious beings can *ever* do with language, if you think about it.

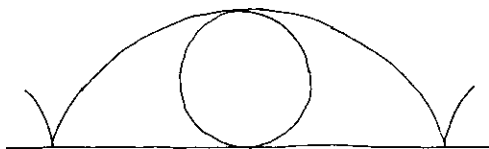
What is the connection between the area of
a parabola and the volume of a pyramid?

Can you solve the differential
equation $2x dy = y dx$?

24

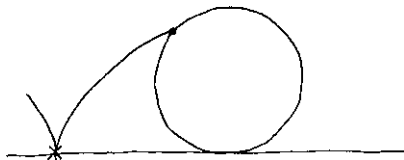
I love the contrast between the ancient and modern approaches to geometric measurement. The classical Greek idea is to hold your measurement down and chop it into pieces; the seventeenth-century method is to let it run free and watch how it changes. There is something slightly perverse (or at least ironic) about how much easier it is to deal with an infinite family of varying measurements than with a single static one. Again, the trick is to figure out a way to get your measurement moving.

Of course, this is particularly easy to do when your problem involves motion—it's not very hard to get things moving if they're moving already. As an illustration of this idea, let's try to measure the length of a cycloid.

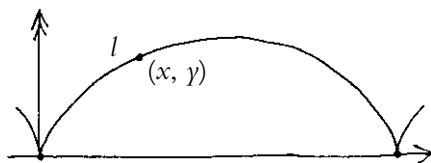


The natural measurement would be the length of one complete arch of the cycloid (let's say compared to the diameter

of the rolling circle). The classical approach would be to chop up the cycloid into tiny pieces, approximate them by straight lines, and try to figure out where the approximate total length is heading. (This approach was in fact carried out by Bernoulli and others in the 1630s.)



If instead we could somehow view the length as a variable, we could then use modern differential methods. But *of course* the length is variable—a cycloid is formed by a rolling circle! So at each moment the moving point has traced out a certain amount of length.

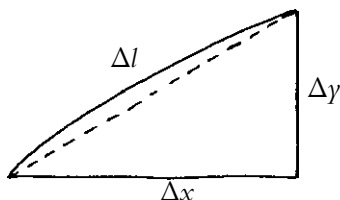


Going back to our coordinate description of the cycloid, we have variables t , x , and y related by the equations

$$\begin{aligned}x &= t - \sin t, \\y &= 1 - \cos t.\end{aligned}$$

At any time t , our moving point is located at the position (x, y) . Let's call the traced-out length l . Our problem is to determine how l depends on t . As usual, the idea is to obtain a differential equation for l .

Let's imagine a very small amount of time going by and consider the small changes in x , y , and l .



When these changes are very small, the length Δl is practically the hypotenuse of the right triangle formed by Δx and Δy . (This is the way the classical idea still comes into play.) The Pythagorean relation then gives us the approximation

$$(\Delta l)^2 \approx (\Delta x)^2 + (\Delta y)^2.$$

Letting the time interval Δt approach zero, we get the desired differential equation,

$$dl^2 = dx^2 + dy^2.$$

It has become customary, by the way, to write dx^2 in place of the more cumbersome $(dx)^2$. We just have to be careful not to get dx^2 confused with $d(x^2)$. Of course, you can always use parentheses if you are worried about it.

So we find a sort of “infinitesimal” Pythagorean relation, which expresses the differential arc length dl in terms of its horizontal and vertical components. Another way to think about it is that since dl/dt measures the rate at which distance is traversed, it must be the same as the *speed* of the moving point; that is, the length of the velocity vector (\dot{x}, \dot{y}) . We get

$$\frac{dl}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2},$$

which is again saying that $dl^2 = dx^2 + dy^2$.

This Pythagorean relation is quite general: it applies to any arc length in the plane, whether it is traced out by a motion or not.

What about arc length in three-dimensional space?

Applying this to our cycloid, we get

$$\begin{aligned} dl^2 &= dx^2 + dy^2 \\ &= (d(t - \sin t))^2 + (d(1 - \cos t))^2 \\ &= (1 - \cos t)^2 dt^2 + \sin^2 t dt^2, \end{aligned}$$

and so

$$dl = \sqrt{(1 - \cos t)^2 + \sin^2 t} dt.$$

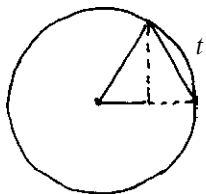
This is the differential equation we need to solve in order to measure the length of a cycloid.

On the face of it, things look pretty bleak. How on earth are we going to integrate such a complicated mess of a differential? The sad truth of the matter is that because of the complexity of the Pythagorean relation (squaring, adding, and then square-rooting), arc lengths almost always lead to integrals that cannot be expressed in the language of elementary functions.

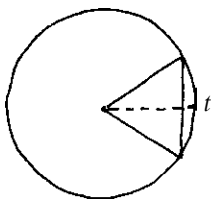
Luckily for us, the cycloid is an exception. It turns out that the expression

$$\sqrt{(1 - \cos t)^2 + \sin^2 t}$$

can be rewritten in a very simple and elegant way. Imagine a circular arc of length t .



We can then view $1 - \cos t$ and $\sin t$ as the sides of a right triangle with hypotenuse $\sqrt{(1 - \cos t)^2 + \sin^2 t}$. In other words, the thing that we are interested in is exactly the length of the chord spanning an arc of length t . (We saw this before when we measured the velocity of the cycloid motion.) Now here's the clever idea: rotate the circle so that this chord is *vertical*.



Now we can see that the chord consists of two halves, each of which is just the sine of an arc *half as long*. That is, the length of our chord can also be written as $2 \sin \frac{t}{2}$. So a simple change of perspective (and isn't that what every great idea comes down to?) leads to the surprising and beautiful result that

$$\sqrt{(1 - \cos t)^2 + \sin^2 t} = 2 \sin \frac{t}{2}.$$

There are many such interrelationships between sine and

cosine, all of them ultimately coming from the symmetry and simplicity of uniform circular motion.

Use this to derive the half-angle formulas

$$\sin^2 \frac{t}{2} = \frac{1}{2} (1 - \cos t),$$

$$\cos^2 \frac{t}{2} = \frac{1}{2} (1 + \cos t).$$

We can now rewrite our differential equation for the arc length of a cycloid as

$$dl = 2 \sin \frac{t}{2} dt.$$

Now, that's more like it! Here is something we have a real chance of being able to integrate. In fact, a reasonable guess would be $l = -\cos \frac{t}{2}$. Now

$$\begin{aligned} d(-\cos \frac{t}{2}) &= \sin \frac{t}{2} d(\frac{t}{2}) \\ &= \frac{1}{2} \sin \frac{t}{2} dt. \end{aligned}$$

So we are off by a factor of 4. Thus we find

$$\int 2 \sin \frac{t}{2} dt = -4 \cos \frac{t}{2}.$$

Of course, we could still be off by an additive constant. Checking the initial values $t = l = 0$, we see that in fact we must have

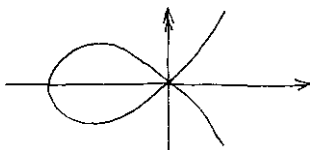
$$l = 4 - 4 \cos \frac{t}{2}.$$

And so there it is! We have successfully measured the arc length variable of a cycloid using the differential calculus (plus a pretty clever idea about circles). In particular, a full arch (from $t = 0$ to $t = 2\pi$) has length

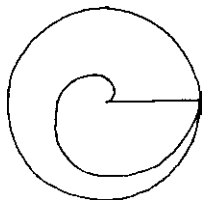
$$4 - 4 \cos \pi = 8.$$

Incredible! The length of a cycloid arch is exactly four diameters. It would be hard to find a more beautiful measurement than that. Unless perhaps it's the *area* of a cycloid.

Show that the area of a cycloid arch is exactly three times the area of the rolling circle.



Find the length and area of the closed loop in the nodal cubic curve $3y^2 = x^3 + x^2$.



Show that the area swept out by one full turn of a spiral takes up exactly one-third of the corresponding circle.

25

I'm going to have to ask you to bear with me once again while I make a few philosophical remarks. The big idea here is that geometry—the study of size and shape—can be subsumed into the study of variables (also known as **analysis**). It is always interesting when seemingly quite different mathematical structures turn out to be the same. As I've said before, the real object of interest to mathematicians is pattern. If you wish to view such a pattern geometrically, that might give you a certain kind of insight; whereas if you think of it as a set of abstract numerical variables, that may lead to another sort of understanding—and certainly the two viewpoints feel very different emotionally.

The curious thing is why history went the way it did, and why the modern approach has been so much more successful. The classical Greek geometers were every bit as brilliant and resourceful as their seventeenth-century counterparts (if not more so). It's certainly not a question of mathematical talent. There are plenty of reasons why the Greeks preferred direct geometric reasoning, aesthetic taste, of course, being one of them. In fact, this prejudice was taken to such an extreme that numbers themselves tended to be viewed geometrically (as lengths of sticks), and numerical operations were thought of as geometric transformations (e.g., multiplication as scaling). This severely hampered their understanding.

The modern approach is almost the exact opposite. Curves and other geometric objects are replaced by numerical patterns, and the problem of measurement essentially becomes the study of differential equations. Why, if these two viewpoints are

equivalent, should one of them be so much more powerful and convenient?

There is no question that as visual animals we prefer a picture to a string of alchemical symbols. I, for one, want to feel connected to my problem on a visceral, tactile level. It helps me understand the relevant issues when I can imagine running my hand over a surface or wiggling part of an object and picturing in my mind's eye what happens. But I know that when push comes to shove, the truth is in the details, and the details are in the number pattern.

Of course, any analytic argument could be painstakingly translated into purely geometric terms, and in fact, this is the way many seventeenth-century mathematicians worked; even then there was still a great deal of prejudice in favor of geometric reasoning. This tends, however, to produce very contorted and artificial explanations in place of concise, almost-too-simple-to-believe analytic arguments.

I suppose what I'm really talking about here is modernism. The exact same issues—abstraction, the study of pattern for its own sake, and (sadly) the resulting alienation of the layperson—are all present in modern art, music, and literature. I would even venture to say that we mathematicians have gone the furthest in this direction, for the simple reason that there is nothing whatever to stop us. Untethered from the constraints of physical reality, we can push much further in the direction of simple beauty. Mathematics is the only true abstract art.

For me, the psychological fact of the matter is that however aesthetically and emotionally satisfying the geometric view may be, the analytic approach is, in the end, far more elegant and powerful. We've already seen a number of instances of

this—the increased descriptive power, the advantages of a uniform language that reveals hidden connections, and the ease of generalization. For instance, the classical geometers (as far as I know) never even *conceived* of four-dimensional space, whereas adding another variable is an obvious and natural analytic extension.

