Making Sense of Calculus

Volume I: The Mathematics of Change and Variation

by
Stephen Monk

University of Washington

Spring 2007

Preface

Calculus is recognized as one of the great mathematical achievements of all time, the culmination of centuries of creative, and rigorous thought. With roots in Ancient Greece, its initial development took place in the 17th century. Yet, it was not completed until the mid-19th century. For those who know the subject well, it is a source of profound mathematical beauty and power. But for many who do not, it is only an object of awe and wonder. They witness it from the outside, as outsiders, from where it seems opaque and mysterious. If they studied calculus, they learned to manipulate symbols and solve arcane problems, but they are unable to penetrate its inner working. Why is this so? What is it about calculus that makes it so opaque, apparently so resistant to human inquiry? What is it that produces this air of mystery?

To be sure, calculus *is*, in ways, a difficult subject. As we know it today, it is the marriage of two very different bodies of mathematical ideas, each with its own set of concepts, principles, and practices. As is often true of the best of marriages, it is the unity, the synthesis, of the two parts of calculus that is the most significant and the most interesting. Nonetheless, each part stands on its own and can best be understood, initially, apart from the other. One part of calculus, called here *The Mathematics of Change and Variation*, is rooted in the subject invented independently and simultaneously by Newton and Leibniz in the 17th century. It consists of a set of conceptual tools for describing and analyzing situations of change found in nature as well as various academic disciplines. Although the new subject was an almost instant success in solving problems and giving new insights into phenomena of change, it was based on a set of concepts that were neither clearly defined nor completely understood. The opacity of these concepts and the absence of a clear framework for defining them brought about a crisis in the foundations of calculus, indeed of mathematics itself, that was not fully resolved until the 19th century. Many of the aspects of calculus that emerged from this crisis became quite independent of the applications of calculus itself and evolved into a subject that was nearly separate from calculus.

In addressing the problem of setting calculus on a sound foundation, both for the sake of communicating the sense and meaning of its concepts and for rigorously establishing their properties, the less well-known successors of Newton and Leibniz, such as Euler, Cauchy, and Weierstrass, realized that entirely new ways of thinking would be required. It is not possible to define concepts like the tangent line to a curve or the area a region with a curved boundary while still using ordinary finite methods, those found in subjects such as geometry, algebra, or arithmetic. New approaches and new modes of reasoning were required. These are the focus of the second part of calculus, called here *Reasoning about Infinite Processes*. It consists of a language, logical framework, and a set of standards for rigorously defining and establishing the properties of the concepts exploited by Newton and Leibniz, but proceeding from an entirely different basis, in an entirely different manner, than that of the two founders of calculus. While this new approach required a rethinking even of our number system, many of its ideas and modes of thinking go back to Ancient Greece.

One of the reasons for the immediate success of calculus, as it was developed by Newton and Leibniz, was that they also invented algebraic machinery that could be applied almost automatically when using the new subject to solve problems. Not only were the ideas of these two geniuses striking in their depth and power, but they could be instantiated in a set of algorithms that can be quickly mastered and used in a vast array of situations. It is this marvelous machine, developed for the quick and routine application of the mathematical ideas of Newton and Leibniz, that is presented in most introductory calculus courses. This in itself is a source of mystery though, because there is little in the use of this machine that suggests the ideas it is based on or why it works.

The air of mystery in calculus is heightened by the fact that a course built around the marvelous machine is inevitably very formal and far removed from ordinary experience. The basic elements of this material are such abstract notions as variables, functions, and operations on them. Despite the presence of problems involving projectiles being launched into parabolic trajectories and farmers' fields being shaped to maximize their area, most beginners have enormous difficulty connecting with the subject. Even a student who has mastered the algorithms and the techniques for solving the problems has little to go on if he or she wishes to get beyond the rules and formalism in order to explore what these things mean and why they work. Rather than being an entry or pathway into the core of calculus, such an introduction leads to very little except more of the same.

This book is intended for a reader who has had some contact with calculus, typically in an introductory course, and wishes to make better sense of the subject, to address its core ideas, to get what a student of mine once called "a behind the scenes look at what is happening when you take a derivative or integral." My purpose in writing it is to enable readers to grapple with these ideas directly, rather than to try to infer what they might mean from the workings of the marvelous machine. I hope this will make it possible for the reader to dispel the mystery that surrounds the subject and replace it by an understanding that evokes a first-hand sense of its beauty and power.

To make sense of calculus, to understand what the core ideas mean and why they work, one must inhabit a world in which these concepts arise explicitly and are the focus of authentic inquiry. But what does this mean in the case of ideas like the derivative and integral which are usually expressed in such formal and abstract terms and can so easily be confused with the techniques that have been developed for computing with them? It means, first of all, that initial versions of the concepts must be presented in a way that is free of algebra. They must be expressed in languages, including number sequences, diagrams, and graphs, that a wider range of readers can work with. These versions of the concepts must also be situated in concrete contexts that readers can readily engage with, so that they can come to feel the puzzling force of the questions these concepts were invented to answer.

This book is a presentation of the main ideas of calculus written with these guidelines in mind. As the title suggests, its chief goal is to help the reader *make sense* of a subject that seems to so many of those who attempt to understand it both obscure and fabulously complicated. Although intended for a wide audience, this book is the result of work over many years with college mathematics majors who have completed most of their mathematical study and are on the way to becoming high school mathematics teachers. The concrete settings used here include the motion of a small car going up and down a straight track, the data for the length of time between sunrise and sunset over the course of a year, and imaginary walks taken by people who move an infinite number of steps, but go only a finite distance. The forms of representation include diagrams, line graphs, bar graphs, and data tables in a computer spreadsheet, as well as formulas.

Calculus means here the study of phenomena of change, in situations in which related quantities vary. This variation can be continuous as well as discrete, meaning that a quantity can assume one value and then another, but not necessarily assume all values in between. Some of these contexts are presented through imaginary stories and others through dialogues. My guiding principal has been to use *any* means that seems likely to genuinely engage the reader in a way that opens onto deeper understandings—even some that are highly unconventional.

A student once described this approach as "doing calculus without calculus," implicitly suggesting a view widely held by students that unless one is manipulating complicated formulas in obscure ways, one is not *really* doing calculus. It's as if one of the defining properties of calculus in the minds of many students is that, at any given time, much of what one is doing is beyond one's comprehension. They even seem to feel on those occasions when they do grasp an underlying idea that something is amiss: either they have actually misunderstood the idea, or what they've understood is only a low-brow version of the idea. I believe that this view is incorrect and needlessly self-limiting, that calculus can be understood by far more people than is presently the case. My goal here is that the reader be able to understand and use calculus in all of its guises and forms, including the version found in textbooks. But I have rejected the widely shared view that the only way to express the ideas of calculus is through algebra, so that a thorough mastery of that subject must serve as a gateway to any contact with, let alone an understanding of, calculus. Thus, an important difference between the material presented here and the contents of standard textbooks is that while the final goal here is an understanding of calculus in the formal and abstract language of formulas, a good deal of the work toward that goal is done in more informal and accessible languages, of number sequences, diagrams, and graphs.

