Calculus sin frontera

by Kathy Davis

Some things just pull so strong
Like the map of the sky is the map of your heart
- Ferron
Table of Contents

Preface: Leaving the Sixties p5

Chapter 0: Why Calculus? p6
    Notes for Chapter 0 p8

Chapter 1: Numbers
    Section 1: Background & History p10
        Notes for Section 1 p17
    Section 2: Building the Numbers p25
        Notes for Section 2 p30
    Section 3: The Problem of Fractions p42
        Notes for Section 3 p47
    Section 4: The Problem of Irrationals p48
        Notes for Section 4 p54
    Section 5: The Real Numbers p57
        Notes for Section 5 p62

Section 6 Part 1: Decimals: Accuracy p65
    Notes for Section 6 p73
    Section 6 Part 2: Decimals: Geometry and Decimals p70
        Notes for Section 7 Part 1 p79

Section 7 Part 1: Units & Standards p75
    Notes for Section 7 Part 1 p79
    Section 7 Part 2: Order of Magnitude p81
        Notes for Section 7 Part 2 p88
Section 8: Computer Numbers p91
   Notes for Section 8 p103
Section 9: Data: Introduction p105
   Notes for Section 8 p112
First Expedition: Vision p121
   Notes for First Expedition p134
List of Figures p145
Index p154
Preface: Leaving the Sixties

I left for college in 1967: the sixties, when everything was changing. My dorm was co-ed, the first in the country. Vietnam war protests shut down the University and students smoked weed in the lounge. But also: an entering physics major, I picked up Feynman’s *Lectures On Physics*. Feynman wrote,

> The special problem we tried to get at with these lectures was to maintain the interest of the very enthusiastic and rather smart students coming out of the high schools . . . . They have heard a lot about how interesting and exciting physics is – the theory of relativity, quantum mechanics, and other modern ideas. [Instead] they were made to study inclined planes, electrostatics, and so forth, and after two years it was quite stultifying.

Today, most students take an AP Calculus course in high school; they repeat the course in college, sometimes with the same book. They’ve heard about amazing discoveries in science, engineering, medicine and technology that change the world, but when they enter a calculus classroom, they’re magically transported back to 1967, as though the last fifty years of progress never happened.

This is not a calculus book; it’s a compendium of what gets left out of traditional books. Calculus touches on so much of the world: history, religion, philosophy, literature, psychology and especially the modern scientific and technical world.

Calculus is an amazing intellectual adventure. Let’s start.

Figure 1: Richard Feynman
Nobel laureate, physicist. Textbook author.
Introduction: Why Calculus?

On Jody Foster’s first day at Yale she wrote a friend: ‘My calculus book is three inches thick. I can’t survive three inches of calculus.’ My students are more direct: ‘I’m going to be a doctor. Why do I need calculus?’ Or, ‘I’m gonna be an engineer: we use computers for everything. Why do I need this class?’ Instead of a direct answer, I’ll tell a story about a doctor, and an engineer, and the ECG, a graph of the electrical currents in a heartbeat, used to evaluate the health of a heart (for a quick summary, see p8).

The story begins in the 1790’s, when the Lucia Galvani, the wife of the Italian doctor Luigi, noticed that the muscles of some frog legs she’d hung out for dinner twitched when a spark of electricity touched them. Luigi followed up, experimenting for many years, concluding that the nervous system operated through electricity. But in 1790 there were no instruments to measure electricity; scientists detected electrical impulses in the body using frog legs. In the 1830’s, the neurophysiologist Carlo Matteucci laid a frog leg on a beating heart, and the leg twitched in rhythm with the heartbeats, showing that heart muscles also generate an electric current when they contract.

Opening the chest of a patient and laying a frog leg across the beating heart isn’t an easy or safe diagnostic technique, but the electrical impulses from the heart would be too weak to measure otherwise. It took another seventy years before the Dutch physiologist Willem Einthoven developed a device to measure those weak currents, and then did the extra engineering to record them.

Figure 2 shows an early ECG machine. The electric currents from the heart cause a very thin wire to move; on the right, a light beam shines on the wire, and on the left, a motorized strip of photograph paper records the movements. And . . . the world’s first human ECG recording, Figure 3.
Technology advances: Einthoven’s moving strip of photograph paper was replaced by a pen moving up and down, writing on a moving strip of paper. The paper has lines and squares to make reading the recordings accurate. And the graph is sharper, showing more detail (Figure 4).

Now the diagnosis: the patient is suffering from an erratic heart beat, and an MD can tell by measuring the distance between peaks. In Figure 4 each little box is .04 sec wide, so an old-fashioned MD would count the squares between the peaks, and compute how many beats per minute there were. In this (short) graph, an erratic heartbeat shows up as a varying number of beats per minute (for more details, see the note on p9).

That’s what a 1950’s MD would do; even in 2014, in a visit to a hospital, my ECG was on a piece of paper a lot like Figure 4. But it won’t help in an emergency room: there’s no time to count squares on a nice piece of paper. Fortunately, we’re not in the 1950’s: we have computers.

Technology advances: now it’s the engineer’s turn. To keep records of a patient’s heart condition, there’s no need to store dozens of sheets of paper in a file cabinet. Instead of paper, write the electrical signals directly into a computer.

Once the data is in a computer – well, instead of drawing the graph on paper and counting squares, a microchip draws the graph on a LCD screen . . . and it counts the squares, it computes the beats-per-minute.

Today you can buy a portable ECG unit (Figure 5). An MD can get a quick sense of whether a heart is behaving normally, and then, if needed, run a full ECG (which takes 10 minutes or so, requires a nurse, square counting, and $$. Technology is great.)

But before the chip can count squares, it has to somehow ‘see’ the peaks. Imagine an engineer programming the chip, working with an MD and neither took calculus. Calculus would have told them peaks happen at critical points, where a derivative is zero. Calculus tell us how to find critical points, but how does a computer do it? The software to find critical points, using a microchip, was written by Jiapu Pan in Shanghai and Willis Tompkins in Wisconsin, using a derivative-based algorithm. Figure 6 shows an example of the technique; it relies on successive mathematical operations to strip off the irrelevant parts of the beat.

This engineer and MD did need calculus, and their work is one of the most cited in biomedical engineering.
Notes

Notes for Chapter 0


p6  An ECG records the electrical activity of the heart. Here’s a quick view of how.

The heart is a series of chambers to hold blood, muscles to pump blood, and veins and arteries to channel the blood. Muscle contractions are associated with electrical currents (we’ll go into more detail later in the book) and the current is measured by placing electrodes on the body; the record of the current flow is the ECG; see below.

On the left, the different chambers of the heart, and the major nerves that orchestrate the sequence of contractions of the chambers (for example, when one chamber contracts, the following chamber has to be open to receive the blood pumped; see *fibrillation*, below. Timing these events is crucial). On the right, a printout gives a visual record of how the different parts of the heart contract, the size of the contraction, and the timing between contractions; compare Figure 3. Well over a hundred years of study, matching symptoms with EKG’s allows an MD to diagnose heart problems using this information (you
can practice this yourself; see http://www.healio.com/cardiology/learn-the-heart/ecg-review/ecg-interpretation-tutorial/approach-to-ecg-interpretation).

The ECG serves as a kind of microscope into the heart, with the advantage that the doctor doesn’t have to do surgery to inspect the internal workings of the heart. It’s a simple, cheap way of monitoring basic health.

p7 The erratic heartbeat in Figure 4 is called fibrillation. As we said, the heart is a series of chambers and pumps, and a working heart chamber pumps only after it gets blood from the preceding chamber. In fibrillation, the pumping happens at random, and the blood can slosh around the heart, never getting to the body. Depending upon the chamber that has the fibrillation, this can lead to death within minutes.
Chapter 1: Numbers

Section 1: Background & History

In the Introduction, we looked at an erratic heartbeat and wound up with some numbers and a graph. No-one even blinks, because numbers are already everywhere: time, speed, area, temperature, pressure, heart rate, GPA . . .

What are numbers? How did they take over our lives? They’re not things we can touch, like spoons. But – if ‘number’ is just in our minds, how did it get there?

This kind of question has troubled philosophers from all the way back, to now. For a bit of background, see p17. Although philosophy has been important in the development of modern science and mathematics, for now we’ll use modern science. Starting with a study of animal intelligence (see p19).

A scientist makes holes in a log and randomly places worms in those holes. A robin watches, and when the scientist leaves, the robin immediately flies to the log to munch on yummy worms (Figure 7). What’s surprising is the robin starts with the holes containing the most worms.

The study suggests that these birds can “count” at least to the extent of perceiving ‘more’ and ‘less’ and understanding how to exploit the perception. Other studies show this ability is cross-species: robins, rats, newly hatched chicks, new-born babies. They won’t pass an algebra test, but something in these tiny brains ‘gets’ number (this view is actually too simple, philosophically and scientifically; see p19 for a detailed discussion).

So: rats, robins and babies all count, but they don’t worry about being a size 6, affording $70K for a BMW, or counting change at the local Starbucks. Humans use numbers differently than animals do: how is this?

Part of the answer relates to information processing in our brain: we’re not very good at it. Studies show we can focus on only a very few items; we can keep only a few facts in immediate memory, and we can switch tasks only very slowly. Compared to our computers, our hardware is slow, limited and obsolete (a good read on these limitations is David Eagleman’s Incognito: The Secret Lives of the Brain, Vintage 2012).
Number is the same: we can perceive the objects in Figure 8, but manipulating them in our mind (number) is hard. Some recent research suggests that the part of our brain which constructs number is limited and simply runs out of room (p21).

But: we can count. Counting is a form of external storage, whether with fingers or number words or marks on paper, sticks or bones (see Figure 9). Making marks on wood is in fact, cross-cultural, and persisted into the twentieth century in places where paper is rare.

In Figure 8, the mind retains the number eleven but loses the image of eleven green objects, their position, size and orientation. Retaining some but not all of the characteristics is a symbolic representation of our perception. Symbols add a new dimension: external counting makes numbers public, and even allows us to make stored records. The ability to take numbers (something in our brain) and translate them into physical, public form, is one difference between human and animal number use. Research suggest symbolic representations happen by connecting two different areas of the brain, possibly pre-linguistic; see p21.

Still: fingers, words, and attention: we run out quickly. What humans do then is almost universal: anthropologists have found many societies in which people count higher than ten by referring to other parts of their bodies such as thighs, arms, etc (see p23). Figure 10 shows a counting system used in markets in the Middle East. The joints of one hand mark one to twelve; the other hand counts how many groups of twelve we have. A written system needs markers for $1 \cdot 12$, for $2 \cdot 12 = 24$, up to $5 \cdot 12 = 60$.

Systems like this work well for personal transactions, but are no use for long distance trade – or for collecting taxes, or for ruling a country:

Since the rules for collecting and manipulating numbers are widely shared, they can easily be transported across oceans and continents and used to co-ordinate activities or settle disputes. Perhaps most crucially, reliance on numbers and quantitative manipulation minimizes the need for intimate knowledge and personal trust. Quantification is well suited for communication that goes beyond the boundaries of locality and community.

Theodore M. Porter, Trust In Numbers (Princeton University Press, 1996)

Denise Schmandt-Besserat established the origin of written numbers in Mesopotamia and from this, the development of all writing. Figure 11 depicts clay objects used for record keeping, from Susa, Iran circa 3300 BCE (see p23). Schmandt-Besserat writes:
The early city states still used tokens to control the levy of dues. When individuals could not pay, the tokens representing the amount of their debts were kept in a round clay envelope. In order to be able to verify the content of the envelope without breaking it, the tokens were impressed on the surface before enclosing them. A cone left a wedge-shaped mark and a disc a circular one. It was the invention of writing.

This innovation allows number symbols to be more than public: they can be institutionalized in economic transactions and in government. A chain of innovations made the system efficient, making it more useful in governing large cities or states. Figure 12 shows and example of one innovation: an early Mesopotamian (modern Iran/Iraq) symbolic system, partially based on tens (see p23). Each round dot represents a ten; each vertical line a one. It’s as though we were to wrote 31 as $10,10,1,1,1$. This notation has its advantages; we could just as easily write $1,10,1,1,1,10$.

It’s harder to write something like 99 in this system. We could do like the Romans: introduce new symbols like $V$ for five and $L$ for fifty, then use what is called a subtractive system, writing $IV$ to mean $5-1$. But this gets complicated: you start to need lots of extra symbols (as in the Roman numerals $I, V, X, L, C, M$ . . . ).

Figure 13, from the Chinese Qin and Han dynasties, shows an improvement over the Mesopotamian notation. Scribes used vertical and horizontal lines for the numbers one through nine, they then arranged these using small boxes. The position of each box from the right determines $9$ vs. $90$ vs. $900$. The change from Figure 12 to Figure 13 is the use of positional notation, which introduces standards as to which number goes where, making interpretation and use of the system more efficient. It also employs a further development, the use of empty boxes to denote zero.

This system, known as rod arithmetic, allowed the operations of addition, subtraction and multiplication to be performed by arranging numbers on a rod, aligning boxes, and performing the operations as we’d do today (Figure 14):

\[
\begin{array}{c}
54 \\
\hline
23 \\
\hline
31
\end{array}
\]
It’s a very fast method for doing arithmetic, and scholars in the Qin Dynasty needed it for massive engineering projects, like the Great Wall. So this is another factor in the spread of numbers into the many corners of life: the development of good notation and fast algorithms for performing computations.

Historically, though, much of the Chinese system remained in China. The writing at the top of Figure 16 shows how numbers were written in early Indian/Pakistani mathematics. The bottom shows the same numbers in Hindu-Arabic notation, the system used in modern Western science and technology.

By 662 CE, use of the ‘Indian’ notational system had spread west to ancient Iran/Iraq, leading to ‘Hindu-Arabic’ numerals. The ‘Arabic’ part comes from work at the Caliph’s court in Baghdad; a typical example is The Book of Addition and Subtraction according to the Indian Calculations, by the court mathematician Muhammad ibn Musa al-Khwarizmi (780 - ?850), Figure 17. The Caliphate ruled a vast commercial and political empire: "Arabs ...had been using [Hindu numbers] for centuries to calculate interest, convert currencies, and solve other problems of trade" (see The Crest of the Peacock: Non-European Roots of Mathematics by George Gheverghese Joseph, Princeton University Press; Third edition, 2010.)

Work in Baghdad took the Hindu system and added the decimal point and the zero. Now decimals could be used for fractions, and fast algorithms handled addition, subtraction, division and multiplication (the term ‘algorithm’ is a European mis-translation of al-Khwarizmi’s name). Now everything needed to carry out Islamic commerce was easily available.

The Hindu-Arabic system came to Europe from contact with Mediterranean Islamic culture, particularly in Al-Andalus, modern Spain. The European scholar Gerbert of Aurillac (946-1003) studied in northern Spain; he used Hindu-Arabic numbers and an abacus much like the Heng/Zhong system of arranging columns of numbers. The quotation at the head of this Chapter is from a monk who used these techniques to compute the dates of Christian holidays.

Gerbert’s decimal numbers were an important contribution to church affairs (and he eventually became Pope!). An Italian, Leonardo of Pisa, was the son of an Italian merchant who traded in the Islamic world. Leonardo worked in customs, and learning the Hindu-Arabic algorithms was essential for his work. His The Book of the Abacus introduced Arabic algorithms to European merchants and bankers.
Mathematics is more than commerce; its origins go very far back in the human story. Figure 18 is a sketch taken from an inscribed antler bone, with markings showing the phases of the moon: a lunar calendar (see p. 23). The bone may date to 32,000 BCE; however accurate the date, humans were clearly thinking about the cycles of the heavens for a very long time.

The constellations – Scorpio, Gemini, Capricorn, etc, come to us from Sumeria (modern southern Iraq), about 2000-3000 BCE. Similarly, the 360 degrees of a circle come from Babylonian base sixty counting systems.

These cultures also had sophisticated computational techniques. Very early, numbers and writing were used for simple account keeping: amount received, amount taxed, number of cattle sold, etc. Soon, though, these kinds of records became "hypothetical", recording not, say, the amount of barley in storage, but the amount of grains required to make a specified quantity of beer; see Figure 19 and Hans J. Nissen and Peter Damerow, Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East, University of Chicago Press, 1994.

The flood of documents […] contains in growing numbers not only such running accounts recording all the assets (primarily including arable land, raw materials and laborers) and liabilities (maintenance, labor costs, and so on) of the central administration, but also a standardized method of calculating the expected performance of laborers and of achieving comparable units of value of labor.


They could predict what should be happening, and note any difference between that and actual performance: in other words, they had a system of accounting for errors and shortages.

This is a different kind of use for numbers: not simply recording what is, but projecting future needs. Figure 20 shows a modern version: insurance estimates for repair of a house after fire damage, one of several hundred pages, but in 2015, generated by a computer.
Projecting into the future was not an invention of Mesopotamia; every sacrifice or prayer to the gods was an attempt to control an uncertain future. Professor Ulla Koch-Westenholz writes (see p23):

**stars and planets were the celestial manifestations of gods, but also seem to have been gods in their own right . . . Sometimes evil omens from a planet were seen as the expression of anger of the god whose celestial image the particular planet was (e.g. Jupiter = Marduk), so that particular god had to be appeased. In this way, messages could be sent directly from a god to the king, . . . auspicious Venus omens [are seen] as an expression of the love Ishtar holds for the king.**

To gain the king’s ear, astrologers needed to predict the appearance of constellations, or of eclipses, or any planetary sign: otherwise the best they could do was interpret events after they’d happened. The ability to predict arose from writing, which allowed them to compile lists of heavenly events. Figure 21 shows one such compilation; in this example, the tablet records twenty-one years of the rising and setting of the planet Venus (modern astronomers can work backwards to date the tablet: one goes back to 1581 BCE; see p24).

Ancient Sumerian and Babylonian astronomers wrote a vast number of these texts; these relate positions of planets and constellations to earthly events. In other words, these ancient cultures practiced divination, the art of reading the future from omens. Omens could be patterns in the innards of sacrificed animals, or they could be the patterns of the heavens: astrology.

Beginning in the Babylonian period (about 1800 BCE), astronomers used the recorded position of planets over many years to develop systems for predicting future events in the heavens. Asger Aaboe (see p24) called these Systems A and B. Each divided the sky into several regions, and had the planets moving with different speeds through the different parts of the sky. The ratios of speeds was always that of numbers like 3:2, small consecutive integers (see p24).

Figure 22 gives an idea of how the predictions worked. The dots on the graph show the changes of position of Jupiter when it first becomes visible on the horizon (it’s a good place to make the observation, because ‘horizon’ is an easy measuring guide, and you can even measure the angle from the observer to the planet). The straight blue lines represent the formula used to predict planetary appearances. Graphs were not developed for another two thousand years, but Babylonian astronomers did know lists, the same kinds used for accounting.
But there’s something strange going on: the straight lines aren’t the true planetary positions. So won’t the predictions be off? Why use lines?

i) It’s astrology. Does anyone expect it to be accurate?

ii) They used the techniques they had. What else could they do?

iii) Sure there were errors, but the errors were small.

iv) Although the predictions had errors, the numbers they really wanted turned out OK.

The very odd thing is – we are going to see these excuses continue to be used for at least another two thousand years (except for i).

We’ll follow the history, and the philosophy, of making these ‘wrong but good enough’ computations. We’ll see that this kind of issue is fundamental to the way mathematics is used in making predictions about the world.

As a side remark, Babylonian astronomy may even have given us the idea of ‘music of the spheres’. Greek theories of musical harmony also relied on ratios of small consecutive integers, just as above. They knew Babylonian tables (see p24) and may have seen the analogy.

In any case, we can see an historical answer to the question at the beginning of the chapter. By 2000 BCE, numbers were an essential part of trade, government, scientific astrology, and even religion.

Now, this explanation only works if mathematics is useful. A deeper question was raised by Nobel laureate Eugene Wigner, in The Unreasonable Effectiveness of Mathematics in the Natural Sciences. How is it that mathematics tells us about – well, everything? Does math have some special relationship to the way the universe works? For contemporary thinking about Wigner’s article see R. W. Hamming, at http://www.calvin.edu/ scofield/courses/m161/materials/readings/Hamming.pdf

For a bit more background on the philosophical issues raised by Wigner, see p18.
Notes for Chapter 1 Section 1: Introduction


The Christian holy days, such as Easter, were determined by the appearance of a new moon, hence Easter as a lunar holiday. This doesn’t translate easily into solar calendars, based on cycles around the sun. To an early Christian, the date of the Resurrection of Christ clearly had cosmic significance – hence, without numbers, all would be lost.

The religious importance of calendars goes very far back in human history. As one example, a Jewish sect called the Essenes or Yahad, living in the desert from about 200 BCE to 70 CE, developed an alternative to the traditional Jewish lunar calendar (Figure 24). They divided the year into 364 days, so that each of the four seasons, and all religious holidays, would fall on the same calendar day every year. Contrast this with a lunar calendar, where the holidays have to be determined by the appearance of the moon. Since this has to be determined by a human, the Yahad considered it less perfect, less reflective of the divine.

p10 Here’s a typical expression of what constitutes mathematics, in the view of Newton and his contemporaries:

Mathematics is the science of number, extension, and measure in abstraction from material things.


We’ve suggested that ideas like extension, measure and number derive from perception. But – if we draw a circle, we can recognize it looks circular, even though nothing we draw can ever be a perfect circle. So, how do we ‘abstract’? And, if real circles are all imperfect, how can we know facts about abstract circles, like $A = \pi r^2$?

These are old questions in philosophy: ‘How do we know?’ ‘What can we know?’ ‘What does it mean to understand?’ These questions are part of the study of epistemology, the philosophy of how we an know. Here, we’re asking: ‘If an abstraction isn’t a physical object, how can we perceive it or understand it?’

Plato conjectured that there was something like an idea of Circle (a form in his words). The form would be unchanging (eternal), unlike the various pictures of circles we could draw. Plato argued that the
only way we could come to know eternal ideas is for our our souls to be eternal; in the passage between one death and rebirth, the soul experiences Circle directly. In life, individual drawings can only ‘partake’ of the form Circle, and our living senses can only grasp a small part of the reality. That grasping can only be done by ignoring the world of the senses, which will always have errors. Only by the use of reasoning and logic can a person come to apprehend something of eternal truths. True understanding, then, requires intellectual discipline. Plato was likely following in a long tradition of Greek spiritual/intellectual disciplines; see the discussion of Pythagoreanism on p18.

Aristotle took a different view of abstraction. We can see his approach in the example of Greek theories of proportion. To give a modern version: we know that if \( \frac{a}{b} = \frac{c}{d} \), then it’s also true that \( \frac{a}{c} = \frac{b}{d} \). Aristotle remarked that this is first established for lengths, then for areas and times. Abstraction comes from from our experience: we take what we know and assume it is true more generally.

For Plato’s theory see Francis MacDonald Cornford’s Plato’s Cosmology: The Timaeus of Plato, Hackett Publishing Company, Inc, 1997. Cornford presents Plato’s dialogue, with section by section comments to guide the reader. Later Platonic dialogues discuss some of the difficulties with his theory, for example: what does it mean ‘to partake of Number’? See the Parmenides dialogue at, say, Wikipedia: https://en.wikipedia.org/wiki/Parmenides_(dialogue)


Wigner’s question (how is it that mathematics can tell us about the universe) has a long history. The earliest we can reasonably go back is to Pythagoras (about 500 BCE) and the Pythagorean school or religion or sect, depending. This has been called the “most controversial subject in all Greek philosophy” because everything we know comes from sources after his death (for a close examination of the evidence, see W. K. C. Guthrie, A History of Greek Philosophy: Volume 1, The Earlier Presocratics and the Pythagoreans, Cambridge University Press; Revised ed. edition February 28, 1979).

Pythagoras believed that people have souls which migrate across lifetimes; to him, everything in the world was sacred, ordered, and beautiful. The world may seem chaotic and often brutal, because we are tied down by our bodies. If we could transcend bodily experience, we could see the true order. We would no longer require a body, but exist as soul alone.
How is this relevant to Wigner’s question? Pythagoras believed that the study of the properties of number is the study of the universe.

"His most striking discovery, and the one which is said to have exercised the widest influence over this thought and to have been the foundation of his mathematical philosophy, was in the field of music. [...] The octave is produced by the ration 2:1, the fifth 3:2, and the fourth 4:3. [...] The discovery lay in the existence of an inherent order, a numerical organization within the nature of sound itself, and it appeared as a kind of revelation concerning the nature of the Universe."


This discovery is often referred to as the idea of the *kosmos*: a universe of beauty and order. Finding this order was linked with the goal of merging with the divine: "For Pythagoras then the purification and salvation of the soul depended [...] on philosophia; and this word, then as now, meant using the powers of reason and observation in order to gain understanding."

For both Plato and the Pythagoreans, the practice of reasoning and logic, and that of rejecting sense information, can only be accomplished by highly trained individuals. It seems analogous to spiritual disciplines for knowing God or achieving Nirvana. In fact, Gautama Buddha and Pythagoras were roughly contemporary. Karl Jaspers suggested these kinds of revelations were characteristic of what he called the Axial Age. See Wikipedia, also Karen Armstrong’s *The Great Transformation: The Beginning of Our Religious Traditions*, Anchor Press 2007.


Figure 7 and information on the robin experiment are from Simon Hunt, Jason Low and K. C. Burns, *Adaptive numerical competency in a food-hoarding songbird*, Proc. R. Soc. B (2008) 275, 2373.


We said that the explanation ‘robins understand number’ is scientifically and philosophically simplistic. What would a ‘not-simplistic’ discussion of the science and the philosophy be? The robin experiment isn’t decisive: the robin might not even notice worms. Maybe the robin has been conditioned: when humans put
things into logs, it’s edible. After all, thousands of experiments with dogs, rats, etc, lead to the idea of conditioned behavior. Give a rat a bit of food for pressing a large button, and an electric shock for pressing a small one and it’ll quickly learn to press that large button.

This critique is from a theory of psychology known as behaviorism. The idea is that all we can observe is what animals do, and perhaps there’s nothing more to it than associations like ‘food:humans’. You don’t need to say the robin can count, or even have any mental ideas like ‘number’. Maybe robins don’t ‘get’ number; maybe all that happens is conditioning. For that matter – maybe humans don’t get it either.

Noam Chomsky famously pointed out this is certainly is too simple a story for, say, our use of numbers. Presented with three chickens, we’re supposed to be conditioned to say the number: ‘three’. We’re far more likely to say: ‘Whoa! You cooking for a party, there?’ Or ‘Are chickens on sale this week?’ And neither of these are, in any reasonable sense, a conditioned reaction.


Again: is the robin is ‘really’ counting? Perhaps the robin doesn’t count at all; maybe it notices the human stays longer at one hole. Or returns to one hole. And that means there’s more food in that hole. To be serious, you’d have to consider alternative ideas about what the robin is doing, then run an entire collection of experiments, varying the way the worms were presented – for example, staying near a hole for a long time but not placing any worms there.

For chimps, and human babies (Figure 25), that large number of experiments has been done: for a non-technical summary, and references to the research literature, see Gwen Dewar’s article, What babies know about numbers at https://www.parentingscience.com/what-babies-know-about-numbers.html. These experiments rule out a large number of alternatives to the explanation ‘babies can count.’

To go more deeply than ruling out alternatives: If number is perceived by the brain, where exactly is that done in the brain? And, if we can ‘localize’ where number is perceived, what happens if that area is damaged by trauma or a stroke? The general idea is this:

Humans and many other animal species have evolved a capacity to represent approximate number. This ‘number sense’ is at the heart of the preverbal ability to perceive and discriminate large numerosities and relates to the intraparietal sulcus [Figure 26], a brain area which contains
neurons tuned to approximate number . . . and which is functionally active already at 3 months of age in humans. Children discriminate numerosity long before language acquisition and formal education, as early as at 3 hours after birth.

Manuela Piazza et. al., Cognition 116(2010) p33


So – does damage to this area of the brain affect the perception of number? Brian Butterworth describes an extreme case, in an individual code-named ‘Signora Gaddi’. Shown a piece of paper with three marks on it, most people could immediately say ‘three’. Signora Gaddi cannot do this, though she can count the items: ‘one, two, three’. When paying for groceries, she simply opens her purse and asks the clerk to take the right amount of money. Yet in all other areas of life, she appears to be an educated lady who leads a normal life. See Butterworth, What Counts: How Every Brain is Hardwired for Math, Free Press 1999. Apparently, the number deficit is separate from other mental functions – consistent with the theory that brains recognize number as such, and that it is not a consequence of other mental systems.

The research on number recognition is in Harvey, B. M., et. al, Topographic representation of numerosity in human parietal cortex, Science, 341, p. 1123- (2013). Figure 27 shows how the localization is arranged. The higher numbers seem to have less and less area of the brain devoted to their recognition, suggesting this is why recognition fails at high numbers.

We’ve just seen some of the argument for where the brain perceives/processes/recognizes small numbers. The next question is how that processing rises to consciousness, or, to (verbal or physical) symbolic numbers. That transition is likely to be something learned, rather than something present at birth:

[...] the acquisition by children of the first number words and their matching to numerosities appears to be a long and hesitant process which does not seem to lead on naturally from the preverbal skills that are already in place in infants.

Michael Fayol and Xavier Seron, About Numerical Representations, in Handbook of Mathematical Cognition, Jamie Campbell, ed., Routledge 2005
Now we have to ask, can this transition, or linking, be localized in the brain?

*One of these non-numerical circuits is in the left frontal lobe, which is associated with linguistic representations, in this case, representations of exact numerical values. The other is found bilaterally in the parietal lobes, a part of the brain associated with visuospatial functions in general, and, by Dehaene et al. in particular, with representations of approximate quantities in the form of a number line.*

These findings are compelling and provocative and they provide further support for the view that humans have at least two means of representing and processing quantity. One is the ability to make perceptually based judgments and comparisons. In this, degree of accuracy varies with set size. The other allows precise quantification through the use of symbols, concepts and rules.


And in more recent research,

*Converging behavioural, brain-imaging, and neurophysiological results suggest that knowledge of number is an evolved competence of the animal and human brain, with a cortical basis in bilateral intraparietal cortex. The number sense hypothesis postulates that this cerebral system is available early on during development, possibly during infancy, and guides the learning of numerals and arithmetic in childhood. Indeed, an association of number processing tasks with intraparietal areas has been demonstrated in 4- and 5-y-old children.*

Veronique Izard et. al, Distinct Cerebral Pathways for Object Identity and Number in Human Infants, Plos Biology, February 2008, Volume 6, Issue 2

The distinction between the internal processing of number, and the public naming of numbers, is at the core of the philosopher Ludwig Wittgenstein’s *private language* argument: is it possible for a human to have a language that no-one else has? Or does the idea of language necessarily involve rules, rules which only make sense if they are public? See [http://plato.stanford.edu/entries/private-language/](http://plato.stanford.edu/entries/private-language/) or the Wikipedia article on private language: [https://en.wikipedia.org/wiki/Private_language_argument](https://en.wikipedia.org/wiki/Private_language_argument)

Wittgenstein’s critiques are definitely ‘not-simple’ philosophy. Before we go into anything this deep, there’s a preliminary case. We mentioned the behaviorist view, that numbering might be nothing more
than a conditioned reaction: you see three objects, hear the word ‘three’ and that’s all that’s going on in the brain. But of course, it isn’t all; to begin with, there’s the whole issue of ‘seeing three objects’. What in our brain knows that those objects are – well, distinct objects, as opposed to blending into one large blur of landscape? How does our brain know ‘objects’? We’ll return to this in First Expedition: Vision. See in particular p122, p136 and p127-128.

The hypothesis of one ancestor for all languages is called the ProtoWorld hypothesis; see the Wikipedia article https://en.wikipedia.org/wiki/Proto-Human_language. Reconstructed works are preceded by an asterisk, for example, *tik. This is a reconstructed word, meaning a finger; related to indicate, to point, and also to digit, as well as to the number one.

The use of ‘hand’ to represent ‘five’ is from Stanislaw Dehaene, The Number Sense: How the Mind Creates Mathematics, Oxford University Press, 2011.

The links between basic number words and body parts may have a neurological explanation. Fayol and Seron, supra, remark: “It could indeed be that the linkage between preverbal number knowledge and language is in fact mediated by the relations children establish between number concepts and the use of their fingers and hands. As rightly noted by Butterworth [The Mathematical Brain, Macmillan, 1999] in all human cultures, children use their fingers to count before they are systematically taught arithmetic in school.”

Figure 12 is from T. Cuyler Young, Jr., of the Royal Ontario Museum, Toronto, Canada, who excavated Godin Tepe in the 1960’s. It is provided from Denise Schmandt-Besserat of the University of Texas at Austin, in How Writing Came About, University of Texas Press; Abridged edition 1997.

The Schmandt-Besserat quote is in From Accounting To Writing, https://sites.utexas.edu/dsb/tokens/from-accounting-to-writing. She showed written numbers evolved from ancient accounting circa 8000 BCE; government uses came later.

This interpretation is due to Alexander Marshack; see The Roots of Civilization: The Cognitive Beginnings of Man’s First Art, Symbol and Notation, Moyer Bell Ltd, December 1991.


Chapter 1: Numbers

Section 2: Building The Numbers

We’ve argued that numbers gained their place in human life because of their public nature. Professional mathematicians form a separate community who use, and think about the uses of, number. If number were a branch of government, or a religion, they’d be the lawyers and priests. Like lawyers, they have their own understandings, which are not necessarily the everyday ways most people use numbers. One of the issues we’ll have to deal with is whether the professionals add any value to everyday understandings.

We’ll start with the numbers \{1, 2, \ldots\}, called the counting, or natural numbers. These came so early in human history that, like ‘1 + 1 = 2’ they seem as unshakeable as the foundations of the earth. For mathematicians, it’s more complicated: it can take a semester to prove ‘1 + 1 = 2’. First, you have to define ‘1’, ‘+’ and ‘=’.

Why would mathematicians bother? It’s part of the attempt to discover true knowledge about the world, and is related to ideas of Pythagoras and Plato (p18) – the use of a strict discipline of reason alone, to avoid being confused by sense experiences (such as pictures, rumors, etc). What is this discipline?

Mathematicians use a system based on two foundations: axioms, and deductive reasoning. Historically,

"The word axiom comes from the Greek word αξιωµα (axioma), meaning "considered worthy".  *ag-ti at American Heritage Dictionary of Indo-European Roots, https://www.ahdictionary.com/word/indoeurop.html#IR000300

In turn, this is from αξιος (axios), meaning "being in balance". Axios itself comes from Proto-Indo-European *ag-ty-o- "weighty"; the images here are an object weighed on a balance; checked for accuracy, thus ‘worthy’ of trust. These words all imply public nature of ‘axiom’

In Greek philosophy, an axiom was a claim which could be seen to be true without further checking. This raises issues: be seen by whom? In general, Aristotle suggested we base reasoning about the world on truths that could be observed by most men, most of the time (the sun rises very morning – unless there’s an eclipse). However, for more subtle phenomena, the axioms should be given by the most pure, or most enlightened, or most knowledgeable individuals: in short, an elite who have had special training; you can see this all the way from Pythagoras to Plato to Aristotle; see p30."
These same kinds of issues also arose in the development of science; we'll discuss that later.

The ‘deductive reasoning’ part (from *duk-a-, to draw out or lead) was codified by Aristotle in the text *Prior Analytics*, where he listed the forms of correct reasoning (for ‘forms’ see p31). One form is the *syllogism*; the classic example is:

All men are mortal
Socrates is a man
therefore Socrates is mortal

(In a syllogism, ‘therefore’ means ‘it is correct to reason’)

One of the first books to follow the axiom/deductive method was Euclid’s *Elements*, a book of geometry and number theory. Euclid set definitions, axioms, postulates, common notions, etc. He then gave deductive proofs for the mathematical results.

Examples from Euclid:
Definition: "A straight line is a line which lies evenly with the points on itself."
Postulate: "[One can] draw a straight line from any point to any point."
Common Notion: "If equals be added to equals, the wholes are equal."
And the historically important 'parallel postulate':
"That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles."
Proposition: "A straight line falling on parallel straight lines makes the alternate angles equal to one another, the exterior angle equal to the interior and opposite angle, and the interior angles on the same side equal to two right angles."


Why all this ... machinery and technical terminology, and where did it all come from? It’s a rather long story; see p30.

Let’s see how this approach is used for numbers. First, a bit of notation. The natural numbers are denoted by $N$, and intuitively given as a list:

$N = \{1, 2, 3, \ldots\}$

If we want to include the negatives and zero, we have the *integers*,

$Z = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$
The \( \mathbb{Z} \) here is from the German word \( \text{Zählen} \), meaning, d’oh, ‘numbers’. None of this, though, tells us what numbers are, and none of this tells us what’s actually on the infinite list \( \{1, 2, 3, \ldots \} \).

Gottlob Frege gave a definition based on the meaning of ‘counting’.
Frege (a German mathematician working in the mid nineteenth century) thought back to a child pointing at toys one at a time and saying the numbers ‘one, two, three’: the process of counting is that of matching one collection with another, like toys to fingers, or number words. If we match the two collections \( \{ \Sigma, \Delta, \Gamma \} \) and \( \{ \infty, +, \int \} \) we’d say they have the same number of items. Matching up different collections partitions the world into chunks which have the same number, and we could \textit{define} three to be the collection of all collections that have the same number as \( \{ \Sigma, \Delta, \Gamma \} \); see p.40.