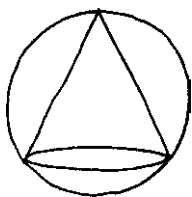
Which is not to say that I am advocating the abandonment of the geometric viewpoint. Obviously, the greatest mathematical pleasure is to be had by *synthesizing* different approaches—to be fluent and comfortable with as many as you can and to inform each part of your mathematical self via the others. Think geometrically when a visual image is helpful (usually to get a big idea or an intuitive connection) and work analytically when that seems appropriate (usually to make a precise measurement).

Maybe it all comes down to this. There are lots of beautiful patterns out there. Some, such as a triangle taking up half its box, can be easily seen and felt; others, like $d(x^3) = 3x^2 dx$, are not so immediately available to our visual imagination. So be it; I myself want to be open to *all* forms of beauty. For me, that's what being a mathematician is all about.

26

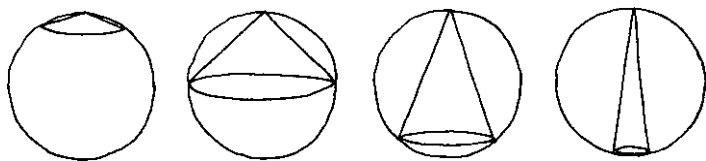
Now I want to tell you about another fantastically beautiful and powerful application of the differential calculus, possibly the most useful in practical terms.

Imagine a cone sitting inside a sphere.



If I knew the measurements of this cone—its height relative to the diameter of the sphere, for instance—it would then be a relatively simple matter to determine its volume. But what if I wanted the largest such cone? Then I don't know its measurements; I only know I want the volume to be as big as possible. Instead of the shape being fixed and wanting to measure its size, here the shape itself is variable.

In fact, we can imagine an entire spectrum of possible cone shapes, from a tiny flat cone snuggled up at the North Pole, all the way to a pointy icicle through the center.

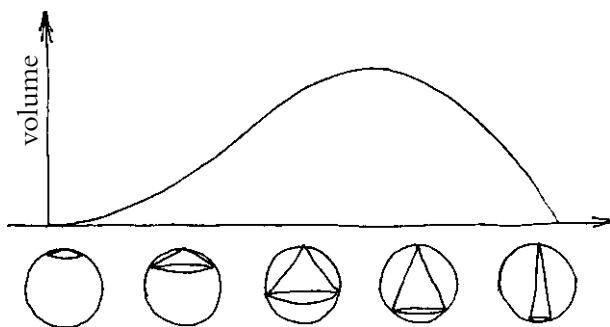


Clearly the best cone (in the sense of maximizing volume) lies somewhere in between. Intuitively, it feels to me like the base of the cone should be slightly below the equator of the sphere, but it's certainly not obvious exactly where.

These kinds of questions—where we're trying to maximize (or minimize) a particular measurement—have a long history and are known as *extremal problems*. For example, the Babylonians knew that among all rectangles with a given perimeter, the square has the most area. Here is a related problem for you to think about:

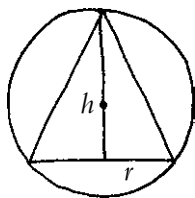
**Among all three-sided rectangles of a given
length (like a fence built against a wall),
which one encloses the most area?**

In the abstract, an extremal problem concerns the dependence of one variable (the measurement) on another (the shape). In the case of the cone inside the sphere, we can imagine a graph of this dependence.

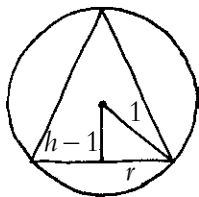


As the height of the cone increases, so does its volume, until it starts becoming detrimental to be so tall and thin, and then the volume decreases down to zero again (I am including the extreme cases of a single-point “cone” of zero volume on the left and a stick of zero volume on the right). Anyway, somewhere in the middle is the cone we want—a little to the right of the middle, if my intuition is correct.

To make this more precise, let’s construct a “variables and relations” model of the situation. (As always, this is the hard part.) Let’s begin by taking the radius of the sphere as our unit (at least *that’s* not changing!), and we’ll denote the height and radius of the cone by h and r respectively. Slicing the sphere vertically through the center, we see this cross-section:



The geometrical constraint on our cone is that it fit snugly in the sphere. This means that h and r must be related somehow. In fact, we can see that the distance from the center of the sphere to the base of the cone is just $h - 1$. (I guess I'm tacitly assuming the base lies *below* the equator, otherwise it would be $1 - h$.)



The distance from the center of the sphere to the edge of the cone is 1, so Pythagoras says that

$$(h - 1)^2 + r^2 = 1.$$

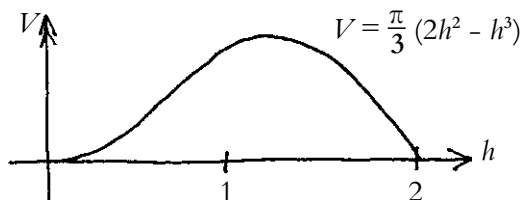
Note that (because of the squaring) we would get the same equation if the base of the cone were above the equator. I love it when that happens. So in either case we get

$$\begin{aligned} r^2 &= 1 - (h - 1)^2 \\ &= 2h - h^2. \end{aligned}$$

This tells us how the radius of the cone varies with the height. Now the volume of the cone is given by

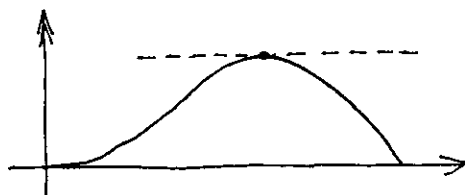
$$\begin{aligned}
 V &= \frac{1}{3} \pi r^2 h \\
 &= \frac{\pi}{3} (2h^2 - h^3).
 \end{aligned}$$

Now we can make our picture more precise.



Our question about cones has become an abstract numerical one: What value of h makes V the largest?

Imagine for a moment that this were the space-time picture of a motion (that is, h represents time and V the height of a ball, say). We are asking at what time the ball reaches its maximum height. The answer, of course, is *when its speed is zero*. Alternatively, we could say it is when the tangent line to the graph is horizontal.

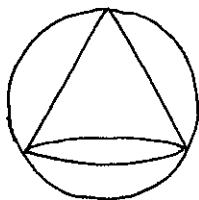


What I'm saying is that when a variable reaches an extreme (either a maximum or a minimum), the rate at which it changes must be zero. Otherwise, it would either still be climbing or already on its way down. So when a variable peaks, its differential must vanish. This is undoubtedly one of the simplest and most powerful discoveries in the history of analysis. Let's see what it says about our cone.

At the precise moment when h hits the right value to maximize V (the so-called *critical point*), we must have $dV = 0$. Now

$$\begin{aligned} dV &= \frac{\pi}{3} d(2h^2 - h^3) \\ &= \frac{\pi}{3} (4h - 3h^2) dh, \end{aligned}$$

and we see that this becomes zero precisely when $4h = 3h^2$. (Notice that we don't have to worry about the differential dh being zero, since at that moment the radius and height are still changing.) Thus we conclude that $h = 4/3$. So the largest cone is attained when the base is one-third of the way below the equator.



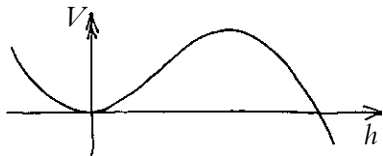
Is that great, or what! So the analytic approach to extremal problems is to find the exact moment when the differential vanishes. There are a couple of technical points here, actually. First of all, who says that there will be only *one* such moment? For example, we might have a dependence that looks like this:



Every one of the marked points is a place where the tangent is horizontal. So the vanishing of the differential occurs not just

at the extremes—the so-called global maxima and minima—but at the local ones as well, where the variable momentarily changes direction.

Going back to our cone problem for a second, let's take a slightly wider view of the situation.



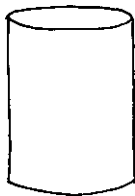
The relationship $V = \frac{\pi}{3}(2h^2 - h^3)$ is completely abstract. Sure, *we* know that we are talking about volumes and heights of cones, but V and h don't know that (nor do they care). In particular, there are values of the variables that correspond to no geometric situation (e.g., $h = -1$, $V = \pi$). There may even be moments where $dV = 0$ that have no bearing on the original problem and are merely artifacts introduced by the abstract viewpoint (there's probably a great modern art analogy here).

In our case, we can see that in fact there is a point other than $h = 4/3$ where $dV = 0$, namely $h = 0$. This happens because the relationship between V and h has a “bend” in it at that point, meaning that as h moves from positive to negative values, V moves toward zero and then rebounds. Of course, this is meaningless geometrically, since negative heights don't make any sense. Or do they?

Can you make geometric sense of a cone with negative height? How about negative volume?

The equation we got when we set $dV = 0$ was $4h = 3h^2$. Notice that $h = 0$ is a solution—one that we blissfully ignored. It is amusing that this point (corresponding to the single-point cone) is a local minimum of volume, but the other extreme (the vertical stick corresponding to $h = 2$) is not. Nevertheless, the stick is a point of minimum volume in our original problem. This annoying asymmetry is due to the fact that only the values $0 \leq h \leq 2$ make geometric sense, and in that range both $h = 0$ and $h = 2$ happen to be minima. In other words, there may be *boundary values* that are potential maxima and minima, as well as local extrema where the differentials vanish.

As another example—and one with obvious practical applications—let's try to determine the best shape for a soup can. By “best” I mean that it holds the most soup among all cans of a given surface area, using a fixed amount of metal. Of course, I'm not really talking about soup and metal, but *cylinders*.



The shape of a cylinder is determined by its radius r and height h , and its volume and surface area are given by

$$\begin{aligned}V &= \pi r^2 h, \\S &= 2\pi r h + 2\pi r^2.\end{aligned}$$

(I'm including the top and bottom lids of the can, of course.) The meaningful range of variation runs from a flat can ($h = 0$)

to a stick ($r = 0$). All the while, the surface area S is being held fixed. This means that r and h are connected. If I wished, I could even write

$$h = \frac{S - 2\pi r^2}{2\pi r},$$

and express everything in terms of the single variable r . But it so happens that I do not wish. Instead, I want to show you another way to proceed that I feel is more elegant.

The idea is this. Since S is constant, we must have $dS = 0$ at all times. Since we want the moment when volume is maximized, we must have $dV = 0$ at that instant. In particular, at the moment of interest, we will have both $dV = 0$ and $dS = 0$. Thus we get *two* differential equations for r and h :

$$\begin{aligned} d(\pi r^2 h) &= 0, \\ d(2\pi r h + 2\pi r^2) &= 0. \end{aligned}$$

Expanding these using the differential calculus (and dividing by constants), we get the system of differential equations

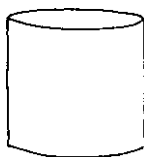
$$\begin{aligned} 2rh \, dr + r^2 \, dh &= 0, \\ (2r + h) \, dr + r \, dh &= 0. \end{aligned}$$

This is what must be true at the precise moment the cylinder shapes pass through the critical point. (Note that neither dr nor dh themselves can be zero, because the can is still narrowing and lengthening at that moment.)