This book is divided into two volumes, each addressed to one of the two main parts of calculus. Volume I is called *The Mathematics of Change and Variation*, and Volume II is called *Reasoning about Infinite Processes*. Reflecting the separateness and distinctiveness of the two parts of calculus, the two volumes have been written in such a way that they can be read in either order. Volume I begins with a brief conceptual history of calculus, in which the main ideas of the subject are first presented as solutions to problems going back to the Ancient Greeks, but seen as urgently pressing in the 17th century as a result of the emergence of the natural sciences. Proceeding with the further elaboration and development of the subject in the 18th century, it culminates with the crisis in calculus caused by its foundational difficulties and the reform of the subject arrived at in the 19th century. Although the historical development of the subject relates more directly to the issues taken up in Volume I, readers of Volume II may also find it of some value.

The initial goal of Volume I, *The Mathematics of Change and Variation*, is to give the reader a strong sense of what this subject called calculus is all about. What problems does it attempt to solve; what are the principle means by which it goes about doing this? The first chapter of this volume, on the history of calculus, is an important part of working toward that goal. Chapter 2 consists of a different approach to addressing the same question, of what calculus is about, presenting a set of cases or vignettes in which students struggle with versions of some of the problems calculus is designed to solve, but couched in a language an outsider can not only understand, but use to solve these problems. Chapters 3 through 5 then present the conceptual structure of calculus in a series of increasingly formal languages, with each chapter building on its predecessor. The first language, of

Spreadsheet Calculus presented in Chapter 3, is of sequences, tables, and bar graphs, as they can be manipulated in a computer spreadsheet. The second language, of *Visual Calculus* presented in Chapter 4, is of graphs, as they can be manipulated using informal approaches to the tangent line and area under a graph. The third language, of *Symbolic Calculus* presented in Chapter 5, is of algebra, the standard language of calculus.

Volume II, *Reasoning about Infinite Processes*, also consists of a set of linked chapters in which a few central concepts are slowly built up. In this case, the concepts are developed in one context, as it relates to an important historical problem, and then another. Chapter 1 of Volume II begins with the well-known Dichotomy Paradox of Zeno, about the impossibility of reaching a wall by repeatedly taking steps each of which is half the length of its predecessor, and it then moves to the more general topic of infinite sequences of numbers. Chapter 2 takes up the problems, first studied by the Ancient Greeks, of determining the area and perimeter of a region with a curved boundary, especially that of a circle. Chapter 3 explores the issues around the notion of an infinite series first used by mathematicians in the 14^{th} century, but central to the development of calculus in the 17^{th} and 18^{th} centuries. Chapter 4 presents the slow evolution of the concepts of continuous function and limit that proved to be so essential to full development of calculus and of its foundations in the 19^{th} century. Beneath these disparate problems, each in its own particular context, a small set of issues slowly emerges, along with ways of reasoning and communicating about them, that ultimately can be seen as forming a single unified whole. This is the small set of core concepts and tools for reasoning about infinite processes, for rigorously defining and working with the central ideas of calculus.

This book does not "cover" calculus. Even the thousand-page tomes that students buy don't begin to do this. Many aspects of the subject are not mentioned here at all. Rather, the goal of this book is to enable the reader to grasp and to work with the basic ideas of the subject, to make sense of it, and to directly experience its beauty and power.

CHAPTER 1. The Historical Development of Calculus

§1.1. Overview

The simple story of the development of calculus is that it was invented in the last part of the 17th century by two men, Isaac Newton (1642–1727) and Gottfried Leibniz (1646–1716), that they did it independently and simultaneously, and that this fact led to an enormous battle between their supporters as to who invented the subject first and whether or not Leibniz stole Newton's ideas. But the history of calculus is far more complicated and far more interesting than that. The development of the subject starts with the Ancient Greeks, in the period 585 BC to 200 BC, but it does not end with Newton and Leibniz. When their work on calculus was finished, the subject consisted mainly of methods for solving particular problems, the basic ideas of the subject were at an early stage of development, and the system for making logical claims about it was extremely shaky. The subject we know today is the final result of several stages of development that took nearly two hundred years to complete, the work of generations of brilliant, but less well-known, mathematicians.

In studying the details of the history of calculus, as it emerged in the writings of Newton, Leibniz and many of their successors, one sees that the mathematical ideas of the subject are barely recognizable to a modern reader. This is not simply a matter of presentation, notation, or terminology. The fact is that earlier mathematicians saw the world very differently than we do, they understood basic concepts in a different way, and had different questions in mind. Fortunately, there has been a great deal of fruitful work in the history of calculus in recent decades giving us much greater access to the development of ideas in the subject[*]. Using the work of these authors as a guide, one finds enormous diversity and complexity within the ideas of the great mathematicians. Although this makes studying the history of calculus much more of a challenge than might have been imagined, it provides a route into the inner workings of the subject,

---

[*] After Boyer, we have Baron, Edwards, Grattan-Guinness, Toeplitz, and others, in addition to those who write for a wider audience, such as Berlinski and Gleick. A full bibliography is given at the end of each volume of this book.

while at the same time giving it far more sense of life than most people imagine it has.

Several important themes begin to emerge as we explore the history of calculus, each providing us with its own pathway into the subject. Among these are:

- There are enormous differences between what Newton and Leibniz accomplished and how they conceived of the subject they were inventing. Examining these differences and their sources, as well as their consequences, is a way of breaking basic ideas apart and seeing how they fit together.

- They did not create this subject from nothing. Much of what Newton and Leibniz accomplished had clear and strong antecedents. They were answering questions that were very much in the air in the previous hundred years, and they were building on partial solutions developed by an array of famous mathematicians. As is the case of almost all inventors of new subjects and theories, Newton and Leibniz were, in the old phrase, standing on the shoulders of giants. Each of them took the ideas that were circulating at the time and built general systems for dealing with a great many diverse problems. Understanding the background and sources of their work gives us a stronger sense of underlying issues calculus addresses and why the subject plays the central role it does.

- Even if we organized and wrote out everything these two men accomplished in what we call calculus—something that neither man ever did—then the subject we would come up with would still be almost unrecognizable to us today. There was a great deal of work to be done by their successors. Trying to understand why the work of Newton and Leibniz's needed to be reconceived and reorganized and then how this was accomplished by their successors is a way of opening up the tightly bound conceptual system at the heart of the modern subject.

- Neither man ever came close to placing his set of ideas on a firm foundation. Ordinarily, we think of mathematical ideas as being presented through a strictly logical procession of theorems built from basic definitions and axioms. This is not at all the case of early versions of calculus. Not only was such a tight deductive style not universally accepted at the time, but there were no common agreements as to what beginning assumptions should be made in order to arrive at the main claims about these new ideas. But the problem is not simply one of logical reasoning. There is a serious question

> as to the sense we are to make of much of what Newton, Leibniz, and many of their successors did. Delving into their versions of calculus, one still has difficulty trying to clarify and make more precise, in one's own terms, how they thought about fundamental issues. Although frustrating at times, such study can illuminate the pathways into some of the most difficult and gratifying aspects of the subject.

The goal of this chapter is to examine the complexities of the invention, emergence, and establishment of calculus, to present a brief thematic history of the development of the core concepts of the subject. But this is not a history of calculus for its own sake. It is a history intended to reveal the development of the main ideas of the subject and the principles and points of view linking them. As will become apparent, the structure of this book is an unusual one, breaking the study of the subject into two main parts that can be read in either order, one exploring the conceptual apparatus of the subject and the other exploring its foundations. The history contained within this chapter should give a sense of what the goals of these two parts are and why such a partition of the subject makes sense.