A different approach follows an idea older than Aristotle, common in early Greek mathematics. The number ‘1’ was considered the father of all numbers, because all numbers could be generated from him. And we wouldn’t consider ‘1’ as a number at all, as a father is not one of his own children.

A thousand years after Aristotle, the mathematician al-Khwarizmi put it this way:

Because one is the root of all numbers, number is nothing but a collection of ones.

Intuitively, \( \mathbb{N} = \{1, 1+1, 1+1+1, \ldots \} \). The Italian mathematician Giuseppe Peano used this approach to define the natural numbers through four axioms \( P_1, \ldots, P_4 \).

\( P_1 \): If \( n \) is a natural number, then \( n \) has a successor. Peano denoted it as \( n' \); we’ll write it and think of it as \( n+1 \).

\( P_2 \): If \( n+1 = m+1 \), then \( n = m \).

\( P_3 \): There’s a unique natural number \( n \) which is not the successor of any other natural number. We denote this as the number ‘1’.

Call \( P_1, \ldots, P_3 \) the ‘plus one’ property. These axioms aren’t enough. A collection like \( \{ \ldots, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, 2, \ldots \} \) has the plus one property, it’s easy to make more, but they all have extra stuff. The way to eliminate the extras is by defining the natural numbers as the smallest collection with the ‘plus one’ property. Then the trick to define ‘smallest’ is to require that there are no smaller collections with the ‘plus one’ property:

\( P_4 \): Assume \( S \) is a collection of natural numbers that contains 1. Assume it also has the ‘plus one’ property: whenever \( n \) is in \( S \), \( n+1 \) is also in \( S \). Then \( S \) is actually all of \( \mathbb{N} \).
Axiom P₄ has another name, the *Axiom of Induction*. But it still isn’t enough: we don’t know if there are any actual collections where all four axioms are true.

Back to al-Khwarizmi: \( \mathbb{N} = \{1, 1 + 1, 1 + 1 + 1, \ldots\} \). Here, it’s the ‘…’ that trips us up. Of course, ‘…’ means ‘continue doing the same thing forever,’ but we’d have to understand ‘forever’ – that is, we’d need a grasp on infinity. That’s the fifth axiom:

P₅: An infinite set exists.

This is called the Axiom of Infinity; see p₄₀ for more on ‘infinite’.

Now we know what the counting numbers are, what do we really know? From Frege’s construction, we’d have the number 2 as the super-collection of all “two” sets: \( \{1, 3\}, \{\pi, e\}, \{\Gamma, \sigma\} \ldots \). Yeah, that really helps. Or, we have it as \( \{1, 1 + 1, 1 + 1 + 1, \ldots\} \). Again, not helping.

Here’s the issue: if I cash a check for $432, the cashier could give me 432 one dollar bills, which is a major pain to use, or even to count. I’m more likely to get four hundreds, three tens and two singles. This, I can easily work with. The idea is very old; see Figure 29. For us, instead of repeating three tens, we use multiplication to write repeated addition as a product: \( 30 = 10 + 10 + 10 = 3 \cdot 10 \). Thirty three is then \( 33 = 3 \cdot 10 + 3 \). We also use exponents to make it easier to write multiplications: ten tens is \( 10 \cdot 10 = 10^2 \). This series of notations is the decimal system; written as a decimal, our cashed check becomes \( 4 \cdot 100 + 3 \cdot 10 + 2 = 4 \cdot 10^2 + 3 \cdot 10^1 + 2 \cdot 10^0 \).

With the Peano axioms, we can prove that every natural number is a decimal, but it’s tricky. The Axiom of Induction suggests a way: show that the decimals obey the plus one axioms. First, we’d show the number ‘1’ is a decimal (easy: \( 1 = 1 \cdot 10^0 \)). We’d also show that if you can write any number \( n \) as a decimal, you can also write \( n + 1 \) as a decimal. That’s easy for numbers like 43 and 44: if we know 43 is a decimal, then \( 43 = 4 \cdot 10 + 3 \), and so \( 44 = 43 + 1 = (4 \cdot 10 + 3) + 1 = 4 \cdot 10 + 4 \) is also a decimal.

This doesn’t work so nicely when we have to carry. It’s hard to write a general carrying formula, because numbers like 99, 999, 3989 etc, are all different; there’s no general pattern. So it’s all a little harder: the right technique here is called *strong induction*; see p₄₁.

Decimals group numbers by size: in daily life a burger is less than ten dollars, local concerts less than a hundred, books for a semester less than a thousand, and so on. These gradations are called *cognitive reference points*. A car repair bill of $326 becomes, in my mind, ‘around three hundred’ – see p₄₁.
Translating that into algebra, grouping by size looks like this: I take $n = 326$, and note $100 \leq n$, but also that $n < 1000$. We’d write that as $10^2 \leq n < 10^3$, and in general:

The Archimedean Principle: If $n$ is a natural number, then there is always a power of ten, $10^m$, with $m \geq 0$ and $10^m \leq n < 10^{m+1}$.

The powers of ten measure all the natural numbers.
Notes for Chapter 1 Section 2: Building The Numbers

p26  We’ll start with a brief discussion of the origins of the axiomatic/deductive way of mathematics, though much is likely "lost in the mists of time." The Danish historian Jens Herup (In Measure, Number, and Weight: Studies in Mathematics and Culture (SUNY series in Science, Technology, and Society) SUNY Press 1994) describes a ‘pre scientific’ mathematics: problems and ideas shared across cultures, spread by trade. Figure 30 shows an example. We know ancient arithmetic could be carried out by arranging stones, then made into patterns such as squares and triangles. Building the patterns would suggest relationships we’d call algebraic, but were treated by Euclid as geometric.

Once you see these patterns, it’s natural to develop ideas such as even/odd numbers, Pythagorean triples \((a, b, c \text{ with } a^2 + b^2 = c^2)\). Even/odd and appear in the proof that \(\sqrt{2}\) is irrational (see p55).

But, why? Why does adding an odd number to a square number give another square number? Aristotle believed that deductive proofs were explanations, and that an explanation is only good if it is of a certain kind. To explain a complicated idea, you’d want to refer back to simpler ideas that are already understood – which is the point of having axioms, common notions, etc.

Where do axioms and common notions come from? This is the problem of knowledge – epistemology; see the discussion on p17. As a pupil of Plato, Aristotle would have been familiar with the theory of eternal Forms; see p18. His own view built on and reacted against Plato:

"Aristotle’s account of knowledge of the indemonstrable first premises of sciences is found in Posterior Analytics II.19 […] what he says there is that it is another cognitive state, nous (translated variously as "insight", "intuition", "intelligence"), which knows them. […] What he is presenting, then, is not a method of discovery but a process of becoming wise. He says that in order for knowledge of immediate premises to be possible, we must have a kind of knowledge of them without having learned it."


As we suggested on p26, knowing the axioms requires some special kind of – wisdom. Compare the research on number recognition, p21: partly built in, partly learned.

Figure 30: Picture Proofs
You can get from one square number to the next by adjoining two sides, and one extra number, and the picture is the proof.
To a modern, the explanation of the picture is algebraic:
\((n + 1)^2 = n^2 + 2n + 1\). 
Euclid was trained in the school of Plato, in addition to that work, he must have known Aristotle’s ideas about proof: he certainly used those. In addition, Euclid had many predecessors; several wrote texts also called Elements. A great deal of modern research has shown how Euclid incorporated and built on and to some extent systematized earlier work. For the earlier work, see Morris Kline, *Mathematical Thought from Ancient to Modern Times Volume 1* Oxford University Press 1990, Volume One. For the influence on Euclid, see for example David Fowler’s *The Mathematics of Plato’s Academy: A New Reconstruction*, Clarendon Press 1987.

Hørup remarked that once deductive mathematics began, some sort of organizational system would have been essential, to avoid proofs that depended on each other (‘circular reasoning’). He gives an example, the proof that the sum of the interior angles of a triangle is 180° degrees.

Figures 31-32-33 provide a ‘picture-proof’, which seems clear enough. But pictures can have problems:

i) Figure 32 involves drawing a parallel line; how do we know a parallel line exists? How do we know there’s only one parallel?

ii) Figure 33 uses the result that complementary angles are equal. Aristotle, in the Prior Analytics, remarks “those persons who […] are drawing parallel lines […] do not realize they are making assumptions which cannot be proved unless parallel lines exist.”

The point of preliminaries like axioms, common notions, etc) is to eliminate all these concerns.

The next few pages are a discussion of the influence of formal reasoning on Western culture; to skip this, head to p40.

The syllogism is called a form of reasoning. Aristotle used the noun *eidos* (*είδος*), from the Proto Indo-European verb *weyd*, to see. Eidos was then originally a thing seen, thus, the shape of the thing, shape.

"[…] Aristotle focuses primarily on living things, characterized by hierarchical orderings of complex arrangements of material. So the horse […] is a complex arrangement of tissues and organs with emergent properties that make it a horse rather than a cow, for example."


For Aristotle, then, form implied a structural arrangement. In the classic "All men are mortal. Socrates is a man. Therefore Socrates is mortal’, we can see the structure If we write the syllogism in symbols (as Aristotle did): "All A are B. x is A. Therefore x is B’.
Forms of reasoning have had a deep effect on European intellectuals. The idea of formal reasoning, and the deductive structure and certainty of Euclid, has appealed to many philosophers and scientists, among them Galileo, Descartes, Hooke, Newton, Leibnitz, and Hobbes (see Figure 34). It even influenced Thomas Jefferson’s Declaration of Independence.

Axioms: We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed.

Theorem: That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government.

Aristotle wrote down many (but not all) of the forms of reasoning, hence, formal logic. When you couple that with Plato, who set standards for education, you get an approach to reasoning that influenced the next two thousand years of Western thought.

For example, in the European middle ages, the trivium (the ‘place where three roads meet’), grammar, rhetoric, and logic, would give the student the equivalent of a modern B.A.. The quadrivium (‘place where four roads meet’), arithmetic, geometry, music and astronomy, would be comparable to graduate study. Geometry, of course, meant Euclid’s geometry. This educational system guaranteed that all educated European men knew Euclid’s work.

The mathematician Gottfried Leibnitz was also influenced by Euclid; Leibnitz believed that “The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate [calculemus], without further ado, to see who is right.” To accomplish this translation of all thought into symbols that could be manipulated formally, Leibnitz proposed that ideas are formed from a small number of simple ideas; when we find those, we can express more complicated ideas as algebraic-like combinations of the simpler ones.

The British mathematician George Boole took a significant step towards making mathematics into a Leibnitz-like language. Logic uses abstract symbols; Boole showed it can be translated into algebra, Boolean Algebra.
Here’s an example, from the logician Charles Dodgson. (Lewis Carroll): The Judaic Old Testament, Leviticus 11, proscribes certain dietary rules; for example:

The LORD said to Moses and Aaron, "Say to the Israelites: ‘Of all the animals that live on land, these are the ones you may eat: You may eat any animal that has a split hoof completely divided and that chews the cud. There are some that only chew the cud or only have a split hoof, but you must not eat them. The camel, though it chews the cud, does not have a split hoof; it is ceremonially unclean for you.

Let’s change this into Boolean algebra. If \( c \) represents "chews the cud" then \( 1 - c \) will represent "does not chew the cud". Similarly, if \( h \) represents "has a cloven hoof", then then \( 1 - h \) will represent "does not have a split hoof." An animal that doesn’t have a cloven hoof and doesn’t chew the cud would be represented as \((1 - c)(1 - h)\). Then the unclean animals would be

\[(1 - c)(1 - h) + (1 - c) + (1 - h)\]

Boole even allows us to use algebra to express more complex forms of reasoning like "Socrates or Homer is human. All human are mortal. Therefore Socrates or Homer is mortal." (for the history of logic, see The Development of Logic, Mary and William Kneale, Oxford University Press 1962).

Boole titled his book The Laws of Thought, and for many intellectuals, formal logic seemed to offer a standard for all thought. It sounds like a happy ending, a resolution of all the problems of reasoning.

Does it work?

Even if we limit it to reasoning in mathematics, we can still ask: Was this the right way to think? Any approach to all of mathematics should do three things:

i) All known math can be fit into this approach.

ii) Any statement in this approach can be either proved or disproved (this is called completeness).

iii) This approach shouldn’t allow us to prove things that are false (this is called consistency).

As it happens, it is not possible to prove all these things, though mathematicians put in a very good try. The German mathematician David Hilbert working at the end of the 1800’s began by showing that Euclid’s geometry could be reduced to algebra/arithmetic, raising the issue of consistency/completeness for arithmetic. This was
partly motivated by work on Euclid’s parallel postulate; two mathematicians had developed geometries that denied the postulate, raising the question whether their geometries were consistent. Hilbert’s work showed that all were equally consistent, by reducing the question to the consistency of arithmetic.

The British mathematicians-philosophers Bertrand Russell and Alfred North Whitehead began an ambitious project to show that all mathematics could be reduced to logic alone. Their project failed; Russell was able to show that their approach led to a statement much like ‘This statement is false’, which if true is false, and if false is true. Thus it’s either unprovable, so that the system is incomplete; or it is provable, making the system inconsistent. Russel and Whitehead were forced to add extra, non-logical axioms.

The Austro-Hungarian mathematician Kurt Gödel, working in the early 1900’s, was able to show that Boolean logic was both complete and consistent; this left open Hilbert’s reduction of geometry to arithmetic. In 1931, Gödel showed that arithmetic was either incomplete or inconsistent. Gödel’s proof realized some of Leibnitz’s dream, translating ideas into a system one could compute with. Gödel could assign numbers to mathematical statements (which is where arithmetic comes in). Then the construction was, like Russel’s, something like: "#2106: Statement #2106 is false."


The British mathematician Alan Turing wrote of a simple (hypothetical) machine that could perform the operations of Boolean logic and arithmetic, and therefore could do all formal (or syntactic) reasoning. There are called Turing machines. In his 1936 paper, On Computable Numbers, with an Application to the Entscheidungsproblem, he showed that Gödel’s results hold for these machines as well. Turing’s applied equally to any system of thought complex enough to do arithmetic.
Boole’s ideas have applications beyond mathematics. The American researcher Claude Shannon showed in 1937 that existing machines of that era could be wired to perform all the operations of Boolean algebra and arithmetic. The work of Turing and Shannon demonstrated that anything that can be computed can be computed by electronic circuits (Figure 35). It’s not much of an intellectual stretch, then, to argue that brains are actually performing some kind of neural computation, and therefore we can understand thought through these computer analogies. For some of the history, see Howard Gardner, *The Mind’s New Science: A History of the Cognitive Revolution*, Basic Books, 1987. To quote the philosopher Jerry Fodor,

"The key idea [...] is that cognitive processes are computational; and the notion of computation thus [borrows] heavily from the foundational work of Alan Turing. A computation, according to this understanding, is a formal operation on syntactically structured representations. Accordingly, a mental process, qua computation, is a formal operation on syntactically structured mental representations."


What does this mean? A sense impression or a thought would need to be some kind of numerical-logical complex or structure (vagueness warning!) that computation could work on. Moreover, that complex would have to be purely formal (syntactic): the computation would have to work without looking at the meaning of the sense impression. These hypothetical complexes are called *representations of* the sense impression or thought to which Fodor refers.

Not all scientists believe all of this; for a discussion of how far these ideas may take us, and may not be able to take us, see Jerry Fodor, *Reply to Steven Pinker ‘So How Does The Mind Work?’*, Mind & Language February 1, 2005.

There is some evidence against the idea that logic is at the base of human thought. In the 1930’s, the Russian psychologist Aleksandr Romanovich Luria studied cognitive development of indigenous Central Asian peoples. If definitions and logical deduction from them were basic to human thought, then they should appear in Central Asia too. But when asked to give a definition of a tree, his subjects might typically respond,

"Why should I? Everyone knows what a tree is; they don’t need me telling them. There are trees everywhere; you won’t find a place that doesn’t have trees. What’s the point of my explaining?"
Luria then examined logical reasoning. He posed the question, "In the far north, where there is snow, all bears are white. Novaya Zemlyya is in the far north and there is snow there. What color are the bears there?" The several responses, below, were not encouraging:

"There are different sorts of bears. [...] I’ve seen a black bear, I’ve never seen any others. [...] we always speak only of what we see; we don’t talk about what we haven’t seen. [...] If a man was sixty or eighty and had seen a white bear and told about it, he could be believed, but I’ve never seen one and hence I can’t say. [...] Those who saw can tell and those who didn’t can’t say anything!"


Luria’s informants lived in an entirely oral culture: no reading, no writing. The culture has its own rules for acquiring reliable knowledge, which are an odd echo of Aristotle’s "what most men see, most of the time." There’s also an echo of ‘only the wisest’ when the informant says "If a man was sixty or eighty [...] he could be believed,” Aristotle himself lived at the end of an oral culture; perhaps his rules reflect that.

There’s another oddity: these indigenous people refused to even accept the premises of Luria’s syllogisms about white bears. It’s almost as though Luria had been playing some game, a game whose rules these people didn’t know.

This idea arises in the work of the Austrian philosopher Ludwig Wittgenstein (Figure 36). To use formal algebra for reasoning, you need algebra, the ability to manipulate symbols like \((1 - c)(1 - h)\). But what is algebra? We have to follow certain rules, and use those rules consistently. Who decides that consistency?

One of the issues most associated with the later Wittgenstein is that of rule-following. Wittgenstein [gives] an example: " . . . we get [a] pupil to continue a series (say +2) beyond 1000 – and he writes 1000, 1004, 1008, 1012”. What do we do, and what does it mean, when the student, upon being corrected, answers ‘But I did go on in the same way’? Wittgenstein proceeds [to ask] "How do we learn rules? How do we follow them? [...] Are they in the mind, along with a mental representation of the rule? Do we appeal to intuition in their application? Are they socially and publicly taught and enforced?"

Wittgenstein asks if following rules in logic or algebra is nothing more than training, no different from learning the correct pronunciation of ‘Dumbledore.’ See Wittgenstein’s early work, Remarks on the Foundations of Mathematics, MIT Press 1983, or his later Philosophical Investigations, Pearson, 1973.

If logic really is social training, then it seems unlikely that it’s behind the way actual brains think. On one hand, we believe that most animals have some kinds of built-in logic. If a mouse saw a snake, it would try to avoid the snake. We believe the process is something like this: “That thing over there is a snake. All snakes are dangerous. I’d better climb/run/crawl/fly/Hide.” That’s the kind of logical inference we believe goes on in brains.

On the other hand, perhaps seeing a snake causes a fear response, which releases the hormone cortisol, which activates all kinds of other chemical reactions, resulting in our trying to avoid the snake. All this business of ‘snakes are dangerous’ might be some verbal story we make up to explain a reaction we actually have little control over.

A paramecium (Figure 37) is a clearer example. The paramecium, swimming happily along, might bump into something. Roughly, the bump stretches the cell membrane, causing certain pores (ion channels) to open, allowing K⁺ (potassium) ions to leave the cell. The change in K⁺ concentration allows an electrical signal to spread through the cell, ending in chemical changes causing the cilia to beat. A bump in the front causes the paramecium to move backwards; and a bump in the back causes a movement forwards.

There’s no logic here, no “Uh oh, I bumped into something. That could be a predator. Predators could eat me; I better get moving.” No logic, just millions of years of changes in DNA.

Humans, however, are verbal; thinking could use formal logic. Here’s one kind of experiment: if our brain is built on logic, then two sentences that are logically equivalent should be processed equally well. Human experiments show this doesn’t always happen. Here’s one example: look at the following three sentences

i) ”If that jacket is over a hundred, I won’t buy it.”
ii) “Hey, it’s only $90. So you’re buying it then?”
iii) ”Yo! I see you bought that jacket. So you FOUND it for less than a hundred”

Many people have a hard time seeing which of the three are the same. See Women, Fire and Dangerous Things: What Categories Reveal About the Mind, University Of Chicago Press 1990.
The same kinds of concerns arise in the study of ethics. One view of ethics is that we have moral rules, like ‘don’t harm other humans’, and that these rules act as axioms for determining moral behavior. Contemporary research suggests that we seem to have two kinds of ethical systems. One system is based on ideas about being treated fairly, respecting elders and people in authority, keeping ourselves clean (especially cleansing before worship), and adhering to the social norms of our group. We make decisions about being treated fairly very quickly: a fast reaction is to get angry when someone cuts in front of us in line. Only after we make that judgement do we rationalize it. For this research, see Daniel Kahneman *Thinking, Fast and Slow*, Farrar, Strauss and Giroux, 2013, as well as Jonathan Haidt, *The Happiness Hypothesis*, Basic Books 2006.

Does it really matter whether logic, or Boolean logic, is intrinsic to our brain, or is cultural and has to be learned? For some philosophers or brain researchers, it seems to matter very much. For others, it could be just a useful tool, like a calculator or microscope. We’re going to look at a classic experiment in molecular biology, the discovery that DNA is the molecule where genetic information is located. We’ll find that logic played a very important role in THAT discovery. For a non-technical introduction, see Horace Freeland Judson, *The Eighth Day of Creation*, Simon and Schuster 1979.

Mendel showed heritable traits (in peas) were governed by discrete units called *genes*. What are genes, and if they’re objects, where are they located? It wasn’t long before genes were located on the chromosomes of the nucleoid or nucleus, but chromosomes are extremely complex structures – so much so that even at the time this book was written, the structure is still being explored; Figure 38.

A key experiment leading to the answer was a virus, *Streptococcus pneumoniae*. It comes in two forms easily distinguished in a culture, R(rough) and S(smooth). the S form is virulent: it kills lab animals; the R form does not. In 1928, Frederick Griffith of the Ministry of Health in London performed an essential experiment: he ‘killed’ S forms so that they did not affect lab animals, and then mixed the dead S with live R. The mix was injected into mice, and when autopsied later, their blood contained living S. The change in S was permanent and inherited. Something in the dead S transformed the R into virulent S.

The something that changed R into S was called ‘the transforming principle’. It was clearly related to hereditary. But, again, what was it?
In the 1940’s biochemistry was advanced enough to identify the main components of a chromosome: several proteins, some sugars, and a chemically simple molecule, deoxyribonucleic acid: DNA.

Proteins were known to be very complex molecules; it was thought that only they would be able express all the variability it took to build something as complex as an animal. How to show the transforming factor is protein?

Enter Oswald Avery of the Rockefeller Institute. With the molecular technology available in the 1940’s, Avery was able to strip off either the proteins or the DNA from the transforming factor. He did a series of experiments, of the form, ‘If the transforming principle is protein, then eliminating proteins will eliminate transformation. Therefore, if transformation still occurs, the transforming principle is not protein’. Simple logic.

The very difficult lab work was in eliminating all proteins. This required many many careful steps of purification and testing the purified substances.

Many have said that the conceptual and laboratory work was of Nobel prize quality. Avery never received it, though he was nominated many times.

Eliminating protein did not eliminate transformation. Proteins do not carry heredity; DNA carries heredity.

Boolean logic has other applications. Since it’s easy to implement on computers, and mathematics uses logic, computers could be programmed to ‘do mathematics.’ Or, mathematicians could use computers to check proofs. There’s an outstanding application of this idea: the solution of the four-color problem (Figure 40).

To draw a map, you color it so that countries with common borders have different colors. The theorem shows that you need at most four different colors. It was proved in 1976 by Kenneth Appel and Wolfgang Haken of the University of Illinois; for their proof, see Appel, Kenneth; Haken, Wolfgang (October 1977), "Solution of the Four Color Map Problem", Scientific American, 237 (4).

The idea was to find a minimal collection of maps that could disprove the conjecture (about 1,900 such) and then use a computer to color them with only four colors, thus eliminating all the examples where four colors might not work. But – are you sure the computer is bug-free? How can you even check? This is ‘program verification; see http://www.cs.princeton.edu/courses/archive/spr16/cos217/lectures/24_ProgramVerif.pdf
We’ve had several pages of opinions from economists, computer designers, mathematicians, linguists and philosophers. Shouldn’t at some point there be work by scientists? What use is a philosopher in all this?

We now return to notes on the main text.

p27 Frege’s construction is based on the idea of equivalence of sets (that is, he introduced ways to find whether two collections had the same number). Sets $A$ and $B$ are equivalent if there exists a mapping $f : A \rightarrow B$ which is one to one and onto; that is, for each element $b \in B$, there is one element $a \in A$ with $f(a) = b$ (onto), and there is only one such $a$ (one to one). The collection of all sets equivalent to $A$ is called the equivalence class of $A$, denoted $[A]$ (again, the idea being that $[A]$ is the collection of all collections of things that have the same number of objects). The collection of all equivalence classes defines numbers.

Frege wrote *Die Grundlagen der Arithmetik (The Foundations of Arithmetic)*; soon after, the British philosopher Bertrand Russell noted that phrases such as ‘the collection of all collections’ leads to paradoxes (see Russell’s Paradox at Wikipedia https://en.wikipedia.org/wiki/Russell%27s_paradox). These paradoxes make Frege’s approach untenable. A modification would see that ‘the collection of all’ is too large; something needs to be done to restrict how many or what kinds of objects are in the collection.


p28 The axiom that infinite sets exist is phrased a bit oddly. It states that there exists a set $S$ and a map $f : S \rightarrow S$ such that $f$ is one to one but not onto. Here’s the idea: let the set be our collection of natural numbers, $\mathbb{N} = \{1, 2, 3, \ldots\}$, and define $f$ as $f(n) = n + 1$. By the Peano axiom $P_1$, every $n$ has a successor $n + 1$, so $f$ is defined for every $n$. To show $f$ is one to one, let $f(n) = f(m)$; then $n + 1 = m + 1$. By Peano $P_2$, $n = m$. To show $f$ is not onto, assume there’s an $n$ with $f(n) = 1$. But this means $n + 1 = 1$, or, in the terms of the Peano axioms, ‘$1$’ is the successor of $n$. But Peano $P_3$ says ‘$1$’ is not the successor of any $n$.

The Axiom of Infinity was introduced by the Hungarian mathematician John von Neumann, as part of a very concrete construction of the natural numbers; he used set theory to construct standard meanings for $1$, $2$, etc. If $\emptyset$ denotes the empty set, then $\emptyset$ is a good candidate to be zero, as $\emptyset$ has zero elements inside. Then $\{\emptyset\}$ is a
candidate for the number one, as it has one element in it, and:

\[ 1 = \{0\} = \{\phi\}; \quad 2 = \{0, 1\} = \{\phi, \{\phi\}\}; \quad 3 = \{0, 1, 2\} = \{\phi, \{\phi\}, \{\phi, \{\phi\}\}\} \]

Now we see how to continue, and the Axiom of Infinity allows us to ‘keep on going’ to yield the infinite set \( \mathbb{N} \). In the study of infinities, \( \mathbb{N} \) can be proved to be the ‘smallest’ infinite list. It’s a bit tricky to define how one infinity can be smaller than another, the idea is to exploit again one-to-one and onto maps. These matters were addressed by the German mathematician Richard Dedekind, who wrote the influential book *Was sind und was sollen die Zahlen?* (roughly, "What are numbers and what should they be?"). The individual who did most to clear up issues about infinities of different sizes was the mathematician Georg Cantor, who developed the theory of transfinite numbers in the late nineteenth century. For a popular, non-technical presentation, see Rudy Rucker’s *Infinity and the Mind*, Princeton University Press 2004. A basic technical introduction is in Michael J. Schramm’s *Introduction to Real Analysis*, Dover Publications 2008.

In cognitive science, numbers like like \( 10, 100, 1000 \ldots \) are called cognitive reference points. Eleanor Rosch did a series of experiments consistent with the idea that these numbers were preferentially referred to when subjects thought about a collection of random numbers such as \( 102, 173 \). See Eleanor Rosch, *Cognitive Reference Points*, Cognitive Psychology 7, 1974 p532.

Chapter 1: Numbers

Section 3: The Problem of Fractions

Compared to the integers, fractions are more like a necessary evil. To a modern mathematician, they’re easily defined, but historically have been difficult to compute with.

Much of what we think of as modern mathematics was developed in civilizations, which used bureaucracies as an organizational tool in what is now India, China, the Middle East and Egypt. All of these developed a class of administrators – individuals who could read, write, and perform mathematical computations. The mathematics was important; the bureaucracies regulated land use, paid or forced labor, wages, livestock, agriculture, military affairs, construction, trade . . . . Often, these professional administrators would be concerned with what we think of as science – for example, astronomy.

All of this involved a great deal of counting, measuring, and computing; administrators developed techniques for doing computations efficiently (and techniques for training the next generation). These are the computational techniques we’ll examine.

We’ll start with Egypt, where our literate administrators were called scribes. Workers (or forced laborers) were paid in standard-sized units of grain, or bread, or beer. The difficulty was in dividing, say three standard loaves among four workers. Each worker gets \( \frac{3}{4} \) of a loaf, but how? Do you cut each loaf in half, then half again, to get fourths? Then each worker gets three little slices? Let’s not even try jugs of beer. The scribes needed more than the ability to write fractions like \( \frac{2}{3} \); they needed to compute with fractions, and relate those computations to other kinds of numbers. We have much the same problem, for example, in food rationing during wars or other crises. Let’s say everyone gets two and a third ounces of cooking oil per week. Multiply that by 23 million people and you need .... well, you need an efficient way to do that kind of computation.

Figure 41 shows Egyptian fractions. They were were rewritten as sums of unit fractions. Unit fractions are fractions with one as the numerator, so, instead of writing \( \frac{3}{4} \), the Egyptian system was to write \( \frac{3}{4} = \frac{1}{2} + \frac{1}{4} \). This method also makes the division of the loaves of bread more practical: everyone gets a half loaf, then a quarter loaf.

Going further back in history, numbers, written language, and accounting seem to have originated in the Middle East (historical Mesopotamia – see p47). The professional administrators developed mixed systems (like our \( \frac{5}{4} \), \( 2 \frac{1}{4} \) and 2.25), as well as mixed units (pints,
quarts, liters). Mesopotamian scribes also used base sixty 'decimals'; they’d write \( \frac{1}{12} \) as \( \frac{5}{60} \), and \( \frac{4}{45} = \frac{1}{12} + \frac{1}{120} = \frac{5}{60} + \frac{20}{600} \). Or, in a modern decimal-like notation 5,20.

Base sixty is very convenient for dealing with unit fractions: \( \frac{1}{2} = \frac{30}{60} \); \( \frac{1}{3} = \frac{20}{60} \) (try that with decimals!). The outcast here is \( \frac{1}{7} \); in decimal notation, \( \frac{1}{7} = 0.142857142857142857 \) and so on. The "and so on" originally had no acceptable definition, though it can be thought of as a shorthand for the phrase "if you continue to divide, you will continue to get blocks of 142857". This is one of the main problems with fractions.

Historically, Mesopotamian scribes wrote something like \( \frac{1}{7} = 0.142857 \) and then warned "approximation given since 7 does not divide". This leaves quite a bit out: for example, if \( \frac{1}{7} \) is the amount of tax on a piece of land, and you’re a government, you want the largest number you can get away with (rounding up). If you’re the one paying that tax, you want the smallest (rounding down). Writing \( \frac{1}{7} = 0.142857 \) doesn’t say where you are.

You could argue that this is a basic deficiency of decimal notation, and for this reason, fractions simply aren’t the same as decimals. Mesopotamians took a different view: at some point, an unknown scribe wrote

\[
8,34,16,59 < \frac{1}{7} < 8,34,18
\]

The decimal version is

\[
0.14285640 < \frac{1}{7} < 0.1428611
\]

Writing \( 0.14285640 < \frac{1}{7} < 0.1428611 \) tells you the largest and smallest value you could take, but now the problem is, it doesn’t leave you with just one number. Our practical scribes had a solution: take the average, 0.14285875. Now we have a single number to use, and we know the largest and smallest variations.

Moderns think about this differently: we’d say 0.14285875 is an approximation to the real value of \( \frac{1}{7} \), but that it isn’t the real value. The way to talk about approximate versus real is to introduce the idea of error: error = (real value - approximation). Here, the error is \( \frac{1}{7} - 0.14285875 \). This doesn’t seem to help, because we don’t know the real value of \( \frac{1}{7} \). But we do know how large and how small \( \frac{1}{7} \) could be:

\[
0.14285640... < \frac{1}{7} < 0.1428611
\]

Now subtract the average (the approximation) from all three sides:

\[
0.14285640... - 0.14285875 < \frac{1}{7} - 0.14285875 < 0.1428611 - 0.14285875
\]
Rewriting, and noticing the middle is now real-approximation = error,
\[-0.00000235 < \text{error} < 0.00000235\]
or as we’d write it today,
\[|\text{error}| < 0.00000235\]
Actually, we’d write \(|\text{error}| < 2.35 \times 10^{-6}\). This might look familiar; it hints at the modern idea of a limit. What’s missing is the epsilon \(\epsilon\), which controls how small the error gets. We’ll deal with that in Section 3, on real numbers.

How does this this help our scribe? Imagine some lowly scribe presenting the taxes to his boss. The boss remarks, “You have taken the seventh part; there is an error. Perhaps the tax is too small?” But now our scribe can bow low and say, “Oh Shining One, the tax on this land is ten bushels of barley, and the error is but a part of one grain.” I’d probably hate being a scribe.

There’s point to this silly story: when we talk about whether fractions are really decimals, or whether infinite decimals are limits, we’re exporting our own twenty-first century beliefs back thousands of years, to a place they don’t belong. The historian Eleanor Robson makes this point explicitly:

On the constructivist historical view, the emphasis is on difference, localism, and choice: why did societies and individuals choose to describe and understand a particular mathematical idea or technique one particular way as opposed to any other? How did the social and material world in which they lived affect their mathematical ideas and praxis?


Robson takes a specific example from a Mesopotamian "problem set": A square is \(\frac{1}{3}\) cubit and \(\frac{1}{2}\) finger on each side; what is its area? (the answer should be \(9\frac{1}{3}\) grains).

Here we have to deal with conversion of units and mixed decimal/fraction notation. The scribe first converted the numbers to base sixty notation, squared the number (using the same kinds of techniques we’d use to do a multiplication), and then converted the answer back to decimal/fraction notation – in different units. The actual answer the scribe gave, though, was not \(9\frac{1}{3}\); the scribe converted the true area to one simpler to write in mixed notation. Base sixty, to the scribe, was just a computational tool to make certain conversions and computations easy. There was no issue of whether fractions were "really" decimals.
But, historical anachronisms aside, we are after a modern understanding of fractions. Let’s start. First, notation: fractions are quotients of integers, \( \frac{p}{q} \), so we write the collection of all fractions as \( \mathbb{Q} \) (Quotients). Getting these into decimals takes work.

Let’s take a simple fraction, \( \frac{23}{10} \) and convert it to 2.3. To start, the fraction \( \frac{23}{10} \) has a piece, 2, to the left of the decimal point. What’s left over is the fractional part, \( \frac{3}{10} \). You can access the fractional part by taking away the first part, to get \( 23.7 - 23 = .7 \). Now multiply by 10: \( 10 \cdot .7 = 7 \), and you have the part to the right of the decimal point.

How do I know to not multiply by 100 to get 70? Why does multiplying by 10 seem to be just right? If we’d had \( \frac{2307}{100} \), multiplying by 100 = 10^2 would have been ‘just right’, and multiplying by 10 would be ‘not enough’. Let’s translate this into mathematics: the part to the left of the decimal point is in \( \mathbb{N} \); the part to the right of the decimal point is the fraction \( f \), where \( 0 \leq f < 1 \).

The mathematical way to say 10 is ‘just right’ for the fractional part \( f = .7 \) is that \( \frac{1}{10} \leq .7 < \frac{1}{10} \). In contrast, for \( f = .07 \), the fact that 100 = 10^2 is ‘just right’ and 10 = 10^1 is ‘too small, gets rewritten as \( \frac{1}{10^2} \leq .07 < \frac{1}{10} \). The general idea is:

**The Archimedean Principle for Rationals:** If \( r = \frac{p}{q} > 1 \) is a rational number, then there is always a power of ten, \( 10^m \), with \( m \geq 0 \) and \( 10^m \leq n < 10^{m+1} \). If instead \( 0 < \frac{p}{q} < 1 \), there’s a negative power of ten, \( 10^{-m} \), with \( m \geq 1 \) and \( 10^{-m} \leq \frac{p}{q} < 10^{-m+1} \).

To check, multiply both sides of the inequality by \( 10^n \), and then you get \( 1 \leq 10^n \cdot \left( \frac{p}{q} \right) < 10 \). You’ve now got a number between one and ten; that’s the next decimal place of \( \frac{p}{q} \) (see p47 for a sketch of the proof). For a finite decimal, repeating the process will bring out all the decimal places, one by one.

Repeating decimals don’t fit this scheme very well – there’s always a fractional part left over. And again we have to ask: does this mean that decimals are the wrong idea to understand fractions?