Multiplying the second equation by r and subtracting the first, we get (after division by dr)

$$(2r^2 + rh) - 2rh = 0.$$

This means that the best cylinder is attained when $2r^2 = rh$. There are two solutions to this equation, namely $r = 0$ and $2r = h$. The first is clearly an artifact at the boundary, and the second is our maximum. Thus the best-shaped soup can has a height equal to its diameter.



In other words, it's a rotated square. How beautiful! Maybe not entirely unexpected, but still. I never cease to be impressed by the simple economy of this technique.

What if we wanted the best *open* soup can?

Find the largest cylinder that fits inside a given cone. How about in a given sphere?

Among all cones of a fixed surface area,
which has the largest volume?

27

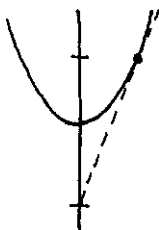
One of the best illustrations of the contrast between the classical and modern viewpoints is the measurement of the

conic sections. Historically, conics have always been a natural test case for geometers, being (apart from straight lines) the simplest curves there are. From a classical perspective, conic sections are literally just that—cross-sections of a cone. These fall naturally into three categories—ellipse, parabola, and hyperbola—depending on the slantedness of the slicing plane. All the classical results (e.g., the focal and tangent properties) follow from this description. Then we have the projective viewpoint, where the conics can be seen as the various projections of a circle. Perhaps simplest of all is the algebraic perspective, which reveals the conics to be those (nondegenerate) curves given by quadratic (i.e., degree 2) equations of the form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey = F.$$

The point being that in whichever structural framework you wish to operate, the conics appear as the simplest nontrivial objects. So it is very natural for us to want to measure them.

The classical geometers certainly wanted to. In fact, one of the most important and influential works in all of Greek mathematics was the *Conics* of Apollonius. This eight-volume masterpiece represented all that was then known about the conic sections and their fascinating properties. Of particular interest was the behavior of their tangents. For instance, Apollonius shows that for any point on a parabola, the tangent line at that point intersects the line of symmetry exactly as far below the bottom of the parabola as the point is above it.



Can you show this using the differential calculus?

We can think of results of this type as being essentially *angle* measurements. As far as lengths and areas go, the classical geometers had much more limited success. The method of exhaustion worked fine for areas bounded by sections of ellipse or parabola but failed miserably for the hyperbola. And the lengths of conic sections proved to be utterly intractable (of particular annoyance was the circumference of an ellipse).

The trouble with the classical exhaustion technique is that it requires us to be too clever. There are infinitely many ways to break something into pieces and approximate it, but it must be a truly ingenious scheme for us to tell where these approximations are heading. That's what the classical geometers were unable to do here. The analysis will not only show *why* they were doomed to failure but will also reveal several beautiful underlying connections that they seem to have missed.

To begin with, we'll need coordinate descriptions of the conic sections, and the simpler the better. The ellipse is easily handled by dilating a circle.



An ellipse with long radius a and short radius b can be viewed as a unit circle stretched by factors of a and b along the coordinate directions. Since a unit circle can be described by the equations $x = \cos t$, $y = \sin t$, we see that to make an ellipse we only need to modify these to

$$\begin{aligned}x &= a \cos t, \\y &= b \sin t.\end{aligned}$$

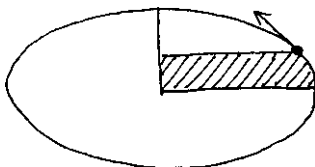
Alternatively, if you don't like carrying the parameter t around, you could instead write this as

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1,$$

which is another way of describing a dilation of the circle $x^2 + y^2 = 1$.

Why do the coordinate variables get divided by the stretch factors? Is this true of dilations in general?

Now that we have an equation for an ellipse, what do the integrals for length and area look like? Of course, we expect the area integral to be elementary (i.e., explicitly describable) since it is just a dilated circle, but let's see.



If we imagine a point moving along the ellipse according to the pattern $x = a \cos t$, $y = b \sin t$, we can see that the collected area A satisfies the differential equation $dA = x dy$, so we are interested in the integral

$$\int x dy = \int ab \cos^2 t dt.$$

This can be handled using the half-angle formula

$$\cos^2 t = \frac{1}{2}(1 + \cos 2t)$$

to give

$$\frac{1}{2} ab \int (1 + \cos 2t) dt = \frac{1}{2} ab \left(t + \frac{1}{2} \sin 2t \right),$$

which, as expected, is describable in elementary terms. In particular, when $t = 2\pi$ we get πab for the area of an ellipse. So in some sense the “reason” the classical geometers could handle the area of an elliptical section is that $\int \cos^2 t dt$ is elementary.

The circumference, however, is another story. The relevant integral is

$$\int \sqrt{dx^2 + dy^2} = \int \sqrt{a^2 \sin^2 t + b^2 \cos^2 t} dt.$$

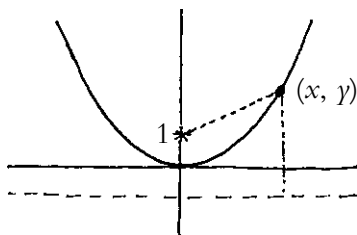
Integrals of this form (known as *elliptic integrals*, naturally) arise fairly often in analysis and are now known to be generically nonelementary. Of course when $a = b$, for example, we get $\int a dt = at$, corresponding to arc length along a circle, but in general the circumference of an ellipse is a nonelementary transcendental function of a and b , so there is no chance of an

explicit description. We might have hoped that since a circle has a circumference of 2π , the circumference of an ellipse might possibly look like

$$2\pi \times (\text{something simple depending on } a \text{ and } b).$$

Well it doesn't. So no wonder the Greeks had a hard time. It's not that they weren't clever enough, it's that the thing they wanted to say isn't sayable in the language they wanted to say it in.

As for the parabola, we have been using the equation $y = x^2$. In case you haven't derived this yourself, let me show you why it makes sense. Suppose we have a parabola, and we choose our units and orientations so that it is symmetrical with respect to the y axis and the focal point is at $(0, 1)$.



The focal property of the parabola says that the distance from any point on the curve to the focal point is the same as the distance to the focal line (which in this case would be the line $y = -1$). So if (x, y) is a point on the parabola, we must have

$$y + 1 = \sqrt{x^2 + (y - 1)^2}.$$

Squaring and rearranging this, we get

$$x^2 = (y + 1)^2 - (y - 1)^2 = 4y.$$

So our parabola has the equation $4y = x^2$. Rescaling if we want (so that the focal distance becomes $\frac{1}{4}$), we get $y = x^2$ as usual. Since every parabola is similar to every other, we may as well use the simplest equation we can.

We have already dealt with the area integral

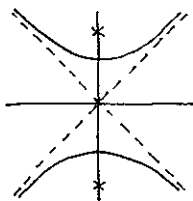
$$\int y \, dx = \int x^2 \, dx = \frac{1}{3}x^3,$$

which is not only elementary, but *algebraic* (no trigonometric functions are involved). This is why Archimedes was successful. By contrast, we have the arc length integral (here I prefer to use the equation $y = \frac{1}{2}x^2$)

$$\int \sqrt{dx^2 + dy^2} = \int \sqrt{1 + x^2} \, dx.$$

Any way of measuring the length of a piece of parabola will be equivalent to evaluating this integral—that's what I mean by a uniform language. Although it may look fairly innocent, this integral is actually quite serious.

Before discussing it further, let's take a look at the hyperbola. Now, since all hyperbolas are dilations of a right hyperbola (i.e., one with perpendicular tangents at infinity), we may as well start by getting an equation for a right hyperbola. If we choose our coordinate directions along the lines of symmetry, we get this picture:



We can then use the focal property of the hyperbola (and rescaling if necessary) to get the equation

$$x^2 - y^2 = 1.$$

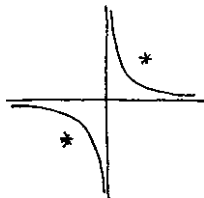
Can you derive this equation?

In particular, this means that all hyperbolas can be described by equations of the form

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1.$$

Notice how this is the same as the equation for an ellipse, only with a minus sign. Of course, this is related to the difference in their focal properties.

On the other hand, choosing our axes to be the tangents at infinity gives us a different view of a right hyperbola:



With this choice of orientation, we get the equation $xy = 1$, which in some ways is simpler. Of course, the two forms must

be related, and in fact this is connected to the Babylonian difference of squares formula:

$$x^2 - y^2 = (x + y)(x - y).$$

The rotated coordinate system is equivalent to choosing $x + y$ and $x - y$ as our new coordinates.

By the way, we can now see that the graph of the reciprocal relation $y = 1/x$ happens to be a right hyperbola.

Can you derive the equation $xy = 1$ for the right hyperbola? What is its focal distance?

For the sake of simplicity, let's restrict our attention to the right hyperbola $xy = 1$. Of course, there are lots of other hyperbolas out there for us to measure, but all the difficulties are present in this special case. The relevant integrals are

$$\int y \, dx = \int \frac{dx}{x}$$

and

$$\int \sqrt{dx^2 + dy^2} = \int \sqrt{1 + \frac{1}{x^4}} \, dx.$$

Once again, we have two perfectly harmless-looking integrals, which are in fact quite thorny. It turns out that the second one (the one for arc length) is provably nonelementary and can be rewritten in terms of (modified) elliptic integrals. So there is at least an abstract sense in which hyperbola length is related to ellipse length. It is the area integral, however, that is the real

surprise. What, we can't integrate dx/x ? What a scandalous state of affairs! Are we really going to stand for this?

Before we deal with this disturbing development, let's go back to the arc length integral for the parabola, $\int \sqrt{1+x^2} dx$. It turns out that this is intimately connected to the hyperbolic area integral $\int dx/x$. I want to show you this for two reasons. First, because I think it is surprising and wonderful that the length of one conic section is related to the area of another, and second, because it is a great example of analytic technique—the power of symbol jiggling.

So we are interested in the integral

$$\int \sqrt{1+t^2} dt.$$

(I've changed notation so there will be no confusion with any of our earlier symbol choices.)