In order to organize the history we are about to study, I have divided the story of calculus into four distinct segments. However, as most of those who study history understand, whether the focus of their study is a nation or a body of ideas, the division of history into separate and distinct chapters is always arbitrary to an extent, a product of the author's intentions and views. People do things, events happen, and there are consequences. When we look back on what has taken place, we organize what we see in order to make sense of it. In doing this, we naturally break these stories into segments or chapters to help us in our sense-making. The main segments of the history of calculus, as it is about to be told, and the sections in which they are discussed, are:

- §1.2. *Problems in the Air* (1600–1650)
- §1.3. *Solutions to the Three Famous Problems: Newton & Leibniz* (1650–1720)
- §1.4. *The Emergence of Calculus as a Subject: the Age of Euler* (1720–1780)
- §1.5. *The Rigorous Reform of Calculus: Cauchy to Weierstrass* (1780–1870)

§1.2. Problems in the Air

The 16th and 17th centuries mark the beginning of the modern era in Western mathematics. The mathematics of the Ancient Greeks, of 1500 years earlier, had been completely recovered and absorbed, and the algebra developed in the Arab world 600 years earlier had been taken in, mastered, and developed further. The period 1600 to 1650, just before Newton and Leibniz arrived on the scene, was one of enormous activity in mathematics and science, with great progress in geometry, algebra, and the theory of numbers, made by such men as Napier, Galileo, Kepler, Cavalieri, Wallis, Pascal, Fermat, and Barrow*. Most important for the history of calculus, the subject we now call *analytic geometry* was developed by Descartes. This made it possible to link geometric questions about curves with algebraic questions about equations, a tool now taken for granted in calculus. Many of the questions worked on in the 17th century had their origins in the Ancient Greek mathematics, while others arose with the re-emergence of science starting with the Renaissance. For our present purposes, we will focus on three of these problems.

The Tangent Problem

This problem goes back to the Ancient Greeks, who had constructed what they called the tangent line to a circle at a point P on the circumference of the circle. They did this by drawing a line, called a radius, from the center of the circle to the point P, and then constructing the line perpendicular to the radius at P (Figure 1a). This line captures the direction of the circle at the point P, although what is meant by the direction of a curved line is not entirely clear, since a line

* The term "mathematician" is problematic for historians of mathematics. It is a modern term, referring generally to someone who spends most of his or her time adding to or applying a certain body of concepts and practices. Usually, this is a professional activity for which this person has been accredited by advanced degrees and for which he or she is paid. None of these criteria makes sense when applied to those who lived in the 17th and 18th centuries. But one still wants to say that people like Newton, Leibniz, Fermat, and Descartes were mathematicians, because they contributed so much to the subject we now call mathematics. But Newton's work in physics is much more important to his reputation than his work in mathematics. Leibniz and Descartes are both better known for their contributions to philosophy than to mathematics. Fermat earned a living as a banker and did mathematics in his spare time. Nonetheless, for the sake of simplicity, I will refer to these men and to all the other members of the historical community from which calculus emerged as "mathematicians."

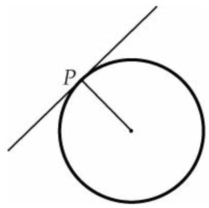

Fig 1a–Tangent to a circle

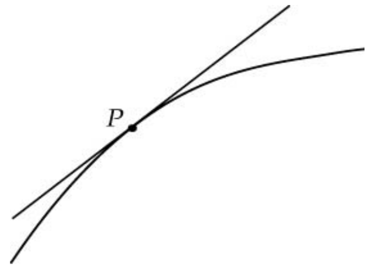

Fig 1b–Tangent to a curve

that is curved changes direction (in whatever way one defines direction) as one goes from one point to another.

One of the important questions for Greek mathematicians was about how the notion of a tangent line to a circle might be extended to other standard curves. (Figure 1b). They assumed that such a line must exist, but could not devise a simple method for constructing it, since there is no counterpart for such curves to the center or radius of a circle. They observed that one key property possessed by the tangent to a circle is that it touches the circle, but does not cut through it. Most of us find it intuitively appealing to say that a line with this relationship to any curve at a point *P* (of touching and not cutting through) captures the direction of the curve at the point *P*. While this is true for most curves, it is not true for all. Moreover, constructing a line with this property is still a difficult task in most cases.

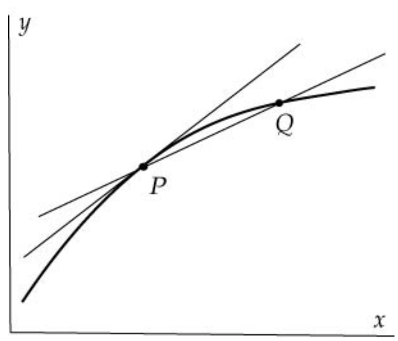

Fig 2–Secants & Tangent

By the 17th century, mathematicians had come upon a different approach to the problem of the existence and construction of a tangent line to a curve at a point *P*. Here, the curve is drawn in connection with coordinate axes, as is done in Descartes' analytic geometry, and a second point *Q* is drawn on the curve near the point *P*. A line, called the *secant line*, connecting the two points *P* and *Q* is then drawn. If the point *Q* is pulled in very close to the point *P*, then the slope of the secant line is very close to the slope of the desired tangent line. But what does "very close" mean? Although there is no difficulty in drawing such a line, there are enormous questions about the relationship between the slopes of such lines, of which infinitely many can be drawn, and the slope of the single tangent line they are supposed to be related to.

The Area Problem.

This problem also goes back to the Ancient Greeks. As is generally known, we measure the area within a figure that has a boundary made up of straight sides by determining how many squares of a standard size would fit into the region. Determining the number of such squares might require cutting and pasting of the figure and the squares, but it can be done, eventually. But what about a plane figure with curved sides (as in Figure 3)? Surely, such a figure has an area. The intuitive sense most of us have of the notion of area of a region is that it indicates something like how much paint we would need to cover that region. If I outlined a region on a floor that has a curve as its boundary, no matter how bizarre the curve, the job would require a certain amount of paint. The more area I have, the more paint I would need. So, any region I could make on my floor would have an area. But how are we to imagine fitting squares into such a region? Squares have straight lines for their sides; how can we place such a figure along a curved boundary?

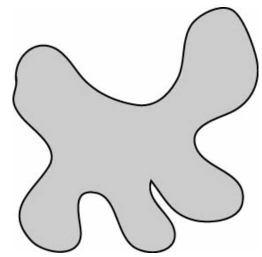

Fig 3–Area of a region

One of the outstanding mathematical problems in Greek mathematics was the case in which the given region is a circle. This is such a basic figure that it's difficult to imagine that we can't somehow or other determine its area. They observed early on that they could draw a regular polygon inside a circle (starting with a square, as in Figure 4a) and that its area can be computed and that it is relatively close to the area of the circle. They can then draw another polygon inside the circle containing the first one (as in Figure 4b) that has an area even closer to that of the first. They knew that this process could be continued indefinitely in order to approximate the area of the given circle, but they could not use it to determine to their own satisfaction the precise value of the area of the circle.