The Chinese scholar Lui Hui (Figure 42) expressed similar ideas when calculating the value of \( \pi \): he used the approximation \( \pi \approx 3 \) and warned that this was not the true value, but was good enough for most practical purposes (see p47).
Lui Hui also said how he’d estimated π: he computed the area of a 96-agon inscribed inside a circle (a 96-agon is a 96-sided figure; for comparison, a triangle is a 3-agon). He also gave a formula for going from one approximation to a better one: see Figure 43, where he goes from the area of a 6-agon to that of a 12-agon. He then goes to a 24-agon, a 48-agon, and finally a 96-agon.

Something new happened: for \( \frac{1}{7} \), we got a decimal approximation. But Lui Hui generates not just one number, but a whole collection of numbers. Technically, the collection of numbers Lui Hui gave for π is called a sequence (see p47). The sequence is a collection of better and better approximations. Again, we think ‘limit’ and ‘numbers like \( \frac{1}{7} \) are limits of actual finite decimals.’ This is teleology: we know how all these issues turned out, so we’re imagining that the path mathematics took was somehow preordained. Limits, etc, may not be at all what Lui was thinking. And perhaps at some point, it might have been argued that decimals really were a poor choice, because, like the ancient scribe wrote, in \( \frac{1}{7} \), ‘7 does not divide.’

These kind of ideas were used in the work of Ptolemy, an astronomer living in Alexandria, Egypt about 100 CE. Like the Babylonian mathematicians (p15), he wanted to predict the position of the planets; the difference was that he believed the planets moved in circular orbits. The mathematics available at that time was Euclidean geometry, and he used trigonometry to relate lengths along circular orbits to angles. Which involved computing the sines of all the angles, or at least a large number of them. He was also aware of trig identities, like \( \sin(2\theta) = 2\sin(\theta)\cos(\theta) \). If he could compute the sine of half a degree, this would give him the sine of a degree, and with other identities, the sine of all integer angles.

So – how to get the sine of a half-degree from the sine of known angles? He used a result of Archimedes:

\[
\text{If } \theta > \psi \text{ Then } \frac{\theta}{\psi} > \frac{\sin(\theta)}{\sin(\psi)}
\]

Using \( \theta = \frac{3}{4} \) and \( \psi = \frac{1}{2} \), he got

\[
\frac{\sin(\frac{3}{4})}{\sin(\frac{1}{2})} < \frac{3}{2}
\]

Rearranging, \( \frac{3}{2} \sin(\frac{3}{4}) < \sin(\frac{1}{2}) \). Numerically, using modern values (and degrees, not radians) we get \( 0.008726397 < 0.0087265355 \), and a similar inequality on the other side, which is remarkably accurate.
Notes for Chapter 1 Section 3: The Problem of Fractions


p45 The Archimedean Principle for rational numbers begins with the division algorithm (not surprising; \( \frac{a}{b} \) is a division!). Roughly, a number like \( \frac{80}{9} \) can be written as \( 8 + \frac{8}{9} \); the second term is a fraction less than one. In Archimedean terms, this shows that if we have \( 10^0 \leq 8 < 10^1 \), it's still true that \( 10^0 \leq 8 + \frac{8}{9} < 10^1 \).

The second half, dealing with \( 0 < r < 1 \), follows by applying the the Archimedean Theorem to \( \frac{1}{r} > 1 \) and then inverting the inequalities.

p45 The work of the mathematician Lui Hui appeared in commentaries and solutions to the Chinese text *The Nine Chapters on the Mathematical Art*, written in 263 CE.

p46 Technically, a sequence is more than just a collection of numbers; there’s also a sense of one number following another (think of the cognate word *sequel* for the movie that follows the original; similarly the word *second* is also a cognate: it’s the number following the first).

To say that approximations get better and better, we need a sense of the direction to go so we can get get better; the sequence provides that direction. So, a sequence comes with a first number, a second number, etc. For 'first number' we write \( a_1 \), the second would be \( a_2 \), and so on. The sequence is then \( (a_1, a_2, \ldots) \).
Chapter 1: Numbers

Section 4: The Problem of Irrationals

We’ve been talking about mathematics as an extension of our mental perception of number, but number is just one quality we perceive. We also know qualities like size, position, length, area, angle, volume, weight . . . . Very early on, numbers were linked with these other kinds of qualities. That link created an association leading to entirely new kinds of number: those connected to geometry.

How exactly qualities like length and area came to be understood as number is a complex story, one which is still being worked out. We’ll discuss the neurophysiology in detail, in Section 9, Time and Space. See also Geometry and Decimals, p70, and Units & Standards, p75.

Geometric qualities: length, distance, area, volume ... Why? Because bureaucracies always tax, and how do you compute tax? If you just take part of the harvest from a farmer, the farmer can hide the harvest before the tax collector visits. Instead, compute the area of the land planted, then compute how much produce the land should yield, then take part of that (actual Egyptian practices were much more sophisticated than this; see p54). A simple scheme, but ancient inheritance involved subdividing land amongst many children, so taxable lands had complicated shapes. Figure 46 shows an Egyptian computation for the area of a complicated figure.

The area of a triangle is half the base times the height, so geometers had to understand the connection between numbers and lengths. This unites counting with measuring; quantity and geometry. We use yardsticks all the time, but the link was less obvious in ancient times: Figure 47 implies tools for surveying are gifts from the Gods. People took their lengths seriously – and state taxation depended on public trust in those tools (public standards, once more).

Figure 48 shows a problem in geometry from Mesopotamia, about 1700 BCE. It’s a right triangle, with two equal sides; each side has length one. The Pythagorean theorem tells us if \( s \) is the length of the two equal sides, and \( h \) that of the hypotenuse, \( s^2 + s^2 = h^2 \), and for \( s = 1 \), \( 2 = h^2 \). We’d write \( h = \sqrt{2} \), but it’s very unlikely a scribe would write any of the above; see p54). On the tablet, in base 60 On the tablet, in base 60, it’s written 1, 24, 51, 10; in decimal notation, 1.41421. We’d call it \( \sqrt{2} \). The Mesopotamian answer in Figure 48 looks like the kind of approximations we used for \( \frac{1}{7} \): it seems there’s nothing new or surprising here.

How did the Mesopotamian scribes get their approximations? No-
one knows, but here’s the best idea historians came up with (this method gives the same answers the Mesopotamian scribe wrote).

Start with a first guess for $\sqrt{2}$: say, $\frac{3}{2}$ or 1.5. That’s too big, but $2/\frac{3}{2} = \frac{4}{3} = 1.33 \ldots$ is too small. Take the average; that’ll be in-between, so it will be closer to the true value than either guess. So, if $g_1$ is the first guess, then you get a second, better guess with

$$g_2 = \frac{1}{2} \left( g_1 + \frac{2}{g_1} \right)$$

With $g_1 = \frac{3}{2}$, then

$$g_2 = \frac{1}{2} \left( \frac{3}{2} + \frac{2}{\frac{3}{2}} \right) = \frac{1}{2} \left( \frac{3}{2} + \frac{4}{3} \right) = \frac{17}{12}$$

And again:

$$g_3 = \frac{1}{2} \left( \frac{17}{12} + \frac{2}{\frac{17}{12}} \right) = \frac{1}{2} \left( \frac{17}{12} + \frac{24}{17} \right) = \frac{577}{408}$$

How good are these approximations? Square them, and compare with 2:

$$g_1^2 = \frac{9}{4} = 2.25; \quad g_2^2 = \frac{289}{144} = 2.00694; \quad g_3^2 = \frac{332929}{166464} = 2.0000060073 \ldots$$

We’re back in the Lui Hui situation: not just an approximation, but a way to get better and better approximations. And as before, you carry out as many decimal places as you need, get on with your job, report to the Chief Scribe, get your beer ration and go home.

But Greek mathematicians discovered $\sqrt{2}$ is not like $\frac{1}{7}$: it cannot be written as a quotient $p/q$, and it is not a repeating decimal (for the proof, see p55). $\sqrt{2}$ isn’t rational; it’s ir-rational (‘ir-’ means ‘not’ as in ‘ir-relevant’). For some history of irrationals, see see p55.

So: what is $\sqrt{2}$? We could still say it’s like $\frac{1}{7}$, because you can get more and more digits of $\sqrt{2}$. While you can’t get them by long division, at least you can get the numbers $g_1, g_2, \ldots$, whose squares approximate 2.

But this is a kind of fraud. When you do the long division for $\frac{1}{7}$, you can see exactly where the decimal starts to repeat and why. As we saw, that means you can find the error in any one approximation.

With the $g_1, g_2, \ldots$ you don’t know what the $g_1, g_2, \ldots$ are going to do, or why. Could $g_4$ be $\frac{17}{12}$ again? Then $g_5$ would be $\frac{577}{408}$ again, and you wouldn’t get better approximations. How can you rule that out? And how can you show it does get ‘better and better’?

The final answer came something like 2000 years later, so these are hard questions. The most significant answer (speaking historically)
was to avoid the question, which involved rethinking the relation between geometry and number.

The Greek mathematician Eudoxus of Cnidos (408 - 355 BCE) undid the link between quantity and geometry by developing a consistent theory of magnitude. Eudoxus used magnitude as an undefined term; one could think of magnitudes as being the lengths of lines, or areas and volumes of figures; he showed how to manipulate magnitudes as ratios, analogous to manipulation of ratios of numbers. For example, we can define \( \frac{m}{n} = \frac{p}{q} \) to mean \( mq = np \). For certain line segments, such as the diagonal of a square, you can no longer think of the magnitude as being the length (as it’s irrational); you have to think that the diagonal itself is the magnitude. You can then answer questions such as, if you double the magnitudes of the sides of a square, do you double the magnitude of the diagonal (yes).

Ratios were enough to do the geometry Eudoxus and most Greek mathematicians wanted, for example, the construction of figures using a ruler and a compass (see p56).

The Eudoxian theory was influential for centuries; even Newton, in his *Arithmetica Universalis* of 1707 defined numbers as ratios of line segments. The prevailing opinion was stated by the German mathematician Michael Stifel (1486-67), who was critical of using approximations to define an irrational:

> ... considerations compel us to deny that irrational numbers are numbers at all. To wit, when we seek to subject them to [decimal representation] ... we find they flee away perpetually, so that not one of them can be apprehended precisely ... Now that cannot be called a true number which which is of such a nature that it lacks precision ... so an irrational number ... is hidden in a kind of cloud of infinity.

Newton’s refusal to accept irrationals may seem inconsistent with his discovery of limits and calculus. However, Newton had a very classical training, beginning with Euclid, and tended to think geometrically. The idea of expressing an irrational number as a limit of rational numbers might have made no sense to him.

In Europe there were no alternatives to Eudoxus for over a thousand years, which didn’t prevent (some) mathematicians from using algebra to deal with "numbers" like \( \sqrt{2} \). The mathematician Leonardo of Pisa, who wrote under the name ‘Fibonacci’, was aware of Arabic work on algebra; in 1225 he published the solution to a problem mentioned by Omar Khayyam, in his book *Al-jabr*: solve the equation \( x^3 + 2x^2 + 10x = 20 \) (see Figure 49).
Fibonacci showed there were no integer solutions, no rational solutions, and that the solution could not be constructed by ruler and compass. So the number was irrational, but of some unknown kind.

The Mesopotamian scribe on p.47 would probably shrug his shoulders: what did it matter, as long as he could compute three or four digits of these numbers, and keep the Chief Scribe happy?

For us it’s more difficult: am I going to run into new kinds of irrationals each time I solve a new equation?

In an attempt to create some kind of order, mathematicians began to rethink their irrationals. Fractions like \( \frac{1}{7} \), and irrationals like \( \sqrt{2}, \frac{1+\sqrt{5}}{2} \ldots \) and even Fibonacci’s irrational, are all solutions of equations:

\[
7x - 1 = 0, \ x^2 - 2 = 0, \ x^2 - x - 1 = 0, \ x^3 + 2x^2 + 10x - 20 = 0
\]

Solutions to these equations were called algebraic numbers, since they could be obtained by solving algebra equations with integer coefficients. Mathematicians were leaving behind Eudoxean geometric methods, moving to ideas that come from algebra.

The numbers \( \pi \) and \( e \) didn’t seem to be similar to algebraic numbers; the mathematician Euler remarked (1744) that these two seemed to go beyond the the techniques of algebra. As the Latin for ‘go beyond’ is ‘transcend’, Euler suggested that these two were transcendental numbers. In 1878 the number \( e \), and in 1882, \( \pi \), were each shown to be transcendental: that is, they were not solutions of algebraic equations with integer coefficients.

Of course \( \pi \) is a solution to the equation \( \cos(\theta) = -1 \), and \( e \) to the equation \( \ln(x) = 1 \); the functions \( \cos(\theta) \) and \( \ln(x) \) are transcendental functions. Again, new equations, new irrationals. By this time, European mathematicians knew many new kinds of functions, for example the Bessel function \( J_1(x) \), which describes the pattern of rings when light is diffracted through a small hole (see Figure 50). All light through microscopes and telescopes gets diffracted, and the presence of the first dark ring determines how close two objects have to be to blur into one, through the lens. In short, \( J_1(x) \) determines the resolution of the lens. The dark rings occur when the intensity is zero, that is, \( J_1(x) = 0 \). Is this going to involve totally new kinds of numbers?

Just how many kinds of irrational are there?

There’s another issue: we know we can compute more and more decimal places of accuracy for \( \frac{1}{7} \), and we believe we can do that for \( \sqrt{2} \) and \( \pi \), but what about these new numbers? Do we even know they’re decimals?
Answering this question is an invitation to a new world: chains of numbers with no decimal expansion, ascending and descending into the infinitely large and infinitely small.

To enter this world, all you have to do is imagine .9999... doesn’t equal 1. Intuitively, the decimal ‘never gets there.’ So there’s a gap: then $\gamma = 1 - .9999... > 0$ measures the size of the gap: what kind of things are inside that gap? We’re going to discover many kinds of numbers inside there – and none of those numbers are decimals. This is the problem: the world of numbers might be strange, almost beyond imagining.

Let’s look a bit at $\gamma$. How big is it? What’s the decimal expansion?

If the decimal expansion of $\gamma$ starts with something like .276..., then that first decimal place makes $\gamma > .1$. But $\gamma = 1 - .9999... = 1 - .9 - (.9999...) < 1 - .9 = .1$ We can’t have .1 < $\gamma$ < .1, so $\gamma$ has to start with something like $\gamma = .0276...$. The problem is that $\gamma = 1 - .9999... < 1 - .99 = .01$ and again, we can’t have .02 < $\gamma$ < .01, so $\gamma$ has to start with something like $\gamma = .00276...$

The problem is, actually, that this never stops: 0 < $\gamma$ < $\frac{1}{10^k}$ for all $k$. $\gamma$ has the decimal expansion $\gamma = 0.00000...$ - but still $\gamma$ isn’t zero. So we have what we feared: a number that doesn’t have a decimal expansion at all. There goes the beer ration for our scribe.

We could say $\gamma$ is an infinitely small number, but not zero. Then $\frac{\gamma}{2} > \frac{\gamma}{3} > \frac{\gamma}{4} > ...$ are also infinitely small numbers with no decimal expansion. And so are $\gamma > \gamma^2 > \gamma^3...$.

So we don’t have just one infinitely small number, we have whole chains of them, getting smaller and smaller. We’ll see it gets worse – infinitely worse.

Since $\gamma < \frac{1}{10^n}$ for all $n$, then $\omega = \frac{1}{7} > 10^n$ for all $n$. Again, this means $\omega$ can’t start with $\omega = 1.374...$, and it can’t start with $\omega = 1384.732...$ because 1384.732... < 10,000 = $10^4$ but $\omega > 10^4$. So $\omega$ can’t start with any numbers before the decimal point: $\omega$ is infinitely large. And so are $\omega^2 < \omega^3...$

And as if those aren’t large enough, $\mu = \omega^{\omega}$ is larger than all of them. Now we start again and we get $\nu = \mu^{\mu}...$ and we get chains of larger and larger infinities.

Now let’s get more infinitely small numbers:

$$\gamma > \gamma^2 > \gamma^3... > \gamma^\omega > \gamma^{\mu}...$$

The gap contains a nightmare of infinities!
There’s supposed to be a way out, if you know limits. We’re supposed to know \(0.999\ldots = 1\) because the collection \(\{0.9, 0.99, 0.999\ldots\}\) has 1 as a limit. Let’s try that. First, a little notation:

\[0.9 = \frac{9}{10}; \quad 0.99 = \frac{9}{10} + \frac{9}{10^2}; \quad 0.999 = \frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3}\ldots\]

Now we can talk about the limit: to say the limit of the \(\{0.9, 0.99, 0.999\ldots\}\) is 1, is to say that for every \(\epsilon > 0\) there’s a point after which the sequence is at least \(\epsilon\) close to 1:

\[|1 - \left(\frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3}\ldots + \frac{9}{10^k}\right)| < \epsilon\]

But

\[1 - \left(\frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3}\ldots + \frac{9}{10^k}\right) = \frac{1}{10^k}\]

So we’re saying there’s a point after which \(\frac{1}{10^k} < \epsilon\). But, this is the whole thing about numbers like \(\gamma\): if \(\epsilon\) is one of our infinitely small numbers, it’s the other way around: \(\frac{1}{10^k} < \epsilon < \frac{1}{10^k}\).

Not only do numbers like \(\gamma\) mess up ideas about decimals, they mess up the whole theory of limits.

Maybe our logic is bad? We can find a mistake in one of those computations? To tell us no such \(\gamma\) could ever exist?

No.

The British mathematician John Conway, Figure 52, invented a number system called the surreal numbers, \(S\), which has infinitely small and infinitely large numbers (though his construction is nothing at all like our simple intuition for \(\gamma\)). Surreal numbers have a kind of not-really-a-decimal expansion; very very roughly, an infinitely large number might be something like

\[\ldots + 3\gamma^2 - 5\gamma + \ldots + \frac{3}{10^2} + 7 + 2\cdot10^5 + \ldots + 3\omega - 5\omega^3 + \ldots + 3\mu + 2\mu^2\ldots + 5\nu + \ldots\]

But they’re not decimals, or any finite kind of thing; it even requires rethinking what it means to take infinite sums. See the references on p 56.

There is no simple way to eliminate the surreals; you have to define real numbers in a way that will exclude them right from the start. We are making a choice; we are constructing real numbers to be the way we think they should be. The next section shows how it was done.
Notes for Chapter 1 Section 4: The Problem of Irrationals

p48 The amount of tax depends not just the area of a field, but also on the quality: fields were graded by their size and position as well as the presence of canals, trees and wells, and possible damage inflicted by floods. A second assessment was made before the harvest, followed by a final weighing of the harvested and threshed grain. See Corrina Rossi, Mixing, Building and Feeding: Mathematics and Technology in Ancient Egypt, in Eleanor Robson and Jacqueline Stedall, The Oxford Handbook of the History of Mathematics, Oxford University Press 2011.

p48 We found the hypotenuse of a right triangle as $\sqrt{2}$, using a bit of algebra. How did the Mesopotamian scribes think of it? It’s long been assumed that the scribes used a kind of algebra; the of problems they were able to solve were like this: *I totaled the area and (the side of) my square: it is 0; 45.* We assume this means ‘If $x$ denotes the side length, solve $x^2 + x = 0; 45$, and the scribe would solve the equation by completing the square.

This might not be anything like what the scribes were thinking. The historian Jens Hørup retranslated a group of tablets, trying to stay as close to the original as possible. He noted that the scribes don’t mention variables; they discuss lengths, areas and volumes. The word used for subtract, Akkadian *nasāhum*, is more literally rendered "to tear out". Hørup then suggests that the scribes had a kind of ‘geometric algebra’, a set of geometric techniques for solving problems. He further speculated that these techniques developed from the concrete problems of surveys. See Jens Hørup, Lengths, Widths, Surfaces: A Portrait of Old Babylonian Algebra and Its Kin Springer 2002. See also Rahul Roy On Ancient Babylonian Algebra and Geometry at https://www.ias.ac.in/article/fulltext/reso/008/08/0027-0042

Along these lines, Greek philosophers and historians attributed the development of their own geometry to Egyptians and Mesopotamians: Herodotus (about 400-500 BCE) wrote: "They said also that this king [Sesostris] divided the land among all Egyptians so as to give each one a quadrangle of equal size and to draw from each his revenues, by imposing a tax to be levied yearly. But every one from whose part the river tore away anything, had to go to him and notify what had happened; he then sent the overseers, who had to measure out by how much the land had become smaller, in order that the owner might pay on what was left, in proportion to the entire tax imposed. In this way, it appears to me, geometry originated, which passed thence to Hellas."
A photograph of the Yale Babylonian Collection’s Tablet YBC 7289 (c. 1800 to 1600 BCE), showing a Babylonian approximation to the square root of 2 (1,24,51,10 base 60). Babylonian mathematicians knew Pythagoras’ Theorem relating the sides of a right triangle. The photo is by Yale professor Bill Casselman; see http://www.math.ubc.ca/ cass/Euclid/ybc/ybc.html).

The result that \( \sqrt{2} \) is not a rational is contained in the works of the Greek mathematician Euclid. The original proof is ‘lost in the mists of time’, but there’s reason to believe it was more of a plausible picture than a proof. The rough sketch of a proof that we give is a modernized version of the one Euclid gave (and is an algebraic version of that more ancient geometric picture-proof).

The first result we need is that every integer is either even (has a factor of 2: 2, 4, 6, 8, …) or odd (has no factor of 2: 1, 3, 5, 7, …). This is clear enough from the list, and is easy to prove using induction. A little algebra gives that the squares of even numbers are even (4, 16, 36, 64, …), and the squares of odd numbers are odd (1, 9, 25, 49, …). Easy, but important, because it allows us to do things where we’d want to use square-root: normally, if \( n \) had a factor of 2, all we could say that \( \sqrt{n} \) has a factor of \( \sqrt{2} \). But, using even/odd, we can do more: if \( n^2 \) has a factor of 2, then \( n \) has a factor of 2, instead of just a factor of \( \sqrt{2} \).

The next result we need is that you can simplify fractions by canceling out common factors: \( 60/36 = 30/18 = 15/9 = 5/3 \) and now there’s no longer any common factors. This is a bit harder to prove; it needs our strong induction.

Now we can start: say you can find \( \sqrt{2} \) as a fraction \( \frac{p}{q} \). You might as well cancel out common factors. Then

\[
\frac{p}{q} = \sqrt{2} \quad \text{so} \quad \frac{p^2}{q^2} = 2 \quad \text{so} \quad p^2 = 2q^2
\]

This shows \( p^2 \) has a factor of 2, so \( p \) also has a factor of 2; write it \( p = 2r \). Then \( p^2 = 2q^2 \) becomes \( (2r)^2 = 2q^2 \) or \( 4r^2 = 2q^2 \). Divide by 2 to get \( 2r^2 = q^2 \); now \( q^2 \) has a factor of 2, so \( q \) also has a factor of 2. But we already cancelled all the 2’s, and so, no such \( p \) and \( q \) can exist.

The discovery of irrational numbers is attributed to the philosopher Hippasus of Metapontum; the irrational in question was likely \( \frac{1+\sqrt{5}}{2} \), derived from a pentagram. Hippasus lived in the late fifth century BC (that is, from 500 to 401 BCE, closer to 401), and was a member of the Pythagoreans, (see p18). Pythagoreans explained all the world by integers, so the discovery of irrationals was a serious
challenge to their beliefs.

A note on ‘rational’ and ‘irrational’. For much of Greek mathematics, fractions were thought of as ratios. Hence, our quotients \( \frac{p}{q} \) were, to the Greeks, the rational numbers, because they were ratios. Historically, ‘rational’ is from the word ‘ratio,’ from the Latin word ‘to compute’. This in turn is from Proto Indo-European, ∗reh-, to ‘put in order’. None of this relates to our uses like ‘rational thought’, which comes the French, meaning "right, just, fitting, fair ".

New irrationals appeared in Book X of Euclid’s Geometry; these came from ‘ruler and compass’ constructions. For example, a right triangle with side lengths one has hypotenuse \( \sqrt{2} \). Now start using \( \sqrt{2} \) as a base of a right triangle; continuing on, you can get \( \sqrt{1 + \sqrt{2}}, \sqrt{5 + \sqrt{3}}, \sqrt{\sqrt{5} + \sqrt{3}} \). Euclid showed that the theory of magnitudes could generate these; his irrationals all came from geometry.

Both Archimedes and Hero of Alexandria worked on finding the value of \( \pi \), the ratio of the circumference of a circle to its diameter. They approximated the circle by \( n \)-agons, dissected these into triangles, as in Figure 53, and computed areas or lengths of their sides; either led to irrationals. But they needed numbers; both used variations on the estimates

\[
p + r \leq \sqrt{p^2 + r^2} \leq p + r
\]

While the estimate is reminiscent of the Mesopotamian computations for \( \frac{1}{7} \), the computation was motivated, once more, by geometry.

Figure 54 gives a geometric proof that \((p + q)^2 = p^2 + 2pq + q^2\): it shows that if you start with a square of area \( p^2 \), you can get a square of larger area \((p + q)^2\) by adding two rectangles of area \( pq \), and a square of area \( q^2 \).

Here’s how this led to an approximation for \( \sqrt{2} \): \( \sqrt{(p + q)^2} = p + q \), so \( \sqrt{p^2 + 2pq + q^2} = p + q \). If \( q \) is a small number (think \( q = .01 \)) then \( q^2 \) is even smaller (think \( q^2 = .0001 \)). In our approximation, we can ignore it; then \( \sqrt{p^2 + 2pq} \) is about the same as \( \sqrt{p^2 + 2pq + q^2} = p + q \). Now let \( r = 2pq \), then \( q = \frac{r}{2p} \) and we get Archimedes’ approximation, \( \sqrt{p^2 + r} \) is about \( p + \frac{r}{2p} \). We’ll use tangent lines, to get the same result.

Chapter 1: Numbers

Section 5: The Real Numbers

In the mid-1800’s, many European mathematicians were working on irrationals (see p62). There were two issues; we’ve discussed the first: are all irrationals infinite decimals? The second issue is the problem of limits. Answering these questions was part of a movement: the arithmetisation of geometry, eliminating intuitive ideas from geometry, and replacing them with better defined notions of arithmetic and algebra.

Limits, by this time, were important to all kinds of mathematics. For example, we can compute \(\sqrt{2}\). Take \(\{g_1, g_2, \ldots\}\) where \(g_2 = \frac{1}{2} \left( g_1 + \frac{2}{g_1} \right)\), etc, and if we believe that \(\{g_1, g_2, \ldots\}\) has a limit, \(g\), then the limit has to be \(g = \frac{1}{2} \left( g + \frac{2}{g} \right)\). Solving this gives \(g^2 = 2\). We suspected the \(\{g_1, g_2, \ldots\}\) became progressively better approximations to \(\sqrt{2}\); with a theory of limits, we can say the limit of the approximations is \(\sqrt{2}\).

A bit after Newton, calculus was differentiation, integration, approximation by tangent lines, and infinite series. Especially infinite series. These could be treated as three separate techniques, until the work of Cauchy, Figure 56, who in 1821 had published the first calculus book: *Cours d’analyse*. Cauchy gave careful proofs of the main results of calculus (see p62), but he also showed how the three topics of traditional calculus could be done carefully by reducing them to problems about limits. No understanding of limits: no calculus. Cauchy was aware of techniques for approximating numbers like \(\pi\), \(\sqrt{2}\), and also of how one could talk about the accuracy of approximations, using the same ideas we discussed for computing \(\frac{1}{2}\) on p46: the *error* in approximating the true value \(t\) by an approximation \(a\) is given by \(|t - a|\), and we can check how small the error is by writing inequalities like \(|t - a| < 2.35 \times 10^{-6}\), as we did with \(\frac{1}{2}\) (p45).

But, leftover from Newton and Stifel, p50, there was still no idea of what \(\sqrt{2}\) is as a number rather than a ‘magnitude’. We might prove \(\{g_1, g_2, \ldots\}\) converges by showing \(|\sqrt{2} - g_n|\) is small – but to do this, we first need to know \(\sqrt{2}\) is a number. Standoff.

The mathematician Georg Cantor resolved these issues, giving a construction of the real numbers. Why construction? Because it wasn’t as though everyone suddenly hit on the one idea that was out there waiting to be discovered. Nor did everyone say ‘Why of course, that’s what I was thinking all along’. In reality, several other constructions competed for acceptance, and others besides Cantor’s are still used.
The real numbers were built, not handed down.

We’re going to do an overview of Cantor’s construction; a large part of modern mathematics depends on this. I first saw the construction in the senior mathematics courses I took at Cornell in 1970; it hasn’t gotten easier, though fifty years of thinking on it has helped.

Cantor took important ideas from Cauchy; what he needed was Cauchy’s idea of how to talk about convergence without mentioning the limit. That idea starts with a list of numbers (a sequence) 
\((a_1, a_2, a_3, \ldots)\) abbreviated as \((a_k)\); it’s understood that \(k\) goes through all the natural numbers, one after another (sequentially).

If a sequence \((a_k)\) converges, then after a while all the \(a_k\) have to be close to the limit, so they have to be close to each other. The phrase ‘after a while’ translates to the existence of a number \(N\) specifying exactly when that closeness begins to happen; the word ‘all’ translates to ‘all numbers bigger than \(N\).’ In symbols, \(j \geq N, k \geq N\).

‘Close to each other’ means that the distance between them is small; that translates into saying that \(|a_j - a_k|\) is small. How small? Well, to be a limit, it has to get smaller and smaller. Put another way, if you tell me how small you want it, I can make it that small. ‘How small you want it’ translates to ‘for any \(\epsilon > 0\), and ‘I can make it that small’ now translates to \(|a_j - a_k| < \epsilon\). When you put it all together, you have a definition of convergence that doesn’t mention the actual limit:

**Cauchy Condition for Convergence:** If the sequence \((a_k)\) converges, then for every \(\epsilon > 0\), there is a number \(N\) such that, whenever both \(j \geq N, k \geq N\), then \(|a_j - a_k| < \epsilon\).

Cantor introduced a second idea: we believe that the sequence 
\((3, 3.1, 3.14, 3.145, \ldots)\) converges to \(\pi\), but we don’t know what \(\pi\) is. Why not define \(\pi\) to be the sequence \((3, 3.1, 3.14, 3.145, \ldots)\)? In general, a real number is defined to be any sequence? Well, not any sequence: the sequence 
\((1, 0, 4, 0, 9, 0, 16, 0, \ldots)\) doesn’t converge, so we don’t want that; we want only the convergent sequences.

That is, Cauchy sequences! Let’s check:

\[
|a_1 - a_2| = |3 - 3.1| = .1 = \frac{1}{10} \\
|a_2 - a_3| = |3.1 - 3.14| = .04 = \frac{4}{100} \\
|a_3 - a_4| = |3.14 - 3.145| = .005 = \frac{5}{1000}
\]

These are getting smaller, so it’s a nice start on the road to showing the sequence is Cauchy, but it isn’t enough: we need all \(j \geq N, k \geq N\), and here we only have \(j \geq N, k = j + 1\). See p62 for a full proof.

The third idea is this: go back to \((3, 3.1, 3.14, 3.145, \ldots)\). All of these are finite decimals; they’re rational numbers. So: a real number is a
Cauchy sequence of rational numbers.

There’s a fourth idea, and it’s that pesky \( \epsilon \) being infinitely small issue, again. Cantor avoided it by restating the Cauchy condition: instead of saying ‘for every \( \epsilon > 0 \)’, he says ‘for every rational number \( \epsilon > 0 \)’. The Archimedean Principle for Rationals (p45) now eliminates any worries about infinitely small epsilons.

Cantor calls the collection of Cauchy sequences the real numbers, \( \mathbb{R} \).

There are lots of problems:

**Question:** So what are rational numbers in this new definition?

**Answer:** A rational number, say \( \frac{1}{2} \), would be the sequence \( \{\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \ldots\} \).

**Question:** How am I supposed to add, subtract, etc, these things?

**Answer:** This turns out to be amazingly easy: take two real numbers \( R = (r_1, r_2, \ldots) \) and \( S = (s_1, s_2, \ldots) \); \( R + S = (r_1 + s_1, r_2 + s_2, \ldots) \). Same for multiplication, etc. These are sequences of rational numbers, and it’s easy to show they’re Cauchy.

**Question:** What happens to the whole theory of limits?

**Answer:** We’d like to have the usual definitions, but the tricky part is defining statements like \( |R_k - R| < \epsilon \). That’s just \( R - \epsilon < R_k < R + \epsilon \), so what we really need to understand is inequalities, like \( R < S \).

That’s the same as asking for the meaning of \( 0 < S - R \), so in the end, we just need to know what it means for a real number \( R = (r_1, r_2, \ldots) \) to be positive.

Here’s where Cantor’s sneaky fourth idea comes in: using rational numbers to define inequalities. We’d like to just say that \( R \) is positive if the rational numbers making up \( R \) are positive, but that doesn’t work: take \( R = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots) \). Intuitively, \( R \) is the ‘limit’, and that ‘limit’ is zero, not positive. So Cantor defined \( R > 0 \) as follows: \( R = (r_1, r_2, \ldots) > 0 \) means there is a rational number \( r > 0 \) and an \( N \) such that if \( k \geq N \), then \( r_k \geq r \): the rational \( r > 0 \) keeps the ‘limit’ away from zero.

We also put on the extra words ‘if \( k \geq N \)’. Again, it’s about limits: the sequence \( (-1, -0.5, 0, 0.5, 0.59, 0.599, 0.5999, \ldots) \) has limit .6, which is positive: what the first three terms do is irrelevant. That what \( N \) does: if I take \( N = 4 \), then \( n \geq N \) says first three are irrelevant. Now with \( r = .5 \), and \( N = 4 \) for all \( k \geq N, r_k \geq r \).

**Question:** Many sequences of rationals converge to \( \sqrt{2} \), so which am I supposed to use?

**Answer:** We can use any sequence we like. All you need to do is check: sequences \( (a_k), (b_k) \) define the same real number if, intuitively, they have the same limit; see p62 for the technical details.
**Question:** What if I take a limit of sequences of reals? Do I get some weird new kind of number?

**Answer:** Actually, no. If you took a sequence of real numbers \( \{R_1, R_2, \ldots\} \) with limit \( R \), then you can also find rational numbers \( \{r_1, r_2, \ldots\} \) with \( R = \{r_1, r_2, \ldots\} \). Very roughly, the idea is this: if

\[
R_1 = (a_1, a_2, a_3, \ldots) \\
R_2 = (b_1, b_2, b_3, \ldots) \\
R_3 = (c_1, c_2, c_3, \ldots)
\]

Then

\[
R = (a_1, b_2, c_3, \ldots),
\]

which will work.

With this out of the way, Cantor proves three easy results:

One: Every Cauchy sequence of real numbers converges to a real number. This is called the **completeness** of the real numbers.

Two: Every real number is a limit of rational numbers; in fact, if \( R = (r_1, r_2, \ldots) \), then, with sloppy notation, \( \lim r_k = R \). This is called the **density** of the rational numbers.

Three: All the known irrationals were limits of rational numbers, so all the known irrationals are now official real numbers.

Finally, Cantor’s definition of \( R > 0 \), using rational numbers, allows us to show the real numbers \( \mathbb{R} \) don’t contain infinitely small numbers. Here’s why:

Take a supposed infinitely small real number \( \gamma = \{r_1, r_2, \ldots\} > 0 \). \( \gamma > 0 \) means there’s a rational number \( r > 0 \) and an \( N \) such that for all \( k \geq N \), \( r_k \geq r \).

Now we have our rational number \( r = \frac{p}{q} \leq \gamma < \frac{1}{10^k} \) for all \( k \). But this contradicts the Archimedean Principle for Rationals (p45), that for a rational \( r, 0 < r < 1 \), there’s an \( m \) with \( \frac{1}{10^m} < r \). No infinitely small rational numbers, no infinitely small real numbers. And, BTW,.9999 . . . actually is equal to 1.

We used decimal expansions to get an idea of what is in the infinite collection \( \mathbb{N} \); we’d like something similar for \( \mathbb{R} \). We know each real number is a limit of rational numbers, and each rational number is a repeating decimal, but it might be that the limit is some very weird kind of object. For example, we’d like \( \pi = (3, 3.1, 3.14, \ldots) \), but it might instead be more like \( (7, 2.1, 6.3, 2.4 \ldots) \). We’ll look at it in the next section, but for now, it’s worth noting, that \( (3, 3.1, 3.14, \ldots) \) isn’t necessarily the best way to approximate and to compute \( \pi \); see p63.

What we have right now is that the real numbers are all limits of rationals. We also know that every infinite decimal gives rise to a Cauchy sequence like \( (3, 3.1, 3.14, \ldots) \), whose limit is the real num-
ber representing that infinite decimal. What we haven’t yet shown is that every irrational is an infinite decimal; we’ll do exactly that in Section 6 Part 2.

Cantor’s theory raised troubling questions. For example: we can actually compute the decimal approximations for \( \sqrt{2} \) and \( \pi \) with the methods given by the Mesopotamians and by Lui Hui. We can even write computer programs to do the computations. Call numbers like \( \pi, \sqrt{2} \) *computable* numbers. Are all real numbers computable? Cantor showed they are not: **most** real numbers are not computable (see p63).