Let's abbreviate $\sqrt{1+t^2}$ by s (so s is a new variable I've invented to take the place of this more complicated expression—a surprisingly powerful technique as you will see). Now we have

$$s^2 - t^2 = 1,$$

which is the equation of a right hyperbola. And our integral becomes simply $\int s dt$ which is precisely the area integral. So already the connection is being revealed, and all we have done is abbreviate. But we can go further. Writing

$$u = s + t,$$

$$v = s - t,$$

we can rephrase our equation $s^2 - t^2 = 1$ as simply $uv = 1$ (this is the rotated form of the right hyperbola again). Now our integral can be rewritten, using

$$\begin{aligned}s &= \frac{1}{2}(u + v), \\ t &= \frac{1}{2}(u - v),\end{aligned}$$

so that we get

$$\begin{aligned}\int s \, dt &= \int \frac{1}{4}(u + v) \, d(u - v) \\ &= \frac{1}{4} \int u \, du - u \, dv + v \, du - v \, dv.\end{aligned}$$

Since $uv = 1$, we have $u \, dv + v \, du = 0$, and our integral becomes

$$\frac{1}{8}(u^2 - v^2) + \frac{1}{2} \int v \, du = \frac{1}{2}st + \frac{1}{2} \int \frac{du}{u}.$$

The upshot of all this fiddling around is that

$$\int \sqrt{1 + t^2} \, dt = \frac{1}{2}t\sqrt{1 + t^2} + \frac{1}{2} \int \frac{du}{u},$$

where $u = t + \sqrt{1 + t^2}$. So the obstruction to measuring the length of a parabola is exactly the same as for the area of a hyperbola, namely $\int du/u$.

The question is, what sort of function is this? Is it algebraic or transcendental? Does it involve trigonometric functions or is it new? Or is it perhaps sitting right under our noses?

Show that

$$\int \frac{dx}{x^{m+1}} = \frac{-1}{mx^m}$$

for all $m \geq 1$, but not for $m = 0$.



Show that the length of one complete period of a sine wave is equal to the circumference of an ellipse with short radius 1 and long radius $\sqrt{2}$.

28

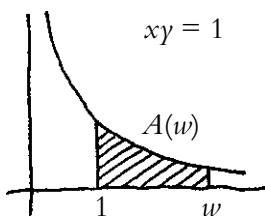
Our attempt to measure the conic sections has put us in a rather awkward and embarrassing position. These are, after all, the simplest possible curves, and they certainly do lead to very elegant and simple-looking differential equations, but for some reason we don't seem to be able to solve them. In particular, both hyperbolic area and parabolic length come down to the same question: What on earth is $\int dx/x$?

Quite apart from the intrinsic interest of measuring conics, this integral is analytically interesting in its own right. What could be a simpler and more natural differential than dx/x ? Surely its integral must be simple and natural as well, mustn't it? So what's the problem?

The obvious way to proceed would be to make a series of highly intelligent (and hopefully lucky) guesses until we find some clever combination of algebraic or trigonometric functions whose derivative is the reciprocal function. Unfortunately, this is a hopeless endeavor. As you have probably guessed, not only is our integral simple and natural, it also represents an entirely new transcendental function.

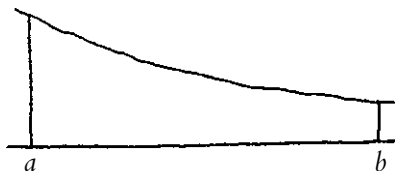
So we are not going to be able to use analytical methods to solve the problem of hyperbolic area. Instead, I want to show you how we can use the geometry of the hyperbola to get information about the integral. This is another nice example of the ongoing conversation between geometry and analysis.

For the sake of definiteness, let's write $A(w)$ for the area collected under the hyperbola $xy = 1$ as x runs from 1 to w .



(Of course, I would much prefer to collect area starting from $x = 0$, but the reciprocal curve is infinite there, so $x = 1$ seems the next best choice.) Now $A(w)$ is exactly the function we seek—that is, $dA = dw/w$.

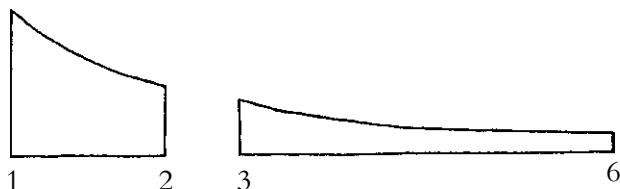
In general, we would want to measure the area between any two points a and b .



If both a and b are greater than 1, then this area can be viewed as the difference $A(b) - A(a)$. (We'll see in a minute how to deal with areas that lie to the left of $x = 1$.) Thus, knowledge of the function A —that is, a precise understanding of exactly how $A(w)$ depends on w —would completely solve the problem of hyperbolic area. Conversely, any information about

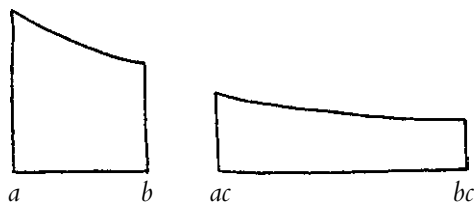
hyperbolic area would tell us something about the behavior of $A(w)$.

As it happens, the reciprocal curve does have a very beautiful area property: *scaling invariance*. To illustrate, let's look at two pieces of hyperbolic area.



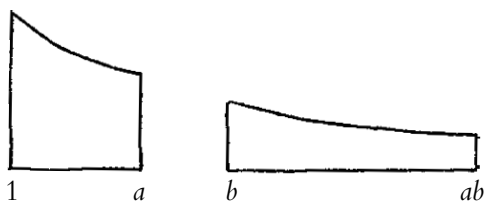
Notice that the second region (running from 3 to 6) is three times wider than the first (from 1 to 2). It is also one-third as high, because we are dealing with the *reciprocal* curve. More precisely, every vertical stick in the first region corresponds to a stick in the second region whose horizontal position is three times as large, while being one-third as tall. If we want, we can think of the second area as a dilation of the first—we stretch horizontally by a factor of 3 and vertically by $1/3$. Does that make sense?

The point here is that these two areas must then be equal. Dilations multiply area by the stretch factor, and we've used two factors that cancel each other out. Of course, there is nothing special about the number 3. The general statement would be that the area from a to b is the same as from ac to bc . Do you see why?



**Can you use scaling invariance to
measure areas to the left of $x = 1$?**

This means that the area of a region under the reciprocal curve depends only on the *ratio* of the endpoints, not on the endpoints themselves. In particular, for any two numbers a and b , the area from 1 to a is the same as from b to ab .



If we express this analytically, in terms of our area function A , it says that $A(ab) - A(b) = A(a) - A(1)$. Since $A(1) = 0$, we can rewrite this in the elegant form

$$A(ab) = A(a) + A(b).$$

How beautiful! The scaling invariance of hyperbolic area tells us something quite unexpected about our mystery function A : it transforms multiplication into addition. That is, if we think of A as a process that converts a number w into $A(w)$, then what we are saying is that if you multiply two numbers and then convert the product, it's the same as converting the two numbers separately and then adding. (Incidentally, this property would not hold true if we chose a different area collection function, starting from a point other than $x = 1$. So this is definitely the right choice.)

As I mentioned before, the function $A(w)$ is known to be transcendental; there is no hope of actually calculating numbers like $A(2)$, $A(3)$, or $A(6)$. But at least we do know that whatever they are, $A(6)$ is exactly equal to $A(2) + A(3)$. This is reminiscent of the situation in trigonometry—although we are usually prohibited from obtaining precise values of sine and cosine, we do know of many beautiful and fascinating interrelationships among them (e.g., the half-angle formulas).

At the moment, we have defined $A(w)$ only for $w \geq 1$. How should we extend the definition of $A(w)$ to include values of w between 0 and 1? Can we do it in such a way that the property $A(ab) = A(a) + A(b)$ still holds?

**Can you think of a way to extend
the definition of $A(w)$?**

While we're on the subject of converting multiplication to addition, I would like to make a brief historical digression. At one point when we were talking about angles and lengths, I mentioned how the advent of long ocean voyages in the late fifteenth century (e.g., 1492) created a need for accurate trigonometric tables—that is, fairly precise approximations of the values of sine and cosine. Such tables had to be painstakingly made by hand, but once completed, the values could then be looked up easily. (I know this sounds terribly prosaic and practical, but bear with me.)

Although these nautical tables spared navigators a fair amount of computational tedium, there was still plenty left in the form of arithmetic. Maybe some good four- or five-digit

approximations could be looked up in a table, but you still had to work with them—to add, subtract, multiply, and divide them, for example.

Now, as you may recall, there is a calculus for doing this sort of thing: digits, place value, carrying, etc. And in fact, when it comes to adding and subtracting, the standard procedures are actually pretty efficient. For example, if we have two five-digit numbers, say 32768 and 48597, we can add them together quite easily:

$$\begin{array}{r} 32768 \\ 48597 \\ \hline 81365 \end{array}$$

The point being that the number of individual steps (one-digit sums with possible carrying) is equal to the number of digits. So adding ten-digit numbers would take only twice as long, even though the numbers themselves are astronomically larger. Subtraction is similar.

Multiplication, on the other hand, is a nightmare (and don't get me started on division!). The trouble is that it takes too long: to multiply two five-digit numbers requires *twenty-five* single-digit multiplications (to say nothing of the necessary adding and carrying). If we want the product of two ten-digit numbers, this entails over a hundred individual calculations. Forgetting about the practical issues facing navigators and accountants, I find it interesting on a purely theoretical level that one operation is so much more costly than the other. Not that this is in any way surprising; multiplication is, after all, repeated addition.

In any case, it came as a great relief to those in the arithmetic business when the Scottish mathematician John Napier invented a better system in 1610.

The idea is this. First, notice how easy it is to multiply by 10: $367 \times 10 = 3670$. This is not due to any special property of the number ten, but rather to our choice of ten as a grouping size. That is, when we write a number like 367, we are choosing to represent that quantity in terms of groups of ten (3 hundreds, 6 tens, 7 ones). So each position in the digit string is worth ten times the next. Multiplication by ten then simply shifts each digit one space to the left so that it counts ten times as much. Of course, we could just as easily use a different grouping size, say seven, and then multiplication by seven would shift the digits. (The advantage of a smaller grouping size would be less memorization—there would be only six nonzero digits, so the multiplication table would be smaller. The disadvantage would be that the representations themselves would then be longer.) The choice of ten as a grouping size thus has no particular mathematical benefit; it is simply a cultural choice stemming from the fact that we happen to have ten fingers. Of course, once such a “decimal” system is in general use, then multiplication by ten becomes especially convenient.

In particular, numbers that are powers of 10, such as 100 or 10000, are especially easy to multiply together: we just count the shifts. Since 100 is the same as 1 with two shifts, and 10000 is 1 with four shifts, their product is simply 1000000 (i.e., 1 with six shifts). The key observation here is that multiplication of powers of 10 is essentially *addition*. That is, to multiply two such numbers we need only to add the shifts: $10^m \cdot 10^n = 10^{m+n}$.

Of course, the same goes for any other number, not just 10. For any number a , we always have

$$a^m \cdot a^n = a^{m+n}$$

because that's what repeated multiplication means. By the way, when one writes something like 2^5 , the number 2 that is being repeatedly multiplied is called the **base** and 5 is the **exponent** (Latin for "on display"). This number would then be referred to as "2 raised to the fifth power" or simply "2 to the fifth," for short.