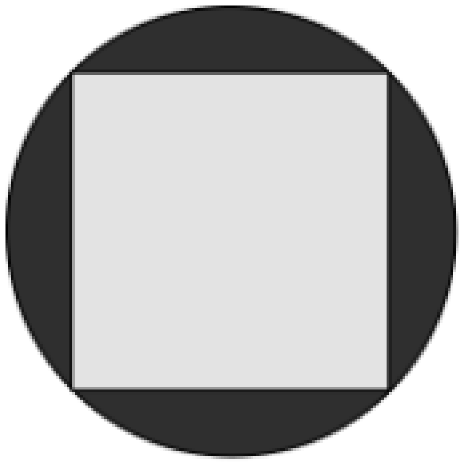

Fig 4a–Square inside a circle

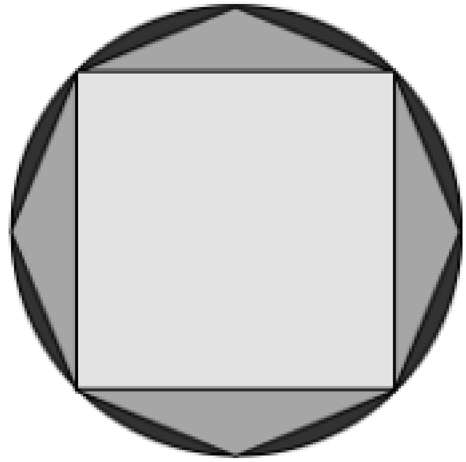

Fig 4b–Octagon inside a circle

It is worth noting, however, a certain similarity between this so-called "method of exhaustion" that the Greeks used to determine the area of a circle and the method for determining the slope of a tangent line by using the secant lines. In each case, we have a geometric object that we might call an *ideal object* because we think we know many things about it, but we cannot quite get hold of it enough to measure it. In both the case of the tangent line and the area of the circle, this is because of the somewhat obscure nature of a curve—which, by definition, *curves*. We then observe that there is a related geometric object (which I call a *discrete object*) which is easily drawn and measured and which also has the property that its measurement is close to the measurement of the given ideal object. But how close is close? And how can we use the closeness of the discrete object to assign a definite measured value to the given ideal object? For all of its complications, this ancient idea turned out to be central to the development of modern calculus.

By the second half of the 17th century, the era of Newton and Leibniz, the area problem was increasingly thought about in terms of a region defined by a curve, as in Figure 5. Given such a curve, one that lies entirely above the horizontal axis, there is a well-defined region between the curve and the axis and between two vertical lines given by the two values $x=a$ and $x=b$. We now call this the "region under the curve from a to b." In some particular cases, this problem was approached in a manner similar to the Greeks' approach to the area of a circle, with the role of the polygons in the circle played by a region made up of rectangles placed side by side in the region (as in Figure 6). The area of the rectangular region is not equal to the area of the region under the curve, because of the mismatches at the top of the curve. But it is not difficult to show that this error can be made small by increasing the number of rectangles used. Once again, we have an ideal object, the region under the curve, which we wish to measure, but can only get at by a discrete object, the rectangular regions, one that can be made as close as we like to the given ideal object.

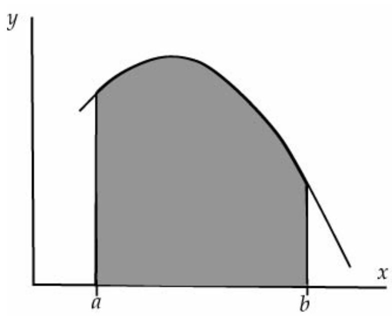

Fig 5–Area under a curve

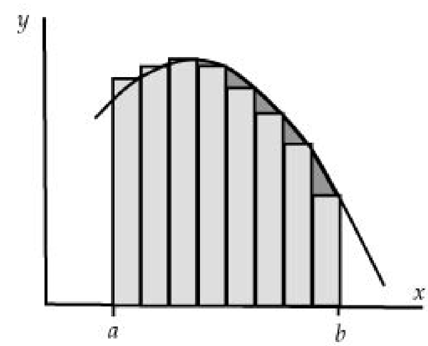

Fig 6–Area approximated by rectangles

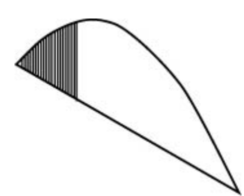

Fig 7–Area by slices

By the first half of 17th century, mathematicians began to take a radically different approach to the Area Problem. They conceived of a region as being made up of small slices that are infinitely narrow, of which there would then have to be infinitely many. This is an entirely different conception of area, related to the widespread use of infinitesimals at that time. The slices used here do not have width measured by an ordinary number and so the area of each slice, whatever that would mean, is not an ordinary number. Only by putting together infinitely many of these slices do we fill the region in question. Nonetheless, the area of this region is to be thought of as a kind of "sum" of the areas of the infinitely narrow slices. Of course, they knew that they could not literally add the infinite number of infinitely small areas, but they came up with ingenious methods for determining, in certain cases, what this sum would be. Although this approach did yield some impressive results, it also led to a great deal of controversy and confusion.

The Problem of Non-Uniform Motion

The problem of describing and analyzing motion goes back to the Ancient Greeks. Aristotle, who lived in the 4th century B.C., set forth the first set of laws of motion in his *Physica*, where he studied what we would call *uniform motion*, the movement of an object at a constant speed*. Although he understood that objects could move in ways in which their speed was not constant, he barely touched on this possibility, let alone analyzed it. This is due, in part, to the limitations of the mathematical tools he had at his disposal for studying such phenomena.

The first studies of non-uniform motion in the West were by scholars at Merton College at Oxford University in the 14th century. They studied the motion of objects that increased their speed, but did so in a very steady manner. A model for such motion would be a situation in which a car accelerates steadily from 20 mph at one time, to 25 mph 1 second later, to 30 mph 1 second after that, to 35 mph 1 second after that, and so forth. Although they did not have the tools for carefully defining such a concept, they did consider what we would now call *instantaneous speed*, by which they meant speed at an instant, in contrast to what we now call *average speed*, the speed of an object over a time interval, which is obtained by dividing the distance covered by the object in this time interval by the length of the time interval.

One of the greatest scientific controversies of the 16th century was around the motion of the sun and the planets in the solar system. While the traditional teachings of the Roman Catholic Church held that all the heavenly bodies including the sun move around the earth, some scientists were beginning to suggest that this is not true, that the sun is at the center of the solar system, and that the earth, just like any other planet, revolves around it. Galilei Galileo

---

* In ordinary everyday language, we do not make a distinction between the terms "speed" and "velocity." They both refer to how fast something is going. But once we get very far into the study of motion, as we will, we will want to make a very important distinction about how fast something is going in relation to its *direction*. This distinction is usually marked by using the term "velocity" in contrast to "speed." For the purposes of this chapter, however, this distinction will not be made. For now, until this distinction becomes important to us, these words will be taken to be synonyms.

(1564–1642) was among those who agreed with this so-called "Heliocentric Theory." He believed further that the question could be illuminated by a greater understanding of gravity and how it pulls objects toward the earth. Thus, he set out to understand the pattern by which a freely-falling object falls to the earth. His writings on this subject and the larger questions about the arrangement and motion of the planets and the sun were regarded by the Church as heretical and he was condemned to death for them. He only escaped being burned at the stake by the intersession of influential friends who succeeded in having his sentence reduced to house arrest for the remainder of his life.