If there are so many (most, actually) real numbers that aren’t even computable, how *real* is \( \mathbb{R} \)? Put another way, we have more numbers than we can understand – did we go too far? It certainly could seem so. After all, we have a very concrete understanding of what a number ought to be; recall the discussion of animal counting from p10. Numbers are likely built into the way our brain sees the world. Cantor’s construction is nothing like our built-in view of a number.

On the other hand, Eudoxus’ theory of magnitude, p50, views \( \sqrt{2} \) as an undefined ‘magnitude,’ whatever that is. It may well be that the basic construct of number that evolution has given us for picking berries is not adequate for understanding a wider and wider universe.

Another answer is this: \( \mathbb{R} \) is the smallest collection of numbers that contains all the limits of numbers from \( \mathbb{Q} \) (and in which the ordinary rules of arithmetic hold). In technical terms, \( \mathbb{R} \) is the smallest *complete ordered field* containing \( \mathbb{Q} \).

And here’s a third answer: we talked about algebraic numbers being solutions to polynomial equations (with integer co-efficients; see p51), but numbers like \( \pi, e \) are solutions to equations with trig or logs.

From this standpoint, \( \mathbb{R} \) is, in a way we’ll make precise, just what we need to get solutions to *all* equations. In technical terms, if \( f(x) \) is a continuous function, and the equation \( f(x) = 0 \) has a solution, then that solution is a real number. Of course, as with \( x^2 + 1 = 0 \), the solution could be a complex number. But the complex numbers \( z \) are all of the form \( z = x + iy \) where \( x \) and \( y \) are real numbers, and we’re back to real numbers.

So – like the story of the three bears – \( \mathbb{R} \) is just right.
Notes for Chapter 1 Section 5: The Real numbers


For a list of contemporary, alternative constructions of the real numbers, see The Real Numbers – A Survey of Constructions by Ittaty Weiss, Rocky Mountain J. Math. Volume 45, Number 3 (2015), 737-762.

p57 Cauchy’s work on convergence wasn’t motivated by the status of the irrationals. He was trying to give the methods of calculus a logical justification by supplying rigorous proofs for all the results. This type of work is ‘working on the foundations’ of the calculus; the analogy being that without good foundations, buildings collapse. Would mathematics collapse without Cauchy?

After Newton and Leibniz developed calculus, mathematicians worked to extend their ideas and give applications to physics, mechanics and engineering; meaning and proof were low priority (see Grabiner, Judith The Origins of Cauchy’s Rigorous Calculus, The MIT Press 1981). This began to change in the mid-eighteen hundreds when it became apparent that some of these applications were actually wrong. Cauchy was one of several mathematicians working to separate out the true results from the false; we’ll meet some of these later on. So, Yes: without foundations, mathematics could have collapsed.

p59 Two Cauchy sequences in Cantor’s construction are equivalent if they have the same limit. Since Cantor was trying to develop a theory of limits, he used a Cauchy sequence idea: Cauchy sequences \((a_k), (b_k)\) are equivalent if for every \(\epsilon > 0\), there is a number \(N\) such that, whenever \(k \geq N\), then \(|a_k - b_k| < \epsilon\). The collection of all sequences equivalent to \((a_k)\) is denoted \([[(a_k)]\]\, and the real number corresponding to \((a_k)\) is actually the equivalence class \([[(a_k)]\].

p58 For Cauchy, it isn’t enough to show \(|a_3 - a_4| < \epsilon\) etc; you also have to show \(|a_3 - a_5| < \epsilon\) and \(|a_3 - a_6| < \epsilon\), etc. But for some sequences there’s a cheat: if \(|a_n - a_{n+1}| < \frac{C}{10^n}\) for all \(n\), then the sequence is Cauchy. We certainly have that with 3, 3.1, 3.14, 3.145, \ldots in fact it’s easy to show any infinite decimal gives a Cauchy sequence. Say you have \(.8769539\ldots\); make a sequence out of it as we did with \(\pi: (.8, .87, .876, \ldots)\). Then you get \(|a_n - a_{n+1}| < \frac{9}{10^n}\), because decimal digits can’t be greater than 9.
What real numbers are ‘computable?’ To understand ‘computable’, look at our basic example, the computation of $\sqrt{2}$. We start with a guess $g_1$, then set $g_2 = \frac{1}{2} \left( g_1 + \frac{2}{g_1} \right)$. The sidenote shows how to make a computer program for this. We could even say that the computable numbers are the numbers that you can get from computer programs. It’s a practical definition of computing!

This means that the collection of computable numbers will be found in the collection of all computer programs (along with a lot of junk, like ‘hello world’). But the collection of all programs is contained in the collection of all finitely long collections of words. We’ll see the collection of all real numbers is the same as the collection of infinite decimals, which has the same size as the collection of all infinite lists. Cantor showed that the collection of all infinite lists is much larger than the collection of finite lists, so the computable numbers are a smaller collection than the collection of all real numbers.

Put another way, if you randomly picked a number from all the reals, the probability that the number is computable is zero. This argument is part of Cantor’s theory of transfinite numbers. For a popular, non-technical presentation, see Rudy Rucker’s *Infinity and the Mind*, Princeton University Press 2004. A basic technical introduction is in Michael J. Schramm’s *Introduction to Real Analysis*, Dover Publications 2008.

On a more hopeful side, we do know that all the algebraic irrationals are computable, and of course this includes all the irrationals known to Euclid.

We have a method to approximate numbers like $\sqrt{2}$ by decimals: the method gives the successive approximations 1, 1.4, 1.41, 1.414, … If you square these numbers, you get 1, 1.96, 1.9881, 1.999396, …, but there are other ways to approximate $\sqrt{2}$.

There’s an old technique (going back to the Greeks), called continued fractions, that gives a very general way to approximate a large class of irrationals. Here’s the idea: $\sqrt{2}$ satisfies the quadratic $x^2 - 2 = 0$. I can’t factor that, but I can factor $x^2 - 1 = 1$ as $(x - 1)(x + 1) = 1$, hence

$$x - 1 = \frac{1}{1 + x} \quad \text{or} \quad x = 1 + \frac{1}{1 + x}$$

Oops, there are $x$’s on both sides. That’s okay, I already solved for $x$; let’s plug that into the right side:

$$x = 1 + \frac{1}{1 + x} = 1 + \frac{1}{1 + \left( 1 + \frac{1}{1 + x} \right)} = 1 + \frac{1}{2 + \frac{1}{1 + x}}$$
If we continue, we get

\[ x = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \cdots}}} \]

Now let’s compute \( \sqrt{2} \). If we start with \( x = 1 \), we get \( x = \frac{3}{2} \), then \( \frac{7}{5}, \frac{17}{12}, \frac{41}{29}, \ldots \). The squares are 2.25, 1.96, 2.006944..., 1.99888109...

For a more serious overview, see http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/cfINTRO.html
Chapter 1: Numbers

Section 6 Part 1: Accuracy

We saw that every real number is a limit of rational numbers, and therefore a limit of decimals. So π could be represented as the limit of (4, 3.3, 5, 2.9, 2.95, 3.28, ...). We’d rather have it simpler, for example, (3, 3.1, 3.14, ...). That is, we’d like to start by picking out the 3 in π, then get the .1, and so on.

The closest example of something like this was back on p29; we showed all natural numbers could be written with decimals. This followed from the Archimedean Principle:

The Archimedean Principle:
If x is a positive integer, then there is always a power of ten, $10^m$, with $m \geq 0$ and $10^m \leq x < 10^{m+1}$

We showed a similar principle for rationals. Because Cantor showed each real is limit of rationals, we can prove:

The Archimedean Principle for Reals: If $x > 1$ is a real number, then there’s an $m \geq 0$, with $10^m \leq x < 10^{m+1}$.

What we’ll try to do next is use this to find the ‘decimal expansion’ for any real number. We’ll make one up: $r = 632.0141596...$ With $m = 2$, $100 \leq 632.141596... < 1000$. Then $1 \leq \frac{632.0141596...}{100} < 10$, and therefore $r/100$ has a leading decimal digit (that is, one of the numbers {0, 1, 2, ..., 9}), plus a fractional part less than 1.

Now subtract the leading decimal digit ($6 \cdot 100$) from $r$, and repeat the above, eliminating the 3 and the 2, until you’re left with just the fraction ($0.0141596...$) We can also prove an Archimedean principle for fractions:

For positive fractions $y$ less than one, $\frac{1}{y} > 1$, so $10^m \leq \frac{1}{y} < 10^{m+1}$, and flipping the inequalities: there is always a power of ten, $10^m$, with $m \geq 0$ and $\frac{1}{10^{m+1}} < y \leq \frac{1}{10^m}$.

In our case, $m = 2$ and $.01 < .0141596... \leq .1$. Again, divide by .01: then $1 < 1.4159... \leq 10$ and once again we have a leading decimal digit plus a fraction. Subtract the 1, and apply the same procedure to .41596...
What we’ve done is to generate a sequence, \(\{600, 630, 632, 632.01, 632.014, \ldots\}\). After the second term, we get:

\[
0 \leq (632.0141596\ldots) - (632.0) = .0141596\ldots \\
0 \leq (632.0141596\ldots) - (632.01) = .0041596\ldots \\
0 \leq (632.0141596\ldots) - (632.014) = .0001596\ldots
\]

These inequalities are enough to show that 632.0141596\ldots is the limit of \(\{600, 630, 632, 632.01, 632.014, \ldots\}\): all real numbers are limits of very simple finite decimals. We can call this the decimal expansion of real numbers (understanding of course, that .1 and .099999\ldots are different expansions of the same number).

We now want to move on to a different topic: many (most) real numbers are not finite decimals, and to use them, we have to try to understand them in finite terms. This brings us back to the problem of \(\frac{1}{7}\) on p43. We talked about approximations and errors: \(\frac{1}{7}\) isn’t a finite decimal, but a number like 0.14285875 is a good approximation. Good because know the error: \(\|error\| = \frac{1}{7} - 0.14285875\). On p43, we saw \(-0.00000235 < error < 0.00000235\), or, \(\|error\| < 2.35 \times 10^{-6}\)

We’ll re-interpret the list of inequalities at the top of this page as errors. We’ll write the real number \(r = 632.0141596\ldots\), and use

\[a_1 = 632, \; a_2 = 632.01, \; a_3 = 632.014, \; a_4 = 632.0141, \; a_5 = 632.01415,\]

as approximations by finite decimals.

Then

\[
0 \leq |r - a_1| < .0141596\ldots \\
0 \leq |r - a_2| < .0041596\ldots \\
0 \leq |r - a_3| < .0001596\ldots
\]

We wrote these inequalities because they give a special kind of error estimate: they tell how many decimal places of accuracy we have. In particular, \(a_1\) has got the first place to the right of the decimal correctly, \(a_2\) has the second, \(a_3\) the third. So we we get more and more accurate decimal places. More than that: once we get, say, the second decimal place right, it stays right.

This is a new way of thinking about error: we could say that \(\pi\) and its approximation 3.145 agree to three decimal places, and this would tell us what finite decimals we could trust. But this, in use, doesn’t work very well. Here’s why.

Let’s take the exact number \(r = 1\), and compare it to approximations \(a = .999\) and \(b = 1.001\). Since \(r = 1.000\ldots\), \(r\) and \(b\) agree to two decimal places, while \(r\) and \(a\) agree to no decimal places at all.

This suggests we want to rethink ‘number of decimal places’, and
think about errors instead. The error in approximating \( r \) by \( b \) is
\[
| r - b | = .001, \text{ and the error for approximating } r \text{ by } a \text{ is }
| r - a | = .001.
\]

So ‘number of decimal places’ and ‘error’ give different standards for accuracy. It gets even trickier: this time take numbers \( a = 3.1451 \) and \( b = 3.1458 \). Each agrees with \( \pi \) to three decimal places. But if we rounded to three decimal places, we’d be ignoring the fact that \( b \) is much closer to 3.146 than to 3.145. So, approximations to three decimal places are not rounding properly.

What we want is an idea that brings together all three ideas: error, decimal places, and rounding. The intuitive idea is to say that \( r \) and \( a \) agree to \( n \) decimal places if they become equal when you round correctly to \( n \) places. Since the cut-off for rounding up or down is 5, we get:

**Definition: Number of Decimal Places of Accuracy**
Numbers \( a \) and \( b \) agree to \( n \) decimal places if
\[
| a - b | < .5 \times 10^{-n}
\]

Repeating the approximations on the previous page,
\[
0 \leq r - a_1 < .0141596 \ldots < .05 = .5 \times 10^{-1}
\]
\[
0 \leq r - a_2 < .0041596 \ldots < .005 = .5 \times 10^{-2}
\]
\[
0 \leq r - a_3 < .0001596 \ldots < .0005 = .5 \times 10^{-3}
\]

These work out just the way we want: \( r \) and \( a_1 \) agree to one decimal place, \( r \) and \( a_2 \) to two decimal places. And for the example where \( r = 1.000 \ldots, a = .999, b = 1.001, \) both \( | r - b | = .001, \) and \( | r - a | = .001. \) As \( .001 < .005 = .5 \times 10^{-2}, r, a, b \) each agrees with \( r \) to two decimal places.

One important note: the idea ‘round to the same number’ is only an intuition. If we were to take \( a = .1234 \) and \( b = .1236, \) then we would have \( | a - b | = .0002 < .5 \times 10^{-2}, \) so they agree to three decimal places. If you round \( a \) to three decimal places, you get .123, but \( b \) rounds to .124: not the same. The point to this example is that ‘round to the same number’ is the intuition, but the real meaning is given by the definition. It’s what we’ve discussed before: these definitions set standards that everyone can implement in the same way; intuitions don’t have this universality – that everyone gets the same answer.

So, we finally have a standard way to find sequence of rational numbers approximating any real number.
Decimals work well with geometric and analytic intuitions, but can be inconvenient for modern science. Two examples:

i) The absorption of light by a photosynthetic molecule occurs in .000000000000015 seconds. This is also the amount of time that electronic stock market transactions take place.

ii) The amount of CO₂ in the atmosphere is increasing, but temperatures on earth haven’t increased as much. A new study shows that from 1865 to 1997, the worlds oceans have absorbed about 15000000000000000000000 Joules of energy: see Figure 58. (a Joule is about the amount of energy released by dropping a tomato from three feet).

Scientific notation helps deal with zeroes. It counts the number of decimal places to the right of the leading decimal, and puts that into a power of ten. It also requires a single (non-zero) digit to the left of the decimal point. Thus, 100 = 1 × 10²; 21060 = 2.106 × 10⁴; 15000000000000000000000 Joules is 1.5 × 10²³ Joules.

For a fractional numbers like .0231, scientific notation counts the number of decimal places to the right of the non-zero leading decimal, so .0231 = 2.31 × 10⁻² and .000000000000015 seconds is 1.5 × 10⁻¹⁶ seconds. Thinking slightly differently, .0231 = \( \frac{2.31}{10^2} \) = 2.31 × 10⁻². In either case, the effect is still to have only a single (non-zero) digit to the left of the decimal point.

Powers of ten are reference points for daily life, but powers of a thousand are more often used in science, engineering, and technology: we have a meter (10⁰ meters), a millimeter (10⁻³ meters), a micrometer (10⁻⁶ meters), a nanometer (10⁻⁹) meters, etc. 10⁻¹⁵ units is a femto-unit, so the photosynthetic reaction above takes place in 15 femtoseconds. Similarly, 10²¹ units is a zetta-unit, so the amount of heat energy released into the atmosphere is 150 zettajoules.

A system using powers of 10³ or 10⁻³ is called engineering notation. In this notation, 21060 would be written as 21.060 × 10³, while .0231 would be 23.1 × 10⁻³.

This notation (engineering) is often used to describe modern computers. Consumers look at processor speed: ‘wow that’s a 4 ghz processor’. Scientists prefer to measure how many computations per second a computer can do. Additions or multiplications of natural numbers are called integer operations; those can be very fast. But most scientific computations are with decimals; these are called floating point operations. Scientists and engineers need to know how many floating point operations a computer can perform per second; this is referred to as FLOPS (the FLOP measurement is not easy to calcu-
late; it’s often a measure of an ideal, rather than what a computer can consistently do on real computations; see p73).

To look at some examples of computer speeds, we need terminology. A mega- is \(10^6\) (a million); a giga- is \(10^9\) (a billion); a tera- is \(10^{12}\) (a trillion); a peta- is \(10^{15}\) (a quadrillion).

The first commercial supercomputer was the CRAY-1 (Figure 59) introduced to Los Alamos in 1976; it computed at 100 megaflops. In comparison, the first Macintosh from 1984 was about 10,000 times slower. Apple’s fastest computer in 2016, the MacPro, runs at a maximum of 436 gigaflops.

And the 2016 world speed champion was the Chinese Sunway Taihu Light, (Figure 60) running at 93 petaflops, though America, Switzerland, Japan and Saudi Arabia also have (slower) petaflop machines.

Extremely fast computers are used to do aerodynamic computations to reduce air resistance in 18-wheelers; to compute Bitcoins; to forecast the weather a week in advance; to design nuclear weapons; to model the motions of molecules as they engage in chemical reactions (for example, to design more effective drugs or to find genetic components of disease).

And, of course, computer gaming, Figure 61. As the author understands these things, the faster the graphics processor, the better a computer game looks. Major game manufacturers are happy to oblige, and some of the most sophisticated computer design and engineering now goes into game processors. Figure 62 shows Microsoft’s 2017 processor, running an astounding 6 petaflops. Personally, the author wasted an hour trying to get Lara out of the first puzzle, and has permanently sworn off games.

We mentioned floating point operations on a computer. We’ll look at computer numbers in Section 8; most modern computers use a different notation, IEEE 754 normalized floating point notation. A number like 3.145 would be written \(3145 \times 10^{-3}\); the number .0023 is \(23 \times 10^{-4}\).

Each of scientific, engineering, and IEEE754 put the decimal point in a different position. It might be nice if there were only one correct way to write real numbers, but — in real life, it all depends on the application.
Chapter 1: Numbers

Section 6 Part 2: Geometry and Decimals

The word ‘geometry’ comes to us from ancient Greek, geo, earth, and metric, to measure. To the Greeks, geometric lengths like the diagonal of a square clearly existed; you could construct them with ruler and compass. Aristotle drew a sharp distinction between number and magnitude (length, area, etc):

Aristotle views the division between number and magnitude to map neatly onto an exclusive division between the discrete and the continuous. Number is said to be discrete, because its parts do not share any boundary, whereas sharing a boundary is one of Aristotle’s criteria for the continuity of one item with another. Lines are continuous, in contrast, because each point may be regarded as a boundary shared between segments.


As we saw on p50, irrational numbers allowed that distinction to remain up until the time of Newton. With the arithmetisation of number, we can reconnect number and magnitude, by showing that every point on a line corresponds to exactly one real number. See also p73

Intuitively, the connection will be like taking a ruler to measure lengths: rulers associate numbers to positions; this association is called the real line. Figure 63 shows an early ruler; humans have been using the idea for millennia.

To make a ruler, take a line, pick a starting point to correspond to the number zero, and choose some fixed length to represent the natural number 1. Mark off the line in units of 1, moving to the right. By the Peano Axioms, p27, this gives us all the natural numbers, arranged along the right side of zero (we can put the negative numbers on the left side of zero; the ideas will be the same).

To account for fractions, a ruler in English units would start subdividing the space between integers: first, cut the space in half, then each piece is cut in half again, etc. This marks off the ruler in halves, eighths, sixteenths, and so on. Again, as we move right we go from $\frac{1}{8}$ to $\frac{1}{4}$ to $\frac{1}{2}$ and so on; yardsticks are like this. In the metric system, we divide the space between units into ten pieces, then each of those into ten again, etc. This yields a yardstick marked off in centimeters, micrometers, millimeters and so on.
So far what we have are just ruler markings on the line; the markings correspond to finite decimals like $\frac{1}{10} + \frac{3}{10} + \frac{4}{10^2}$; we don’t even have $\frac{1}{7}$ yet. For that we’ll need a little more.

Let’s take $\pi$ as an example. Since it’s a real number, it’s a limit of known decimals: $\pi = 3.1459 \ldots$. To locate this with a ruler, the initial digit tells us it’s between 3 and 4. The second digit tells us it’s between the markings for 3.1 and 3.2, and so on: each decimal digit gives us a range of markings on the ruler, and $\pi$ is between all of those. Does this locate $\pi$ as a single point on the line? Let’s try it.

Imagine there was a second number, $\phi$, sharing the space with $\pi$. Then $\pi, \phi$ would each be in the interval between 3 and 4, so they’d have to be no further apart than the length of that interval:

$$-1 \leq \pi - \phi \leq 1$$

And, as $3.1 \leq \pi, \phi \leq 3.2$, again they share that interval, and can’t be further apart than the length of that interval

$$-1 \leq \pi - \phi \leq 1$$

And, as $3.14 \leq \pi \leq 3.15$, we have

$$-0.01 \leq \pi - \phi \leq 0.01$$

We’ll cut this short: $\pi, \phi$ share the same location, so for every $m$,

$$-\frac{1}{10^m} \leq \pi - \phi \leq \frac{1}{10^m}$$

By the Achimedian Principle for reals, neither $\pi - \phi$ nor $\phi - \pi$ can be a positive number; hence $\pi - \phi = 0$. There’s only one number at these successively accurate ruler markings.

If you run this process in reverse, any point on the line sits between markings on the ruler, and these give the decimal expansion.

This construction lets us use geometric and decimal intuitions together, in a consistent way:

i) As we move right on the line, we go from 2 to 32 to 632, and we’d write $2 < 32 < 632$: we associate rightwards motion with growing larger.

ii) Inequalities like those in i) are easy: we just check 632 has more decimal places to the left of the decimal point than 32.

iii) Decimals work well with our notions of limits to infinity: as we move right along the line, we move towards infinity, and, correspondingly, as we add more zeroes immediately to the left of the decimal points, numbers go to infinity (see p71): 98, 870, 7600, 65000 . . . .
iv) For decimal fractions, moving right goes from .125 to .25 to .5; this is easier to see as moving from .125 to .250 to .500, and again the leading decimal place determines the size.

v) Again, decimals work well with the idea of a limit to zero. As we move left to zero, we add more and more zeroes immediately to the right of the decimal point: .1, .02, .003, … (see p71).

The relation between infinite decimals and straight lines relies on intuitions about geometric lines; for example, that we can subdivide the line into smaller pieces, indefinitely. No physical system is like this: if we had a metal ruler, we’d soon come down to the individual crystals composing the metal, Figure 65. No surprise here; we understand atoms, and the atomic theory is quite ancient (the Indian school of philosophy, Vaieika argued for the existence of atoms, too small to be seen, as early as 500-600 BCE: see https://plato.stanford.edu/entries/early-modern-india/#VaiAto).

So, if you believe in a world where only practical issues are important, then subdivisions and rulers stop here. At this point, then, real numbers also stop, and the supposed correspondence between infinite decimals and lines is just philosophy.

But, what about distances inside atoms? How would we measure that? One answer is that quantum theory limits our ability to simultaneously measure position and velocity; does this mean we really can’t measure very small distances?

Scientists studying black holes and quantum theories of gravity are facing exactly these issues. Some theories suggest that space itself is quantized, rather than being infinitely divisible; that there may be a smallest possible length. That length would be comparable to the Planck length: approximately $1.6 \times 10^{-35}$ meters.

There: we don’t need infinite decimals. Just decimals 36 places long. It’s not clear this is a vast improvement.

For a discussion of quantum gravity and Planck length, see the blog entry The Minimal Length Scale at BackReaction (http://backreaction.blogspot.com/2006/05/).
Notes for Chapter 1 Section 6: Decimals

p69 For some of the difficulties thinking about FLOPS computations, see https://devtalk.nvidia.com/default/topic/745504/comparing-cpu-and-gpu-theoretical-gflops/

p70 Aristotle’s idea was this: we say the line is continuous; this means that if you take a single point away from the line, it falls into two pieces. The German mathematician Richard Dedekind used this idea to construct the reals; we’ll sketch his idea.

To begin, we can’t talk about ‘the line’, since that’s exactly what we’re trying to construct. We can talk about the rationals, Q, either as quotients of integers \( \frac{p}{q} \), or as finite and repeating decimals. For the pieces, we have left and right pieces, \( L \) and \( R \). To be the left and right pieces, all the numbers in \( L \) have to be left of all the numbers in \( R \): if \( l \) is in \( L \), \( r \) is in \( R \), then \( l < r \).

To make sure these really are the pieces, the two together have to make up all of \( Q \). Finally, and this is picky but important, you have to actually take away a point: neither \( L \) nor \( R \) can be all of \( Q \). We’ll call the pair \( (L, R) \) a cut, or just a real number.

So a number like \( \sqrt{2} \) would have \( L \) be, intuitively, something like \( L = \{ \ldots, 1, 1.4, 1.41, 1.414, \ldots \} \) and the right piece would be like \( R = \{ \ldots, 2, 1.5, 1.42, \ldots \} \).

Of course, we’d have to define addition, multiplication, and so on, but there’s a very important definition we’ve left out: What does \( (L_1, R_1) \leq (L_2, R_2) \) mean? An easy way to think is with a picture:

![Diagram](image)

If we think of the real number located where the arrow tips meet, then the top is smaller, because \( L_1 \) is smaller than \( L_2 \) – or you could say because \( R_2 \) is smaller than \( R_1 \).

With that settled, we could talk about limits, but Dedekind was after something different: he wanted to think of \( \sqrt{2} \) as being the next number after \( L = \{ \ldots, 1, 1.4, 1.41, 1.414, \ldots \} \). This ‘next number after’ is called the least upper bound of \( L \). Or, you could just as well use \( R \), and say \( \sqrt{2} \) is the first number before \( R = \{ \ldots, 2, 1.5, 1.42, \ldots \} \) – the greatest lower bound of \( R \). In this way, we can make sense of both ‘the next after’ and ‘the first before’,
For example, if we had a sequence like \( \{1, 1.4, 1.41, 1.414, \ldots \} \), the "next number after' would be \( \sqrt{2} \), and this would be the limit of the sequence. Working along these lines would show that Cantor’s construction and Dedekind’s result in the same real numbers.

Dedekind added the ‘picky but important’ requirement that neither \( L \) nor \( R \) is allowed to be \( \mathbb{Q} \). What if \( L = \mathbb{Q} \)? Then we get a very interesting cut \( \omega = (\mathbb{Q}, \phi) \). But, every \( L_1 \) is smaller than \( \mathbb{Q} \). If we look at our definition of ‘\( \leq \)', every number every \( (L_1, R_1) \) is smaller than \( \omega \). So \( (\mathbb{Q}, \phi) \) is, basically, infinitely large.

Dedekind’s picky condition prevents us from constructing infinitely large numbers like \( \omega \).

More decimal places to the left of the decimal makes numbers larger. A decimal-based definition of limit would go something like this:

**Definition** We say \( x \to \infty \) if \( x \) contains more and more decimal places to the left of the decimal point.

We’ll see later how this matches with standard definitions of \( x \to \infty \).

We can use the same ideas of size in decimal numbers to see how to define \( x \to 0 \); the numbers \( .5, .05, .005 \) get smaller and smaller. So,

**Definition** We say \( x \to 0 \) if \( x < 1 \) and \( x \) contains more and more leading zeros immediately to the right of the decimal point.

Each of these definitions can be made to work because of the Archimedean Property: the powers of ten measure the size of real numbers, and powers of ten are expressed as a 1 followed or preceded by zeroes.
Chapter 1: Numbers

Section 7 Part 1: Units & Standards

Woe to those who give short weight!

- The Holy Quran

This section is about standard units for measurements – meters, degrees Celsius, Joules . . . We start learning the standards of our culture shortly after we’re born: the words, the ways to pronounce them, how to behave towards elders, how to dress and act in public. But also, the names of numbers, letters, characters and kanji.

While many of these are universal – all cultures have words for white and red, for example (see p79) but meaning differs amongst cultures. Americans associate white with purity and red with danger, while Chinese associate red with good fortune (Figure 66) and white with death. Researchers now believe that perception of basic colors is built into our brains, but that culture tells us what to do with those colors.

In the same way, both our society and the way we interact with the world tell us what to do with numbers. The Nootka, a fishing culture on Vancouver Island, use month names like Eneecoresamilth, salmon fishing moon (see p79). The name reflects how Nootka culture interacts with the world.

Religions tell a different part of the story. Chinese New Year, Passover, Easter, Ramadan and Holi are all lunar celebrations: all are associated with the appearance of a full or new moon, and all occur on a different days or months as the years change. Even though the date of these holidays changes from year to year, lunar holidays are a sophisticated solution to scheduling public holidays before instantaneous communications. They work because everyone can look and see the moon (see p79).

So, standards have to be accessible: Figure 67 shows British standards for lengths, set in a public square outside Greenwich Observatory. Along these lines, a sixteenth-century German town defined the ‘foot’ as follows:

Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished; then make them put their left feet one behind the other, and the length thus obtained shall be a right and lawful rood to measure and survey the land with, and the 16th part of it shall be the right and lawful foot (see p790).
Polynesian men working together to make a boat measure off distances using the length of their finger joints. In pre-revolutionary France, over 700 local French units existed, with some 250,000 different measures (see p80).

Each of these ways of measuring works in a context: in surveying one town, a local measure of the foot is simple and useable. Boats, held together with lashed reeds, are flexible enough to allow small variations in lengths. We saw this in Babylonian astronomy, p24: to chart the position of the planets, astronomers used an easily accessible reference, the first appearance of a planet on the horizon. Again, in pre-revolutionary France, plots of land might be measured by their productivity or the difficulty in working the land. But these kinds of standards work less well when ships are made of metal plates. They also don’t travel very well: as French villages became connected into an empire, traders needed long tables of equivalents between measures, and the empire had difficulty assessing taxes.

But enforcing standard measure across an empire is difficult: Figure 68 shows the Babylonian Sun God Shamash, holding a standard measuring rod and a coiled rope, both used in surveying land. In the Hebrew Holy Bible, Leviticus 19: 36, we read:

‘You are to have honest balances, honest weights, an honest dry measure, and an honest liquid measure; I am Yahweh your God, who brought you out of the land of Egypt.’

The suggestion in these quotations, and in the header to this Section, is that standards are a basic form of honesty necessary for public order. What better way to enforce standards than to connect them with the law of God?

Even today, standards are a serious issue. In a hospital, Clinical error and negligence are responsible for disabling injuries in about 1 in 25 hospital admissions. Most of these injuries are caused by adverse drug events [...] Converting among ratios, percentages, international units, mols, micrograms, and milligrams causes substantial difficulty (see p79).

Here’s an example anyone might come across (see Figure 69): the illegal psychoactive drug MDMA. It has an average half-life of 8 hours in the body, and drug tests can detect it in blood concentrations as low as 500 nanograms per milliliter. An average amount of blood is 5.9 liters, and average dose is 120 milligrams. How long before MDMA is undetectable?

You could easily spend ten minutes converting these different measures into only one standard unit.
Modern scientific units are based on the metric system, first introduced in post-revolutionary France in the late 1700’s. The standardization of measures was taken up by the (formerly Royal) Academy of Science; revolutionary principles and Enlightenment ideals suggested standards should be based on nature itself, and would then be universal. Eventually, four major principles were set:

i) The unit of length should be based on a fraction of the circumference of the earth – measured on a circumference that passed through France (of course).

ii) The standards should be linked together into one coherent system: for example, length is measured in meters, so the units for area have to be square meters.

iii) Numbers should be expressed in base ten.

iv) Fractions and multiples of basic units should be named systematically, using Greek prefixes: milli-meter, mega-watt, etc. These are to denote units in ranks of $10, 10^2, \ldots$.

Little was natural or universal about these principles (see p80); many scientists argued for using a pendulum to measure lengths: the time for a pendulum to go a full cycle is

$$ t = \sqrt{\frac{L}{g}} $$

Here $g$ is the acceleration of gravity, and $L$ is the length of the pendulum. Now set the meter to that length which lets the pendulum go through a full cycle in exactly one second. Unfortunately, $g$ varies with one’s position on the earth, though this wasn’t well understood at the time (see Figure 70).

Other objections were that a base eight system (instead of a decimal system) would allow shopkeepers to easily compute half, then half again, and half again. A base twelve allows halves, thirds, sixths. And, why Greek names? (“These names are novel and unintelligible to the large majority of our citizens, are not necessary for the maintenance of the Republic.”)

All of these early standards were based on measures of ‘natural’ objects; the kilogram was based on the weight of water at a given temperature, and this standard was converted into an equivalent weight of a platinum-iridium bar kept in a bank vault in Paris. What happens next is told in Rachel Courtland’s article *The Kilogram, Reinvented* (see p80):

*Once every few decades, a scientist plucks the cylinder from its perch with chamois-leather-padded pincers, rubs its surface with a cloth soaked in*
alcohol and ether, and steam-cleans it. Then he puts the prototype in a precise balance that compares it to the bureau’s official copies, which are in turn compared to copies kept by member countries. And thus the prototype mass trickles down to set the standard for the rest of the world.

The system has been far from seamless. When the cylinder was last removed from the vault in 1988, the bureau’s metrologists were disappointed to discover that its mass and those of its official copies had drifted apart by as much as 70 micrograms since 1889.

The metric system was adopted on 23 September 1801 (or, speaking of standards, the Revolutionary government gave the date as 1 vendémiaire an X); many businesses covertly kept the old measures. Government documents, as well as legal, military and engineering documents were required to be submitted in metric units, though for some time dual systems were used in the government. And, of course, public standards needed to be publicly available; the Agency of Weights and Measures printed seventy thousand conversion tables. Mass production of meter sticks was more difficult, and eventually the Committee for Public Safety turned the task over to the Atelier de perfectionnement – an armory specializing in mass production of interchangeable parts for rifles (see p80). In France, use of the old units only died out in the early 1900’s.

Internationally, the United States still uses British units. In 1999, a joint American-British space probe crashed into the surface of Mars because flight controllers had no real idea where the probe was. Software designers in Britain used meters; those in the United States used feet, and the ‘orbiter’ entered too deeply into the Martian atmosphere and broke up. The NASA report (see p80) noted that metric units were specified in the contract; that NASA had a history of reusing old, undocumented code, and that two navigators had reported problems but were ignored.

Standards alone accomplish little; only people can make them work.
Notes for Chapter 1 Section 7 Part 1: Units & Standards

The universality of certain colors, called focal colors, contradicts theories of cultural and physical relativism. Brent Berlin and Paul Kay looked at 20 oral world languages, and

"Berlin and Kay found out that people focus certain points in the color continuum as a kind of orientation. Such reference points or 'best examples' were called 'foci'. Focal colors had not only been detected in English but also in the remaining 19 languages (Berlin & Kay, Basic Color Terms: Their Universality and Evolution, Berkeley: University of California Press 1969).

Eleanor Rosch went further, using experiments to determine the physical and psychological role of focal colors:

Rosch was able to find out that focal colors are more perceptually salient than non-focal ones (Rosch, Eleanor (1973), Natural Categories, Cognitive Psychology 4: 328-350). This cognitive salience is probably not anchored in language but reflects certain physiological aspects of [...] perspective mechanisms. Later, she coined the term 'prototype' instead of 'focal'.

Quotations are from http://www.glottopedia.org/index.php/Focal_Colors

Nootka numbers from William J. Folan

Even in a lunar calendar, deciding the exact day the moon is full or new has to be standardized. In Islam, this is the duty of the Imam. In a small village in Africa, it would be the duty of the Chief Priest. Chinua Achebe, the Nobel prize-winning author, describing the process in his novel Arrow of God:

'The moon he saw that day was as thin as an orphan fed grudgingly by a cruel foster-mother. He peered more closely to make sure he was not deceived by a feather of cloud.'

The quotation defining the foot is from Jacob Koebel, Geometrei. Von künstlichem Feldmessen und absehen.


Part of the issue was that powerful political forces were at work. The Ministry of Finance wanted to compute taxes for the entire country; scientists at the Academy received large grants to perform the difficult and inaccurate job of measuring the circumference of the earth. But, in the end . . . opposition was so intense that the Emperor Bonaparte rescinded the system, and it was delayed for decades. The metric day, the metric week (a ten day week, with nine work days and one day of rest) and the metric year never caught on.


Quotation from Rachel Courtland’s article is from http://spectrum.ieee.org/consumer-electronics/standards/the-kilogram-reinvented


Chapter 1: Numbers

Section 7 Part 2: Orders of Magnitude

I first heard about orders of magnitude in 1968, at a public astronomy seminar, Cornell University. In 1967, the radio astronomer Jocelyn Bell had detected a signal with a peak repeating every 1.33 seconds (see Figure 72 and p88). The signal followed the rotation of the earth, so she and her advisor Anthony Hewish concluded it was extraterrestrial; they called the signal LGM-1 (for ‘Little Green Men’: a signal so regular could be from an extraterrestrial civilisation). The signals were later understood to come from stars, and the stars were named pulsars.

When Frank Drake returned from the Aricebo radio telescope in Puerto Rico (see Figure 82 and p88) with data suggesting a new pulsar in a different part of the sky, Cornell opened a seminar to discuss what might cause these signals.

