Galileo and Einstein

Text for Physics 109, Fall 2009

Michael Fowler, UVa Physics

8/21/2009

TABLE OF CONTENTS

1 INTRODUCTION: WHAT IS THIS COURSE ABOUT? 12

1.1 Some Basic Ideas 12

1.2 Babylonians and Greeks 12

1.3 Greek Classics Come to Baghdad 12

1.4 Monasteries and Universities 13

1.5 Galileo 13

1.6 Newton 13

1.7 From Newton to Einstein 13

1.8 What about Other Civilizations? 14

1.9 Plan of the Course 14

2 COUNTING IN BABYLON 16

2.1 The Earliest Written Language 16

2.2 Weights and Measures: 60s everywhere! 16

2.3 Number Systems: Ours, the Roman and the Babylonian 17

2.4 Fractions 19

2.5 Ancient Math Tables 19

2.6 How Practical are Babylonian Weights and Measures? 20

2.7 Pythagoras' Theorem a Thousand Years before Pythagoras 20

2.8 Babylonian Pythagorean Triplets 21

3 EARLY GREEK SCIENCE: THALES TO PLATO 22

3.1 The Milesians 22

3.2 Early Geometry 23

3.3 Early Geometry According to Proclus 24
3.4 The Pythagoreans: a Cult with a Theorem, and an Irrational Discovery 26
3.5 The Square on the Hypotenuse 26
3.6 Rational and Irrational Numbers 27
3.7 What's so Important about Irrational Numbers? 30
3.8 Change and Constancy in the Physical World 30
3.9 Hippocrates and his Followers 31
3.10 Plato 32
3.11 References 33

4 MOTION IN THE HEAVENS: STARS, SUN, MOON, PLANETS 34

4.1 Introduction 34
4.2 Looking at the Stars 34
4.3 Motion of the Sun 34
4.4 Motion of the Moon against the Starry Vault 37
4.5 Motion of the Planets 38

5 ARISTOTLE 40

5.1 Beginnings of Science and Philosophy in Athens 40
5.2 Plato's Idea of a Good Education 40
5.3 Aristotle and Alexander 40
5.4 Aristotle Founds the Lyceum 41
5.5 Aristotle's Science 42
5.6 Aristotle's Method 42
5.7 "Causes" 43
5.8 Biology 44