The motion of a freely-falling object was in many ways mysterious at the time of Galileo. Many believed that such an object falls at a constant speed. And as we know from the story of the two objects being dropped from the Tower of Pisa, most assumed that a heavier object falls faster than a lighter one. One of Galileo's most important findings was that all objects fall at the same speed and that this speed increases proportionately with the amount of time the object falls. Thus a freely falling object has the same pattern of motion as the type studied by the Mertonian scholars. Using his results, Galileo could also accurately describe how far such an object falls after a given amount of time. Central to Galileo's writings was a clear and strong insistence that the speed being discussed here is instantaneous speed, speed *at an instant* and not average speed, speed *over a time interval*. But he took this notion of instantaneous speed as a given—something that everyone agreed existed and could be measured.

Galileo's results suggested an approach to other kinds of motion in nature, especially the motion of the planets and the sun, as well as a way of framing these problems, which we will call the Problem of Distance and Speed: If we know what the pattern (in time) of the *instantaneous speed* of an object is, how can we determine its pattern of the distance (in time), and *vice versa*, how can we obtain instantaneous speed vs. time information from distance vs. time information?

But beneath this problem is another one that was not yet taken seriously by the scientists and mathematicians of the 17th century, the so-called *Instantaneous Speed Problem*: What does it mean to talk about the speed of an object at a given instant of time and how should we measure it? Most of us would say that to measure speed, you always divide the distance covered by an object by the time the object was moving. But this is the *average* speed of the object over that time interval. When pressed, many people say that the instantaneous speed is just the average speed over a very tiny time interval. But that is not really what Galileo and his contemporaries meant by instantaneous speed and it is not what scientists nowadays mean by this term. They would agree that the average speed over a very tiny time interval approximates the instantaneous speed very well, but that, conceptually, there is a difference. An object falling freely is always speeding up. If you measure its average velocity over a small time interval you will fall short of its actual speed at the end of that interval, because you are still taking into account speeds before that time, which are slower. The instantaneous speed is another ideal object, just like the slope of the tangent and the area under a curve are. One can approximate it by average speed, which is a discrete object (mathematically), but even though one is close, one is only close. One doesn't yet have the real thing.

§1.3. Solution to the Three Famous Problems: Newton & Leibniz

The mathematical scene in the middle of the 17th century was one of enormous ferment, of many partial solutions to a set of long-standing problems, with a number of illustrious mathematicians working on them. As so often happens in such situations, the final complete solutions to these problems were not devised by any of the established mathematicians, but independently by two newcomers, Isaac Newton (1642–1727) and Gottfried Leibniz (1646–1716). Each man seemed to be able to take a new and original approach to these problems that enabled him to arrive at solutions that had eluded his more established, more knowledgeable older colleagues. It seems likely, in the cases of Newton and Leibniz, both brilliant and creative thinkers, that rather than being a handicap, their lack of experience was an advantage, in that they were not the captives of older, less productive ways of seeing the problems.

Newton was a graduate student in mathematics at Cambridge University in the period 1664-1666, living in Lincolnshire, away from all distraction and professional contact. He was living there, where he had grown up, because the university was closed due to an outbreak of the Black Plague that had swept much of the south of England. In this period, when he was 22 to 24 years old, he developed and wrote down many of the ideas in calculus, gravity, and optics that later made him famous. He showed these writings to some English scientists and mathematicians of the time, but he did not make any of his discoveries in calculus public. His reasons for not broadly publishing what he had accomplished, although complicated, seem to be related to a deep (almost pathological) aversion Newton had toward having his work criticized in any way.

Leibniz had been a well-known child prodigy from Leipzig, Germany, who was trained in the law and was on a diplomatic mission to Paris in 1672. His diplomatic duties were apparently not very demanding, since he had the time to study mathematics with some of the leading mathematicians of the period living in Paris. By 1675, at 29, a relative newcomer to mathematics, Leibniz had developed the main ideas of his approach to calculus, which he sketched in two brief papers published in 1684 and 1686. Not regarding himself as primarily a mathematician, Leibniz did not follow up on this work until his ideas were taken up by other mathematicians several years later.

Although Newton's ideas on calculus were known to many people in England and Europe, there was nothing in print about them in 1687, many years after he developed them, when he published his great work, *Principia mathematica philosophiae naturalis*, in which his three Laws of Motion are developed. While many of his ideas about calculus are used there, this is a book about kinematics, the study of motion, and not mathematics—in spite of the fact that its title contains the word *mathematica*. By publishing in this way, Newton was choosing not to present his ideas on calculus in a way that might open them to criticism, but only as a sidelight, as a conceptual and computational tool, as he was presenting his revolutionary ideas on physics, which he was very confident of.

It is likely that Leibniz published his ideas on calculus in response to hearing that Newton would be presenting his ideas on the subject as part of his *Principia*. As soon as Leibniz presented his ideas, Newton's friends in England accused him of stealing Newton's ideas. They pointed to the fact that Leibniz had spoken to Newton about calculus while visiting London early in his career and claimed that he must have learned of Newton's approach at that time. It seems very unlikely though that the highly secretive Newton would have shared much of his thinking with Leibniz. Moreover, Leibniz's approach is sufficiently different from Newton's that it is not clear what he would have learned from such a conversation. But none of this prevented the supporters of both men from engaging in a vicious battle over who deserved the title "*The* Inventor of Calculus."

Neither Newton or Leibniz ever wrote a complete presentation of his ideas in the field that became calculus. Instead, as is common in mathematics, they wrote briefer reports on their solutions to particular problems, in effect, demonstrating their thinking through these solutions. We begin our discussion of their work in a similar manner through their solutions to the three problems discussed in the previous section along with presentations of several other themes that emerge in this work.

Solutions to the Tangent Problem

Newton's solution to the Tangent Problem reveals his very particular approach to a number of issues. Although his focus is on a curve, it is not simply as a static figure lying on the paper, but as a path that a point or an object can move along. Imagine a point (even a ball) moving along the path shown in Figure 8. As it does so, its motion can be thought of

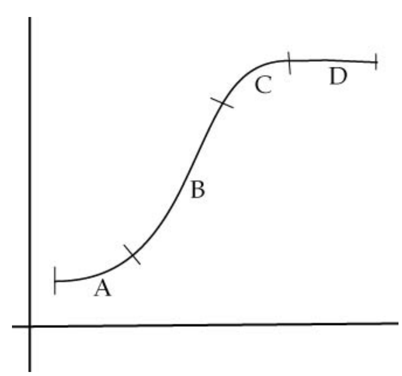

Fig 8–Direction of a path

in terms of two components, one in the horizontal direction and the other in the vertical direction. As the ball moves along in the portion of the curve labeled A, it moves horizontally in time much more than vertically, the horizontal component of its motion is much greater in time than its vertical component. Thus, the horizontal component of its speed is much greater than the vertical component of its speed. As it moves along the B section, it moves vertically in time much more than it had, although it continues to move horizontally at the same time. It's vertical speed has increased greatly, while its horizontal speed has probably decreased from earlier. In section C, the point once again has roughly equal amounts of vertical and horizontal speed, and in section D it has zero vertical speed, while it moves horizontally.