Memory fades after fifty years. I recall Hans Bethe was there, the Nobel laureate who explained the nuclear reactions that produce the energy of stars. Drake opened, discussing how the telescope needed to be modified to receive these signals; he had a slide with a graduate student doing the updates, hanging upside down 150m above the valley floor. Edwin Salpeter was there (he’d explained the production of carbon in stars). He discussed the energy required to generate the signal; for the Crab pulsar (Figure 73) the energy is order of magnitude $10^{32}$ joules/year. Salpeter explained for the non-physicists that a joule was about the order of magnitude of the energy released by a tomato falling one foot onto a floor.

Carl Sagan stood and remarked the signal was about five orders of magnitude greater than the entire energy output of the sun, and therefore, if the signal came from an intelligent civilization, ‘it was a very stupid intelligent civilization’ – there are many ways of sending signals requiring orders of magnitude less energy.

Finally, Tommy Gold (who had developed the theory of the magnetosphere surrounding the earth) stood and proposed that pulsars were highly magnetized rotating neutron stars. As the star rotated, its magnetic field would accelerate particles to near the speed of light, and those would then emit synchrotron radiation like the blue light seen in the Crab nebula, Figure 73. Gold’s explanation is the accepted explanation today (see p89).
It was amazing to see these brilliant scientists doing science, thinking about results, doing computations off the top of their heads. It was clear I had a lot to learn, and not just about orders of magnitude.

The powers of ten \((10^0, 10 = 10^1, 100 = 10^2 \ldots)\) form standards for the way we think about numbers – we think of sandwiches as costing about ten dollars, tickets to a theme park about a hundred (see the research earlier referenced as cognitive reference points, p28 and the references there). The idea of orders of magnitude make this precise.

What we want to do, roughly, is say the pair 4, 7 are the same size, as are the pair 60, 80. In scientific notation, we’d have \(4 \times 10^0, 7 \times 10^0\) versus \(6 \times 10^1, 8 \times 10^1\). The different exponents tell us the two pairs are different orders of magnitude.

But what about 9.9999999 versus 10.00000001? It’s misleading to say one is substantially larger/smaller than the other. Should I be rounding? Where’s the cut off from one order of magnitude to the next? For example, if 9 should belong with 11, what about 8? Maybe 5 is the cut off, since it’s half-way?

Once again, it’s about setting standards: we have to establish an unambiguous system. One way to do this is think about what we said, "different exponents tell us the two pairs are different orders of magnitude”. Exponents can always be accessed as logarithms of the numbers; more precisely, if \(x = 4.768 \times 10^8\), \(\log_{10}(x) = 8.5761 \ldots\), and it’s the 8 we want to pick out (this is called the characteristic of \(x\)). But even this is a bit tricky. \(50 = 5 \times 10^1\) and the log is 1.6989 \ldots, so we’d give it order of magnitude 1. On the other hand, \(.05 = 5 \times 10^{-2}\) but the log is \(-1.3010299957\), so do we give it order of magnitude \(-1\) or \(-2\)?

The standard we set is this (see p89):

**ORDER OF MAGNITUDE:** To find the order of magnitude of a number \(x\), take \(\log_{10} x\), and correctly round it.

For example, \(x = 2.78\) has \(\log_{10} x = 0.4440447959\). Since we have 0.4440447959 < .5, this is rounded down to zero and 2.78 has order of magnitude zero. For \(x = 3.78\), \(\log_{10} x = 0.5774917998 > .5\), so this number is rounded up to one, and 3.78 has order of magnitude one.

Most people can’t compute logs in their head. Without logs, the cut off between order of magnitude \(n\) and \(n + 1\) is \(10^n + ^{\frac{1}{2}} = 10^n \cdot 10^{\frac{1}{2}} = 3.1622776602 \cdot 10^n\). Then \(2.78 < 3.1622776602 < 3.78\), though likely we’d use 3.16 instead of 3.1622776602.
Returning to Carl Sagan and pulsars, a large order of magnitude was used to show something was unlikely or impossible. The next example is from biology: the mode of action of drugs on cells.

This has an antique sound; in fact it’s the title of a book by the British physiologist/mathematician A.V. Hill (Nobel Prize 1922), who started his work in 1909. Some drugs were known to work specifically on certain diseases – for example, quinine on malaria. These were called ‘specifics’, but how they worked was a mystery. In 1797 Caddel and Davies wrote *Medical, Philosophical and Vulgar Errors of Various Kinds Considered and Refuted*:

"...supposing an admiral sent down channel, across the Bay of Biscay, and up to the Mediterranean, with express orders to attack the Maltese, but with the strictest charge not to molest any other state whatever; I cannot conceive any medicine such a specific as to conform most punctually with such orders, to act vigorously against one particular gland or humour of the body, without in the least affecting or disturbing any other."

One theory was that a given drug coated the surface of the cells affected by the drug. Hill worked at the beginning of quantitative physiology and pharmacology; he measured the amount of drug necessary to produce a response, and calculated it was several orders of magnitude too small to coat a cell. Drugs must be doing something else.

We can imaging the drug getting inside a cell and doing its work there, but this again is orders of magnitude off. The cell membrane is composed of phospholipid molecules; Figure 74. But,

"...hormones, being mostly hydrophilic (or lipophobic) substances, are unable to pass through membranes, so that their influence must somehow be exerted from outside. The membranes of cells, although very thin (3 to 6 nm) are effectively impermeable to ions and polar molecules. Although $K^+$ ions might achieve diffusional equilibrium over this distance in water in about 5 ms, they would take some 12 days (280 hrs) to equilibrate across a phospholipid bilayer [...] Likewise, the permeability of membranes to polar molecules is low."

Here the order of magnitude difference is $280$ hours, or $10^6$ seconds, verses $10^{-3}$ seconds. This vast gap tells us drugs cannot work that way. The receptor theory of drug interaction explains these discrepancies. The drug binds to a molecule called a receptor that spans the cell membrane. The chemical interaction on the outside of the cell changes the receptor’s shape on the interior, where it interacts with other molecules to have an effect on the cell. Cells that don’t produce the particular receptor are not affected by the drug, hence the specificity.

This next example is about precision: here, order of magnitude will tell us how accurately we need to be in building machines, in this case, large modern airplanes. The spar of an airplane wing is a support beam; see Figure 75. Many different kinds of structures attach directly onto it – for example, the airplane engines and the ailerons that control turning. One consequence is that the spar has to be a precise size, so that each piece will fit in exactly the right place. For a modern 27 meter long airplane wing, the spar has to be constructed to .3mm accuracy, a difference of five orders of magnitude. Such extreme accuracy is needed because of the strong forces acting on the wing during flight; differences between the wings could cause instability, or help cracks form and spread, causing the loss of the wing during flight.

Our second example comes from square-cube ideas. It’s inspired by recent work on environmental remediation. At the height of the space race in the 1960’s launch complexes were used to clean and degrease rocket engines; the chemical used was trichloroethylene (TCE), which is now known to be toxic and carcinogenic. It seeps into the ground and, over long periods of time, can contaminate groundwater. See Figures 76 and 77.

To de-contaminate the soil, metal particles are injected into the ground; these combine with TCE to form new compounds that are no longer toxic. The question is how the metals should be delivered. If we think of these as small spheres, the TCE will act on the surface of the sphere; for maximum efficiency, the surface area should be as large as possible. We’ll need a couple of formulas: A sphere of radius $r$ has surface area $S = 4\pi r^2$ and volume $V = \frac{4}{3}\pi r^3$. Now let’s compare two sizes of metal balls.

In the first, our spheres are about the size of a cell: the radius is $100\mu$m or $10^{-4}$m. This gives it a volume of $\frac{4}{3}\pi10^{-12}$m, and a surface area of $4\pi10^{-8}$m.

Now, instead of using cell-sized metal balls, let’s use virus-sized par-
articles. Their radius is 100nm or 10^{-7}m, with a volume of \( \frac{4}{3}\pi 10^{-21}m \), and a surface area of \( 4\pi 10^{-14}m \). This is worse – smaller surface area – but, since we’re using smaller metal balls, we’ll have more of them. How many more? Divide the volume of the large balls by that of the small. Cancelling the \( \pi \) and the \( \frac{4}{3} \), it’s \( 10^{-12} \) divided by \( 10^{-21} \), or, \( 10^9 \).

Now I have \( 10^9 \) particles, each with surface area \( 4\pi 10^{-14}m \), so my new surface area is \( 4\pi 10^{-5}m \). The single larger ball had a surface area of only \( 4\pi 10^{-8}m \), deactivation of TCE by three orders of magnitude. It’s *nano-particle* remediation.

Our last example is digestion: we’ll follow a piece of potato as it’s converted into sugars and fats in the body (see p89 for references). We’ll see a range of eight orders of magnitude, from organs like the stomach, that we can easily see, to molecules we can only understand with complex scientific instruments. The lesson here is about the slow progress of medicine; what we now know required hundreds of years of work, and hundreds of years of development of scientific instruments.

We’ll be taking a journey from organs we can see with our eyes and feel with our hands, to smaller and smaller structures, all the way down to single molecules. We can no longer see or manipulate these tiny structures; the closer we zoom in on these micro-structures, the more exact understanding we get on how food is processed.

Digestion actually starts in the mouth, goes on to the esophagus, stomach and small intestine. Some of this is mechanical: food is broken into small pieces. Size does matter here: a variety of enzymes (for example proteases and lipases) are responsible for breaking down proteins and fats. These enzymes attach to the surface of the food, so the larger the surface area presented to them, the faster and more thoroughly they work.

If we look at a piece of food – say a sphere 1cm in radius – the surface area \( S \) is about 12.6 cm\(^2\), order of magnitude 1. If the food is broken down to a thousand smaller spheres, \( S \) is about 270 cm\(^2\), order of magnitude 3. Chewing processes food to be digested three orders of magnitude more efficiently. Snakes and alligators gulp, and it can take days for a snake to digest a mouse.

Our small particles enter the intestine next. An average small intestine is about 6m long – say order of magnitude two. The intestine is lined with *circular folds* shown in Figure 78. Each fold is roughly 8mm or \( 8 \times 10^{-2}m \) high: order of magnitude \(-2\), a jump of five orders of magnitude.

Each circular fold is has a surface coated with thousands of tubes called *villi*, about 1.5mm or \( 1.5 \times 10^{-3}m \) high (see Figure 79); we’ve
jumped down one more order of magnitude. The villi increase the area that can absorb nutrients by a factor of thirty: one order of magnitude.

The villi themselves are coated with a brush border, made up of further wormy tubes called microvilli (Figure 79 again), 1µm or 10^{-6} m high, they serve much the same purpose as the villi. For a decrease of three orders of magnitude in size, they increase the surface are by a factor of 600, or three orders of magnitude.

And there it ends – almost. The folds and villi and microvilli can grab food from the flow through the intestines, but fats still need to be broken down by lipases, proteins by proteases; complex carbohydrates need to be hydrolyzed to monosaccharides. Finally, all these molecules have to get to the bloodstream.

All this occurs in a thin layer covering the microvilli called the glyco-calyx. This is made up of actin filaments, a protein that usually forms the contractile filaments of muscle cells, but in this case forms a protein layer that can contract to keep fluid moving. The actin molecules themselves are about 5 nm or 5 × 10^{-9} m, another two orders of magnitude smaller. The brush border holds digestive enzymes, and the resulting sugars, etc, are transferred from the interior of the intestine to the bloodstream by the cells comprising the villi, shown in Figure 80.

The cells have transport molecules located in their membrane; one transport might carry selected sugars through the cell wall: glucose, for example. This has a molecular diameter (see p.90) of 9 Angstroms or 9 × 10^{-10} m, another order of magnitude. These are small enough to get from the microvilli to the blood through cell pores (actual gaps in the cell wall) about 500 to 800 angstroms or 5 × 10^{-8} m to 8 × 10^{-8} m – certainly large enough to let the glucose flow freely into the blood.

Here’s an example from a neuroscientist:

"Single large neurons have physical dimensions observable at low optical magnification, that of a tenth of a millimeter. That is big enough to be dissected by hand with pins, using a good magnifying glass. Moving just two orders of magnitude down to the micrometer level, which requires a good microscope, one is at the scale of synaptic transmission. One may observe synapses at the union between nerve and muscle, for example. Two orders of magnitude further down, at tens of nanometers, with the aid of electron microscopy, we find the realm of single ion channels and of signal transduction and molecular biology. (Rodolfo R. Llinas I of the Vortex: From Neurons to Self. The MIT Press, 2001.)"
As in biology, some of the largest orders of magnitude in physics arise in dealing with the very small. We’ll look at one example, the discovery of the Higgs boson in 2013. To explain the what and the why, we’ll rely on an analogy by the theoretical physicist Brian Greene (How the Higgs Boson Was Found, Smithsonian Magazine July 2013).

Greene asks us to think about a fish who happens to be a theoretical physicist. The fish would notice that it’s very difficult to push objects. She’d grow up learning about the strange properties of motion. Then one day, she has an amazing insight: the entire fish universe is filled with an invisible quality that resists motion. She calls it the ‘water field’. The water field is generated by elementary particles of a new kind of matter, called ‘water’. It’s very very hard to observe water, but other fish physicists realize that if they could arrange a truly enormous splash, perhaps a particle of water would break off.

We, in our universe, are in a similar position. We notice that it’s hard to start things in motion. Once they are in motion, they have energy. But some objects in motion seem to have more energy that others. We invent a quality called ‘mass’, and with this we can explain the energy of motion: \( E = \frac{1}{2}mv^2 \), where \( v \) is the velocity of the object and \( m \) is its ‘mass’.

The insight of Higgs, Englert (1964) and others was that perhaps mass is a consequence of an invisible field pervading the entire universe. This idea happened to fit in well with current theories of elementary particles, and explained other phenomena. The particle associated with the field would be a new form of matter, which came to be called the Higgs boson, but proving its existence was very hard. It might appear in collisions of particles accelerated to 99.99999% of the speed of light. If so, the particle would only exist for a billion of a trillionth of a second: \( 10^{-15} \) sec. To see it, scientists analyzed 800 trillion \((8 \times 10^{11})\) collisions, and found collisions that had only 3 chances in a million \((3 \times 10^{-6})\) of being due to something else. They announced the result in 2013 – almost fifty years after the initial idea\((4.9 \times 10^3)\).

Notes for Chapter 1 Section 7 Part 2: Order of Magnitude

p81 The paper by Hewish, Bell et. al. is in Nature, 217(1968) p709. The authors remark:

The remarkable nature of these signals at first suggested an origin in terms of man-made transmissions which might arise from deep-space probes, planetary radar, or the reflection of terrestrial signals from the moon. None of these interpretations can, however, be accepted because the absence of any parallax shows that the source lies far outside the solar system. [...] A tentative explanation of these unusual sources in terms of the stable oscillations of white dwarf or neutron stars is proposed.

p81 The Aricebo telescope consists of a reflecting mesh hung above shade-tolerant vegetation near the town of Aricebo, Puerto Rico (also known as La Villa del Capitan Correa); see Figure 82. The mesh is about 1000m in diameter, and reflects radio waves to a detector about 150m above the dish. Construction of the telescope was an enormous engineering and political undertaking.

Aricebo was conceived by DARPA (the Defense Advanced Research Projects Agency) as a means to detect incoming guided missiles: high speed objects entering the upper atmosphere ionize it, and this can be detected on radar. As the upper atmosphere was poorly understood, the telescope was designed to carry out atmospheric research; W.E. Gordon of Cornell Engineering wrote in his proposal:

The discovery that free electrons in the Earth’s ionosphere incoherently scatter signals that are weak but detectable [...] makes possible the exploration of the upper atmosphere [...] that the radar components [...] are all within the state of the art means that the exploration can begin as soon as the radar is assembled. From "Design study of a radar to explore the earth’s ionosphere and surrounding space" by W. E. Gordon, H. G. Booker, & B. Nichols, Cornell.

Ward Low at DARPA realized the telescope could also be used to study the behavior of radio waves, and to intercept Soviet communications. The proposed atmospheric telescope would be built as a general-purpose radio telescope. The perfect location would have to be in the tropics, so that all planets passed over the telescope; a natural valley would save on construction costs ... and Braulio Dueño from the University of Puerto Rico at Mayaguez was studying at Cornell. Aricebo was chosen, and the telescope opened in 1964. The first big scientific result was the discovery that the rotational period of Mercury was 59 days, not 88 days as previously thought. For military intelligence, it detected Soviet radar waves reflected off the moon.
and gave the location of the radar station. For the history, see Daniel R. Altschuler, *The National Astronomy and Ionosphere Center’s (NAIC) Arecibo Observatory in Puerto Rico*, http://www.astro.wisc.edu/ stansimi/Students/daltschuler_2.pdf


‘At a conference in London in 1951 I had argued that dense, collapsed stars would be ideally suited to emit strong radio signals, since their magnetic fields may be enormously strengthened by the collapse and extend out into low density space.

‘The pulsars now seemed to represent just the stellar objects I had discussed then. Calculations existed for the collapsed “neutron stars” that indicated approximately their size, as small as a few kilometers, and their mass, on the order of a solar mass .4 Astronomers generally thought that even if they existed, they could never be discovered. However hot, a star so small could not be seen at astronomical distances. But they had not considered the energy concentration resulting from the collapse: enormous magnetic field strengths and a spin energy quite comparable with the entire content of nuclear energy of the star before its collapse. With these considerations it was not unreasonable to expect them to be observable.

‘I had another clue to the nature of the new objects: While there was some irregularity in the pulse-to-pulse timing, the long-term accuracy of these “clocks” was enormously better than a statistical addition of the pulse irregularities would have allowed.

‘I proposed the model of a rapidly spinning neutron star, which, as a result of some asymmetries, sent out a strong beam of radiation from one region of longitude. This beam would sweep over the Earth once in each rotation period and would be seen as a short pulse. The underlying long-term accuracy would then be produced by the spin of the object, while short-term timing fluctuations would just reflect details of the emitting mechanism. I compared it to the rotating beacon of a lighthouse, whose lamp, hanging from the rotating shaft, could wobble a little; the long-term accuracy would still be that of the rotation of the shaft.’


p85 This discussion of digestion follows Thomas Fischer Weiss’ *Cellular Biophysics: Transport*, MIT Press 1996.
A single molecule is an object in constant motion, spinning, contracting, expanding, etc; it doesn’t have any single size. Instead, quantum mechanics assigns a probability that the molecule has a given size. The idea of ‘molecular diameter’ is a substitute. There are many ways to do the computation; see for example ‘kinetic diameter’ at https://en.wikipedia.org/wiki/Kinetic_diameter; atomic radius at https://en.wikipedia.org/wiki/Atomic_radius; hard spheres models, http://www.kayelaby.npl.co.uk/general_physics/2_2/2_2_4.html, and more.
Chapter 1: Numbers

Section 8: Computer Numbers

This section is about how computers handle numbers. For many people, this is right up there with asking how variable-speed transmissions work in cars: they do their thing, and who cares how? This kind of complacency can lead to trouble. We’ll start at (almost) the beginning: the word ‘computer’ was first used in the 1600’s; it meant a person who does computations professionally. The astronomer Kepler started as a computer for the astronomer Tycho Brahe. By the late 1800’s middle-class women wanted work as calculators; it fit in with social standards of what work a married woman could do, especially if they were not allowed to leave home. See Figure 84.

World War One involved powerful artillery; computing the path of a shell launched at a single angle required about 750 multiplications and about 7 woman-hours of work. The computations had to be done for many angles, and by the end of the second world war an analogue machine reduced the time to 20 minutes. Herman Goldstine and John Mauchley received funding from the US Army to build a digital version, the ENIAC.

The women who performed calculations knew that repeated calculations could lead to errors. Goldstine and Mauchley understood these errors would arise on electronic devices, but without the human understanding the women provided, the machine would not be aware there was a problem. And this lack of intuition and feeling for the numbers can be responsible for all kinds of errors, especially for those who assume computers do their thing and who cares how.

An early error happened on Intel’s Pentium P5 chip, Figure 85. Thomas Nicely discovered that certain inputs of $x$ and $y$ would cause an error in the computation

$$x - \left(\frac{x}{y}\right) \cdot y = 0$$

As an example, for $x = 4195835$ and $y = 3145727$, the chip returned 256 instead of 0 – leading to Figure 86, the autographed chip. We don’t think about a computer making errors like this – after all, it’s just a bunch of wires, which would be either always wrong or always right. That’s why we use computers. In reality, modern computer
chips are very complex circuits, and they have to be programmed to do computations. As the programs are written by humans ... there can be mistakes; see page 103.

Even though errors like this are inevitable (in 2018 the problem was SPECTRE and MELTDOWN bugs, with again caused by errors in programming), these aren’t the main issue. Since 1994, computer and chip manufacturers have systems in place to recall and replace defective units (Intel started this in 1994).

But there’s another level of problem – defective software. It’s estimated that the cost of dealing with bugs ran to over one trillion US$ in 2016 alone. Engineers and scientists typically do not have the same kinds of losses, but there can be losses to reputation for scientists, or loss of life in engineering projects. For example, the scientists C.L.Reyes and G. Chang retracted their paper Structure of the ABC transporter MsbA in complex with ADP:vanadate and lipopolysaccharide, Science, May 13, 2005. The structure was analyzed with a computer program, and an error in ± led to a structure where some portions of the molecule were backwards; see Figure 87. The structure in question was a transporter molecule: it allows harmful bacteria to defend themselves against drug treatment, by transporting the drugs out through the bacterium cell wall. Understanding the mechanism used by ABC transporters might lead to techniques to disable it and eliminate resistance to antibiotics; see page 103.

What went wrong? Structure of MsbA from E. coli: A Homolog of the Multidrug Resistance ATP Binding Cassette(ABC) Transporters Science 7 Sept 2001 vol 293, discusses how the structure of this molecule was determined:

Iterative eightfold noncrystallographic symmetry averaging, solvent flattening/flipping, phase extension, and amplitude sharpening using in-house programs yielded electron density maps of excellent quality for tracing a polypeptide chain.

This involves intensive computer manipulation of the data, and, as it happened, one of the programs they used had an error.

In lab science, errors in procedure are to be expected. A scientist will spend years or decades learning correct lab procedure; what is different here is that scientists typically do not have decades of experience in computer programming. Here, a program was imported from another lab, but, Chang remarks, “you just trust the code to do the right job”. He reports that he now triple checks everything (see Zeeya Merali, Nature 14 October 2010 vol 467, p 775).
Figure 88 gives another example where defective software destroyed (literally) a scientific mission: the $245 million Mars Climate Orbiter. Launched in 1998 to study the climate and atmosphere of Mars, the mission plan was to use onboard thruster rockets to position the spacecraft to enter Mars orbit. Instead, the thrusters sent the craft deep into the Martian atmosphere, where it burned up before crashing.

Here, the problem is more complex: a mission such as this involves thousands of individuals, across multiple corporations and government institutions. On the other hand, the same is true for many modern engineering projects: the construction of highways, skyscrapers, sports arenas, or air-traffic control systems, and even smartphones. Engineering firms taking on these kinds of tasks need to have quality control systems in place. Compare the discussion on p46.

For the Mars Orbiter, two studies of the failure were made: the first, NASA’s internal study; a second in *Why the Mars Probe went off course*, IEEE Spectrum 1 Dec 1999. For details of the reports, see p103.

In these two examples, training and management systems can help reduce errors and losses. But there are mathematical issues.

One of the core issues was inconsistent use of units for the thrust of the orbiter’s rockets (one subsystem used meters, another used feet). The actual numbers generated differed by only a small amount. But two important points emerged:

i) A small initial error, over a million mile trajectory, can result in a large final error. Compared to the million mile trajectory, the final path was off by only a few kilometers – a tiny percentage, but large enough to destroy the orbiter.

ii) Flight engineers made many course corrections; this meant that the small error was made many times. A small error repeated many times can result in a large final error.

We’ll see these two issues again.

There’s a third major issue in using computers: operator error, illustrated in Figure 90. This is an example of a CAC or a computer aided catastrophe. As shown in Figure 89, the oil platform was intended to rest on the sea floor, meaning that the long beams supporting the platform had to carry not only the weight of the platform, but also deal with water pressure and storms. As the supports were lowered to the sea floor, one cracked, causing immediate flooding, dragging the platform deeper, where buoyancy tanks imploded under pressure; see https://wikivisually.com/wiki/Sleipner_A.

The problem was traced to the design of the supports. Large com-
plex structures acted on by many forces are typically designed using computer tools. The standard mathematical technique is called finite element analysis; in this technique, a complex structure is analyzed as smaller parts exerting forces on each other, and, in this case, responding to external forces such as water pressure, wind, waves and gravity. It’s then possible to see how a force on one part of the structure flows through the other parts. NASA developed a program, NASTRAN, used in modeling forces on multi-stage rockets, which is famously successful. In the Slepnir CAC, the designer did not fully understand how to implement the elements of the structure, which resulted in a design that underestimated forces by almost 50%.

The Slepnir designer is not alone. Ivo Babuska, a professor of civil engineering at the University of Texas at Austin, tried a small experiment: he sent a problem to a group of engineers. The problem is the Girkman problem, illustrated by the structure in Figure 91: how are the massive concrete walls to be supported under their own weight and wind forces?

Babuska specified that the engineers had to use professional-level computer software to solve the problem; of the 15 licensed professional engineers who submitted a solution, half were wrong. "They gave us numbers which were completely wrong and they believed in them" said Babuska. He commented, "How is it possible that this happened is a good question. There could be many various reasons. Nevertheless, in this case the reason was only one. Some of the analysts did not have sufficient engineering intuition and mathematical and engineering knowledge and possibly used the software incorrectly."

Operator error is a difficult problem to overcome, as it depends not on project management, or program design, but on the integrity of users. Babuska phrases the issue as 'signing the blueprints.' Only humans can sign blueprints and take financial and legal responsibility for structures.

All of the above examples of computer fail are really examples of human failure. Is the myth really true, that computers don’t make errors? As we said at the beginning: repeated computations can lead to errors on paper or on a computer; these kind of errors are a consequence of how numbers are written down or stored. They can only be avoided by understanding the issue, and planning against it – much like Babuska’s “mathematical and engineering knowledge.” This is the take-away: serious users of computers must develop an intuition about computer behavior.
We’ll start with an example:

\[
\lim_{x \to 0^+} \frac{1 - \cos(x)}{x^2}
\]

With a calculator: we let \( x \) approach \( 0^+ \), by taking \( x = 10^{-1}, 10^{-2}, 10^{-3}, \ldots \). Next, recall \( \lim_{x \to c} f(x) = L \) means the more decimal places \( x \) and \( c \) agree, the more decimal places \( f(x) \) and \( L \) agree. So, with a Texas Instruments TI-85 calculator:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( (1 - \cos(x))/x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>0.45970</td>
</tr>
<tr>
<td>.01</td>
<td>0.50000</td>
</tr>
<tr>
<td>.001</td>
<td>0.50000</td>
</tr>
<tr>
<td>( 10^{-5} )</td>
<td>0.50000</td>
</tr>
<tr>
<td>( 10^{-6} )</td>
<td>0.50004</td>
</tr>
<tr>
<td>( 10^{-7} )</td>
<td>0.49960</td>
</tr>
<tr>
<td>( 10^{-8} )</td>
<td>0</td>
</tr>
<tr>
<td>( 10^{-9} )</td>
<td>0</td>
</tr>
</tbody>
</table>

We believe the closer \( x = 10^{-n} \) is to \( 0^+ \), the better answer we’ll get. Therefore, the correct answer is the zero, not the .5.

It’s a kind of a suspicious answer, because of the sudden jump from 0.50000 to zero. Let’s use our mathematical knowledge and intuition to check it in another way, with L’Hospital’s rule:

\[
\lim_{x \to 0^+} \frac{1 - \cos(x)}{x^2} = \frac{0}{0} = \lim_{x \to 0^+} \frac{[1 - \cos(x)]'}{[x^2]'} = \lim_{x \to 0^+} \frac{\sin(x)}{2x} = \frac{0}{0} = \lim_{x \to 0^+} \frac{[\sin(x)]'}{[2x]'} = \lim_{x \to 0^+} \frac{\cos(x)}{2} = \frac{1}{2}
\]

The L’Hospital answer uses a well-know theorem that’s been around for centuries, with no reported errors. But then – that means something very strange must be happening inside the calculator. And to sort this out, we’re going to have to go inside the calculator.

Computers and calculators are electronic devices – they work by shifting electrical charges around. Just as batteries can store and release charge, devices called capacitors can store and release charges quickly enough for modern computers.

We’ll conceptualize a computer number as a row of capacitors, and the charge in each capacitor represents a number; it would look something like the Heng/Zong system from p13, Figure 92. This picture shows the problem: there are only five boxes to store digits, so numbers larger than 99999 or smaller than .00001 can’t be written. We can add more boxes, but in a computer, there’s a limit; this limit is referred to as the word size of the computer.
There are other limitations. First, there’s an issue of making the best use of the small word size. For example, if I wrote .000003 as $3 \times 10^{-6}$ I could store it as $3, -, 6$. All I need to know is the third place stores the exponent, the second the sign of the exponent, and the first the actual number.

While we can imagine our capacitors storing ten different levels of charge to represent the numbers $0, 1, \ldots, 9$, in fact subtle variations in charge are very hard to detect. Modern computers use only two levels, which are traditionally thought of as the numbers one and zero (though in fact the two charges are more like $\pm .5$). Then a number like $1001$ could represent $1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$, or $9$. That is, we have to use binary arithmetic.

We also have to have standards, to know how to write numbers. Usually, to subtract $5.42$ from $21.06$, we’d align the numbers:

```
  2 1.0 6
- 5.4 2
  1 5.6 4
```

This is called fixed point arithmetic, and was used in the earliest computers (see p104). Of course, we could use many other kinds of representation: $21.06$ would be written as $0.2106 \times 10^2$, or, in scientific notation, $2.106 \times 10^1$. In the 1960’s, when computers began to be used in business (see Figure 93), standards for storing numbers varied. Each representation had its own problems and associated errors; multiplication by $1.0$ could cause loss of the last four decimal places; programmers would use tricks such as replacing $x$ by $(x+x)-x$ to fool the computer into getting the answer right. Programming each individual computer was a craft in its own right, but as long as manufacturers like IBM kept one standard, programmers could adjust, and companies paid for programmer-craftworkers.


To see how we can get into trouble, we’ll invent a silly machine, the Kathytron 5, Figure 94. The machine uses floating point arithmetic: for the number $21.06$ the decimal point floats to the front, to give $0.2106 \times 10^2$. With the bit structure described, the Kathytron has machine numbers of the form $\pm b_1 b_2 \times 2^{\pm e_1}$, where $b_1$, $b_2$, $e_1$ can be $0$, $1$. These numbers are:
Converting binary to a fraction, \(.11 = \frac{1}{2^1} + \frac{1}{2^2} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}\) and then to a decimal:

\[
\begin{array}{cccc}
.00 & .00 & .00 & .00 \\
.01 & .01 & .01 & .01 \\
.10 & .10 & .10 & .10 \\
.11 & .11 & .11 & .11 \\
\end{array}
\]

On a number line, the positive machine numbers look like this:

These numbers are not equally spaced, and, although the Kathytron 5 is a small machine, this unevenness is typical for this standard; we’ll talk about that when we discuss errors in approximating actual numbers by machine numbers.

There’s also a smallest non-zero number the machine can represent, called machine epsilon. In many programming languages, you can access machine epsilon by running the command `eps`; on the Kathytron, you’d get `eps = .125` This book is being written on a MacBook Air running the public domain language Octave; and the `eps` command gives `eps = 2.2204 \times 10^{-16}`, which will give rather more accurate computations than the Kathytron 5.

The Kathytron rounds a number smaller than machine epsilon to zero; this is called underflow, and a given computer may or may not send a message that this has occurred. Any number larger than the largest machine number generates overflow. For many machines, the processing unit will generate an error message; depending on the machine, this can terminate the computation.
Now, we’re not finished. The Kathytron saves one bit by using normalized floating point notation: the initial number can’t be a zero. In doing a computation, the machine will take the $b_1b_2$ from memory, and send it to the processing unit as $.1b_1b_2$; our Kathytron 5 numbers are now $\pm .1b_1b_2 \times 2^{\pm e_1}$. This gives some smaller numbers; our table is now

<table>
<thead>
<tr>
<th>.1 $b_1b_2$</th>
<th>$\times 2^{-1}$</th>
<th>$\times 2^0$</th>
<th>$\times 2^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.100</td>
<td>.100</td>
<td>.100</td>
<td>.100</td>
</tr>
<tr>
<td>.101</td>
<td>.101</td>
<td>.101</td>
<td>.101</td>
</tr>
<tr>
<td>.110</td>
<td>.110</td>
<td>.110</td>
<td>.110</td>
</tr>
<tr>
<td>.111</td>
<td>.111</td>
<td>.111</td>
<td>.111</td>
</tr>
</tbody>
</table>

In decimals

| .25   | .5   | 1    |
| .3125 | .625 | 1.25 |
| .375  | .75  | 1.5  |
| .4375 | .875 | 1.75 |

The numbers are still not equally spaced, especially near zero and machine epsilon, which is now twice as large. This is called the gap at zero, and occurs in machines that use a similar standard. And it is the cause of the original problem: when I subtract the machine numbers .375 and .3125, I get .0625, which is smaller than machine epsilon. This machine rounds this down to zero.

This is what went wrong when I computed $\lim_{x \to 0^+} \frac{1 - \cos(x)}{x^2}$. When $x$ was very close to zero, the subtraction $1 - \cos(x)$ caused the leading terms in the decimal to cancel, and the decimal places that remained were less than machine epsilon. Without notice, the machine rounded down to zero, and my table had a sudden jump from .5 to 0, where $1 - \cos(x)$ becomes less than machine epsilon.

We mentioned ‘intuitions’ and ‘craftworker tricks’ earlier; here’s one, for $1 - \cos(x)$:

$$
\frac{1 - \cos(x)}{x^2} = \frac{1 - \cos(x)}{x^2} \left( \frac{1 + \cos(x)}{1 + \cos(x)} \right) = \sin^2(x) \left( \frac{1}{1 + \cos(x)} \right)
$$
The trick gives us what calculator couldn’t:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$\sin^2 x \left( \frac{1}{1 + \cos x} \right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>0.49958</td>
</tr>
<tr>
<td>.01</td>
<td>0.49999</td>
</tr>
<tr>
<td>.001</td>
<td>0.50000</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>0.50000</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>0.50000</td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td>0.50000</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>0.50000</td>
</tr>
<tr>
<td>$10^{-9}$</td>
<td>0.50000</td>
</tr>
</tbody>
</table>

But: the user has to know this can happen, and plan against it. Here’s why we need programmers-craftworkers, or, in the old days, women calculators. The trick is to eliminate possible subtractions where many leading terms in a decimal will cancel. There’s a rule of thumb: ‘Don’t subtract nearly equal numbers.’ That’s one of our ‘mathematical intuitions’.

Cancellation is one issue; underflow is another. If we try using the Pythagorean Theorem to compute the hypotenuse of a right triangle (sides $x$, $y$ and hypotenuse $h$) we have $h = \sqrt{x^2 + y^2}$. If one or both of $x$, $y$ are small, their square will underflow to zero. For example, on the Kathytron with normalized floating point numbers, choosing $x = y = \frac{1}{4}$ gives $x^2 = y^2 = \frac{1}{16}$, which underflows to zero.

One way to avoid the problem is to make $x$ and $y$ appear larger. If we let $s = x + y$, then

$$h = s \sqrt{\left( \frac{x}{s} \right)^2 + \left( \frac{y}{s} \right)^2}$$

Now $\frac{x}{x+y} = \frac{y}{x+y} = \frac{1}{2}$, so

$$h = \left( \frac{1}{4} + \frac{1}{4} \right) \sqrt{\left( \frac{1}{2} \right)^2 + \left( \frac{1}{2} \right)^2} = \frac{1}{2} \sqrt{\frac{1}{4} + \frac{1}{4}}$$

Each of the squares, and their sum, is a machine number, so the formula produces the correct $\frac{1}{2} \sqrt{2}$. This will have to be rounded to a machine number, in this case $\frac{1}{2}$.

There is another way to deal with underflow: allow numbers in the gap near zero. You can do this by allowing non-normalized numbers, $\pm 0 \times 2^{\pm e_1}$. This brings in .00 again, but also $\pm 0 \times 2^{-1}$ and also $\pm 0 \times 2^0, \pm 0 \times 2^1$, which only gives us the new numbers $\pm \frac{1}{8}$, inserted into the gap near zero.
When computers became available for medical, scientific and engineering research, the lack of a single standard became an issue. Software written for one computer would not run properly on other computers. Estimates of the accuracy of a computation were only good on a specific computer, and could be unreliable on others.

But the real motivation for setting a standard came from the development of microcomputers. IBM might sell a few hundred machines, Cray a dozen, but a chip designer like Intel expected to sell millions. Moreover, the users would not be large businesses with expensive craftworker-programmers, but small businesses with perhaps no programmers, who still expected correct answers.

A group of scientists began meeting to resolve the problem. Intel and Motorola wanted not just a standard, but the best possible standard. Dealing with underflow, overflow, the gap near zero – all of these inserted extra steps into computations. The desired standard would be one that didn’t slow down the chips too much. In 1974, the IEEE 794 standard for representing numbers on a computer came out. In reality it’s a set of standards; see again p104. Initially, more like a set of suggestions, but, as all the large chip-makers followed it, it became a ‘standard’ by default. IEEE 794 includes the idea of using non-normalized numbers to fill the gap near zero, as well as sending notifications for overflow and underflow.