This pattern is so simple and so pretty, that it is often extended to include negative and fractional exponents as well. That is, we can make sense of something like $2^{-3/8}$ by insisting that whatever we choose it to mean, we want the pattern $2^{m+n} = 2^m \cdot 2^n$ to be preserved. This is a major theme in mathematics: extending ideas and patterns into new territory. Mathematical patterns are like crystals; they hold their shape and can grow beyond their original confines. Our extension of sine and cosine to arbitrary angles is one example; projective space is another. Now we're going to do the same thing with repeated multiplication.

Let's start with the powers of 2. Writing out the first few, we notice a simple pattern:

$$2^1 = 2, \quad 2^2 = 4, \quad 2^3 = 8, \quad 2^4 = 16, \dots$$

Each time the exponent goes up by 1, the number itself doubles. Of course, this is patently obvious. But it also means

that whenever the exponent goes *down* by 1, the number gets cut in half. And this allows us to extend the meaning of 2^n . First of all, it suggests that 2^0 should equal 1 ! What is interesting here is that the original meaning of 2^n , namely “ n copies of 2 multiplied together,” no longer makes any sense. Are we really saying that no 2s multiplied together is equal to 1? I guess we could say it if we want to, but what we *really* mean is that we are shifting the meaning of 2^n from “ n copies of 2 multiplied together” to “whatever it needs to be to keep the pretty pattern going.” It would not be much of an overstatement to say that this is how *all* meaning in mathematics is made.

Continuing the pattern, we find that

$$2^{-1} = \frac{1}{2}, \quad 2^{-2} = \frac{1}{4}, \quad 2^{-3} = \frac{1}{8}, \quad 2^{-4} = \frac{1}{16}, \dots$$

and so on. In general, we would be choosing a^{-n} to mean $1/a^n$. Thus $3^{-2} = \frac{1}{9}$ and $(\frac{2}{3})^{-3} = \frac{27}{8}$. (In particular, a^{-1} is an amusing way to write $1/a$.)

Show that $a^{m-n} = a^m/a^n$ for all m, n .

Show that $d(x^m) = mx^{m-1} dx$ for all m positive, negative, and 0.

Let's go a little further. Is there a good way to give meaning to $2^{1/2}$? The pattern, if it can continue to hold in such uncharted territory, would say that

$$2^{1/2} \cdot 2^{1/2} = 2^1 = 2.$$

This means that whatever $2^{1/2}$ is, when we multiply it by itself we get 2. So it must be $\sqrt{2}$. Similarly, $10^{1/2} = \sqrt{10}$, and in general $a^{1/2} = \sqrt{a}$.

Actually, we have to be a little careful here, since \sqrt{a} is slightly ambiguous. There are, after all, *two* square roots of a number a , if a is positive. Which one do we want $a^{1/2}$ to mean? Also, if a is negative then we have an even bigger problem. We don't yet have a meaning for the square root of a negative number, so what are we going to do with something like $(-2)^{1/2}$?

One easy way out is to simply restrict ourselves to positive bases. That is, we will only assign meaning to $a^{1/2}$ when a is a positive number. The other possibility is to extend our number system to include new objects like $\sqrt{-2}$. This can actually be done—and you should do it! Unfortunately, this still doesn't solve our ambiguity problem. We still need to choose the meaning of $a^{1/2}$ (if we want it to have meaning) to be one of the square roots of a . Which one? Well, the usual choice when a is positive is to choose the positive square root. Thus $4^{1/2} = 2$, not -2 . Of course, this is somewhat arbitrary, but at least it makes a nice consistent pattern.

For the time being, let's agree that our base will always be positive and that whenever we need to make a choice, we will choose positive values. So we will say that $a^{1/2}$ only has meaning when a is positive, and its meaning is the (unique) positive square root of a .

Of course, you may find this whole enterprise repulsive and not wish to make any of these choices. You may see no advantage whatever in writing things this way. I personally like it because it illustrates the persistence of pattern. I feel like this is what the pattern wants—to be set free of its shackles. So let's keep going.

How should we define something like $a^{3/4}$? Whatever it is, when we raise it to the fourth power (that is, multiply four copies of it together), we should get a^3 . Do you see why? This means that $a^{3/4}$ must be the fourth root of a cubed, or $\sqrt[4]{a^3}$. The general pattern is now clear: $a^{m/n}$ must be an n -th root of a^m (and, of course, we choose it to be the positive one).

Show that for any fraction m/n , we need to choose $a^{m/n} = \sqrt[n]{a^m}$. Is this the same as $(\sqrt[n]{a})^m$?

This is what the pattern forces. So this is what we as mathematicians tend to accept, because beautiful simple patterns are more important to us than anything else, even our own desires and intuitions. Plus, it's not like we had some a priori idea of what we wanted $2^{-3/8}$ to mean. The point is that if we choose it to mean "the reciprocal of the eighth root of 2 cubed" then the pattern keeps going.

Show that $(a^m)^n = a^{mn}$ for any whole numbers m and n . Does it still work if m and n are fractions?

Show that $d(x^m) = mx^{m-1} dx$ for all fractions m .

So now we have an idea of what a^b should mean when b is a rational number. But what if the exponent is irrational? Can we make sense of numbers like $2^{\sqrt{2}}$ or 10^π ?

Let's do something that mathematicians often do: we'll assume we can and see what happens. (This philosophical method has a long history. The Greeks called it *analysis*, as opposed to *synthesis*, which refers to the building up

of knowledge from first principles.) Anyway, let's say that somehow we have given meaning to the expression a^b for any number b . Of course, we insist that the pattern remain intact (otherwise, what's the point; we may as well define 2^π to be 37). So we assume that not only does a^b make sense, but that it continues to obey the pattern

$$a^b \cdot a^c = a^{b+c}.$$

In particular, whatever the numbers $3^{\sqrt{2}}$ and 3^π are (and they will most certainly be transcendental if anything) we insist that $3^{\sqrt{2}+\pi}$ be their product. (Not that we are in a position to insist on anything; we're simply hoping that this is possible.)

Now here is Napier's idea. Suppose we have some number like 32768. Clearly, this lies between 10^4 and 10^5 . Napier's realization was that there must be *some* number p between 4 and 5 so that $32768 = 10^p$. In other words, *every* number is a power of 10. Since powers of 10 are easy to multiply, this should mean that *all* numbers are easy to multiply. Of course, the hard part is to figure out what power of 10 a given number is. So there are really two problems here. First, is it really *true* that every number is a power of 10? And second, how on earth can we hope to calculate such an exponent? These are pretty serious questions, actually.

On the other hand, for all *practical* purposes all that is needed are approximations. Here the subtle mathematical issues disappear. We don't need to know if a^b makes sense for irrational numbers b , because every number is approximately a fraction. For example, if I want to represent a number like 37 as an approximate power of 10, I only need to find a fraction m/n so

that $10^{m/n} \approx 37$. In other words, 10^m should be roughly 37^n . Let's look at some powers of 37 that are reasonably close to powers of 10:

$$\begin{aligned} 37^2 &= 1369 \approx 10^3 \\ 37^7 &= 94931877133 \approx 10^{11}. \end{aligned}$$

So $3/2 = 1.5$ should be a so-so estimate, and $11/7 \approx 1.57$ a pretty good one. The point is that we don't need an exact value for the exponent in order to navigate a ship or any other mundane purpose like that. If we cared enough, we could even obtain an extremely accurate estimate like 1.56820. Of course, it would require an enormous amount of work to obtain such approximations for every number we might wish to use, but just as for trigonometric tables, the work would only have to be done once. And this is just what Napier set out to do.

For each number N , we are trying to find (at least approximately) a number p so that $N = 10^p$. Napier called p the **logarithm** of N (from the Greek *logos* + *arithmos*, meaning "way of reckoning"). So, for example, the logarithm of 37 is about 1.5682. Let's write $L(N)$ for the logarithm of N . Then a section of Napier's table might look something like this:

N	$L(N)$
35	1.5441
36	1.5563
37	1.5682
38	1.5798
39	1.5911

Now, here's the point. Suppose we wanted to multiply two numbers together, say the ones we had before: 32768 and 48597. Ordinarily, this would be an annoying, multistep procedure. But using Napier's "admirable table of logarithms," we can rewrite these numbers (again, approximately) as powers of 10:

$$\begin{aligned}32768 &\approx 10^{4.5154}, \\48597 &\approx 10^{4.6866},\end{aligned}$$

and since multiplying merely adds the exponents, we get

$$32768 \times 48597 \approx 10^{9.2020}.$$

Consulting the logarithm tables (in reverse), we find that the number whose logarithm is closest to 9.2020 is 1592208727. This means that the true product should be pretty close. In fact, $32768 \times 48597 = 1592426496$, so our estimate is accurate up to the fourth decimal place. In other words, we are off by about one part in ten thousand. But the point is that we only had to do three table look-ups and one addition. So this is a huge time-saver.

**What if you wanted to multiply
three or more numbers together?**

The skeptical reader may find it improbable that such tables exist going up to numbers as large as 1592208727, and the skeptical reader would be right—they don't. In actual practice, one only requires logarithm tables for numbers between 1 and 10. Everything else can be obtained by shifting. For example,

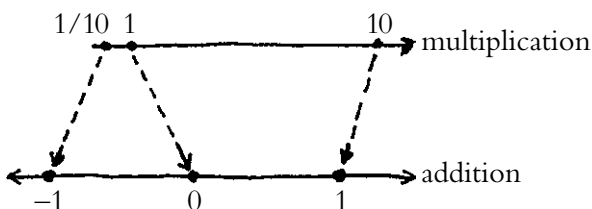
if I wanted $L(32768)$, I would actually be content to look up $L(3.2768) = 0.5154$, and then add 4. This is because multiplication by 10 has the effect of adding 1 to the exponent; that is, the logarithm. Similarly to find the “antilogarithm” of 9.2020, I would look for 0.2020 in the logarithm column and see that it corresponds to the number 1.5922 (assuming my tables are accurate to four decimal places, which is pretty standard). Then I would multiply this by 10^9 to get 1592200000, which is pretty much as accurate as before.

Of course, the practical use of logarithms for arithmetic computation is now obsolete, due to the advent of high-speed electronic calculators. In fact, almost all computation these days is done by machine (as Leibniz himself predicted). My point in bringing up logarithms was not their computational utility—now a historical footnote—but to illustrate a particularly curious example of an unforeseen connection in mathematics: hyperbolic area (the integral of dx/x) turns out to be related to the behavior of exponents (logarithms). How strange that a method intended to speed up practical arithmetic should turn out to be so intimately connected to the classical measurement of conics! Again, the connection is that in both cases multiplication is somehow being converted into addition.

How could you use logarithms to divide two numbers? To take the square root of a number?