The basic insight arising from this way of thinking about motion along a curve is that the relationship, in fact, the *ratio*, between the vertical and horizontal speeds depends on the direction of the curve. But this statement can also be turned around to say that the direction of the curve is determined by the ratio of the two speeds. Although terms like "speeds" and "direction of the curve" have been used as if they are perfectly clear and unambiguous, Newton intended them to mean *instantaneous* speed and the direction of a curve *at a particular point*. The latter term is exactly what would be indicated by the slope of the tangent to the curve. Thus, Newton defined the slope of the tangent to a curve as the ratio of the vertical and horizontal components of the (instantaneous) speed of a point moving along the curve. This worked for Newton, since he saw no difficulties around the notion of instantaneous speed.

The strength of Newton's approach to the Tangent Problem is that it is based on an imaginable physical phenomenon. If you feel comfortable with the notion of instantaneous speed, as many people do, as well as these refinements in terms of components of this speed, then the notion of the slope of a tangent line and, therefore, the tangent line itself, makes sense and has a logical basis. There is a shortcoming in this approach in that it tells us very little about the numerical value of the slope of the tangent line in those situations in which the information we have about a curve is determined by an algebraic formula. In showing how to obtain the slope in this case, Newton made some of the specific arguments that resulted in claims that seemed questionable to others.

For the purposes of computation in the case in which a curve is described by an algebraic formula, Newton used the approach of approximating a tangent line by its discrete counterpart, the secant line, discussed in the previous section and shown in Figure 9. It connects the point *P* and a nearby point *Q*. The slope of this line is close to the slope of the curve at the point *P*. The closer you choose the second point *Q* to *P*, the better the approximation. Except in the case of a curve that is actually a straight line, no one secant line will be just right, though, because changing the point *Q* will give a different line and a different slope. What one wants to do is to choose a secant line as close as possible to the intended tangent line. But, of course, any point *Q* you choose could not be *the* closest possible, because there is another point between it and *P*. There is no single point next to *P* that you can use. If you actually made *Q* equal to *P*, then this approach—of using a secant line—would completely collapse, because there would be no second point and no secant line to base it on. All you can do is get a better and better approximation. It's a never-ending story: you get closer and closer, but you never get there—because you can't.

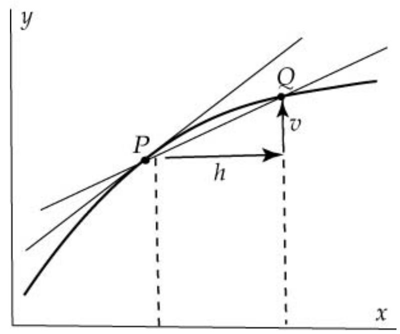

Fig 9–Slopes of secants and tangent

Newton took an algebraic approach to the problem of relating the slopes of the secant lines to the desired slope of the tangent line. He let *h* be the horizontal component of the distance between *P* and *Q*, and used the underlying expression for the curve to write an algebraic expression for the slope of the secant line. This expression, which involves the quantity *h* among others is indicated in line (1) in the diagram in Figure 10. He could then examine what happens to the slope of the secant line as Q gets closer and closer to *P* by examining the behavior of this algebraic expression for the slope as *h* gets closer and closer to 0. The crucial requirement is that *h* can never equal 0. Using standard techniques from algebra, he rewrote the expression for the secant line by breaking the expression for the slope into two parts: those terms that do not involve *h* and those that do involve *h*. This is indicated in line (2) of the sketch in Figure 10.

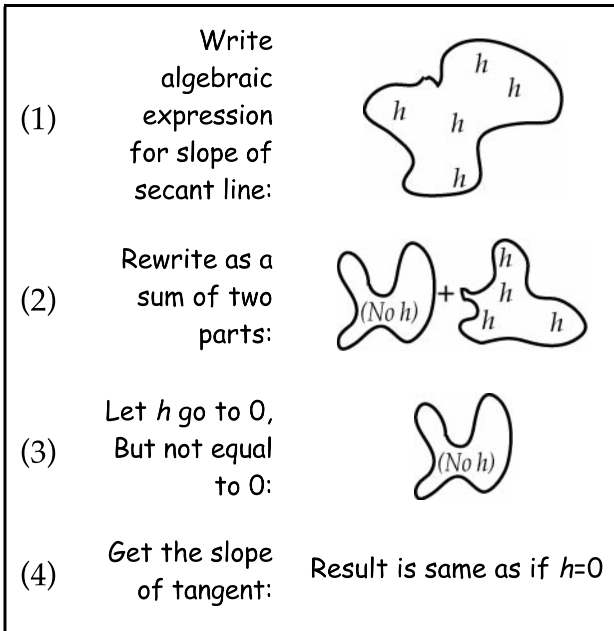

Fig 10–The blob diagram

Newton then reasoned as follows: We are interested in the situation in which *h* is made closer and closer to 0. When this happens the terms in expression (2) that do not involve *h* (the first "blob") are unaffected. On the other hand, the terms that do involve *h* (the second blob) are affected: they become closer and closer to 0. In effect, they disappear. His way of putting this is that they are *negligible terms* that are not 0, but which we can safely ignore. The terms that remain, the first blob, must therefore describe the slope of the tangent line, as indicated in line (3) of Figure 10. In other words, the outcome after letting *h* get closer and closer to 0 is the same as if Newton had actually let h equal 0. But the requirement was that *h* was never to equal 0. Operationally, letting *h* get closer and closer to 0 had the same effect as setting *h* equal to 0—a move that was strictly forbidden. This argument has always struck many people as both amazing and a bit suspect. It starts with a value *h* that is definitely not 0 and it proceeds through algebraic steps to a point at which we can assume that *h* is effectively 0. It disappears and we get a simple answer to the difficult problem of computing the slope of the tangent line.

This extraordinary process certainly has a paradoxical and puzzling quality to it. Through algebra and notions like "negligible terms" Newton is able to tell us what happens as the result of an infinite processes of the secant lines getting closer and closer to the tangent line. It is one of several infinite processes that are crucial to calculus. Anyone who has studied calculus will recognize it, since it is still found, in one form or another, in every introduction to calculus, the thousand page tomes and the comic book versions. Needless to say, Newton had some explaining to do, to convince his readers that this was an acceptable way to establish such an important concept. His explanation, which has a certain plausibility to it, used language like "evanescent increments," "negligible, but not 0," etc. It was not clear how others were to make sense of these terms and use them in their own work. How is one to know when terms disappear and when they don't? It was a source of controversy from the start.

The solution to the Tangent Problem given by Leibniz, like that of Newton, is based on the relationship between the tangent line at a given point *P* and the secant line between *P* and a nearby point *Q*. Whereas Newton's approach is to start with such a secant line and bring the points ever closer and closer, Leibniz's approach is to start with a line similar to a secant line, but one in which the two points are already "infinitely close," by which Leibniz meant that the distance between them is an infinitesimal number, a number that is smaller than any possible positive number, but still not 0, as in Figure 11.

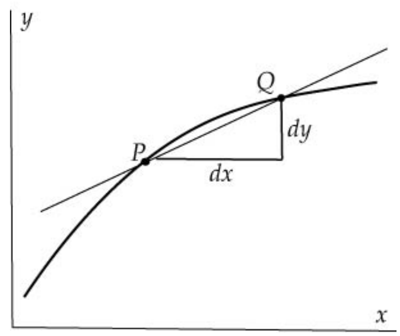

Fig 11–Tangent by differentials

Although the concept of an infinitesimal number was used by many mathematicians of his time, Leibniz used it much more extensively and systematically than anyone else. And even though the concept was slowly dropped from formal proofs by the mid-18th century, we still use the notation that Leibniz developed around the concept. This notation enables us to loosely employ Leibniz's ideas as an informal reasoning tool. It's worthwhile, therefore, to pause, to try to understand Leibniz's view of this concept, especially since it is so easily misunderstood.