Unfortunately, this doesn’t finish the job: every computation must result in a machine number, so almost every computation has to be rounded to a machine number, which results in a ‘wrong’ answer. If we’re going to use computers, we have to understand how much wrong those answers are.

We’ll start with one example of how a roundoff-error can go badly wrong: the 1991 SCUD missile attack on a US Army base at Dhahran, Saudi Arabia. The base was protected from missile attacks by the Patriot anti-missile defense (see Figure 96). In the 1991 incident, the system failed to intercept a SCUD attack (Figure 97), resulting in major loss of life. Investigation showed it was due to a software error caused by rounding.

The Patriot system was designed to run for short lengths of time, to avoid detection – and because the internal clock rounded absolute time to machine numbers, causing an error of $9.5 \times 10^{-8}$. This particular system was protecting soldiers 24/7, so it had been left running for over 100 hours. The accumulated round-off errors caused the system to generate an estimate of the SCUD position that was off by 572 meters. In short, the Patriot missile missed. For a short discussion see
As if this weren’t enough, all kinds of odd algebra can happen with machine numbers. For example, the machine number .3125 squares to .09765625 < \varepsilon, so the square becomes zero. We add numbers in any order; this is the associative law: \((a + b) + c = a + (b + c)\). It doesn’t work in a computer, even using just machine numbers: 
\[1.75 + .25 - .5 \text{ can be } (1.75 + .25) - .5 = \text{overflow. But we could add in a different order to get } 1.75 + (.25 - .5) = 1.75 - .25 = 1.5\]

It almost seems as we’re back in the days of hand-made arithmetic, with craftwomen calculators checking that every step works. In reality, it’s not that bad: we can predict errors.

We’ll start with the error caused by roundoff. Let’s look at an example, again from IEEE 794 on the Kathytron. Our number \(x\) will be .6. The Kathytron follows IEEE, which requires rounding to the closest machine number. That’s .625. There are two different ways to look at error:

- **Absolute Error** = \(|\text{True value} - \text{Approximate value}|\), and
- **Relative Error** = \(\frac{\text{Absolute Error}}{\text{True value}}\).

In this case, the relative error is
\[
\left| \frac{.6 - .625}{.6} \right| = \left| 1 - 1 + \frac{.025}{6} \right| = \frac{1.25}{3} \leq \frac{1}{2} \varepsilon
\]

– since machine epsilon is \(\varepsilon = .25\). In general:

\[
\text{Relative Error} \leq \frac{1}{2} \varepsilon
\]

This is an easy result to understand: the number you round is between two machine numbers, so the error is at most half the length between the machine numbers. The gap between machine numbers increases when the numbers get large (see p97). But, while the absolute error can increase as the numbers increase, the relative error gets divided by the size of the number, which makes the gap \(\varepsilon\), and relative error \(\leq \frac{1}{2} \varepsilon\).

The next thing we have to understand is how errors change when we do basic arithmetic and round. We’ll use a special notation: if we take a number \(a\), the closest machine number is denoted by \(\text{fl}(a)\). What we want to understand, then, is how \(\text{fl}(a + b), \text{fl}(a \cdot b), \text{fl}\left(\frac{a}{b}\right)\), relate to \(\text{fl}(a)\) and \(\text{fl}(b)\).
We’ll follow the discussion in Ward Cheney and David Kincaid, *Numerical Mathematics and Computing*, Brooks/Cole 1994. The assumption is that the machine takes \( a, b \), computes the arithmetic operation correctly, then rounds. This is partly true; many machines use double word lengths to do the computation, and then round to machine numbers. The assumption fails for underflow or overflow, but, aside from this, is reasonable. So we still have Relative Error \( \leq \frac{1}{2} \epsilon \), or, for example,

\[
\left| \frac{fl(a + b) - (a + b)}{a + b} \right| \leq \frac{1}{2} \epsilon
\]

or,

\[
\frac{fl(a + b) - (a + b)}{a + b} = \delta
\]

where \( |\delta| \leq \frac{1}{2} \epsilon \). Rewriting,

\[
fl(a + b) - (a + b) = (a + b) \times \delta
\]

and similarly for multiplication and division. Let’s check this with the special case of \( a = \frac{3}{4} = .75 \) a machine number, and \( b = \frac{3}{5} = .6 \) not a machine number; we chose these numbers to avoid underflow or overflow.

| Operation | Result | \( fl(B) \) | \( |B - C| \) | \( \frac{1}{2} \epsilon \times B \) |
|-----------|--------|-------------|-------------|----------------|
| \( a + b \) | 1.35   | 1.25        | .1          | .16875         |
| \( a \times b \) | .45    | .5          | .05         | .05625         |
| \( \frac{b}{a} \) | .8     | .75         | .05         | .1             |

Column D should always be less than Column E, which works out, though rather closely for multiplication. So, we can in fact control the errors in arithmetic operations, which addresses early concerns about using inaccurate floating point numbers rather than accurate fixed point numbers; see p.104.
Notes for Chapter 1 Section 8: Computer Numbers

p92 The P5 problem arose because of the way computers do division. In a way, it’s very much like we do division. Try dividing 376 by 7. We have a technique – an algorithm. We’d first divide 7 into 37. We remember that $7 \times 5 = 35$ but $7 \times 6 = 42$. So the divisor is 5, with a remainder of 26. Now divide 7 into 26. Again we remember that $7 \times 3 = 21$ but $7 \times 4 = 28$. So the divisor is 3, with a remainder of 5.

When I was younger – a lot younger – I’d look up some of those multiplications, in a multiplication table. The P5 used a division algorithm called the SRT algorithm, named after the inventors, Sweeney, Robertson, and Tocher. It also uses a lookup table, though everything is binary, so it seems rather strange. It’s similar to Figure 98. See An analysis of division and implementations by Stuart F Oberman and Michael J Flynn at http://i.stanford.edu/pub/cstr/reports/csl/tr/95/675/CSL-TR-95-675.pdf

p92 For the importance of the ABC transporters, see Christopher F. Higgins and Kenneth J. Linton, The xyz of ABC Transporters, Science 7 Sept 2001 vol 293 p1782. They remark,

*An ABC cell must selectively translocate molecules across its plasma membrane to maintain the composition of its cytoplasm distinct from that of the surrounding milieu. The most intriguing, and, arguably, the most important membrane proteins for this purpose are the ABC (ATP-binding cassette) transporters. These proteins, found in all species, use the energy of ATP hydrolysis to translocate specific substrates across cellular membranes.

Overexpression of certain ABC transporters is the most frequent cause of resistance to cytotoxic agents including antibiotics, antifungals, herbicides, and anticancer drugs.

p93 In the analyses of the failure of the Mars Orbiter Mission, NASA identified the problem as a software issue: a contractor wrote a software module to determine the thrust of the orbiters rockets. The module gave thrust in $ft/sec^2$; the software to which it reported expected force to be reported in Newtons. NASA blamed its own internal procedures for not finding the units mismatch.

The IEEE Spectrum analysis asserts that the problems in NASA went much deeper: flight controllers believed that something was wrong with the trajectory, but were over-ruled by higher management. The article claims that the engineers were told to “stop thinking like engineers and think like managers.” They claim that the expectation was that all systems worked perfectly, and if problems were suspected, it was up to the engineers to prove there was a problem. This is the opposite of standards for airline safety.
For a brief history, see Michael L. Overton *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM Press 2001. Overton notes: "Von Neumann […] promoted the use of fixed point representation. He was well aware that the range limitations of fixed point would be too severe to be practical, but he believed that the necessary scaling by a power of 2 should be done by the programmer, not the machine; he argued that bits were too precious to be wasted on storing an exponent when they could be used to extend the precision of the significand." See Figure 99, von Neumann with the 1949 EDVAC computer, using both binary arithmetic and stored programs.
Chapter 1: Numbers

Section 9: Data: Introduction

It’s likely that anyone who’s taken a physics, chemistry or biology course will think of an experiment which produces data as something like Figure 100, graphed as in Figure 101, and finally manipulated mathematically, as in Figure 102. But all of what we now think of as science is historically very recent; for over a thousand years after Mesopotamian or Greek scholars, Europeans trying to understand the universe were opposed to experiments, wrote polemics against using numbers, and had no idea of data analysis (see p112).

In this section, and in the notes, we’ll look at how the Western view of science was constructed, and we’ll start with one example: how scientists came to believe that the atmosphere is like an ocean of air, with us at the bottom of that ocean. And like water, air has weight, which exerts pressure.

The idea of atmospheric pressure gave a clear answer to a Renaissance paradox: how do pumps work? In Galileo’s Dialogues Concerning Two New Sciences (1638), one of the characters mentions the problem:

This pump worked perfectly so long as the water in the cistern stood above a certain level; but below this level the pump failed to work. When I first noticed this phenomenon I thought the machine was out of order; but the workman whom I called in to repair it told me the defect was not in the pump but in the water which had fallen too low to be raised through such a height; and he added that it was not possible, either by a pump or by any other machine working on the principle of attraction, to lift water a hair’s breadth above eighteen cubits; whether the pump be large or small this is the extreme limit of the lift.

A modern explanation is that lifting the handle of a pump evacuates air from the tube of the pump, and atmospheric pressure forces water up the tube. Since atmospheric pressure is finite, it can only lift the water so far.

But the term ‘evacuate’ implied a vacuum had been created. Since Aristotle said a vacuum was impossible, the preferred explanation...
was that the eighteen cubit limit was due to a force resisting the formation of a vacuum. The history of these discussions is long and complex; see C. Webster, *The Discovery of Boyle’s Law, and the Concept of the Elasticity of Air in the Seventeenth Century*, Archive for History of Exact Sciences Vol. 2, No. 6 (31.12.1965), pp. 441-502.

Eighteen cubits is over thirty feet, and there are many kinds of pumps; as Webster (above) remarks, “...such evidence was confused and unreliable, since real or imaginary pumps of other designs were in principle capable of lifting more than eighteen cubits of water.”

What was needed was evidence that was not confusing and was reliable; this was provided by Evangelista Torricelli (1608-1647). He knew Archimedes’ work on water pressure: the pressure on a submerged object was proportional to the weight of the water above it, which would be the density of the water times the depth of the water (the proportionality factor was the acceleration of gravity, a quantity completely unknown at this time). Torricelli also knew mercury was denser than water, therefore if he used mercury rather than water, the height the liquid mercury could be drawn would be much less than thirty feet.

In addition, he realized he could get rid of pumps entirely: Figure 103 illustrates his idea. When the tube is tipped over into the mercury, the level of mercury dropped to about 76 centimeters. Torricelli claimed that the weight of the atmosphere, pushing on the mercury in the dish, pushes mercury up the tube. It was a great demonstration (actually carried out by his student, Vincenzo Viviani). No complicated pumps needed, no thirty feet of water. Easily understood; reliable because anyone could repeat the experiment. You don’t have to depend on ancient authority or even Toricelli’s authority.

Interpretations still varied. If the mercury had filled the entire tube, but had now dropped by 24 cm, the top of the tube must contain a vacuum. The mathematician/scientist/philosopher Blaise Pascal argued, as did others, that mercury was not forced up the tube by atmospheric pressure, but drawn up by nature’s resistance to the formation of a vacuum. Standoff.

But Torricelli had one more demonstration in mind. Take a surface submerged in water. Archimedes taught that the higher above the surface one went, the less pressure there would be. If the height of mercury in the tube represented the pressure of the atmosphere at the surface of the earth, the higher above the surface of the earth one went, the lower the mercury in the tube would be. Archimedes expressed this as a proportion: the change in mercury level could be computed relative to the height above the surface of the earth. This
computation is important to the theory, because it eliminates counter-
arguments like ‘maybe something else caused the change.’ Maybe,
but why would the change match Archimedes so well? This kind of
experimental technique was new in its time; see p113.

An apparatus was carried up an actual mountain by, of all people,
Pascal’s brother-in-law, and the change in the height of mercury
was exactly what was expected. Pascal was now convinced: air has
weight and it is that weight which drives mercury up the tube.

Alas, the change in height of mercury turned out to match the
Archimedean prediction too exactly, raising questions whether the
experiment had actually been done. Nonetheless, Torricelli’s work
inspired others across Europe to theorize and experiment with air.
Isaac Beeckman in Holland compared the air surround the earth to
a large sponge; Renee Descartes compared it to the fleece of wool.
More significantly, Marin Mersenne in France actually experimented
on air, finding that it could be compressed to 1/1000 of its original
volume, and then expanded again. See C. Webster, The Discovery of
Boyle’s Law, and the Concept of the Elasticity of Air in the Seventeenth
Century, Archive for the History of Exact Sciences, 2(6) 1965.

This was the situation when Robert Boyle and his assistant Robert
Hooke began work. Although the two published over forty experi-
ments on air pressure, we’ll look at only one:

Divers ways have been proposed to show both the Pressure of the Air, as
the Atmosphere is a heavy Body, and that Air, especially when com-
pressed by outward force, has a Spring that enables it to sustain or resist
equal to that of as much of the atmosphere, as can come to bear against
it, and also to show, that such Air as we live in, and is not condensed by
any human or adventitious force, has not only a resisting Spring, but an
active Spring (if I may so speak) in some measure, as when it distends a
flaccid or breaks a fullblown bladder […]

Robert Boyle, New Experiments, Physico-mechanicall, touching the
Spring of the Air, LONDON, Printed by Miles Flesher for Richard
Davis, Bookseller in Oxford, MDCLXXXII.

Boyle’s ‘spring of the air’, can be compared to an actual spring, an
automobile shock absorber, Figure 104. This shock absorber uses a
metal spring which contracts when pushed down, and returns to its
original shape when left free.

The point of Boyle’s comment is that air behaves in the same way
(see Figure 105). We’ll look at how Boyle and Hooke’s took the issue
beyond analogies with sponges and fleece, to prove the Springe of
Air. (Hooke later found the general law governing the behavior of
springs: Hooke’s Law. We’ll explore this later.)
Figure 106 shows something like what Boyle might have used; a bent tube, sealed off on the left, open to the atmosphere on the right. Also, "That the tube being to (sic) tall that we could not conveniently make use of it in a Chamber, we were fain to use it on a pair of Stairs, which yet were very lightsome, the tube being for preservations sake by strings so suspended, that it did scarce touch the box [...] ." Robert Boyle, *New Experiments, Physico-mechanicall, touching the Spring of the Air*, cited above. His tube was a good deal larger than the one in Figure 106.

But: if you just pour mercury into the tube, it compresses the air on the left. Boyle jiggled the tube to equalize the pressure on both sides. He then poured mercury in on the open side, and noted the height of the air on the left, as well as the mercury on the right, columns A and B in the table below.

These heights should be proportional to the volumes of each; for the mercury, that would be the pressure exerted on the compressed air. So the numbers recorded would really be volume and pressure. With one fudge: the tube was open to the air, so atmospheric pressure needed to be accounted for, in column D.
When Boyle showed these numbers to his friends, several people, including Hooke, thought that pressure and volume seemed to be ‘in reciprocal relationship’; in modern terminology, \( A = \text{Constant}/D \). Boyle himself was not particularly interested in numerical relationships: ironic, as the mathematical relation is known as Boyle’s Law. This may have been Hooke’s contribution; his background and interests were very different from Boyle’s; see p116.

In any case, Boyle added an extra column, \( E \), comparing the measured data in \( D \) with the theoretical \( 1/A \). The columns in the table require us to look at one row at a time; a graph gives us an overview of the difference between experiment and prediction (Figure 107), though graphing data wasn’t used until the 1800’s (see p113).

Although the numbers seemed to be ‘in reciprocal relationship’, it also seems there’s a substantial difference between the data and its reciprocal. This is actually no surprise: Column \( E \) doesn’t take the constant into account. How would we do that? If \( A = \text{Constant}/D \) as suggested, then a graph of \( A \) versus \( 1/D \) should be a line, the constant would be the slope of that line. Figure 108 shows this graph, which really does looks like a (slightly wiggly) line.

But how to find the slope? Again, we have techniques that hadn’t been invented in Boyle’s time. Astronomers had just begun to think about how to choose the best example from a series of different observations of a planet’s position (using the then-new technique of averaging data). To find the best line through a set of two-dimensional had to wait until the 1800’s, when Carl Friedrich Gauss and Adrien-Marie Legendre invented the method of least squares. We’ll cover cover this in Section 10; for now, we’ll accept that it gives the ‘best’ straight-line version of the wiggly line, in the form \( y = mx + b \). Figure 109 shows the two curves together; the green line has intercept \(-2.45353 \times 10^{-5}\) and slope \(0.00071009\). The graph suggests that the red curve is close to being a straight line, except more like \( A = \text{Constant}/D + \text{ExtraConstant} \) instead of \( A = \text{Constant}/D \). Why is the extra constant there at all?

The rulers were marked off in units of \( 1/16 \), it would be easy to make an error reading the true height of the mercury; the error would be somewhere in the \( 1/16 \) gap, so at most an error of \( \frac{1}{2} \times 16 = .03125 \), about 128 times larger than observed. Another issue is “heights should be proportional to the volumes of each”; volume is proportional to height only when the tube is a perfect cylinder; glassblowers in the 1700’s were nowhere near perfect.

For the time, the data is very good.
We’ve talked history and philosophy, but the subject of Section 9 is really the numbers. We’ll look at:

i) In column A of Boyle’s data (p108), measurements were only taken every two units. The process of selecting just a few of the many possibly units is called sampling. How does sampling data affect the results we get?

ii) In column B, heights of mercury were recorded to one $\frac{1}{16}$ of an inch. A modern experiment might talk of “recording to two decimal places”; in either case this process is called quantization. We saw variation in these numbers producing errors; how does this work?

iii) The data Boyle collected was restricted to only a limited number of pressures – or equivalently, heights of mercury: *We were hindered from prosecuting the trial at that time by the casual breaking of the tube. But because an accurate Experiment of this nature would be of great importance to the Doctrine of the Springe of the Air …* [insert about ten subordinate clauses] *the several Observations that were thus successively made, and as they were made set down, afforded us the ensuing Table (p108). Robert Boyle, New Experiments, Physico-mechanick, touching the Spring of the Air, cited above.*

This is called range restriction. At very high pressures, Boyle’ Law no longer works; range restriction can cause problems.

iv) Boyle was dis-interested in using his data; his Law was discovered by others. This was intentional:

*Boyle’s books were among the first scientific writings to embody the principles laid down by Bacon that the story should be told without embellishment or flights of rhetoric, but as a straightforward account of what had actually been done in the experiments, what had been observed as a result, and what the theoretical implications were.*


Francis Bacon’s scientific program was a reaction against Aristotle’s program of explanations; Bacon believed the business of the experimenter was to provide facts; from accumulated facts would grow control over Nature, and bring about a scientific utopia (see p109). For Bacon, and Boyle, experiments produce facts, and facts speak for themselves.

Boyle was unusual for his time in letting the facts speak at great length; he described all he details of his experiments. He knew that very few had the means to repeat those experiments; he wanted his readers to believe the results were as he said.

For modern scientists, this has been problematic; recently, labs have
been unable to repeat some important experiments and get the same results; we’ll discuss this later. There’s a second issue: mathematical manipulation of data is still manipulation; can it be trusted? This is again an issue arising recently: is the mathematics being applied correctly?

v) There’s a hidden issue: Boyle chose to record only pressures, but temperature is important (see p120). The data we leave out can lead to errors; we’ll see this in other fields; again later.
Just before the Elizabethan period, the mathematician/astrologer John Dee was arrested on charges of "calculating", "conjuring" and "witchcraft," all of which, along with alchemy, were considered equally evil. Why? With Kepler’s laws, the orbits of the planets could be computed. This meant that knowing the day and hour of an individual’s birth, the position of the planets could be computed (retrospectively), and so their horoscope would be known. Compare the early development of mathematics in Mesopotamian astronomy, p15. If the horoscope was for the king, or anyone in high office, such knowledge was dangerous, especially if the planets told of impending disasters. Elizabeth herself, and her court, believed in magic – and again, with the right knowledge, spells might be cast against her. Knowledge of mysteries such as calculation was dangerous. In the English Civil War, pamphlets and almanacks proclaimed the fall of kings was near; people of the time said these prognostications likely caused the fall of Charles I.

Galileo was in a different position. He worked as a mathematician, a very low-status job compared to philosophers who followed Aristotle. To give a bit of context, Galileo worked three centuries after Dante Alighieri composed his *Comedia* (The Divine Comedy); by Galileo’s time, a common joke was "What circle of Hell contains the mathematicians?" (Today, probably dot-com billionaires) For mathematicians, the answer, of course, is the circle of the fraudulent. Right next to Judas, and ever-so-slightly above Satan.

Why fraud? Babylonian astronomers, Ptolemy, Copernicus and Kepler all devised mathematical systems to predict the positions of the known planets and the known moons of the planets. The last four followed Greek ideas, that objects in the heavens must move in perfect circles, with constant speed. Since they do not appear to do so, mathematical astronomers used all kinds of variations to get the answers to come out right: circles moving on circles, off center circles, and eventually ellipses. This was criticized as "saving the appearances" – that is, adding more and more unjustified assumptions, just to get the right answer. True physicists, even as late as Galileo’s time, were supposed to start from known truths, and then deduce from those how nature behaves. The techniques these astronomers used were, therefore, frauds.

---

Figure 110: Hell
Gustav Doré’s Satan, at the very last circle of Hell, farthest from Heaven.
“The dismal Situation waste and wilde,
A Dungeon horrible, on all sides round
As one great Furnace flam’d,
yet from those flames
No light, but rather darkness visible”
— Milton, *Paradise Lost*

Figure 111: Ptolemaic System
Illustrating Ptolemy’s ‘circles on circles’, or epicycles, to save the appearance of circular motion for the planets.
Graphs had been used as early in the 1300’s, for scholarly and scientific purposes. The French mathematician Nicola Oresme . . . conceived of the idea of using rectangular coordinates (latitudo and longitudo) and the resulting geometric figures (configurationes) to distinguish between uniform and nonuniform distributions of various quantities, such as the change of velocity in relation to time . . . In the discussion of motions the base line (longitudo) is the time, while the perpendiculars raised on the base line (latitudines) represent the velocity from instant to instant in the motion . . . Stefan Kirschner, Nicole Oresme Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/nicole-oresme/#Mat

For a constant (uniform) velocity, distance = speed × time. The graph would be a horizontal line, and the speed would be the area under the graph. If the velocity increases uniformly, the graph is a slanted line; the area underneath is a trapezoid, whose area was known, and this again represents a distance (Figure 112). This result was also known to a group of scholars at Merton College, the ‘Oxford Calculators’.

Graphing actual data instead of philosophical concepts seems to have been the invention of William Playfair, in the early 1800’s; see Figure 113. Playfair even remarked that the graph allowed one to comprehend complex patterns more easily than a list of numbers – much the same reason we still use graphs today.

Torricelli used a form of investigation we’d call empiricism:

One of the distinctive features of modern science is a commitment to empiricism – a fundamental expectation that theoretical hypotheses will survive encounters with observations. Those that comport with the theory’s explanations and predictions confirm the theory. Anomalous observations that do not fit theoretical expectations disconfirm it. Moreover, experiments can be contrived to generate observations that might serve to confirm or disconfirm a theory.


This is how we think science should be done (see Avery’s work on DNA, p39), but for over a thousand years, it wasn’t what scholars in the West thought was right, and this was partly due to Aristotle.
Aristotle’s interest was not to discover new facts, but to explain why — why things happened and why they had to happen that way. As we discussed on p30, his explanations had to be of a certain type: deductions from well-known facts.

For Aristotle, who was to become the preeminent “ancient authority,” phenomena were, literally, data, “givens.” They were statements about how things behave in the world, and they were to be taken into account when discussing topics concerning nature. The immediate sources of phenomena were diverse: common opinion and the assertions of philosophers, as well as sense-perception. Given these statements, a system of syllogistic reasoning yielded, in principle, a theoretical description and explanation of them.

Peter Dear Totus in Verba: Rhetoric and Authority in the Early Royal Society, ISIS 1985, 76 145-161

There’s something off about this approach: had humans been concerned with ‘understanding’ rather than discovering, there wouldn’t be flint tools, pigments mixed for rock-painting, roots boiled to make them edible, firing of clay to make pots, planting of crops, smelting of metals, . . . . Apparently science, pursued as ‘understanding’, had little to do with how real people made real progress.

All that was to change. But what caused the change from deduction to empiricism?

This is a technical question, and not a simple one; historians still disagree. Some scholars were thinking empirically as early as the 1500’s:

The experiential approach to an understanding of the physical world was, to some extent at least, always promoted in the medical faculties. The Italian universities, Montpellier in France, and even the highly traditional Paris Medical Faculty expected medical students to study practical aspects of medicine by a kind of apprenticeship to a local practitioner, while undertaking their more theoretical studies in the university. From the sixteenth century medical schools became the prime sites for a number of facilities essential for the promotion of observational and empirical science: anatomy dissections, botanical gardens, and in some cases chemical laboratories.


It’s not obvious how this influenced individuals thinking about physical science, but the early 1600’s was a time when scholars in general began to communicate and form societies. Others communicated by letters; the French priest Marin Mersenne maintained a web of con-

The philosophy of knowledge was also changing. Some of this was quite old – going back to Galen and even further to Hippocrates; see Wallace, Galileo’s Pisa Studies in Science and Philosophy, in Peter Machemer ed, The Cambridge Companion to GALILEO, Cambridge University Press, 1998. Without getting lost in the complexities of medieval scholastic philosophy, the change was a form of deductive reasoning called regressus (from the Latin, to return).

As we saw, true knowledge was understanding of the causes of phenomena. If you observe an eclipse, you understand the cause is a planet coming in front of the sun. The disc-like shadow results from the spherical shape of planets.

This doesn’t leave room for true knowledge about more complex phenomena, like illnesses; regressus helps with these. For example, when you observe a fever, the first step is to find a cause – say an imbalance of of hot/cold, wet/dry in the body. In the second step you use deduction to establish that the cause you found (guessed) really does result in the effect (this it the regressus: return to the original).

The difficulty here is in guessing causes from effects. It’s similar to the problem of finding first premises in mathematics; Aristotle attributed this to a different cognitive state: insight, intuition, etc (p30). Throughout the 1500’s, scholars debated exactly what this extra function was (and whether it was needed). One term used was negotiatio intellectus – roughly, the work of the intellect:

Yet regressus also inserted an ambiguity into the understanding of this relationship, in the form of the negotiatio intellectus. Though an essential step in regressus, there was no consensus as to how the negotiatio was supposed to proceed. The discriminatory use of observations could be seen as a natural way to resolve this ambiguity. Thus, the regressus method suggested a novel methodology in natural science that admitted observations as epistemic grounds for accepting and rejecting theories.


Where does Torricelli’s work come in? He worked under Galileo, and may have learned the technique from him, then applied these ideas to atmospheric pressure, as we saw on p106.
We’ll spend serious time, later, on Galileo. For now, what is known with some certainty is that Galileo studied at the University of Pisa (a center of experimental botany and medicine); some of the scholars he is known to have worked with were using regressus. Galileo himself used regressus in his very early work (again, see Wallace, above, p96).

To emphasize the point, though: this is what may have helped lead to empiricism and a new way of doing science; it is not the final word; see the literature quoted above.

As an example of what careful research can achieve, it had been accepted that Galileo never did careful experiments. Part of the justification was that in the late 1890’s, the mathematician/physicist Antonio Favaro in Italy published the Edizione Nazionale of Galileo’s papers, and no evidence of experimental work was to be found in them.

In the 1970’s, the historian Stillman Drake realized that Favoro had heavily edited the papers, and had not published miscellaneous sketches and random notes at all. Returning to the original papers, Drake found

>This unpublished material includes at least one group of notes which cannot satisfactorily be accounted for except as representing a series of experiments designed to test a fundamental assumption, which led to a new, important discovery. In these documents precise empirical data are given numerically, comparisons are made with calculated values derived from theory, a source of discrepancy from still another expected result is noted, a new experiment is designed to eliminate this, and further empirical data are recorded. Stillman Drake, Galileo’s Experimental Confirmation of Horizontal Inertia: Unpublished Manuscripts, Isis Vol. 64, No. 3 (Sep., 1973).

p109 Hooke and Boyle were influenced by the cultural, political and religious conflicts of their time. Even before Elizabethan times,

>Mathematics itself came in many guises both institutionally and extra-institutionally. Certainly geometry was taught at the universities, but also there were the mathematical sciences of astronomy, geography and sometimes mechanics. Outside the sanctioned institutions mathematics reigned quite lively in the realms of natural magic, astrology and hermetic practices, and the cabala, as well as in the more mundane, pragmatic spheres such as the principles of painting, construction of fortification and the design of machines.

Peter Machamer, Galileo’s Machines, his mathematics and his experiments, in Peter Machamer ed, The Cambridge Companion to
Elizabethan London was an international port, trading with most of the continent. At the very least, traders would need to convert currencies and keep records of complex deals. There was another financial concern: Elizabeth inherited a bankrupt throne. She was actively interested in supporting engineering or craft schemes that could enrich her subjects and the crown; see Deborah E. Harkness, *The Jewel House: Elizabethan London and the Scientific Revolution*, Yale University Press 2008.

For a trading country, navigation was crucial, and navigation meant geometry. Even for Charles Darwin, on the Beagle. His bunkmate Midshipman Stokes had "main responsibility to look after and redraft the navigational charts which were the object of the voyage. To his chagrin, Darwin found his Cambridge education a poor substitute for Stokes' practical expertise. 'After looking at my 11 books of Euclid, & first part of Algebra (including binomial theorem?) I may then begin trigonometry after which I must begin Spherical?"” Janet Browne, *Voyaging*, Princeton University Press 1966. An English translation of Euclid helped those who had no Oxford or Cambridge training in Latin; John Dee wrote the preface.

This was only one of many mathematics texts published; another was Edward Wright's *Certaine errors in navigation* (1599):

> Wright wanted to do all he could to put reliable and verifiable information in the hands of England's navigators and mariners. His meticulous accounts of observations set a new standard for accuracy and implicitly encouraged replication of results by recounting details about the instruments used and the precise locations where the observations were made. Wright’s early attention to the precise location and instruments used to make observations made him a trailblazer of verifiable, reproducible experimental knowledge.


This was the complex London in which Hooke worked. There was also the political-religious London of the Commonwealth and Protectorate. The Civil War was partly a religious and partly an economic war. The Anglican Church was the established church of England, and every loyal citizen had to be baptized into it – and pay tithes of 10% for upkeep of the church and the clergy. King Charles I was also returning to elaborate rituals, more characteristic of Catholicism than Protestantism. After Charles lost the war and his head, the Anglican church lost its status as the official State religion, as well
as its tithes, and many of the vicars, bishops and other officials were turned out of their jobs. For centuries, bright young men had found good careers through the Church or the State (even Darwin considered being a clergyman). For Hooke, growing up in this period, there would be no clear path to employment. Fortunately, some Cavaliers (adherents of the monarchy) still held University positions, understood the economic and scientific changes happening in England, and believed that the new order needed practical men, who could measure and compute:

*Through John Wilkins’ efforts, a handpicked group of mathematically inclined and scientifically able men was assembled in Oxford in the early Commonwealth years. On the whole the were men of ‘cavilier’ persuasion – moderate supporters of the monarchy whose hopes for the future had been dashed by the violent termination of the reign of Charles I, and who now found themselves with no prospect of political or clerical preferment, constrained to make their living outside the established Church and the Government.*


Hooke, though poor, was looked after by friends of his father. He worked his way through Oxford (likely as a servant to wealthier young men) and was noticed and taken up by Wilkins, who actually recommended Hooke to Boyle. Boyle employed Hooke to run his experiments, giving Hooke a base from which to explore Nature.

Boyle’s path was rather different. His father married into money, accumulated more, mostly in the form of land in Ireland, became an Earl, and by the time of the Civil War was called the wealthiest man in England. A good deal of this land was redistributed after the fall of the monarchy, though Boyle was still quite well-off. He was, however, subject to different kinds of influences.

While it may seem that England was undergoing some kind of scientific rational enlightenment, that was only just starting. Even Francis Bacon could write “I . . . understand [magic] as the science which applies the knowledge of hidden forms to the production of wonderful operations; and by uniting (as they say) actives with passives, displays the wonderful works of nature.” Bacon, *De Augmentis*. Some wanted to construct a ‘science of prophecy’; Christopher Hill wrote:

*Sir Walter Ralegh, Sir Francis Bacon, Sir Kenelm Digby and many other members of the future Royal Society, believed in sympathetic magic: ...John Locke believed in it too. We cannot separate the early history of science from the history of magic. ... Giodano Bruno, John Dee, Johannes*
Kepler, Tycho Brahe were all magi [magicians]. John Wilkins, future secretary of the Royal Society, in 1648 still quoted Dee and Fludd as authorities on 'mathematical magic.'


A problem arose when these ancient beliefs mixed with the end of an official religion. With that came the lifting of censorship of books, allowing many new religious sects to spring up: Ranters, Levellers, Anabaptists, Familists, Quakers, Diggers, Muggletonians . . . . As one example, belief in a world pervaded by spirits willing to assist magicians was consistent with beliefs that traditional religion was in error; everyone could contact the spirit of God:

...the Ranters embraced the concept of the Indwelling Spirit, but went further by claiming that anyone who had made a personal relationship with God was no longer bound by conventional society and that whatever was done in the Spirit was justifiable. This encouraged a sense of liberation from all legal and moral restraint. Organized forms of religion could be rejected, the concept of sinfulness dismissed and the Bible itself disregarded. Free love, drinking, smoking and swearing were regarded as viable routes to spiritual liberation.


Along with this, there was the sense that the nobility were no more noble or deserving than commoners, that a worldly paradise of equality among people was at hand – as was the Second Coming of Christ. Besides the breakdown of what the wealthy considered decent society, groups such as these had no use for Christianity, nobles, kings or government; the end of the world was in sight, bringing a heaven on earth, where everyone would be equal. While Boyle stood to lose his estates and social position, more important for him was Christianity and the Anglican Church. He saw heaven on earth in a Baconian way, as the continued advancement of science and control over nature – not through magic, but though knowledge of the natural world:

*During the 1650s the reformers – Boyle, Walter Charleton, and others- modified their philosophy in the face of the radical threat: in the place of the now discredited occultism they adopted what Boyle called the corpuscular philosophy. This amounted to a Christianized Epicurean atomism treated as a hypothesis to be tested by experiment. The corpuscularians held with Epicurus that the world was made up of lifeless atoms colliding in the vacuum of space. But the Puritan philosophers departed from Epicurus by denying that the world as we know it was the product of a long*
succession of random atomic collisions. Rather they held that a providential God was responsible for all motion in the universe. He determined the paths the atoms took and hence maintained the order of the universe. Not only was this a workable scientific hypothesis capable of being refined and elaborated by a Baconian program of experiment, it was also an attractive candidate for adoption because it was applicable to social issues.

What united them all was the belief that rational explanations could be arrived at for everything in the natural world, and that such form of explanation were confirmation of the existence of an all-knowing God, whose representatives on earth – the Anglican clergy – were the custodians and guides on behalf of those unable to rise to full understanding on their own.


These ideas pervaded even Newton’s work – as did a more sophisticated ‘science of prophesy’.

We’ll write Boyle’s Law in modern terms: let $P$ be the pressure exerted on a gas and $V$ the volume of the gas. Then $PV = c$ where $c$ is a constant. This is true only when the temperature $T$ of the gas is constant, which, during compression or expansion, isn’t the case. The Ideal Gas Law states $PV = cT$, and for simple gases at low pressures and temperatures, this is more accurate. In Boyle’s time, temperature was not understood, and this law could not even be stated.
The eye ... has fire within it
- Theophrastus, c 300 BCE

First Expedition: Vision

We began the book with a question: how can we know number? The more general question is how can we know anything? The philosophy of knowledge is epistemology, and though we’re here for science not philosophy, a small break will help set out the issues.

Written work on epistemology traces at least as far back as Plato and Aristotle, about 300 BCE. Plato compared us to people living in a cave (Figure 114), chained so that we can only see shadows of the world, projected in front of us. Plato is examining the gulf: on the one side, senses, which tell us about the world, and on the other, understandings of the world. For Plato, all that the senses can tell us about reality is plausible myths: mere stories. To get to the core of reality, an individual needs to use reason. Mathematics is an essential part of that reason:

... the Platonic classification of existence [has] two orders. The higher is the realm of unchanging and eternal being ... [containing] the objects of rational understanding ... namely, arguments of mathematics and dialectic which yield a securely grounded apprehension of truth and reality. The lower realm contains ‘that which is always becoming’ – passing into existence, changing, and perishing, but never having real being. This is the world of things perceived by our senses. ... sense can only state a fact ... The reason why can only be apprehended by the higher faculty of understanding.


We’ll be after something different: not the chasm between senses and reason, but that between reality and the senses. Yet, Plato’s ‘plausible myths’ will throw a shadow over our work.