29

From a modern perspective, Napier's logarithm can be viewed as an isomorphism between two apparently different algebraic structures. On the one hand, we have the system of positive numbers under multiplication, and on the other hand, the system consisting of all numbers (positive and negative) under addition. Napier's logarithm provides a "dictionary" between these two worlds:



Under this correspondence, a positive number w is sent to its logarithm $L(w)$. For instance, the number one million (10^6) would translate to its base 10 exponent, in this case 6. The number 1, which is multiplicatively inert ($1 \times w = w$ for all numbers w), corresponds to the number 0, which does nothing additively ($0 + w = w$). Similarly, division (that is, un-multiplication) corresponds to subtraction (un-addition). The point is that the two systems are structurally identical, and it is the logarithm that allows us to see that. More precisely, for any positive numbers a and b , we have

$$L(ab) = L(a) + L(b).$$

Thus Napier's logarithm behaves exactly like hyperbolic area: it converts products to sums. That was, of course, the whole

point of Napier's discovery: addition is fast; multiplication is slow. But now we see that they are in fact *the same*.

That is, if there really *is* such a thing as a logarithm. It is one thing to crudely approximate numbers by fractional powers of 10; it's quite another to prove that this can be done exactly, with infinite precision. Is pi really an exact power of 10? If so, what kind of number is that exponent? In other words, how do we know that pi (or 2 for that matter) really has a logarithm? Do you get what I'm saying?

The other issue is *ten*. Napier's logarithm is based (quite literally!) on this particular not-so-interesting number. That's all very well and good for a calculation system designed for a decimal culture, but as mathematicians, we should be looking for something more intrinsically beautiful and natural. Once again, it's the whole *unit independence* issue. Choosing a base for our exponents is exactly the same as choosing a measuring unit—essentially, we are measuring the size of a number by how many decimal digits it has. So that's pretty arbitrary, and to me arbitrary means *ugly*.

On the other hand, would a different base be any better? What if we designed a base 2 logarithm? This would assign to each number the exponent needed to represent it as a power of 2. This logarithm would also convert multiplication to addition, and everything would work fine. We could make binary logarithm tables with no problem. So 10 is completely irrelevant. If what you want is to convert multiplication to addition, then any base is as good as any other. (It has become traditional, by the way, to use the notation $\log_a x$ for the base a logarithm of a number x . In particular, the Napier logarithm $L(x)$ is commonly written $\log_{10} x$. Thus $\log_2 8 = 3$, and $\log_5 \frac{1}{25} = -2$, for instance.)

The simplest (and most abstract) way to frame these ideas is to call any process that converts multiplication to addition a logarithm. That is, any (continuous) function whatsoever that satisfies

$$\log(xy) = \log(x) + \log(y)$$

for all positive numbers x and y qualifies as a logarithm (I'm using the generic symbol \log to represent any such activity). Thus Napier's function L , the binary logarithm \log_2 , and the hyperbolic area function A are all logarithms in this abstract sense.

Given any such function \log , let's call the reverse process \exp (short for exponentiation). Then

$$\log(\exp(x)) = \exp(\log(x)) = x,$$

because that is what *reverse* means. In particular, if you choose \log to be the Napier logarithm, then \exp will simply be base 10 exponentiation, $\exp(x) = 10^x$. In general, the function \exp inherits the property

$$\exp(x + y) = \exp(x) \cdot \exp(y).$$

Why must \exp behave this way?

Show that for any logarithm,

$$\log(1) = 0 \text{ and } \log(1/x) = -\log(x).$$

Now here is an idea that I find very clever and pretty. The property of being a logarithm implies that

$$\log(x^m) = m \log x, \text{ for } m = 1, 2, 3, \dots$$

Do you see why? Applying \exp to both sides of this equation, we get $x^m = \exp(m \log x)$. This makes good sense for any positive whole number m . But the right-hand side is in fact meaningful for *any* number m —rational, irrational, whatever. So the existence of a logarithm allows us to *define* what it means to raise any positive number a to any power b :

$$a^b = \exp(b \log a).$$

All the properties that you want a^b to have follow directly from the properties of \log and \exp .

Show that with this definition,

$$a^{b+c} = a^b \cdot a^c, (ab)^c = a^c \cdot b^c, \text{ and } (a^b)^c = a^{bc}.$$

So, given any logarithm (or what I like to think of as a \log/\exp pair), we get a corresponding definition of a^b . Luckily, as we will see, it turns out that its value does not depend on which logarithm we choose.

Now we have to be a little careful here about circular reasoning. Our problem with Napier's logarithm, you may recall, was that we didn't quite know what 10^x should mean (at least when x is irrational). Now that we have a satisfactory definition of a^b , it might seem as though our logarithm problems are solved. The trouble is that our clever definition of a^b requires that a well-defined logarithm already be in place. So we can't then turn around and use this to *define* our logarithm. On the face of it, this looks pretty bad. We seem

to need a definition of exponentiation to define a logarithm, and vice versa.

But wait—our hyperbolic area function is a logarithm! And luckily, it requires no notion of exponentiation to get it off the ground; it is simply the collected area under the reciprocal curve. This means that we can base our entire theory of exponents and logarithms on the integral of dx/x .

So here's the plan: we will define, once and for all, the **natural logarithm** of a positive number x to be $A(x)$, the area under the reciprocal curve from 1 to x . Since this particular logarithm is the only one mathematicians ever use (and we will shortly see why), we will do it the honor of being written simply as $\log x$. (This convention varies somewhat, actually. Some people—scientists, engineers, calculator manufacturers—prefer to use the symbol \log for Napier's decimal logarithm; others—mostly computer scientists—like to use it to denote the binary logarithm. The natural logarithm is then given the unappetizing name \ln .)

Now that we have a well-defined logarithm, we likewise define the **natural exponential** to be the corresponding exponential function, which we will write simply as \exp . So $\exp(3)$, for instance, refers to the number w with $A(w) = 3$. We can then define a^b as $\exp(b \log a)$ without any circularity in our reasoning. Thus, the number 2^π can now be seen as that number that gives us π times as much collected area as 2 does.

As bizarre as this sequence of ideas may initially seem, the point is that we get a precise definition of exponentiation that satisfies all the properties that we want it to have.

In particular, now that we have a precise notion of what a^b means, it's not hard to determine the base a logarithm of a number x to be

$$\log_a x = \frac{\log x}{\log a}.$$

Can you derive this elegant logarithm formula?

As a special case, we see that the Napier logarithm of a number is simply its natural logarithm divided by a certain constant, namely $\log 10 \approx 2.3$. Thus Napier's logarithm (and indeed *any* logarithm) is just a constant multiple of the natural logarithm. In other words, all logarithm functions are proportional to each other. This is why you never need more than one logarithm—they're all essentially the same.

Except they're not. The natural logarithm is nicer than the others, and here's why: it has the simplest differential. In fact, from the very definition of the natural logarithm we have

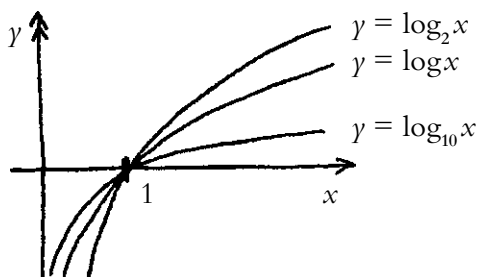
$$d(\log x) = \frac{dx}{x}.$$

This means that any other logarithm, being a constant multiple of $\log x$, will have a differential that is some multiple of dx/x . For instance, the differential of the Napier logarithm would be

$$d(\log_{10} x) = \frac{1}{\log 10} \frac{dx}{x}.$$

But who wants some ugly constant like $1/\log 10$ cluttering up the place? If all logarithms are more or less equivalent, why not go with the one whose differential is as nice as possible?

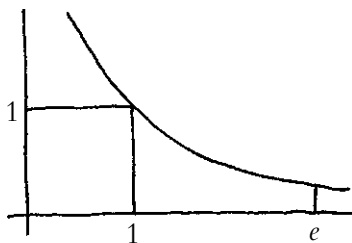
Another way to think of this is to look at the graphs of the various logarithm functions.



Being proportional, these curves all behave pretty much the same way (in particular, logarithms are famous for their exceedingly slow growth). But notice how their tangents at the point $x = 1$ vary from nearly horizontal to nearly vertical. The natural logarithm is the one whose slantedness is exactly half-way between these extremes, making a nice 45-degree angle with the axes.

So the natural logarithm is the simplest, and it therefore is the only logarithm that mathematicians ever use. It also deserves its name, since it arose naturally from our attempt to measure conics, as opposed to making some arbitrary choice of base. But that raises an interesting question. What is the base of the natural logarithm?

Since the base of a logarithm is just the number whose logarithm is equal to 1, we are asking how far we have to go along the reciprocal curve to collect exactly one unit of area.



This number, usually denoted by the letter e (for exponential), stands out from all other numbers as the most aesthetically pleasing base. So what number is it? Well, it turns out that $e \approx 2.71828$, and I don't suppose it would surprise you very much to learn that it is transcendental. (As a matter of fact, e was the *first* naturally occurring mathematical constant to be proved transcendental, by Hermite in 1873.)

This means that just as we did for the trigonometric functions and for π , we will need to enlarge our language to include \log , \exp , and e . Isn't it funny how every time we run across an interesting number it turns out to be inexpressible? Maybe numbers like e and π are simply too beautiful to be captured by something as prosaic as a fraction or an algebraic equation. If e were rational, for instance, what numerator and denominator could possibly be good enough? In any case, we have no choice but to simply give names to these things and then incorporate them into our vocabulary. (In particular, it is customary to include \log and \exp in the category of elementary functions.)

Let's step back a bit and try to figure out what has really happened here. We began with a problem: What is the integral of dx/x ? Did we solve this problem? In some sense it seems like we cheated—all we did was to name it $\log x$. (Similarly, we just named the proportion of circumference to diameter π and then walked away.) What kind of "solution" is that? Are mathematicians just a pack of namers and abbreviators?

No. The words and symbols are irrelevant. What matters are the patterns and our ideas about them. (As Gauss famously quipped, what we need are *notions*, not notations.) Maybe we didn't solve our problem in the sense of expressing the integral of dx/x in algebraic terms (which we now know to be

impossible), but we did discover that whatever it is (and we may as well call it $\log x$), it satisfies the surprising and elegant property $\log(ab) = \log(a) + \log(b)$. If the names and abbreviations help us to understand the pattern, then they're worth it. Otherwise, they're in the way. We should give names to things only when we need to and in such a way that it helps us to reveal and to distinguish more clearly the patterns that obtrude themselves upon our imaginations.

Speaking of which, there is one more pattern that I would like to show you. We saw how the natural logarithm distinguishes itself from the other logarithms by having the nicest differential. Shouldn't the natural exponential have a similar property? What is the differential of an exponential a^x ?

Let's start with the natural exponential $\exp(x)$ (which, if you want, you can also write as e^x). The simplest way to proceed is to give this a name, say y . Then $y = \exp(x)$, so $x = \log y$. Taking differentials, we get $dx = dy/y$. This means that $dy = y \, dx$. In other words,

$$d \exp(x) = \exp(x) \, dx.$$

Or, if you prefer, you can express this as

$$d(e^x) = e^x \, dx.$$

What a beautiful discovery! The natural exponential has the property that its derivative is itself. Geometrically, this means that the *slantedness* of the graph of $y = e^x$ at a point is always equal to its *height*.