For Leibniz and his contemporaries, an infinitesimal number is a radically different kind of number, one that is smaller than any ordinary positive number. It is quite different from the very, very small *variable* quantity used by Newton that can be made as close as one likes to 0, so that certain algebraic expressions become "negligible." Whereas Newton's numbers were in the process of going to 0, and not getting there, Leibniz's numbers were already there—but still not 0.

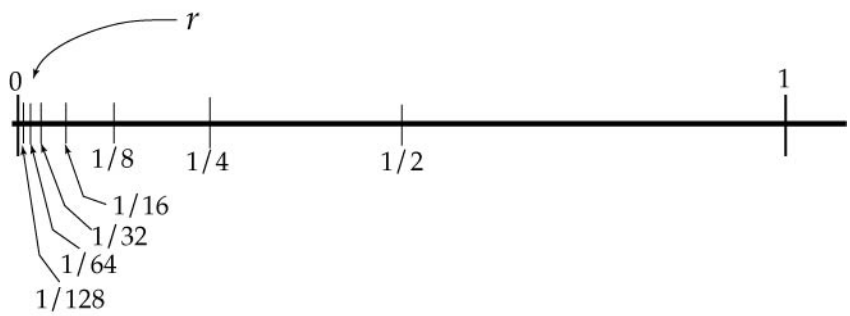

Fig 12–Where are the infinitesimals?

We have already encountered the notion of an infinitesimal in the discussion of area of region made up of infinitely thin slices. Since the slices have no width, they can't have ordinary area. Nonetheless, the slices can be combined to form a region that does have ordinary area. One way to think about infinitesimals is in terms of a process of particular numbers we think of as "very small." For instance, suppose we started with a number line marked off by the numbers 0 and 1, and suppose we began a process of repeatedly marking off numbers closer and closer to 0, by first marking off the number 1/2, then 1/4, then 1/8, and 1/16, and so forth, as in the diagram in Figure 12. Most of us implicitly assume that these numbers we are marking off, by halving the value each time, not only get closer and closer to 0, but that, if you select any other number, say *r*, to the right of 0, then this sequence of marked-off points will eventually move past *r*.

Another way of putting this is that if we are given any positive number *r*, then raising the power of 1/2 high enough, we will produce a number smaller than *r*. But this only holds if the number *r* is an *ordinary* positive number. The infinitesimal numbers are all smaller than all of these powers of 1/2. These powers move further and further to the left, but never go past the infinitesimals. It is not only the existence and location of these infinitesimals that are strange to a modern reader, however. The mathematicians of the 16^{th} and 17^{th} centuries seemed to do a kind of arithmetic with them which is even more ingenious and elusive—to us.

A casual look at the diagram in Figure 11 used to display Leibniz's approach to the Tangent Problem suggests that it is not really all that different from Newton's approach. This view is supported by the fact that both men produced the same formulas for the slope of the tangent line when we are given a formula for the underlying curve. It is worth repeating, then, that in Newton's diagram, we are looking at an ordinary distance, between P and Q, that is in the process of getting smaller and smaller. In Leibniz's diagram, we are looking at an infinitely small distance, which has been visually enlarged, in order to be seen. So, we are to imagine, in Leibniz's diagram, moving an infinitesimal distance to the right of the given point. Leibniz denoted this by dx, and called it a *differential*. In changing the value of x by this amount, there is a corresponding infinitesimal change dy in the value of y, as long as we are to remain on the curve (as in Figure 11). There is a relationship between the two kinds of change, dx and dy, which depends on the shape of the curve. The quotient, dy/dx, indicates the direction of the curve.

There is a shortcoming in the views both Leibniz and Newton had of the Tangent Problem in that, while both had ways of thinking about what the tangent line *is* in terms of the direction of a curve, each needed to develop a separate approach to the task of actually computing the direction of a curve given by a specific formula. While Newton used algebra and claims about terms that are negligible, Leibniz worked out rules and practices for dealing with combinations of his differentials, dx and dy. They both derived the same formulas and routines however—those that are taught to all beginning calculus students today.

When examined closely, the approaches of both Newton and Leibniz have a certain strange, almost fantastical, quality to them. In Newton's approach, we consider what happens to an algebraic expression when certain terms get closer and closer to 0, but never get there. This seems to involve making claims about the outcome of a process that never ends. In Leibniz's approach we start off by talking about numbers whose existence we can hardly grasp and then doing a kind of arithmetic with them. Those who know some calculus may argue that Newton's is the more believable, or natural. This is supported by the fact that it is Newton's approach that is codified in the rigorous foundations of calculus developed in the 19th century. One might argue, however, that most of us are simply more familiar with Newton's approach than Leibniz's. This view is supported by the fact that Leibniz's notion of infinitesimals has also been codified by an alternative approach to the foundations of calculus, one based on the work in the 1960's of the American logician Abraham Robinson. This work which is accepted by mathematicians as completely sound, gives an alternative basis for our number system. Thus we now have the "standard basis," in which the powers of 1/2, described above get closer and closer to 0 and there is nothing between 0 and all of them, and what is called the "non-standard basis" in which the infinitesimal numbers stand between 0 and all positive ordinary numbers, including all the powers of 1/2, and can be used to give interesting and new insights into many aspects of calculus. In fact, a calculus book has been published based on the non-standard foundations of the number system. Although it has not been widely adopted, it seems to work well for many students who have the patience to learn about an entirely new concept of number*.

***The Algorithm for taking the derivative*.**

One of the most striking aspects of the work of both Newton and Leibniz is that the solutions they gave to the Tangent Problem were completely general. Although one needs to know certain facts about the expression for a given curve in order to proceed, the overall approach is always the same, regardless of the particular expression. This was a great departure from what their predecessors

---

* See Davis & Hersh, *The Mathematical Experience*, (1981) for an excellent exposition of these matters. This account of a little-known point of view that rejects the conventional assumptions of a subject might sound like some of the current controversies around Darwin's theory of evolution in biology. However, there is no connection. Mathematics can be seen as the study of formal systems, not of the world as it *is*. All mathematicians accept the view that there are many alternative axiom systems for describing various aspects of the world, as was well established in the 19th century by the existence of several alternative versions of geometry, both Euclidean and non-Euclidean.

had done. Newton found his secant line and let the point Q get closer and closer to P, and Leibniz went his infinitely small distance away from the given point P. But each could do this very generally. Each followed the same pattern in all cases.