We’ll start with vision: what can vision tell us about the world? Figure 115 shows how a flower looks to us, in sunlight; then to a bee, in ultraviolet light. It’s conjectured that flowers evolved markings to direct bees to nectar; the payoff for the plant is pollination.

Yes, our vision is limited – yet, we use UV lights to do ‘bee.’ Again: Athanasius Kircher used a microscope to examine the blood of plague victims; he noted ‘little animals’ which he believed caused the disease; see Figure 116. Having a microscope to extend his vision allowed Kircher to guess at the germ theory of disease.
Before Kircher, though, there were parts of reality we couldn’t experience. Will there be whole chunks of science and medicine we can’t do, because we don’t have the right ’microscopes’? The issue reverberates through culture. Ghosts: helping and vengeful spirits are common across cultures; if we could see them we’d know whether they’re real. There’s a similar issue in modern cosmology: some theories posit universes parallel but unconnected to ours – is there a way to detect them? (see p135)

We’ll begin with experiments on vision – from the 1950’s. The article What the Frog’s Eyes Tells the Frog’s Brain (p135) discusses experiments presenting different kinds of visual stimuli to frogs, then recording which stimuli cause a particular cell in the eye to respond. The article posits these stimuli determine what a frog sees. See, however, p130. On this hypothesis, what the frog can see is: differences in contrast (possibly representing an insect standing out from the background), convexity (possibly representing the shape of prey), a moving edge and its direction (possibly a moving insect) and dimming of light (possibly indicating a predator in back). Figure 117 gives an artist’s interpretation of what a pond might look like to a frog: it’s nothing like what we would see. Is our vision also hiding much of the world?

We need to understand how vision works. Johannes Kepler in the 1600’s based his theories of planetary motion on the observations of Tycho Brahe. He knew the atmosphere distorted light, and his computations took that into account. He wondered whether the eye also distorted observations. Kepler believed that the lens of the eye focuses light on the lining at the back of the eye, opposite the lens: retina, (figure 119). To Kepler, this may have been by analogy to the artists tool for drawing in perspective, the camera obscura (Figure 118).

We now know the retina contains cells responsible for converting light to electrical charges, which, interpreted by the brain (and the retina itself!) constitute vision. Using mammalian retinas, Nobel laureate Santiago Ramon y Cajal (Figure 120) used a microscope and developed innovative cell staining techniques to elucidate the cellular structure of the retina, diagrammed in Figure 121. He wrote:

[…] the retina is a genuine neural center, a sort of peripheral cerebral segment whose thinness, transparency and other qualities render it particularly favorable to histological analysis. In fact, though its cells and fibers are essentially similar to those of other centers, they are arranged in a more regular fashion, different types of cells being distributed in distinctly different layers.

Cajal, La retine des vertebres, 1892.
Cajal is saying the retina is a protrusion of the brain. He identified ordered layers of cells (Figure 121), suggesting that the eye doesn’t just gather light: the layers he discovered process the light, sending the results on to the brain. Find the limitations in the processing and we may find the limitations of vision.

To help think about processing in the retina, we’ll compare it to something simpler: a modern high-megapixel digital camera. Perhaps camera vision will help explain human vision. These kinds of analogies are behind many kinds of attempts to understand the brain: even today, people compare our brains to computers. And indeed, a modern camera is very much a computer. However: the eye-brain system is not like a camera; vision is much more like an odd kind of movie.

In reality (and this is very obvious) human vision is video, not photography. Even when staring at a photograph, the brain is taking multiple “snapshots” as it moves the center of focus over the picture, stacking and assembling them into the final image we perceive. Look at a photograph for a few minutes and you’ll realize that subconsciously your eye has drifted over the picture, getting an overview of the image, focusing in on details here and there and, after a few seconds, realizing some things about it that weren’t obvious at first glance. Roger Cicala, The Camera Versus the Human Eye, PetaPixel Nov 17, 2012.

What vision does, then, is a complex interaction between eye and brain; scenes aren’t recognized all at once; the brain will move the eye to detect other parts of the visual field, and then will put all those together, somehow, to construct perception.

However, the individual ‘snapshots’ could be thought of as analogous to a camera picture, and we’ll continue with the comparison.

Both the camera and the eye have a lens to focus light on a region that can recognize the incoming light and translate the color and intensity of the light into an electrical response; we’ll call the response a signal. This is analogy, and by itself it can distort how we think about vision; see p136.

In modern cameras, the region that takes light and changes it into an electrical signal is called a CMOS sensor (CMOS refers to both the design and the materials of a chip; these kind of chips are resistant to noise and consume little power). The chip has a rectangular array of photodiodes shown in Figure 122. When light hits the photodiode, it generates an electron. Photodiodes store electrons; the number of
electrons stored is proportional to brightness of the incoming light. Photodiodes only detect brightness, so colored lenses overlay each photodiode, and the color has be be generated by the camera processor. This makes a mosaic; Figure 123 shows a mosaic from a simple sensor, with only $180 \times 80$ photodiodes. Figure 124 shows what we’d see from a slightly better sensor. It looks blurred because the mosaic limits resolution, that is, how much detail we can see.

The retina, on the other hand, uses cells to recognizes light: rods and cones. The rods do black and white vision; cones detect red, green and blue colors. Light energy causes a molecule to decompose; in the rod, the molecule is rhodopsin. Even one photon of light causes rhodopsin to decompose within picoseconds, triggering a chain of reactions that result in a change of the charge across the cell membrane.

Back to the camera. After collecting the electrons that will be used for the picture, the electrons are led out of the sensor, and converted to a voltage. If you think of millions of electrons, you can have millions of possible voltages – far too much information for the camera. At this point, voltages are assigned to one of a small numbers of different levels. The number of levels is determined by the number of bits used in the camera circuitry; a twelve-bit sensor can handle $2^{12} = 4096$ levels. The assignment of numbers to a limited collection of levels is called quantization; see p140. We met up with quantization in Sections 8 and 9, when we discussed how computers represent numbers, and how scientific data is recorded.

If we think of a photodiode as a small box for collecting electrons, we can imagine a very bright light could overfill the box. In this case, the electrons stored in the photodiode overflow into nearby diodes. Figure 126 shows overflow; the effect is called camera bloom.

In contrast, retinal cells respond logarithmically to light: even one photon can cause a response, but brighter lights cause smaller responses. This allows us to see a wider range of dark/light than a camera; correspondingly, the output from cameras has to be logarithmically adjusted to match our eyes; the adjustment is called gamma correction.

But the retina too can also overload – at the molecular level. Rhodopsin decomposes quickly in response to light, but takes longer to rebuild. This causes what’s called the ‘theater effect’: on a very bright day, light saturates the rhodopsin; if you leave sunlight to enter a dark room (like a movie theater), the delay in rebuilding rhodopsin means there’s not enough available, so there’s a short time when you can’t see very well.

Size is another limitation: neither we nor the camera can see molecules
or even viruses; Figure 125 suggested why: the size and number of
the photodiodes limit the amount of detail. Our 20 megapixel camera
has 5384 (H) x 3752 (V) photodiodes, each is 1.12µm by 1.12µm (µm
is a micrometer, 10^-6 meters). Each photodiode gives rise to a 1.12µm
pixel or picture element, and these are equally distributed across the
picture frame. In the high-rez picture (Figure 124), light changes very
quickly as we go from one tiny pixel to the next; we call these quick
changes high-frequency. In contrast, the low-rez sensor (Figure 125)
has large pixels, or big blocks where the light is constant. This is
called low-frequency information.

The smallest object we, or our camera, can see is also an issue of
how we process high-frequency information. Each pixel in Figure
125 comes from one photodiode; we can think of it as a little box
to hold light. The camera pushes incoming light into these boxes;
information processing theory calls this sampling. Sampling is at the
foundation of all scientific/medical data gathering, as we saw in
Section 9. For now, the question is what happens when the camera
or eye tries to push high frequency information into low frequency
boxes.

We’ll give a simple analogy; for the precise details, see Chapter 1 Sec-
tion 9.3. When you try to push light into boxes and run out of boxes,
the light has to be placed in boxes that have already been used. You
can see an example of this in Figures 127 and 128: the left side of
Figure 128 shows the bricks in a castle wall, as they should look. The
right side shows the effect of putting the information into boxes al-
ready used. The effect is called aliasing; it distorts the original picture,
and can add the appearance of patterns that were never in the orig-
inal. Cameras avoid aliasing by adding a layer of material over the
sensor, blurring the picture slightly. This leaves low frequency infor-
mation alone, but limits high frequency information: tech people use
the term ‘low-pass’ filter (a filter, e.g. a coffee filter, allows small things
to get through but blocks big things. It’s a ‘small thing pass’ filter).
Since the high frequency information contains the small changes, the
filter reduces the resolution of pictures.

Unlike a CMOS sensor, rods and cones not equally distributed. The
retina has about 120 million rods and some 7 million cones. It seems
to avoid aliasing. Research suggests this is because the cells recogniz-
ing light are not arranged in a regular pattern like a CMOS sensor;
they randomly deviate from that pattern (Figure 129). See John I Yel-
lo tt Jr, Spectral Analysis of Spatial Sampling by Photoreceptors: Topological
There are also limitations in how quickly cameras and eyes can respond. For a camera to take a picture, all the accumulated electron charges in the photodiodes of the sensor are dumped, using the transistor circuitry shown in Figures 131 and 130. The dump is parallel, and goes to the camera’s central processor, which is limited in speed.

A modern camera can take 22 pictures per second; faster than that, one picture blurs into another. Engineers design delays to stop taking too many pictures too quickly. Very fast changes are thus invisible to the camera.

We have the same issue: experiments show we can’t see a difference between two pictures if they appear for less than 16ms to 13ms. So the eye could process about 60 frames per second, if it acted like a camera.

To begin with,

\[\text{at } 20/20 \text{ vision, the human eye is able to resolve the equivalent of a 52 megapixel camera (assuming a 60º angle of view). However, such calculations are misleading. Only our central vision is } 20/20, \text{ so we never actually resolve that much detail in a single glance. Away from the center, our visual ability decreases dramatically, such that by just 20º off-center our eyes resolve only one-tenth as much detail. At the periphery, we only detect large-scale contrast and minimal color. \ldots \text{ a single glance by our eyes is therefore only capable of perceiving detail comparable to a 5-15 megapixel camera (depending on one’s eyesight). However, our mind doesn’t actually remember images pixel by pixel; it instead records memorable textures, color and contrast on an image by image basis. In order to assemble a detailed mental image, our eyes therefore focus on several regions of interest in rapid succession. This effectively paints our perception. The end result is a mental image whose detail has effectively been prioritized based on interest.}
\]

For now, let’s do some comparative numbers.

i) Camera: \(2 \times 10^7\) pixels, at 12 bits, and 22 pictures per second; we get about \(1.8 \times 10^{12}\), or two trillion bits per second.

ii) Eye: 127 million, or \(1.27 \times 10^8\) receptors. At 16 bits (too low, but for comparison \ldots\), and 60 images per second, we get about \(1.2 \times 10^{11}\), or 120 billion bits per second.

In the camera, these bits either go directly to memory, or to a processor; for example, one which uses jpeg compression. Figure 132 shows that the information from rods and cones go through several layers.
of cells. The round cells at the top of the figure are ganglion cells; the ganglia are the final stop before the signal is transferred to the areas in the brain responsible for vision.

In the layers before the ganglia, the light signal is represented by changes in the charge of cellular membranes; this can be transmitted very quickly. The ganglia, however, transmit the signal as a series of pulses. A fast (mylenated) nerve fiber can transmit about 2500 pulses a second. But this is misleading, for several reasons.

First, even resting cells put out spikes; the issue is not how many spikes there are, but how much information the spikes give. Second, not all ganglion cells are the same; the experiments on the frog eye discussed on p122 were done on ganglion cells. Further experiments identified many different kinds of ganglion cells, with different rates of spiking; see the survey article Functional Architecture of the Mammalian Retina, Heinz Wässel and Brian B. Boycott, Physiological Reviews Vol. 71, No. 2, April 1991.

The computation that balances information and firing rates was performed by Kristin Koch et. al in How Much the Eye Tells the Brain, Current Biology 16, 1428 -1434, July 25, 2006. For the million ganglia of the retina, the amount of information that can be transmitted is about 875,000 bits per second. As the retina is putting out 120 billion bytes per second, there’s an enormous mismatch. Some of the information from the eye has to be eliminated before it reaches the brain.

The layers of cells between the rods and cones and the ganglion cells (the middle layer in Figure 133) seem to be like the jpg compression software in a camera: they reduce, or compress, what comes from the rods and cones:

…the world that you see is not the world that exists – it has been heavily retouched by your retina. The modified image uses less computational power than the raw form because, before being sent to the brain, it is packaged into more than 30 representations that emphasize specific features of the visual scene. The content of these messages is partially understood.


Researchers are now deciphering those partially understood messages, though much work remains. The first layer of cells that respond to changes in rod/cone membrane charge is the layer of horizontal cells. These cells connect a rod or cone to those nearby. One message they can carry is, roughly ‘Wow, it sure is bright. Let’s tone things down.’ More accurately,

The horizontal cell […] measures the average level of illumination falling
upon a region of the retinal surface. It then subtracts a proportionate value from the output of the photoreceptors. This serves to hold the signal input to the inner retinal circuitry within its operating range, an extremely useful function in a natural world where any scene may contain individual objects with brightness that varies across several orders of magnitude. The signal representing the brightest objects would otherwise dazzle the retina at those locations, just as a bright object in a dim room saturates a camera’s film or chip. . . .


Another function that we understand is, roughly, ‘Hey! I’m all lit up. If you guys would quiet down, everyone could see me.’ Masland again:

[…] objects neighboring a bright object have their signal reduced […] in the extreme, the area just outside a white object on a black field is made to be blacker than black. This creates edge enhancement […] .

Edge enhancement allows us to pick out objects from a background. In this way, the eye can construct the appearance of individual objects (for us, an important step in counting individual objects!).

The next layer consists of bipolar cells, then amacrine cells which connect directly to the ganglia. There are back-and-forward connections among these cells, and their interaction is complex:

In the inner retina, roughly 42 types of mostly inhibitory amacrine cell modulate bipolar cell output. Although some amacrine cell circuits have been studied in depth, we still understand little about the general principles by which amacrine cell circuits help to decompose the visual scene into the parallel channels carried by the bipolar cell.

Katrin Franki *et. al.*, Inhibition decorrelates visual feature representations in the inner retina., Nature 542, February 23, 2017

Again, researchers use the language of signal processing: the different kinds of information generated from retinal ‘circuits’ are referred to as ‘channels’. The action of these channels ‘decorrelates’ features.

A correlation between two signals determines how much they have in common: for example, Are they simultaneous? Do they respond to the same stimulus? TV channels, for example, are decorrelated: football doesn’t blend in and out of cooking shows. The idea of uncorrelated channels carrying different kinds of information goes back to the original experiments on amphibian vision. The experimenter provides a stimulus, say a moving dot projected on a screen, and
records what shapes or sizes or speeds or directions cause a cell to respond. Different stimuli cause different cells to respond; the conclusion of this work is (roughly) that one stimulus is ‘represented’ by one group of neurons in the visual cortex; see Figure 135. Later work showed that this stimulus/response specificity begins in the retina: as Masland remarked above, compression begins in the retina.

For a discussion of ‘channels’, see p136.

But: Franki et al. point out, "a deeper understanding of the functional diversity of bipolar cells and its origin is lacking." In consequence, experimenters don’t know what stimuli activate which channels. Masland again: "The challenge is how to choose test stimuli, and how to interpret the bipolar-cell responses to them. It is unlikely that naturalistic responses can be achieved using spots of light, striped patterns... ."

Franki et al. chose a very complex visual stimulus: they used a light flickering at different rates (frequency) and changing intensity (amplitude); see Figure 136. They then used statistical techniques to detect clustering of responses. We don’t yet know what these clusters ‘mean’, so don’t know what the eye is telling the brain. The channels are statistically decorrelated, at least in the retina, but what this means is up in the air. We know the eye is withholding some kinds of information, because vision presents too much information for the system to handle. But we don’t know what; see p140.

This kind of issue arose in signal processing when pictures, video and music became digital. A single CD recording could hold only about 12 or so songs (using the ‘RedBook’ recording standard, Figure 137). To get more, engineers developed compression, like jpeg, mpeg, mp3. These lose some of the original; what we can afford is the subject called compressed sensing.

Figure 136: Chirp!

One of the signals Franke et al. used to analyze response of retinal cells to patterns of light. Vertical axis represents intensity, horizontal time.

These kinds of signals are called chirps. The sound of the word imitates the kinds of sounds made by birds; it also applies to the ultrasonic sounds made by bats.

Figure 137: CD Standard

When CD’s were developed, the standards were contained in a large red folder. These became the RedBook Standard, determining how CD’s would be made and how CD players would read them. Every RedBook standard CD has the words shown above printed on the label (BlueRay, of course, uses a different standard, which CD players can’t read).

Unsurprisingly, capturing music on a CD involves quantization, sampling, and aliasing.
Charles Darwin puts in a word here. His brother Erasmus reminded him of Plato’s allegory that we understand ideas like numbers because, between an endless series of deaths and rebirths, our soul has directly experienced abstract concepts. Darwin wrote in his notebooks that instead of a preexisting soul, we have pre-existing monkey ancestors. That is, Darwin contended perception evolved to help organisms survive and reproduce; in evolutionary terms, the way we see the world reflects what was useful as our species evolved. The flowers and the bees seem to confirm this: what we don’t need, we can’t see.

Is it true?

...it has long been assumed that neurons are adapted, at evolutionary, developmental, and behavioral timescales, to the signals to which they are exposed. Because not all signals are equally likely, it is natural to assume that perceptual systems should be able to best process those signals that occur most frequently. Thus, it is the statistical properties of the environment that are relevant for sensory processing. Such concepts are fundamental in engineering disciplines: Source coding, estimation, and decision theories all rely heavily on a statistical ‘prior’ model of the environment.


The experiments on frog vision relied on an artificial stimulus – a moving black dot, for example. A cell in the brain firing in response was considered to recognize dots moving in that direction. A model of the frog’s environment, but not a statistical one. The chirp of Franki (p129) isn’t even an attempt to represent the environment. Moreover, one cell may take part in recognizing many other motions and will stimulate other neurons throughout the brain. As Peiran Gao, Eric Trautmann et. al. note,

...how can we record on the order of hundreds of neurons in regions deep within the brain, far from the sensory and motor peripheries, like mammalian hippocampus, or pre-frontal, parietal, or motor cortices, and obtain scientifically interpretable results that relate neural activity to behavior and cognition? ...we could be completely misleading ourselves: perhaps we should not trust scientific conclusions drawn from statistical analyses of so few neurons, as such conclusions might become qualitatively different as we record more.


Figure 138: CHARLES DARWIN

Darwin in 1854, at the time he was writing The Origin of Species.
The issue of which neurons react is the coding problem; as it could be as many as $10^9$, the issue is complicated. One way to reduce the complexity is this: if a large numbers of neuron respond in a similar way (are correlated) then deciphering the coding is simpler. This is called a low-dimension response; it is the type of response that investigators found in the frog.

If the neurons all react in a unique way, the code is said to be high-dimensional; this is the dimensionality problem. Very low-dimension codes can’t distinguish different visual scenes, like a child to whom every moving vehicle is ‘car’. High dimension codes are highly sensitive to differences in a scene, as though the presence of an extra leaf could obscure a rattlesnake. This is a typical difficulty for artificial intelligences, Figure 139.

If we assume that visual systems are highly adapted to the environment, we’d believe in something like low-dimensionality. The scientific problem is in providing natural images to an animal, and recording simultaneously from thousands of neurons. An experiment recording from over 10,000 neurons was recently reported in High-dimensional geometry of population responses in visual cortex by Carsen Stringer, Marius Pachitariu et. al., Nature Vol 571, 18 July 2019.

Here we recorded the simultaneous activity of approximately 10,000 neurons in the mouse visual cortex, in response to thousands of natural images. We found that stimulus responses were neither uncorrelated (efficient coding) nor low-dimensional. Instead, responses occupied a multidimensional space . . . . These findings suggest that the population responses are constrained by efficiency, to make best use of limited numbers of neurons, and smoothness, which enables similar images to evoke similar responses.

There’s another issue. After the experiments on ‘one direction of motion=one neuron’ in the frog, it was understood that neurons recognizing direction grouped together physically:

The modular organization of nervous systems is a widely documented principle of design for both vertebrate and invertebrate brains of which the columnar organization of the neocortex is an example. The classical cytoarchitectural areas of the neocortex are composed of smaller units, local neural circuits repeated iteratively within each area. Modules may vary in cell type and number, in internal and external connectivity, and in mode of neuronal processing between different large entities; within any single large entity they have a basic similarity of internal design and operation. Modules are most commonly grouped into entities by sets
of dominating external connections. [...]. The set of all modules composing such an entity may be fractionated into different modular subsets by different extrinsic connections. Linkages between them and subsets in other large entities form distributed systems. The neighborhood relations between connected subsets of modules in different entities result in nested distributed systems that serve distributed functions.

Vernon B. Mountcastle, The columnar organization of the neocortex, Brain (1997), 120, 701-722

It’s an attractive picture: small groups of neurons clump together to process certain features of the visual environment; the brain has many different groups, each recognizing a single feature: the recognition of features is ‘distributed’. In cognitive science, there are extensions of this idea: the functions of the brain are carried out by modules, each of which performs a particular cognitive function. See Jerry Fodor, Modularity of Mind: An Essay on Faculty Psychology, MIT Press 1983, as well as Marvin Minsky, The Society of Mind, Simon & Schuster 1988.

But this picture leads to a question: if the visual field is distributed, where and how is the information re-assembled? This is called the binding problem. One way to say it is this: you’re playing tennis, and a ball is coming towards you. Part of the brain recognizes the motion, another part the spin of the ball, another part the color, another the shape.

Why do all of these stick together? Why doesn’t the color peel off from the shape? There’s some recent research that hints at what may be happening – and, once again, it has to do with dimensionality. The experiment records neuron behaviors in the visual cortex of the rat.

Figure 140 shows the idea: the eyes detect objects; the visual response is sent through nerve pathways to a specific area of the brain, where the modules responsible for recognition sit, the visual cortex. An experiment testing this was run by Carsen Stringer, Marius Pachitariu et. al. (Spontaneous behaviors drive multidimensional, brainwide activity, Science 364, 255 19 April 2019). They recorded from roughly 10,000 sites in the visual cortex of a rat moving about in a completely dark room. There was tremendous activity, of high dimension – but the activity must have certainly been noise. Was it?

In part two of the experiment, infrared cameras recorded the rats as they ran, groomed themselves, moved their heads or ears, etc. As it happened, these non-visual behaviours correlated with what had been assumed to be ‘noise’.

Recording more than 10,000 neurons in mouse visual cortex, we observed
that spontaneous activity reliably encoded a high-dimensional latent state, which was partially related to the mouse’s ongoing behavior and was represented not just in visual cortex but also across the forebrain. Sensory inputs did not interrupt this ongoing signal but added onto it a representation of external stimuli in orthogonal dimensions. Thus, visual cortical population activity, despite its apparently noisy structure, reliably encodes an orthogonal fusion of sensory and multidimensional behavioral information.

Not only was the visual cortex aware of non-visual information, areas outside of it were aware as well. It seems that visual and non-visual activities are linked from the very beginning, rather than being processed separately. How this linking works, and what it means, will need further research.

Regrettably, this makes traditional kinds of experimentation – ‘change only this, record only that’ – much more difficult; one has to analyze system-wide changes across many dimensions. An interesting question is whether our minds are capable of understanding this complexity. Some see artificial intelligence as a solution, but this also has problems: see pp 131 and 140.
Notes for First Expedition: Vision

If we think back to the Babylonian Systems A and B, p15, these mathematical models are a bit like Plato’s ‘myths’. These ancient astronomers didn’t expect their equations to capture a complete reality; they considered it a useful approximation, and hoped it will explain or lead to new truths– though these again may be only approximations.

p121 The difference between our view of a flower and that of a bee shows our eyes aren’t like those of other organisms; Figure 141 shows a more extreme example, a sea urchin. The red areas are the ‘eyes’ – more accurately, cells that respond to light. The urchin crawls about on the sea bed, detecting light without lens or eyeball, as though the whole organism were an all-directional eye.

Color vision is always trickier. Newton showed that light passing through a prism (Figure 142) could be broken up into an entire range of colors. He used the term spectrum for this range of colors (from the Latin ‘to look’, cognate to ‘spectacles’, ‘inspect’, etc). But this is a vast range of colors; most animal eyes can detect only a few of them. For humans, the main colors are Red, Green and Blue: RGB. The human brain then does a bit of mathematical juggling: it takes the relative amounts of R, G and B, and constructs new colors, as an artist would mix paints; see Figure 143.

The mantis shrimp (Figure 144) has receptors for twelve distinct colors. As an analogy, in Figure 142 the prism breaks light into its different colors, and the black dot is a photoreceptor for each color. These 12 receptors form a representation of the colors present; called a spectral analysis. “Why use 12 color channels when three or four are sufficient for fine color discrimination? Behavioral wavelength discrimination tests … revealed a surprisingly poor performance, ruling out color vision that makes use of the conventional … coding system. Instead, our experiments suggest that stomatopods use a previously unknown color vision system based on temporal signaling combined with scanning eye movements, enabling a type of color recognition rather than discrimination.” Hanne H. Thoen et. al., A Different Form of Color Vision in Mantis Shrimp Science vol 343 24 January 2014; see also Michael F. Land and Dan-Eric Nilsson, Animal Eyes, Oxford University Press; 2nd edition 2012. Although the mantis vision system is not fully understood, these authors suggest that the shrimp scans across its receptors one at a time, rather than merging them as the human brain does. This requires less co-operation amongst the receptors; it also doesn’t require a layer of cells to integrate the responses of the different receptors. This makes it likely that the shrimp can...
respond more quickly to prey.

These exotic kinds of visions are studied as *Visual Ecology*; species take basic structures like light sensors (photoreceptors), and then those structures become adapted towards detecting the parts of its environment assisting reproductive fitness. Since species have different needs, it’s no surprise our unaided eyes can’t detect the reality of other organisms; see the article in Science News, Vol. 190, No. 2, July 23, 2016, p 35. For a detailed study, see Thomas W. Cronin *et. al.*, *Visual Ecology*, Princeton University Press 2014.

Many cultures believe in an invisible world of ancestors, of helping and harmful spirits. The Japanese festival of *obon* is a contemporary survival of such beliefs. During the festival, families visit the gravesite of their ancestors and symbolically carry the spirits back home, where they are offered food.

We dismiss the belief in invisible worlds as superstition, but some modern cosmological/quantum theories theorize many universes. The *many worlds* interpretation of quantum mechanics posits that every possible outcome of an experiment actually occurs in some world; new universes are constantly being birthed (see for example Max Tegmark, *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*, Vintage, 2015). The inflationary theory of the origin of this universe (see Alan Guth’s *The Inflationary Universe*, Basic Books, 1988) suggests many universes sprouted at the same time as ours, with different laws of physics. See also Yasunori Nomura, *The Quantum Multiverse*, Scientific American June 2017.


For an introduction to how the brain processes visual information, see the book *From Neuron to Brain*, by John G. Nicholls and A. Robert Martin, p126

The frog work is from the late 1960’s; we now know a great deal more about vision.

*In the vertebrate visual system, all output of the retina is carried by retinal ganglion cells. Each type encodes distinct visual features in parallel for transmission to the brain. How many such output channels exist and what each encodes are areas of intense debate . . . we show that the mouse retina harbours substantially more than 30 functional output channels.* Tom Baden *et. al.*, *The functional diversity of retinal ganglion cells in the mouse*, Nature(2016) 529 p345.
Baden et. al. note that some of the visual features (or ‘output channels’) extracted for presentation to the brain are: local motion, direction of motion, and illumination. Most of the channels, however, are kinds of information we don’t yet understand, and perhaps can’t conceptualize; see the note on p140, below.

Similarly,

*The retina actually performs a significant amount of preprocessing right inside the eye and then sends a series of partial representations to the brain for interpretation. We came to this surprising conclusion after investigating the retinas of rabbits, which are remarkably similar to those in humans.*


The retina sends a series of images to the brain; since the images change over time, the authors call these ‘movies’ (Figure 146). They identify twelve different kinds of movies that the retina generates and sends on: some show the edges of a scene, some show brightness, or reflectance, and, as in the mouse, some show information we have no name for. The brain then integrates these movies into what we call vision.

In using the term *signal* here we’re using the language of modern information processing. Much of it was developed in Bell Labs, a research unit of the AT&T corporation. The research was concerned with very general properties of electrical *signals*, how those could *encode* voice, or pictures, or ... *information*, and *transmit* it over wires, undersea cables, satellites, microwave systems ... *channels*. Initially developed for telegraph and telephone systems, it was extended to computers. This language is currently used to describe both computers and nervous systems; we’ll use terms like *signal*, *information*, *channel* and *signal processing*. We also will be careful: nerves and wires, brains and computers, are not the same. *Signal processing* sounds very scientific but is just an analogy. Nerves are much more complicated than wires. Here’s more detailed discussion about the use of the word ‘channel’ in talking about the retina and ganglion cells. Those uninterested can skip directly to p140.

The word ‘channel’ comes from the Latin *canalis*, ‘pipe’. Information-theory channels take an input, and with some probability produce an output (see Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc, 1991). In a telegraph: you key the input ‘dot-dash’, the channel is miles of wire, and, 90% of the time, ‘dot-dash’ is the output. You could think of the body’s system for regulating blood pressure as a channel; see Figure 147. A *baroreceptor*
on a vein or artery senses pressure; there’s your input. The receptor translates this to an electrical signal, which is sent to the brain. The brain controls pressure by signals to the heart through sympathetic or parasympathetic nerves. These release hormones which either slow (norepinephrine) or increase (epinephrine) the heartbeat; those hormones are the output of the channel.

‘Channel’ as ‘pipe’ may be too simple an idea for the above. Information theory had in mind something more like a headphone cable. An mp3 player emits varying voltages representing sounds; the cable transmits these as electromagnetic waves, and the headphone changes this to sound. In this example, it’s very clear what input and output mean.

Clear, but inaccurate. The easiest way to see the problem is by thinking about a prism (Figure 148). The low frequency red light travels more slowly through glass than the high-frequency blue light, so the two frequencies are bent slightly differently by the glass, and the prism separates them out.

Earphone cables act like prisms: bass and treble sounds travel at different speeds, and arrive at the ear at different times. In severe cases, this produces a mushy, blurred sound. People who are serious about music (the author) spend tons of money (not the author) on cables that minimize blurring. Even then, you’re advised to 'burn-in' your cables, by playing a hundred or so hours of the music you like. Apparently, the cables can adapt (or de-adapt) themselves to your music.

If you think of a channel as a pipe, the idea ‘water in one end, water out the other’ is appealing and simple. Take something more complicated, even just voltages traveling through wires, and the simplicity is lost.

In fact, the simplicity was never there. Water travels more slowly nearer the pipe, and more quickly though the center. And, as most water contains dissolved solids, those can wind up encrusting the inside of the pipe, slowing things further. So even the simplest kind of pipe adapts to its inputs.

We can think about ‘channels’ made of neurons, and use information theory, as with ganglion cells in the retina, but these channels are complex. We’ll start with the stereotype of a brain cell, shown in Figure 149. There’s the purple cell, with its nucleus in black, and little finger-ish extensions called dendrites, and little yellow blobs called glial cells. Let’s start with the blobs.

*Originally, scientists didn’t think they did anything. Until the last 20 years, brain scientists believed . . . that glia were kind of like stucco and...*
mortar holding the house together. They were considered simple insulators for neuron communication.


Recent research shows glia have their own energy and signaling systems and can direct or destroy the growth of neurons (see Darran Yates, *Glial messaging*, Nature Reviews Neuroscience 18, 2017).

Of course ‘brain cells’ don’t float around in splendid isolation, making occasional connections with other brain cells. The figure below is from Kasthuri et al., *Saturated Reconstruction of a Volume of Neocortex* 2015, Cell 661 July 30, 2015. It demonstrates the complex mesh of support systems for cortical cells; some of the support cells are shown in isolation below.

Not only is the ‘brain cell’ surrounded by support cells, it interacts with them and is influenced in turn. Dendrites (see Figure 149) are a good example. In our simplified picture, dendrites are extensions from the main cell; they can interact with other neurons. Older theories viewed nerve cells as wired to each other through dendrites.
This ‘wiring diagram’ is a bit simplistic, as the figure above illustrates. On the left, a section of the cortex of a rat brain; all kinds of cells, indicated by colors, are packed tightly together; on the right, the cell types. They interact with the neuron, and other neurons, and with the other kinds of cells, in ways that are not understood. ‘Wiring’ may be too simple a term for these interactions.

We’re interested in the retina. Recall Franki’s comment p128 "amacrine cell circuits help to decompose the visual scene into the parallel channels carried by the bipolar cell."

We could read Franki and think of simple channels that can only process certain fixed kinds of information. Then what the eye can see is limited by the channel, and there are kinds of things we simply can’t see. Again, rather like a television news channel, where see only what the politics of that station wants us to know. But we could think instead of more complex channels, which are self-adapting. In that case, the retina itself might be capable of learning, of seeing new things.

Recent work on the retina and its interactions with adjacent cells suggests something like that. Cells in the retina require energy to convert light to changes in cell potential; that energy is carried in the highly reactive element, oxygen, which is supplied to the cells by blood vessels. As it happens, the network of blood vessels is partly controlled by the amacrine cells. Yoshihiko Usui et. al. examined the relationship between these. Their results:

i) "Amacrine and horizontal cells form neurovascular units with capillaries . . ."

ii) "Amacrine cell and horizontal cell derived VEGF is essential for neurovascular-unit formation . . ."

VEGF is Vascular Endothelial Growth Factor, a protein promoting the growth of blood vessels. See Yoshihiko Usui et. al. Neurovascular crosstalk between interneurons and capillaries is required for vision, The Journal of Clinical Investigation, Volume 125 Number 6 June 2015.

It seems the neurons construct their own environment. It stands as a warning: words like pipes, channels, signals and information suggest a fixed system, which may not exist. The kinds of metaphors we choose influence our science – another example of what we can and cannot see! All this brings us back to epistemology and the discussion on p134, as well as Plato’s remark that all our senses can tell us about the world is a kind of myth.
p124 The number of electrons that a photodiode captures could be anywhere from one to millions; much of this is meaningless, as the eye can’t perceive the difference between 1,000 versus 1,001 electrons. Circuitry assigns data from the photodiode to one of $2^{10}$ different levels; the process is called quantization. We won’t estimate the exact number of electrons, but we’ll be off by at most $2^{-10}$.

p129 The discussion above, and that on p129, asks us to think: can we understand what the different channels of the retina are telling the brain? As we saw, the ‘channels’ are complex self-constructing systems of their own. They also appear to compress visual signals, before passing them to the brain. The engineering equivalent would be jpeg or mp3 compression. But those were designed by engineers, using well-known techniques from signal processing. jpeg, mp3, are cross-platform, designed for use in many different kinds of electronics. So engineers need to know, in advance, how well they work. This is typical engineering: there are specifications and you design to meet them. And, if there’s a failure, it can be traced.

Vision is radically different: it evolved, over hundreds of millions of years, and performance was shaped by reproductive success (including survival). System failures are diseases and blindness, but even today, after years of medical progress, there are kinds of blindness we don’t understand and can’t treat.

Now we’re trying to look inside, see what makes it go. Should we expect to understand it? We suspect the very simplest mathematics of signal processing and decorrelated channels might not apply. What else do we have?

In contemporary technology, the closest analogy to evolution is artificial intelligence, AI, which uses non-classical engineering. An example is Google’s work on what it calls deep learning. Google engineers taught a computer to play the game of Go as well as the other video games on early Atari machines (see Figures 150 and 151).
The engineers applied deep learning in neural networks – brain-inspired programs in which connections between layers of simulated neurons are strengthened through examples and experience. It first studied 30 million positions from expert games, gleaning abstract information on the state of play from board data, much as other programmes categorize images from pixels (see Nature 505, 146-148; 2014). Then it played against itself across 50 computers, improving with each iteration, a technique known as reinforcement learning.


These techniques were evolved: the computer reprogrammed itself through its experience of what worked and what didn’t. Engineers can experiment with the program, but can’t look at the code to ‘fix’ anything.

Getting computers to play games isn’t the real goal of deep learning; the goal is to program self-driving cars, web browser page-ranking schemes and speech recognition. And to build ‘an M.D. in a box’.

Will we trust computer doctors?

Paul Voosen writes of a programmer who worked on deep learning to help diagnose pneumonia (The AI Detectives, Science 357, 9 July 2017 Issue 6346). “In general, sending the hale and hearty home is best, so they can avoid picking up other infections in the hospital. But some patients, especially those with complicating factors such as asthma, should be admitted immediately. […] disturbingly, he saw that a simpler, transparent model […] suggested sending asthmatic patients home.” The programmer wonders what other mistakes the computer might be making.

Apple’s Siri, Amazon’s Alexa, and Netflix’s movie recommendations all use deep learning. Some credit companies also use it to determine who is worthy of a loan. What if there’s a mistake?