Show that in general, for any base a ,
we have $d(a^x) = a^x \log a \, dx$.

The natural exponential thus emerges as the unique exponential function whose derivative is itself, as opposed to some unpleasant constant multiple of itself.

What is special about the tangent to the graph
of $y = e^x$ at the point $x = 0$ compared to
the other exponential functions?

Finally, to put the last finishing touches on our differential calculus, we can now extend our formula for $d(x^m)$ to arbitrary exponents m :

$$\begin{aligned} d(x^m) &= d(\exp(m \log x)) \\ &= \exp(m \log x) \, d(m \log x) \\ &= x^m \cdot m \, dx/x \\ &= m x^{m-1} \, dx. \end{aligned}$$

In particular, this means that

$$\int x^m \, dx = \frac{x^{m+1}}{m+1}$$

for all numbers m , *except* for $m = -1$, where the expression on the right becomes meaningless. In this latter case, as odd as it may seem, the pattern is broken, and we get, of all things, the natural logarithm.

Show that for any two variables x and y ,
we have $d(x^y) = yx^{y-1} dx + x^y \log x dy$.

Show that as n increases indefinitely,
 $(1 + 1/n)^n$ approaches e .

Can you determine $\int \log x dx$?

30

What a wild and amazing place mathematical reality is! There is just no end to its mystery and beauty. And there is so much more I want to tell you about it—so many more delightful and surprising (and *scary*) discoveries. Nevertheless, I feel that the time has come for me to put down my pen. (Perhaps you have had that feeling for quite some time!)

Not that we've done much more than scratch the surface. Mathematics is a vast, ever-expanding jungle, and measurement is only one of its many rivers (though certainly a major one). But my goal was not to be exhaustive, only illustrative (and hopefully entertaining). I suppose what I really wanted to do was to give you a feeling for what it is we mathematicians do and why we do it.

I especially wanted to get across the idea that mathematics is a quintessentially human activity—that whatever strange product of evolutionary biochemistry our minds are, one thing is for sure: *we love patterns*. Mathematics is a meeting place for

language, pattern, curiosity, and joy. And it has given me a lifetime of free entertainment.

There is one small issue I feel I should address before I go: *reality*. Why haven't we talked about the real world at all? What about all those wonderful applications of geometry and analysis to the problems of physics, engineering, and architecture? What about the motions of the heavenly spheres, for crying out loud? How can I claim to have written a book about measurement when I take such a dismissive view of the very reality in which my brain is located?

Well, first of all, I am me, and I write about what I am interested in, which is the nature of mathematical reality. What else can anyone do? Second, it's not like there is a tremendous shortage of books about the physical universe. They're all over the place, and many of them are quite good. I felt the need to write a book about mathematics, because, quite frankly, there really aren't very many. Not many that are *honest* and *personal*. Not many that feel like real books with a point of view. Also, I didn't want to talk about the applications of mathematics to the sciences (which are fairly obvious anyway) because I feel that the value of mathematics lies not in its utility but in the pleasure it gives.

Which is not to say that reality isn't interesting and exciting. Don't get me wrong, I'm very happy to be here. There are birds and trees and love and chocolate. I have no complaints about physical reality, only a much deeper intellectual and aesthetic attraction to pattern in the abstract. Maybe the bottom line is that I don't have that much to say about the real world. Maybe part of it is that I'm not altogether entirely *here* a lot of the time.

Maybe the point of this book is to give you a glimpse of what it is like to live a mathematical life—to have the better part of one’s mentality off in an imaginary world. At any rate, I know that I am by nature permanently isolated from reality—my brain is alone, receiving only the (possibly illusory) sensory input that it does—but mathematical reality is *me*.

Which brings me to *you*. This mathematical reality we have been talking about—although it certainly feels like it’s “out there” somewhere—I don’t want you to feel shy about entering it, as if it were located in some restricted government facility and being worked on by experts in lab coats. Mathematical reality is not “theirs”—it’s yours. You have an imaginary universe in your head whether you like it or not. You can choose to ignore it, or you can ask questions about it, but you cannot deny that it is very much a part of you. Which is one of the reasons why mathematics is so compelling: you are discovering things about yourself and the way your personal mental constructions behave.

So keep exploring! It doesn’t matter how much experience you have. Whether you are an expert or a beginner, the feeling is the same. You are wandering around in the jungle, following one river and then another. The journey is endless, and the only goal is to explore and have fun. Enjoy!

ACKNOWLEDGMENTS

I want to express my sincere gratitude to my editor, Michael Fisher, whose tireless support and enthusiasm for the book are matched only by his unlimited patience with its author.

If this book has a claim to visual elegance and readability, it is due to the talents and expertise of Tim Jones, Peter Holm, and Kate Mueller. I feel fortunate indeed to have had the design, typesetting, and copyediting in their capable hands.

Thanks also to Lauren Esdaile, Donna Bouvier, David Foss, and the rest of the team at Harvard University Press for their important behind-the-scenes work.

I am indebted to all my students, colleagues, friends, and family who contributed so many wonderful problems, suggestions, and helpful criticism.

Finally, special thanks to my dear friend Keith Goldfarb for his invaluable advice and encouragement, and for suggesting to me, back when we were sixteen years old, that I should someday sit down and write a math book that really tells it like it is. I hope it was worth the wait.

INDEX

- Acceleration, 300
Algebra, 56, 60, 116
Algebraic, 341, 366, 370
Alphabet soup, 113
Analysis, 17, 349, 381
Analyst, The, 272
Angle, 24, 29, 119, 128, 130, 166, 213, 218, 241, 362
 acute/obtuse, 135
 between curves, 149
 inside, 27, 109
 outside, 27, 108, 239
 right, 36, 66, 86, 127, 180, 242
Angle bisector, 136
Annulus, 96–98
Antilogarithm, 385
Apollonius, 161, 361
Arc (of a circle), 66, 99, 241, 242, 290, 346
Arc length, 344–345, 366
Archimedes, 15, 31, 67, 82, 86, 89, 90, 117, 190, 310, 320, 326, 329, 366
Archimedes's solid, 90
Area, 38, 52, 324, 331
Argument, 7, 18, 26, 50, 148
 See also Proof
Arithmetic, 203, 316, 375–377, 385
Arithmetization of geometry, 335
Artifact, 357, 360
Average, 57, 97, 102, 104, 332

Ball, 217, 271
Barber pole, 195
Base, 112, 378, 380, 387, 392

Beauty, 10, 19, 45, 107, 164, 351
Berkeley, George (Bishop), 272, 275, 310
Bernoulli, Johann, 343
Billiard ball, 224, 225
Blah, 222, 315
Blueprint, 35, 108
Boundary value, 358
Box, 41
 diagonal of, 48
 volume of, 41, 84, 218
Brahmagupta's formula, 136

Calculus, 316–317, 341, 376
 differential, 317, 320, 330, 335
 integral, 340
 multiplicative, 317
Can of worms, 190
Cavalieri principle, 82, 84, 86, 102, 335
Cavalieri, Bonaventura, 83, 86
Center, 13, 14, 26, 98, 100, 102–104
Center of mass, 104
Centroid, 104–105, 107
 of perimeter, 106–107
Chaining, 315
Change (in a variable), 297, 303, 344
Chord, 290, 346
Circle, 62, 70, 140, 148, 217, 232, 261
 area of, 65, 69, 336
 circumference of, 63, 66, 69, 217, 252, 365
 flat, 231
Circular reasoning, 389

- Classification, 217, 229
- Clock, 219, 244
- Coefficient (of time), 264, 267
- Collinear, 164–165
- Complexity, 192
- Cone, 78, 107, 159, 172
 - double, 89
 - infinite, 160
 - generalized, 85
 - surface of, 82
 - volume of, 78, 82, 218, 352
- Conic section, 161, 163, 174, 188, 262, 361
 - degenerate, 262
- Conics* of Apollonius, 361
- Conjecture, 17, 19
- Constant speed, 194, 222, 223, 246, 254, 266
- Coordinate geometry, 259
- Coordinate system, 208, 214, 246, 257, 260
 - circular, 232, 234–238, 240, 242
- Cosine, 123, 124, 126–130, 238–243
- Cosine wave, 294
- Counterexample, 17, 19
- Critical point, 356, 359
- Cross-section, 80, 84, 141, 159, 191
 - equal, 81–83
- Cube, 76, 77, 82, 85, 94
- Curvature, 300
- Curve, 206, 226, 229, 279, 331
- Curved, 230–231
- Cusp, 196
- Cycloid, 195, 232, 247, 254
 - area of, 291, 348
 - length of, 342, 345, 348
 - velocity of, 286–290
- Cylinder, 70, 86, 95, 101, 105, 140, 172, 358
 - generalized, 72, 142
 - surface of, 73, 358
 - volume of, 72, 358
- d. See* Leibniz *d*-operator
- Dandelin, Germinal Pierre, 144
- Dandelin spheres, 145, 159, 177, 183
- Degree two equation, 113, 261, 262, 361
- Derivative, 279, 295, 299
- Descartes, René, 256, 258, 291
- Description, 95, 98, 107–108, 137–139, 158, 191–193, 221–222, 259, 263, 341
- Diagonal, 30, 36, 52
- Diameter, 66, 67, 183
- Diamond, 35, 179, 180
- Dictionary, 127, 259, 386
- Difference of squares formula, 57, 97, 116, 368
- Differential, 304–307, 314, 316, 391, 394
- Differential equation, 306, 323, 330, 332, 341
- Dilation, 39, 140–142, 363, 373
 - effect on measurement, 73–75, 77
- Dimension, 214, 216–218, 226
- Discrete variable, 337
- Disk, 101, 193, 217
- Dissection, 110
- Dodecagon, 52
- Doughnut, 93, 101, 106, 191
 - See also* Torus
- Driving around, 26, 108
- e* (base of the natural logarithm), 393, 396
- Elementary, 340, 345, 364, 393
- Ellipse, 73, 139, 148, 158, 163, 362

- area of, 75, 189
 - perimeter of, 74, 189, 362, 364, 365, 371
- Elliptic integral, 364, 368
- Epicycloid, 196, 254
- Etch A Sketch, 292, 294
- Eudoxus, 65, 311
- Event, 220, 224, 257
- Exponent, 378, 385
- Exponential, 394–395
 - natural, 390, 394–395
- Exponentiation, 388, 390
- Extremal problems, 352–360

- Fermat, Pierre, 324
- Figure eight, 191, 192
- Fluent, 296
- Fluxion, 299, 303, 337
- Focal constant, 179–181, 183
- Focal line, 185, 261
- Focal point, 143, 148, 176, 182, 183, 261
- Focal property, 143, 159, 176, 183
- Frequency, 295