But even more striking was that patterns quickly developed in the relationship between the underlying expression for the curve and the expression for the slope of the tangent line. If we begin with a curve given by an expression, say x^4, then we obtain another expression for the slopes of tangent line to the curve, in this case $4x^3$. If we begin with a curve given by the expression x^6, then we obtain, as the expression for the slopes of tangent line, the expression $6x^5$. The same is true for curves defined by more complicated algebraic expressions, as well as expressions using other functions, such as the *sine* and *cosine* functions. In general, one begins with any expression that describes a curve and by the rules of calculus determines the derivative of this expression which then gives the slope of the tangent to the given curve at any point. When studying motion, which was Newton's main focus, one begins with an expression that defines the position of a moving object in time and one takes the derivative of this expression. The result is an expression for the speed of that object in time. Because of the power and generality of these rules, the attention of the subject shifts slowly from particular tangent lines, even particular functions, to this operation of going from a given expression to the expression for its derivative is called in calculus courses, "taking the derivative." For many beginning students, learning the rules for this operation dominates their early study of calculus.

Arriving at general rules for writing out the derivative of a given expression was an enormous achievement, and was quickly recognized as such at the time of Newton and Leibniz's work. This is the source of the word *calculus* itself. It comes from the Latin word for "stone," which takes on the meaning of our words to "reckon," or to "calculate," since, in earlier times people used stones as an aid in calculating. The rules for taking derivatives and other such processes were so compelling that the subject was, in effect, named for them.

The Area Problem

Both Newton and Leibniz solved the Area Problem, and, as in the case of the Tangent Problem, they did so in very different ways. The differences in the case of area between the views they had of the concept itself and the resulting approaches they took to the problem, were so different as to bring to the surface a deep issue within calculus as a body of ideas.

To most mathematicians in the period before Newton's work, the Tangent and Area Problems were viewed as unrelated mathematical problems to be approached with very different tools. One of the biggest surprises to those who learned of Newton's solution to the Area Problem is that it depended heavily on his solution to the Tangent Problem. Area and tangent seem such different concepts; how could knowledge of one provide a means for obtaining knowledge of the other? Newton's basic idea was that the Area Problem could be converted into a problem about rate of change, and that this kind of problem was made solvable by the concept of the derivative that he had developed as a result of his work on the Tangent Problem. This insight of Newton's is so creative and so central to calculus that it is well worth the effort of trying to understand it here. We will touch upon it several times in this book in several different forms; it is captured in what is known as the Fundamental Theorem of Calculus.

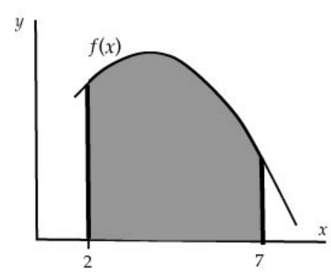

Fig 13–Area under curve between *x*=2 and *x*=7.

In order to understand Newton's insight, let's consider a more specific case in which we have a curve given by a specific formula, which we will denote by $f(x)$. We wish to find the area of the shaded region in Figure 13 which is under the curve, above the axis, and between the two vertical lines at x=2 and x=7. Newton's first move is to tell us that our question is too specific. Rather than asking for the area under the given curve between the two given vertical lines, at x=2 and x=7, we should ask a more general question: What is the area under the curve, above the axis, to the right of the line at x=2 and to the left of the line at x=b, where b is *any* possible value? Thus, instead of considering the *fixed* right hand edge of the region, we are to consider a *moving* right hand edge. Newton's idea is that we can solve the more general problem, and then use our solution to obtain an answer to the specific problem, by substituting the value 7 for the variable b.

Thus, rather than think about the static diagram in Figure 13, with 7 replaced by b, Newton tells us to think in terms of a more dynamic diagram, as in Figure 14, where the right-hand edge of the region moves back and forth. As it does so the size of the region and the numerical value of its area change. This is typical of Newton's approach. He has converted a question about geometrical figures into a question of motion. We move the right-hand edge of the shaded region and this produces a region that gets larger and smaller.

In order to discuss the resulting relationship between b and the area of the region that results from any particular choice of b, we denote the value of the area of the region by $A(b)$. The value of b increases and decreases, and the area $A(b)$ increases and decreases, although in ways that can be quite complicated. Our new goal is to understand this relationship, the one between the value of b and the value of $A(b)$. In more concrete terms, we would like to find some kind of expression in terms of the varying quantity b that tells us what the corresponding value is for $A(b)$.

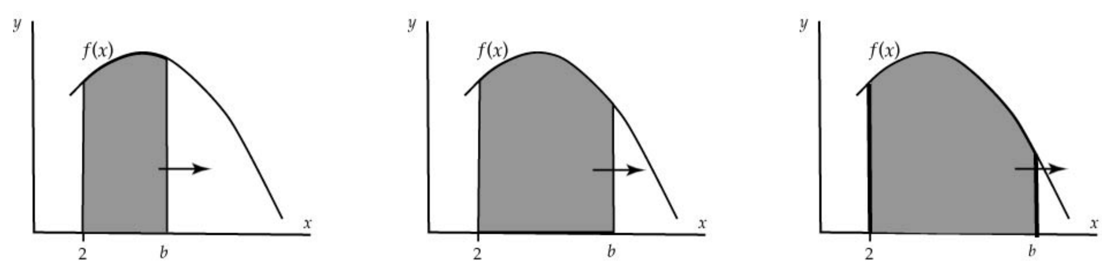

Fig 14–Areas under a curve as a variable quantity

Then Newton redirects our attention a second time. Although we wish to find an expression $A(b)$ for the relationship between the area and the right-hand endpoint b, he points out that we can obtain this information more indirectly, by working with its *rate of change* with respect to b. He shows that:

(*) The expression for the *rate of change* of $A(b)$ is the expression $f(x)$

that describes the underlying curve*.

This is worth repeating. What Newton tells us here is that even though we don't have direct information about this expression $A(b)$, between the area and the right-hand end-point $x=b$, we know about it indirectly: we know what its rate of change is. It is just the expression $f(x)$ for the curve we began with.

However, in his work toward the Tangent Problem, Newton showed that whenever you have one quantity varying with another, the rate of change of the first with respect to the second is given by this automatic calculating machine called the derivative. We can therefore simplify the statement in the box (∗), given above, by saying that :

(∗′) The *derivative* of the expression for $A(b)$ is the expression $f(x)$ that describes the underlying curve†.

This is very surprising. It says that even though we don't know what the expression for $A(b)$ is, we do know what its derivative is. So, in order to determine what $A(b)$ is, we need to figure out what expression has the given expression $f(x)$ as its derivative.

Thus, starting with the expression $f(x)$, the expression for the area $A(b)$ is the one we would begin with in order to arrive at $f(x)$ through the process of taking the derivative. Or, putting it another way, in order to determine the expression for $A(b)$, we must guess backwards: "What was the expression we began with when we arrived at $f(x)$ in taking the derivative?" This says that the Area Problem is solved by guessing backwards in the process used for solving the Tangent Problem (see Figure 15). In some way that must be made much clearer, "determining area is the inverse or reciprocal of determining the slope of the tangent."

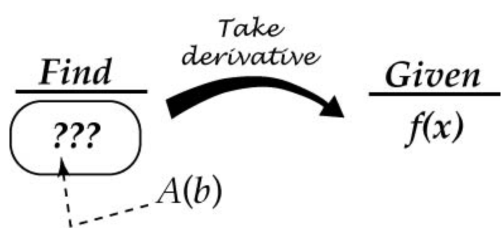

Fig 15–Guessing backwards

* There is a complication about the names of the variables x and b that we will not go into now. While it is not a serious problem, it does get in the way of trying to understand this aspect of calculus.

† Again, except for the substitution of the letter x for the letter b.