In 2018, Article 13 of the European Union’s General Data Protection Regulation law will take effect; it offers specific legal protections against deep learning:

**Paragraph 1 […] the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing.**

**Subparagraph (f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.**
That is, the computer can be required to explain its decisions. This is an active field of research; see the above article by Voosen in Science. Our question is whether we can use these techniques to understand vision, or more generally, use deep learning to do science.

Eventually, some researchers believe, computers equipped with deep learning may even display imagination and creativity. “You would just throw data at this machine, and it would come back with the laws of nature,” says Jean-Roch Vlimant, a physicist at the California Institute of Technology in Pasadena.

The Black Box of AI Nature, 538, October 6, 2016.

Here’s a thought experiment: what if we’d invented statistics and deep learning before we invented calculus? A deep learner has centuries of data on the positions of all the planets. Would we get all the laws of physics? Or perhaps, there wouldn’t be concepts or laws, just statistical correlations.


The example concerned the motion of the planet Mars, as seen from the earth. As shown in Figure 152, the apparent motion of Mars, as seen from the earth, is complex: Mars appears to stop and turn around. This is called retrograde motion. The figure gives a modern interpretation: as the Earth and Mars circle the sun, the difference in their speeds makes one catch up, then surpass the earth, giving the illusion of retrograde motion. This interpretation relies on the theory that the planets circle the sun: the heliocentric model of planetary motion.

Raban Iten, Tony Metger, et. al. took the observed position of the Sun and Mars, gave it to the AI, and

. . . our network learns to compress experimental data to a simple representation and uses the representation to answer questions about the physical system. [...] For example, given a time series of the positions of the Sun and Mars as observed from Earth, the network discovers the heliocentric model of the solar system - that is, it encodes the data into the angles of the two planets as seen from the Sun. Our work provides a first step towards answering the question whether the traditional ways by which physicists model nature naturally arise from the experimental data without any mathematical and physical pre-knowledge, or if there are alternative elegant formalisms.
The following critique is from the article *Are Neural Networks About to Reinvent Physics?* by Gary Marcus and Ernest Davis, in NAUTILUS November 21 2019:

*The trouble is that it is entirely misleading to say that their neural network infers that ‘the Earth and Mars revolve about the sun.’ The neural network doesn’t actually understand that anything is revolving around anything, in a geometric sense; it has no sense of geometry and no idea what it would mean to revolve. All the neural network does is to extract the two numerical parameters involved; it has no idea that these represent angles from some fixed central point. As far as the network is concerned, these could be time-varying masses, or electric charges, or angles from two different central points. Correlations between data sources were extracted, but the system made no inferences about how those data sources related to the world; it is the human scientists who identify these as the angles of the Earth and Mars as measured from the sun, and who abstract the facts that such parameters are best interpreted as orbits. All of the real work of Copernicus’ actual discovery is done in advance; the system was a calculator, not a discoverer.*

*Further, in the synthetic data the authors generated, the Earth and Mars move in constant-velocity circular orbits in the same plane. In the real solar system, things are notably trickier…*

There are other issues: data on the relative positions of Mars and the Sun were chosen out of the vast collection of data on the positions of celestial objects visible in the sky. No AI would know this; the Sun, Mars, planets, orbits, angles are all human constructs, added ontop of the pure data.

How much does it matter that ‘the system was a calculator, not a discoverer’? Calculators are great tools, though, as we saw in Section 8 on computer numbers, one has to have an intuition about where calculators might get into trouble. That kind of intuition is often lacking in AI: we don’t know what the calculator is doing.

There’s a second issue: humans have reasons for their computations. The planets, their positions, their risings and disappearances, were part of religions and astrologies. Today, the position of the moon can be used to date holidays: Easter, Holi, Yom Kippur, Chinese New Year, Ramadan and Eid-al-Fitr: all of these depend on the appearance of the moon. Additionally, once we had a theory of gravity and understood that tides were caused by the motions of the moon, we use that to predict the tides.
Figure 153 shows a collection of words written in one of the three writing systems used in modern Japan. An AI might ‘learn’ to identify these words after being exposed to thousands of variants. It might even pronounce the word correctly. But then what?

A human, in this case, the author, could use these words to read, or converse, to write, to purchase items from a grocery or clothing store. And the human would notice very quickly that the two characters at the end of each word are identical, are pronounced ‘mono’, and mean, roughly, ‘things’. Thus, drink things, things cooked in a nabe pot, clothing things, pickled things. The human could then go on to read, and construct, new words.

It brings us back to Wittgenstein (p36): the meaning of a word is embedded in the cultural, human context of the uses of that word; the meaning is our interactions. Is the same true for laws of nature, physics, biology: are they just embedded in human culture? Can they mean anything outside that?

 [...] two aspects of human conceptual knowledge have eluded machine systems. First, for most interesting kinds of natural and man-made categories, people can learn from just one or a handful of examples, whereas standard algorithms in machine learning require tens or hundreds of examples to perform similarly ...

Second, people learn richer representations than machines do, even for simple concepts, using them for a wider range of functions, including creating new exemplars, parsing objects into parts and relations, and creating new abstract categories of objects based on existing categories. Brenden M. Lake et. al., Human-level concept learning through probabilistic program induction Science 350, 11 December 2015 Issue 6266.

As with Plato, this challenges what we mean by "understanding." And, with Plato, we have to ask: if understanding doesn’t come from our senses, where does it come from?
List of Figures

Chapter 0

p6 Figure 2 ECG machine picture from

p6 Figure 3 The first human ECG. From Einthoven W. The galvanometric registration of the human electrocardiogram. Pflug’s Arch f.d. ges Physiol 1903;99:472-480.

p7 Figure 4 Picture of ECG indicating irregular heartbeat, from

p7 Figure 5 Picture of portable ECG unit from

p7 Figure 6 is from Alfonso, Valentino X ECG QRS Detection, in Tompkins, Ed. Biomedical Digital Signal Processing, C-Language Examples and Laboratory Experiments for the IBM PC, Prentice Hall PTR, 1995.


p8 The drawing linking the regions of the heart with their electrical signature is from Frank H. Netter, Heart (CIBA Collection of Medical Illustrations, Volume 5, CIBA 1978.

Chapter 1

p10 Figure 7 from Simon Hunt, Jason Low and K. C. Burns, Adaptive numerical competency in a food-hoarding songbird, Proc. R. Soc. B (2008) 275, 2373-2379.

p11 Figure 10 is from http://www.gregapodaca.com/numerography/files/020.html

p11 Figure 9 is from http://afrolegends.com/2013/08/29/the-ishango-bone-craddle-of-mathematics/

p11 The image in Figure 11 is by Ben M. Harvey, Utrecht University.

p12 Figure 12 is from T. Cuyler Young, Jr., of the Royal Ontario Museum, Toronto, Canada.

p12 Figure 13 is licensed under Public Domain via Commons https://commons.wikimedia.org/wiki/File:Zero_in_Rod_Calculus.png#/media/File:Zero_in_Rod_Calculus.png
Figure 14 is by Gisling, licensed under Public Domain via Commons https://commons.wikimedia.org/wiki/File:Zero_in_Rod_Calculus.png#/media/File:Zero_in_Rod_Calculus.png

Figure 15 of Qin era imperial tomb is from http://www.chinaescapade.com/travel-guide/xian/emperor-qin-terracotta-army.html

Figure 16 of Gwalior numerals is by the author.

Figure 17 of Al Khwarizimi is from https://www.fisikaislam.com/2015/05/muhammad-ibn-musa-al-khwarizmi.html.

Figure 18 is an illustration by Alexander Marshak, and is from https://www.awn.com/animationworld/animators-belong-ages

Figure 19 showing a cuneiform tablet is from https://www.christies.com/lotfinder/Lot/an-important-mesopotamian/proto-cuneiform-clay-administrative-tablet-4585441-details.aspx

Figure 20 of an insurance estimate is from the author.

Figure 21 is from http://www.livius.org/articles/person/kidinnu-the-chaldaeans-and-babylonian-astronomy

Figure 22 showing the Babylonian Systems A and B is from Mathieu Ossendrijver, Babylonian computational astronomy, Altorientalisches Seminar Univ. Tübingen 2006, https://www.mps.mpg.de/4556833/sss-2006-ossendrijver-print.pdf


Figure 24 showing translation of a Yahad scroll, is from Eshbal Ratzon and Jonathan Ben-Dov, A Newly Reconstructed Calendrical Scroll from Qumran in Cryptic Script, Journal of Biblical Literature, Volume 136, Number 4, 2017, pp. 905-936

Figure 25 showing infant counting is from https://www.newscientist.com/article/dn20142-toddlers-know-counting-rules-at-18-months/

Figure 28 The page from Euclid’s Elements is from Sir Thomas Little Heath’s The Thirteen Books of the Elements, Vol. 1: Books 1-2, Dover Publications; 2 edition 1956).

Figure 30 The picture proof involving squares is from https://en.wikipedia.org/wiki/Square_number#/media/File:Square_number_9.png, public domain.
p32 Figure 34 is the frontispiece of the book Leviathan by Thomas Hobbes; engraving by Abraham Bosse, public domain.

p36 Figure 36 photo of Wittgenstein is by Moritz Nühr, an Austrian photographer also known for his portraits of the composer Gustav Mahler and the artist Gustav Klimt. From the archives of the Austrian National Library, and is public domain.

p37 Figure 37 of a paramecium is from https://commons.wikimedia.org/wiki/File:Paramecium.jpg, by Barfooz from the English Wikipedia. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License.

p38 Figure 38 of a chromosome is from https://www.genome.gov/about-genomics/fact-sheets/Chromosomes-Fact-Sheet

p39 Figure 39 of Oswald Avery is cropped from a Rockefeller Institute for Medical Research staff photograph. Source=http://profiles.nlm.nih.gov/CC/A/A/L/P/_/ccaalp_.jpg Date=1937 Public Domain

p39 Figure 40 of a coloring of a map is from https://commons.wikimedia.org/wiki/File:Four_Colour_Map_Example.svg#file based on a this raster image by chas zzz brown on en.wikipedia. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License.

p42 Figure 41 of Ahmes is from the University College, London, http://www.ucl.ac.uk/museums-static/digitalegypt/lahun/uc32159.html

p45 Figure 42 of Lui Hui is from http://www.slideshare.net/sbrierton/liu-hui

p46 Figure 43 of Lui Hui’s construction for $\pi$ is from CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=4322998

p47 Figure 44 Painting of Euclid from Galleria Nazionale delle Marche, Italy; public domain. This picture is meant to represent the famous mathematician Euclid of Alexandria, who was, in medieval times, wrongly identified with Euclid of Megara, the disciple of Socrates.

p48 Figure 47 of Ashurbanipal is from https://commons.wikimedia.org/wiki/File:Table_of_Shamash_relief.jpg licensed under Creative Commons.

p48 Figure 48 of the Pythagorean theorem is from the personal collection of Bill Casselman; see http://www.math.ubc.ca/ cass/Euclid/ybc/ybc.html. The original tablet is in the Yale Babylonian Collection; see http://nelc.yale.edu

p51 Figure 50 of diffraction rings is from https://commons.wikimedia.org/wiki/File:Laser_Interference.JPGdiffract

Figure 51 illustrating surreal numbers is by Lukas Lansky

Figure 52 photo of John Conway is from Princeton University Office of Communications, by Denise Applewhite

Figure 53 on Archimedes’ estimates for π is from http://mrstevesscience.blogspot.com/2011/03/help-6th-graders-figure-out-for.html

Figure 54 on a geometric proof of the binomial expansion is from http://www.mathaware.org/mam/oo/master/essays/B3D2/binomial.html

Figure 56 of Cauchy is from http://www-groups.dcs.st-and.ac.uk/history/PictDisplay/Cauchy.html

Figure 57 is from a portrait of Archimedes in the Museo del Prado, Spain. This is from https://www.cs.drexel.edu/crorres/Archimedes/contents.html

Figure 59 of the Cray computer is from macobserver.com/tmo/article/The_Fastest_Mac_Compared_to_Todays_Supercomputers

Figure 60 of the Sunway Taihen Light is from http://www.gizmag.com/supercomputer-china-taihuilight-processor-petaflops/43982/

Figure 61 of Lara Croft is from https://www.flickeringmyth.com/2015/06/e3-gameplay-trailer-for-rise-of-the-tomb-raider/rise-of-the-tomb-raider-3/

Figure 62 Xbox is from http://www.xbox.com/en-us/xbox-one-x

Photo of ancient ruler, 63, is from http://www.factknow.com/2016/03/some-of-greatest-discoveries-made-by.html

Figure 64 of Japanese money is taken from http://sputniknews.com/business/20150903/1026525213.html

Figure 65 of crystalline structure is taken from http://www.ebsd.com/introduction

Figure 66 Picture of child with money envelope from https://www.sassandbelle.co.uk/blog/2015/02/19

Figure 67 Picture of British length standards from The Metrology Handbook, American Society for Quality Measurement, by Bennett, Keith Bucher, Jay L., ed.

Figure 47 Image of Sun God from https://en.wikipedia.org/wiki/Measuring_rod

Figure 69 Picture of MMDA tablets is copyrighted by Erowid.org. See https://erowid.org/chemicals/mdma/mdma.shtml
Figure 70 Gravity map from http://www.nasa.gov/vision/earth/lookingatearth/earth_drag.html

Figure 71 Mars Orbiter picture from NASA JPL.

Figure 72 Photo of Jocelyn Bell in 1967 from http://campus.pari.edu/radiosky/lessons/pulsars/02.shtml

Figure 73 Crab Nebula photo from http://imagesci.com/img/2013/11/crab-nebula-1613-hd-wallpapers.jpg

Figure 74 of the phospholipids in a cell membrane on is from Wikipedia, derivative work: Dhatfield (talk)Cell_membrane_detailed_diagram_3.svg;
*derivative work: Dhatfield (talk)Cell_membrane_detailed_diagram.svg;
LadyofHats Mariana Ruiz - Cell_membrane_detailed_diagram_3.svg,

Figure 75 Diagram of airplane spar from https://www.quora.com/What-is-aircraft-wing-spar

Figure 76 Launch Complex photo from http://www.hq.nasa.gov/office/pao/History/alsj/misc/apmisc-SAT-3-24.jpg This file is in the public domain in the United States because it was solely created by NASA.

Figure 77 Deserted Launch Complex photo from http://zimmer.csufresno.edu/ fring-wal/ksc.html

Figure 78 Image of circular folds in the small intestine from https://s3.amazonaws.com/classconnection/699/flashcards/3329699/png/plicae-14C4570528C65150E72.png

Figure 79 Sketch of villi from https://commons.wikimedia.org/wiki/File:Esquema_del_epitelio_del_intestino_delgado.png

Figure 80 Sketch of single villus from http://biology4ibdp.weebly.com/61-digestion.html

Figure 81 Photo of Peter Higgs receiving Nobel Prize is from https://www.nobelprize.org/prizes/physics/2013/higgs/photo-gallery/

Figure 82 Overhead photo of the Aricebo Radio Observatory by H. Schweiker/WIYN and NOAO/AURA/NSF. Original source : NSF - Download Images, Public Domain, https://commons.wikimedia.org/w/index.php?curid=16005203


Figure 85 CPU Intel Pentium A80501 with GoldCap, 66 MHz,
Vcore = 5V, SX837 = with FDIV-Bug. Konstantin Lanzet - CPU Collection Konstantin Lanzet. Permission details This work is free and may be used by anyone for any purpose. If you wish to use this content, you do not need to request permission as long as you follow any licensing requirements mentioned on this page. Wikimedia has received an e-mail confirming that the copyright holder has approved publication under the terms mentioned on this page. This correspondence has been reviewed by an OTRS member and stored in our permission archive.

p91 Figure 86 of the autographed Pentium chip is from https://www.viva64.com/en/b/0454/.

p92 Figure 87 of the (erroneous) structure of an ABC transporter molecule is from https://bl831.als.lbl.gov/mcfuser/publications/New%20Journal%20Papers/Reyes_Science.pdf

p93 Figure 88 of the Mars Orbiter flyby is from http://www.teoti.com/funny-stuff/130568-great-miscalculations-the-french-railway-error-and-10-others.html

p93 Figure 89 of the Slepnir project is from https://wikivisually.com/wiki/Sleipner_A

p94 Figure 91 of the Sydney Opera House is from https://www.ices.utexas.edu/about/news/254/

p96 Figure 93 of the IBM System/360 computer is from https://www.theregister.co.uk/2014/04/07/ibm_s_360_50_anniversary

p100 Figure 96 of the Patriot Missile System is by is by Anthony Sweeney of the U.S. Army

p100 Figure 97 of the SCUD missile is a work of a U.S. Army soldier or employee, taken or made as part of that person’s official duties. As a work of the U.S. federal government, the image is in the public domain. https://commons.wikimedia.org/wiki/File:AlHussein-Strike.JPEG

p103 Figure 98 of a lookup table for performing computer arithmetic is from http://i.stanford.edu/pub/cstr/reports/csl/tr/95/675/CSL-TR-95-675.pdf

p104 Figure 99 of the mathematician John von Neumann is from https://www.eldiario.es/turing/John-Neumann-revolucionando-computacion-Manhattan_o_380412943.html
Figure 103 of Torricelli’s apparatus is from French wikipedia, original upload 21 Septembre 2004 by Jean-Jacques MILAN fr:Image:Baro 1.png. This work is in the public domain in its country of origin and other countries and areas where the copyright term is the author’s life plus 70 years or less.

Figure 104 of a shockabsorber is by Avsar Aras and is posted at https://en.wikipedia.org/wiki/Shock_absorber. This file is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

Figure 105 is a picture of an air-based shock absorber manufactured by KYB. It is at https://www.kyb.com

The table of numbers is a page from Robert Boyle’s book New Experiments, Physico-mechanickal, touching the Spring of the Air, LONDON, Printed by Miles Flesher for Richard Davis, Bookseller in Oxford, MDCLXXXII, taken from https://www.e-rara.ch/zut/collections/content/titleinfo/5038753.

Figure 112 of the mean speed theorem is by ElApro, at https://commons.wikimedia.org/wiki/File:MertonRuleOresme.jpg This file is licensed under the Creative Commons Attribution Share Alike 4.0 International license.

Figure 113, the graph of a time series. is from William Playfair’s Time Series of Exports and Imports of Denmark and Norway, at https://en.wikipedia.org/wiki/William_Playfair This media file is in the public domain in the United States.

Figure 114 of Plato’s cave is from http://www.philosophyzer.com/the-allegory-of-the-cave-by-plato-summary-and-meaning/

Figure 115 is from http://www.howplantswork.com/2008/11/30/flowers-what-you-see-versus-what-the-bees-see

Figure 116 Picture of protozoa, etc from https://delftschoolmicrobiology.weblog.tudelft.nl/files/2015/01/Henry-Baker-s.jpg

Figure 117 is a still from the movie ‘What the Frog Sees’ by Skip Battaglia. Video at http://www.visualnews.com/2012/10/16/the-frog-does-not-see-what-does-not-move-to-the-frog-to-move-is-to-exist/

Figure 118 is from http://mvslim.com/islam-and-science-ibn-al-haytham-the-eye-of-physics/

Figure 119 Sketch of the retina by Helga Kolb, from http://webvision.med.utah.edu/imageswv/Sagschem.jpeg, Creative Commons Noncommercial license. Copyright Webvision.
Figure 120 Picture of Santiago Ramon y Cajal from http://www.robotspacebrain.com/tag/santiago-ramon-y-cajal/

Figure 121 Cajal’s drawing of the retina from from https://commons.wikimedia.org/wiki/File:Cajal_Retina.jpg

Figure 122 Photo of CMOS chip by Natural Philo - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24805309

Figures 123 to 124: The picture Colorful spring garden.jpg is by Anita Martinz from Klagenfurt, Austria. See the additional attributions at https://commons.wikimedia.org/w/index.php?curid=27188384

Figure 126 is by the author, taken with a Leica M8 and the very sharp Summicrom 50mm f2 lens. Detail of the seawall in Vancouver, British Columbia.

Figures 127 and 128 is are from http://www.thephotoforum.com/threads/having-the-anti-aliasing-filter-removed.178121/aliasing

Figure 129 of the distribution of rods and cones in the retina is by Mark Fairchild, available at https://en.wikipedia.org/wiki/Retinal_mosaic#/media/File:ConeMosaics.jpg This is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

Figure 132 is from https://en.wikipedia.org/wiki/Retina#/media/File:Retina-diagram.svg, and is a version of Cajal’s diagram, by Anka Friedrich.

Figure 135 is from http://braintour.harvard.edu/archives/portfolio-items/hubel-and-wiesel

Figure 136 is taken from Figure 1 of Katrin Franke et. al., Nature 542, February 23, 2017

Figure 137 on the RedBook standard is from http://www.testedtechnology.co.uk/datasheet-recorded-audio-formats/

Figure 138 of Charles Darwin is by Henry Maull (1829Ð1914) and John Fox (1832Ð1907), derivative work: Beao - Charles_Darwin_seated.jpg, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11264065

Figure 139 of a stop sign is from "Deep Trouble for Deep Learning: The Flaws of Neural Networks" by Douglas Heaven, Nature vol 574 10 October 2019.

Figure 140 of the visual cortex is by Madsøo, at https://upload.wikimedia.org/wikipedia/commons/4/4d/Neural_pathway_diagram.svg. This file is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

Figure 142 of a prism is by By Cmglee and is posted at https://en.wikipedia.org/wiki/Prism#/media/File:Comparison_refraction_diffraction_spectra.svg
See the licensing at https://creativecommons.org/licenses/

p134 Figure 143 of RGB space is by SharkD, website at https://commons.wikimedia.org/wiki/User:Datumizer
This file is licensed under the Creative Commons Attribution-Share

p134 Figure 141 Sea urchin from E.M. Ullrich-Luter et. al., /Integr. Comp. Biol., 2013

p134 Figure 144 of the mantis shrimp eye is by Alexander Vasenin, from
https://commons.wikimedia.org/wiki/File:Mantis_Shrimp_Eyes.JPG
"I, the copyright holder of this work, hereby publish it under the
following license: This file is licensed under the Creative Commons
Attribution-Share Alike 3.0 Unported license"

p135 Figure 145 of Obon is from
https://en.wikipedia.org/wiki/Bon_Festival#/media/File:Osurasma,_or_praying_a_soul_out_of_purgatory-J._M._W._Silver.jpg

p136 Figure 146 is from The Movies In Our Eyes, Scientific American, April 2007.

p137 Figure 147 is from http://www.printablediagram.com/printable-nervous-system-diagrams/

p137 Figure 148 of a prism is adapted from and image by D-Kuru,
and is released under the Creative Commons Attribution-Share Alike
3.0 Austria license.

p137 Figure 149 is by Quasar Jarosz at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=7616130

p140 Figures 150 and 151 Picture of Go board and Space Invaders screen, are from Google Masters Go Nature, 529, January 28 2016 and

p142 Figure 152 of the retrograde motion of Mars is by Brian Brondel - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=1652425,
from https://en.wikipedia.org/wiki/Apparent_retrograde_motion#/media/File:Retrograde_Motion.bjb.svg

p144 Figure 153 of Japanese words written in Hiragana is from the author.
## Index

**A**
- abacus, 13
- ABC molecules, 92
- abstraction, 18
- accounting, 11, 42, 47
- Achebe, Chinua, 79
- aerodynamics, 69
- AI, 131, 140
- Al-Andalus, 13
- alchemy, 112
- Alexandria, 46
- algebraic numbers, 51, 61
- algorithm, 13
- aliasing, 129
- amacrine, 128
- amacrine cells, 139
- amplitude, 129
- Anglican Church, 117
- anthropologists, 11
- approximation, 43
- approximaton, 43
- Archimedean Principle, 29, 45, 65, 74
- Archimedes, 56, 106
- Aricebo, 81, 88
- Aristotle, 18, 25–27, 31, 32, 36, 70, 73, 105, 113–115, 121
- arithmetisation, 57, 70
- Armstrong, Karen, 19
- artificial intelligence, 140
- astrology, 15
- astronomy, 81
- AT&T, 136
- Atari, 140
- Atistotle, 105
- atom, 72
- Avery, Oswald, 39
- Axial Age, 19
- axiom, 25, 27, 47

**B**
- base sixty, 11
- Babuska, Ivar, 94
- Babylonian, 47, 76
- Babylonian Astronomy, 15
- Bacon, Francis, 118, 119
- Baghdad, 13
- Bakhshali Manuscript, 13
- baroreceptor, 137
- bat, 129
- bee, 121
- Beeckman, Issac, 107
- beer, 14, 49
- Behaviorism, 20, 22
- Bell Labs, 136
- Bell, Jocelyn, 81
- Berlin, Brent, 79
- Bessel function, 51
- Bethe, Hans, 81
- binding problem, 132
- binomial expansion, 56
- bipolar cell, 139
- Bitcoins, 69
- bloom, 124
- Boole, George, 32
- Boyle’ Law, 109
- Boyle’s Law, 110, 120
- Boyle, Robert, 105, 107–110, 117, 119, 120
- Brahe, Tycho, 91, 122
- Buddha, Gautama, 19
- calender, 14
- Caliphate, 13
- camera obscura, 122
- Cantor, 57, 61–63, 74
- Cantor, Georg, 41, 62
- carbohydrates, 86
- carcinogenic, 84
- Carroll, Lewis, 33
- Catholicism, 117
- Cauchy Sequence, 58
- Cauchy, Augustin, 57, 62
- Cavalier, 117
- CFIT, 78, 93
- channel, 136, 137, 140
- characteristic, 82
- Charles I of England, 112, 117
- Chinese, 12, 45, 69
- Chinese New Year, 143
- chirp, 129, 130
- Chomsky, Noam, 20
- Christ, Jesus, 119
- Christian, 13
- circular folds, intestinal, 85
- Civil War, English, 112, 117, 118
- CMOS, 123
- coding problem, 131
- cognitive reference points, 41
- Cognitive Science, 41
- Commonwealth, 117
- completeness, 33, 60, 61
- complex numbers, 61
- compress, 140
- compressed sensing, 129
- compression, 127, 129
- computable, 61, 63
- conditioned behavior, 19
- consistency, 33
- constellations, 14
- constructivist history, 44
continued fractions, 63
Continuity, 61, 73
Conway, John, 53
Copernicus, Nicolaus, 112
cortex, visual, 129
cortisol, 37
counting, 27
Crab Nebula, 81
Cray, 69, 100
cuneiform, 15
cut, Dedekind, 73

Dante, 112
Darwin, Charles, 130
Dead Sea Scroll, 17
Dedekind, Richard, 62, 73, 74
deduction, 25
deductive reasoning, 26
Dee, John, 112
deep learning, 140, 142
degrees, 15
Dehaene, Stanislaw, 21, 23
density of rationals, 60
Derivative, 113
Descartes, Renee, 107
diffraction, 51
dimensionality problem, 131
disease, genetic, 69
divination, 15
Divine Comedy, 112
DNA, 38
Dodgson, Charles, 33
Dore, Gustav, 112
Drake, Frank, 81
Drake, Stillman, 116
drug design, 69
drug resistance, 92
drugs, 83
Dumbledore, 37

E
Easter, 17, 75, 143
ECG, 6, 8
echolocation, 129
Ecology, Visual, 135
EDVAC, 104
Egypt, 42, 49, 54
Eid-al-Fitr, 143
eighteen-wheelers, 69

Eighth Day of Creation, 38
Einthoven, 6
EKG, 105
Elements of Euclid, 30
Elements, Euclid’s, 26
Elizabeth I, 112
empiricism, 112
Engineering, 13
Englert, Francois, 87
ENIAC, 91
Enlightenment, 77
Enuma Anu Enlil, 15
environmental remediation, 84
epinephrine, 137
epistemology, 17, 30, 63, 121, 139
epsilon, 44
equivalence, 40
equivalence class, 40, 62
error, 43
Error, Relative, 101
Essene, 17
Euclid, 26, 30–32, 55, 56, 63
Eudoxus, 50
Euler, 51

F
Fibonacci, 50, 51
fibrillation, 9
Field, 61
filter, digital, 7
filter, lowpass, 125
finite element analysis, 93
fixed point, 96
floating point, 69, 96
FLOP, 69
focal colors, 79
Fodor, Jerry, 35
form, 31
formal reasoning, 31
forms, 26
Foster, Jody, 6
foundations of calculus, 62
Four color problem, 39
fractions, 42, 45
fractions, unit, 42
France, 77
Frege, Gottlob, 27
frequency, 129
frog brain, 122

Galen, 115
Galileo, 105, 112
Galvani, Luigi, 6
ganglion, 127
gap at zero, 98
Gauss, Carl Friedrich, 109
genesis, 38
geometric, 48
glial, 137
gigaflop, 69
glycocalyx, 86
Go, 140
God, 19
Godel, Kurt, 34
Godin Tepe, 12, 23, 28
Gold, Thomas, 81
Google, 140
graph of data, 109
graphing, 113
gravitational constant g, 77
gravity, 77
Great Wall of China, 13
greatest lower bound, 73
Greek, 49, 56, 63
Greek language, 25
Greene, Brian, 87
Greenwich, 75
Greenwich Mean Time, 75
Griffith, Frederick, 38
Gwalior Numerals, 13
Goldstine, Herman, 91

H
Hamming, R. W., 16
Han, 12
Han Dynasty, 95
Hebrew, 76
heliocentric, 142
Hell, 112
Heng/Zong, 95
Heng/Zong notation, 12
Hero of Alexandria, 56
Herodotus, 54
Hewish, Anthony, 81
Higgs, Peter, 87
Hilbert, David, 33
Hindu-Arabic numerals, 13
Hippasus, 55
Hippocrates, 115  
Hiraganna, 142  
histology, 122  
Holi, 75, 143  
Hooke’s Law, 107  
Hooke, Robert, 108, 109, 117  
horoscope, 112  
ydrolized, 86  
retinal horizontal cells, 127  
Hooke, Robert, 107  

I  
IBM, 100  
Ibn-al-Haytham, 122  
Ideal Gas Law, 120  
IEEE, 93, 98, 100  
Indian numerals, 13  
induction, 55  
Induction, Axiom of, 28  
Inferno, 112  
infinite, 28, 41  
information theory, 136  
integers, 26  
inegral, 113  
Intel, 91, 100  
intraparietal sulcus, 21  
ion channel, 37  
Iraq, 14, 44  
irrational, 56  
Ishango Bone, 11  
Islam, 13  

J  
Japanese, 142  
Jaspers, Karl, 19  
Jefferson, Thomas, 32  
Jewish, 17  
jpeg, 126, 129, 140  
Judas, 112  
Judson, Horace Freeland, 38  

K  
al-Khwarizmi, 13, 27, 40  
Kathytron 5, 96  
Kay, Paul, 79  
Kepler, 91  
Kepler, Johannes, 112, 122  
Kircher, Athanasius, 122  
kosmos, 19  
KYB shock absorber, 151  

L  
LAGEOS, 77  
Least Squares, 109  
ext upper bound, 73  
Leavitt, Henrietta Swan, 91  
Legendre, Adrien-Marie, 109  
Leibnitz, Friedrich Wilhelm, 34  
Leibnitz, Gottfried Wilhelm, 32  
Leibnitz, Gottfried Wilhelm, 32, 62  
Leonardo of Pisa, 13, 50  
Leviticus, 76  
Lilnas, Rodolfo, 86  
limit, 44  
lipase, 86  
logarithm, 82  
logic, Boolean, 32  
logic, formal, 32  
longitude, 75  
Lovelace, Robert, 81  
Lui Hui, 45, 61  
Lunar holiday, 17  
Luria, Aleksandr, 36  
Luria, Aleksandr Romanovich, 35  

M  
Mac, 69  
machine epsilon, 97, 98  
machine numbers, 96  
magnetsphere, 81  
magnitude, 50, 61  
many worlds, 135  
Mars Orbiter, 78, 93  
Marshack, Alexander, 23  
mathematical modeling, 134  
Mediterranean, 13  
MELTDOWN, 92  
Mendel, Gregor, 38  
Mersenne, Marin, 107, 114  
Mesopotamia, 12, 42, 47, 48, 54, 61  
meter, 77  
metric system, 77  
microvilli, 86  
Milton, John, 112  
modeling, 69  
module, 132  
molecular diameter, 86, 90  
monosaccharides, 86  
month, 75  
moon, 75  
mosaic, 124  
Motorola, 100  
mp3, 129, 140  
music of the spheres, 16  
mylenated, 127  
myth, 134  
Mauchley, John, 91  

N  
nanoparticle, 85  
NASA, 93, 94  
Nasa, 78  
natural numbers, 25, 26  
neuroscience, 86  
Newton, Isaac, 17, 50, 62, 120, 134  
nightmare, 52  
Nirvana, 19  
Nobel, 122  
Nootka, 75  
norepinephrine, 137  
normalized floating point, 98  
nous, 30  
von Neumann, John, 40  

O  
obon, 135  
Octave, 97  
Old Testament, 105  
omens, 15  
orders of magnitude, 81, 82  
Oresme, Nicola, 113  
overflow, 97  
Oxford Calculators, 113  

P  
Pakistan, 13  
Pan, Jiapu, 145  
Pan-Tompkins, 145  
Paradise Lost, 112  
parallel postulate, 26  
Paramecium, 37  
Parmenides, Platonic Dialogue, 18  
Pascal, Blaise, 106  
Passover, 75  
Patriot, 101  
Peano axioms, 28  
Peano, Giuseppe, 27, 70  
pendulum, 77  
Pentium, 103  
Pentium chip, 91  
petaflop, 69  
pharmacology, 83  
philosophy, 121
photodiode, 123
photosynthesis, 68
Pinker, Steven, 35
pixel, 125
Planck, 72
Plato, 17, 25, 30, 32, 121, 130, 144
Platyfai, William, 113
Pong, 140
positional notation, 12
Prior Analytics, 26, 31
prism, 134, 137
program verification, 39
proportion, Greek theories, 18
protease, 86
Protectorate, 117
proto-Indo-European, 25
ProtoWorld, 23
psychologists, educational, 140
Ptolemy, 46, 112
Ptolemy, Claudius, 112
Puerto Rico, 88
pulsar, 81, 83
pump, 105
Pythagoras, 18, 19, 25, 55
Pythagorean, 17
Pythagorean Theorem, 48, 99
Protestantism, 117

Q
Qin, 12
quadrivium, 32
quantization, 110, 124, 129, 140
quantum mechanics, 90, 135
Quran, 75

R
radar, 88
Ramadan, 75, 143
range restriction, 110
Ranters, 119
rational numbers, 56
real line, 70
real numbers, 59
reasoning, deductive, 25
receptor theory, 84
receptors, 83
RedBook, 129
regressus, 112, 115
relativism, 79
remediation, 85
representation, 11, 35
retina, 122, 137, 139
retrograde motion, 142
RGB, 134
rhodopsin, 124
Robin, 10
Robson, Eleanor, 44
Roman numerals, 12
Rosch, Eleanor, 41, 79
rounding, 67
ruler and compass, 56
Russell, Bertrand, 34

S
Sagan, Carl, 83
Salteter, Edwin, 81
sampling, 110, 125, 129
Satan, 112
Schmandt-Besserat, Denise, 23
scholastic philosophy, 114
scientific notation, 68
scribe, 42
SCUD, 101
sea urchin, 134
Second Coming, 119
Seleucid, 47
sequence, 46, 47, 58
Shamash, 48, 76
Shannon, Claude, 35
Shannon, Claude, 35
Shining One, 44
shrimp, mantis, 135
signal, 123, 136
signal processing, 128, 136, 140
Skinner, Verbal Behavior, 20
Soviet, 88
Space, 48
Spain, 13
spar, of a wing, 84
spectral analysis, 135
SPECTRE, 92
spectrum, 134
square-cube, 84
SRT algorithm, 103
stock market trading, 68
stored program, 104
Streptococcus pneumoniae, 38
strong induction, 28, 41, 55
surface area of sphere, 84
surreal numbers, 53
surveying, 76
sylllogism, 26, 31
symbolic, 11
synapse, 86
System A and B, 15, 24

T
tax, 48
TCE, 84
teleology, 46
theater effect, 124
Timaeus, 18
Timaeus of Plato, 121
Time, 48
tithes, 117
Tomkins, Willis, 145
Torrlicelli, Evangelista, 106
transcendental functions, 51
transcendental numbers, 51
transfer learning, 140
transfinite numbers, 41, 63
trivium, 32
Turing machine, 34
Turing, Alan, 34, 35
Tyler, G. Leonard, 81

U
ultraviolet light, 121
underflow, 97
understand, 144

V
vacuum, 105, 106
Vaieika, 72
villi, 85
visual cortex, 132
Visual Ecology, 135
volume of a sphere, 84
von Neumann, John, 104

W
Wars, World One and Two, 91
warming, global, 68
weapons, nuclear, 69
weather forecasting, 69
Weinberg, Stephen, 19
Whitehead, Alfred North, 34
Wigner, Eugene, 16, 18
Wilkins, John, 117
Wisdom 11:23, 105
Wittgenstein, Ludwig, 36
word size, computer, 95
Y
Yahad, 17

Yayeh, 76
Yom Kippur, 143