- Galilei, Galileo, 83
- Gauss, Carl Friedrich, 15, 393
- Generalization, 14
- Generic solution, 329
- Geometric quantity, 301
- Geometry, 42, 139, 155, 157, 226, 229, 349
- Golden rectangle, 59
- Graph, 223, 227, 243, 261, 279, 333
- Guessing, 326, 329

- Half-angle formulas, 347, 364
- Heading, 267, 283, 290
- Height, 49, 112, 114, 120, 135, 357

- Helix, 193, 195, 197, 245, 298
- Hemisphere, 91
- Heptagon, 52
- Hermite, Charles, 393
- Heron of Alexandria, 117
- Heron's formula, 117, 132, 136
- Hexagon, 28, 52, 67, 82
- Higher derivatives, 300
- Homogeneous, 113
- Hyperbola, 161, 163, 173, 176, 180
 - area of, 369, 371
 - length of, 368
 - right, 180, 181, 366–370
- Hyperbolic area, 372–375, 385, 386, 390
 - scaling invariance of, 373–374
- Hypocycloid, 196, 254
- Hypotenuse, 120, 344, 346

- Iceberg, 82
- Icicle, 352
- Incommensurable, 44
- Infinite sum, 68
- Infinitesimal change, 306, 310, 337
- Initial condition, 328, 329, 333, 337
- Initial position, 264, 267
- Integral, 339
- Integral sign, 338, 339
- Integral tables, 340
- Integration, 339
- Intersection, 100, 105, 166
- Invariant, 164–166, 171, 216
- Irrational, 45, 48, 59, 129, 381
- Isometry, 204, 205, 229
- Isomorphism, 204, 211, 386

- Jungle, 2, 9, 16, 17, 20, 396, 398

- Kinematics, 226
- La Géométrie*, 259
- Ladder, 197
- Lambert, Johann Heinrich, 67
- Lamp shade, 184
- Language, 45, 55, 255–256, 325, 340–341, 393
- Later, 274
- Law of cosines, 134
- Law of sines, 134, 238
- Leg (of a right triangle), 123, 236
- Leibniz, Gottfried Wilhelm, 68, 303, 304, 306, 310, 317, 332, 337, 338, 339, 385
- Leibniz d -operator, 304, 307, 313, 316, 337
as enzyme, 314
- Leibniz's rule, 309–312
- Lindemann, Ferdinand, 68
- Line at infinity, 170, 174
- Line of symmetry, 11, 177, 361, 366
- Linear, 261, 307
- Literacy, 256
- Logarithm, 383–385, 388
base a , 387
binary, 387, 390
Napier's, 386, 387
natural, 390–392, 394, 395
- Logarithm table, 384
- Log/exp pair, 389
- Long radius, 182, 363
- Lunatic, 169
- Map, 201, 207, 216, 244, 256
- Mathematical reality, 1, 37, 155, 396, 398
- Mathematics, 17, 26, 50, 396
- Maxima and minima, 355–358
See also Extremal problems
- Measurement, 32, 49, 95, 98, 192–193, 340
- Mechanical curve, 195, 227
- Mechanical relativity, 253, 285
- Method of exhaustion, 65, 70, 79, 84, 273, 334, 362
- Metric, 229, 230, 232
- Mixing board, 296, 302
- Model, 295, 296, 299, 353
- Modernism, 350
- Moment, 219, 274, 306
- Mosaic, 23, 28, 29, 47
- Motion, 94, 201, 219, 224, 226
helical, 246, 285, 291, 298
parabolic, 319
spiral, 257, 318, 319
uniform circular, 244, 247, 269, 293
uniform linear, 264–267
- Musical notation, 255
- Napier, John, 377, 382–383, 387
- Newton, Isaac, 270, 272, 275, 276, 296, 299, 310, 337
- Newtonian methodology, 297
- Nodal cubic, 348
- North Pole, 87, 352
- Now, 274, 275
- Number line, 203, 207
- Number circle, 227
- Octagon, 52
- Octahedron, 77, 128
- Orientation, 202, 209, 214, 219, 295
- Origin, 202, 208, 219

- Pappus of Alexandria, 93, 95, 97,
 103, 104, 107
 Pappus's theorem, 100, 103, 105, 332
 Parabola, 161, 163, 172, 188, 261,
 263, 365
 area of, 321, 326, 329, 342
 length of, 366, 369, 371
 Parabolic mirror, 188
 Parabolic rectangle, 189
 Parabolic section, 190, 320, 329
 Parabolic sector, 189
 Paraboloid, 187, 335
 Parallel, 142, 163, 167, 168
 Parallelogram, 37, 40, 213
 Path, 94, 97–99, 104
 shortest, 151, 153–155
 Pattern, 126, 138–139, 192, 349, 381,
 396
 Pentagon, 53, 132
 area of, 62
 diagonal of, 53, 56, 59, 61, 120
 Perimeter, 26, 33, 41, 64, 106, 112
 Perpendicular, 100, 142, 150, 213,
 287
 Perspective, 163, 164, 175, 176, 249,
 346
 Physical reality, 1, 5, 7, 24, 137–139,
 155, 397
 Pi, 67, 130, 393
 approximations to, 67
 transcendence of, 68, 131
 Pineapple ring, 89, 95
 See also Annulus
 Pitch space, 256
 Pitch-time, 257
 Place shifting, 317, 377, 384
 Plane, 141, 162, 168, 207, 261
 at infinity, 170
 Plato, 65
 Platonic solids, 31, 86, 128
 Point at infinity, 170–174
 Polygon, 23, 40, 107, 109, 110, 123
 area of, 64, 110
 closed, 109, 110
 description of, 108
 regular, 23, 29, 30, 51, 63, 107
 simple, 109
 Polyhedron, 30, 91, 136
 regular, *See* Platonic solids
 symmetrical, 31, 86, 128
 Pool table, 148, 150, 155–157, 183
 See also Tangent property
 Position number, 220, 222, 223, 264
 Position vector, 253, 267, 270, 283
 Powers of ten, 377, 382
 Projection, 141, 164, 171–172, 175
 central, 162, 171–172
 Projection point, 162, 163, 167, 172
 Projective geometry, 164, 165, 175
 Projective line, 170
 Projective plane, 170
 Projective space, 170, 171, 175
 Proof, 12, 20, 176
 Proportion, 32–34, 124, 133, 134
 Pyramid, 75, 78, 80, 136
 volume of, 75, 218, 342
 Pythagoras, 48, 131
 Pythagorean theorem, 48, 120
 generalized, 121–127
 infinitesimal, 344–345
 Quadratic equation. *See* Degree two
 equation
 Radius, 63, 70, 87, 150, 233
 Railroad tracks, 168
 Rate, 279, 301, 302, 322, 330
 Rate-time, 279

- Reciprocal, 312, 368, 372, 373, 392
- Rectangle, 37, 59, 95, 96, 102, 105
 - area of, 39–40
 - diagonal of, 46, 48
- Reduction strategy, 110, 123, 136, 286
- Reference point, 202, 208, 227, 232, 234
 - See also* Origin
- Reflection, 152, 204
- Relativity, 248, 249, 251
 - See also* Mechanical relativity
- Representation, 55, 60, 257, 258, 377
- Research, 11
- Rhombus, 35, 179
- Ridiculous solid, 104
- Ring, 88, 96, 102, 336
 - See also* Annulus
- Rocks, 316, 317, 335
- Rolling, 195, 196, 252, 343
- Rotation, 13, 105, 107, 193, 287, 346
- Running total, 338
- Sandwich, 25, 26
- Scaffolding, 308, 329
- Scaling, 33, 54, 180, 186, 205, 211
 - effect on measurement, 34, 41, 218
- Scaling independence, 34, 124, 134
- Scanning, 331, 333, 334
- Semicircle, 107
- Semiperimeter, 117, 136
- Shadow, 126, 162, 293
- Shift, 203, 204, 210, 248, 283
- Short radius, 182, 363
- Similar, 33, 34, 37, 53, 54, 180, 186, 366
- Sine, 123, 124, 127–130, 238–243
- Sine wave, 294, 371
- Sinusoidal arch, 333, 335
- Slantedness, 265, 271–274, 276, 279, 394
- Slider, 296, 302, 305
- Sliver, 99, 100, 103, 310, 322, 332
- Snowplow, 96, 101
- Solar oven, 188
- Solid, 218
- Soup can, 358, 360
- Space, 128, 168, 201, 206, 213–214, 230–231, 266
 - four-dimensional, 214–216, 226, 351
- Space-time, 224–227
- Speed, 264, 267, 283, 292, 344
- Sphere, 86, 147, 217, 246, 337
 - surface of, 91, 92
 - volume of, 90, 218
- Spherical cap, 93
- Spiral, 194, 348
- Spirograph, 196, 254, 291
- Spread, 57
- Square, 35, 36, 40
 - diagonal of, 42, 45, 55, 129, 236
- Square root, 56, 325, 326, 338, 380, 385
- Square root of two, 44, 55, 68, 131
- Star, 29
- Stick, 5, 6, 32, 95, 100, 126, 353, 358
- Stopwatch, 219, 221
- Straight line, 150, 224, 230, 232, 361
- Straightness, 165, 230–231
- Structure-preserving transformation, 153, 229
- Stuck, 6, 14, 230
- Surface, 217, 218
- Symmetry, 10–11, 13, 23, 26, 107, 112, 116, 122, 135, 170, 177
- Synthesis, 351, 381

- Tangent, 146, 147, 149, 150, 165,
 271–274, 355, 361, 395
 at infinity, 178
 Tangent property, 150, 154, 156,
 183, 187
 Ten, 377, 387
 Tetrahedron, 78, 85, 92, 128
 The, 340
 Theorem, 17
 Thing of fries, 32
 Time, 201, 219, 227
 Time line, 219, 223, 226
 Toric section, 191
 Torus, 94, 101, 197
 surface of, 106
 volume of, 103
 Transcendental, 68, 130–131, 371,
 393
 Triangle, 5, 110, 132, 164, 276
 angle sum of, 25, 133
 area of, 40, 111, 112, 117, 118, 128
 equilateral, 10, 14, 26, 49, 51, 52
 isosceles, 14, 136
 labeling scheme for, 119
 right, 34, 107, 113, 120, 123, 124,
 128, 236–238, 344
 Trigonometry, 111, 132, 136, 375
 Un-*d*, 325, 327, 328, 337, 338
 See also Integration
 Unit, 32, 38, 202, 208, 219, 233,
 243, 308, 387
 Unit circle, 181, 290, 363
 Unit hyperbola, 182
 Unit independence, 51, 301, 329, 387
 Unit square, 38
 Unit vector, 212, 250
 Unknown, 114
 Vanishing point, 168
 Variable, 258, 296, 307, 319, 321,
 349, 352
 Vector, 210, 213, 248, 249, 266, 287
 Velocity, 267, 269, 270–272,
 283–284, 319
 approximate, 274, 276, 280, 297
 instantaneous, 272, 277, 298, 303
 of a sum, 282, 285
 Volume, 41, 335, 357
 Z shape, 25