5.9 Elements 44

5.10 Dynamics: Motion, And Why Things Move 45

5.11 Natural Motion and Violent Motion 45

5.12 Aristotle's Laws of Motion 46

5.13 Planetary Dynamics 47

5.14 Aristotle's Achievements 47

6 MEASURING THE SOLAR SYSTEM 48

6.1 How Big is the Earth? 48

6.2 How High is the Moon? 49

6.3 How Far Away is the Sun? 51

7 GREEK SCIENCE AFTER ARISTOTLE 53

7.1 Strato 53

7.2 Aristarchus 53

7.3 Euclid 54

7.4 Plato, Aristotle and Christianity 54

7.5 Archimedes 55

7.6 Archimedes' Principle 55

7.7 Archimedes and Leverage 57

7.8 Apollonius 59

7.9 Hypatia 59

8 BASIC IDEAS IN GREEK MATHEMATICS 60

8.1 Closing in on the Square Root of 2 60

8.2 Zeno's Paradoxes 61

8.3 Achilles and the Tortoise. 61

8.4 The Arrow ... 62

8.5 Instants and Intervals ... 62

8.6 Speed at an Instant ... 63

8.7 The Beginning of Calculus ... 64

8.8 Archimedes Begins Calculating Pi ... 64

8.9 Squaring the Circle ... 66

8.10 Eudoxus' Method of Exhaustion ... 67

8.11 Archimedes does an Integral ... 69

8.12 Conclusion ... 69

9 HOW THE GREEKS USED GEOMETRY TO UNDERSTAND THE STARS ... 70

9.1 Crystal Spheres: Plato, Eudoxus, Aristotle ... 70

9.2 Measuring the Earth, the Moon and the Sun: Eratosthenes and Aristarchus ... 71

9.3 Cycles and Epicycles: Hipparchus and Ptolemy ... 71

9.4 Ptolemy's View of the Earth ... 74

10 HOW CLASSICAL KNOWLEDGE REACHED BAGHDAD ... 76

10.1 The Classical Achievement in Mathematics and Science ... 76

10.2 Why did Mathematics and Science Grind to a Halt? ... 77

10.3 But Some Christians Preserved the Classical Knowledge… ... 78

10.4 How the Nestorians Helped Science Survive ... 78

10.5 On into Persia ... 80

10.6 The Advent of Islamic Rule ... 80

10.7 The House of Wisdom: al-Khwarismi ... 81

11 LATER ISLAMIC SCIENCE ... 83

11.1 The Islamic World ... 83

11.2 Omar Khayyam 83

11.3 Al-Tusi 84

12 GALILEO AND THE TELESCOPE 87

12.1 Copernicus Challenges Ptolemy's Scheme 87

12.2 The Evolution of the Telescope 88

12.3 Mountains on the Moon 90

13 LIFE OF GALILEO 92

13.1 Books 92

13.2 Like Father, like Son 92

13.3 Pendulums and Pulses 93

13.4 The Roof of Hell 94

13.5 Venice: Wine, Women and Dialogue 95

13.6 The Telescope: Heaven Abolished? 95

13.7 Galileo Wins Over Some Jesuit Astronomers… 96

13.8 …but Alienates Some Others 97

14 SCALING: WHY GIANTS DON'T EXIST 100

15 GALILEO'S ACCELERATION EXPERIMENT 104

15.1 Summarizing Aristotle's View 104

15.2 Two New Sciences 104

15.3 Naturally Accelerated Motion 105

15.4 Galileo's Acceleration Hypothesis 106

15.5 Slowing Down the Motion 106

15.6 Galileo's Acceleration Experiment 107

15.7 Actually Doing the Experiment 108

16 NATURALLY ACCELERATED MOTION 109

16.1 Distance Covered in Uniform Acceleration 109

16.2 A Video Test of Galileo's Hypothesis 110

16.3 Throwing a Ball Upwards 111

16.4 Speed and Velocity 111

16.5 What's the Acceleration at the Topmost Point? 112

16.6 The Motion of Projectiles 113

16.7 Compound Motion 114

17 USING VECTORS TO DESCRIBE MOTION 115

17.1 Uniform Motion in a Straight Line 115

17.2 Uniform Motion in a Plane 116

17.3 Relative Velocities: a Child Running in a Train 118

17.4 Aristotle's Law of Horizontal Motion 119

17.5 Galileo's Law of Horizontal Motion 120

17.6 Galileo's Law of Vertical Motion 120

17.7 Describing Projectile Motion with Vectors 121

17.8 Acceleration 121

18 TYCHO BRAHE AND JOHANNES KEPLER 122

19 ISAAC NEWTON 124

19.1 Newton's Life 124

19.2 Projectiles and Planets 126

19.3 The Moon is Falling 128

20 HOW NEWTON BUILT ON GALILEO'S IDEAS 130

20.1 Newton's Laws 130

20.2 Acceleration Again ... 130

20.3 An Accelerating Body that isn't Changing Speed ... 130

20.4 Finding the Acceleration in Circular Motion ... 132

20.5 An Accelerating Body that isn't Moving ... 133

20.6 Galileo's Analysis of Motion: Two Kinds ... 133

20.7 Newton Puts Them Together ... 134

20.8 Force is the Key ... 134

20.9 Newton's First Law: no Force, no Change in Motion ... 134

20.10 Newton's Second Law: Acceleration of a Body is Proportional to Force ... 135

20.11 What About Same Force, Different Bodies? ... 135

20.12 Falling Bodies One More Time: What is Mass? ... 135

20.13 Mass and Weight ... 136

20.14 The Unit of Force ... 136

20.15 Newton's Third Law: Action and Reaction ... 137

20.16 Newton's Second Law in Everyday Life ... 138

20.17 Gravity ... 139

20.18 The Law of Gravity ... 139

20.19 Weighing the Earth ... 141

21 THE SPEED OF LIGHT ... 142

21.1 Early Ideas about Light Propagation ... 142

21.2 Measuring the Speed of Light with Jupiter's Moons ... 142

21.3 Starlight and Rain ... 143

21.4 Fast Flickering Lanterns ... 144

21.5 Albert Abraham Michelson ... 144

21.6 Sailing the Silent Seas: Galilean Relativity ... 145

21.7 Michelson Measures the Speed of Light ... 146

22 THE MICHELSON-MORLEY EXPERIMENT ... 147

22.1 The Nature of Light ... 147

22.2 The Wavelike Nature of Sound ... 147

22.3 Is Light a Wave? ... 148

22.4 If Light *is* a Wave, What is Waving? ... 149

22.5 Detecting the Aether Wind: the Michelson-Morley Experiment ... 149

22.6 Einstein's Answer ... 155

23 SPECIAL RELATIVITY ... 155

23.1 Galilean Relativity again ... 155

23.2 Generalizing Galilean Relativity to Include Light: Special Relativity ... 157

23.3 You *Really* Can't Tell You're Moving! ... 157

23.4 Truth and Consequences ... 158

24 SPECIAL RELATIVITY: WHAT TIME IS IT? ... 160

24.1 Special Relativity in a Nutshell ... 160

24.2 A Simple but Reliable Clock ... 160

24.3 Looking at Somebody Else's Clock ... 161

24.4 Fitzgerald Contraction ... 164

24.5 Experimental Evidence for Time Dilation: Dying Muons ... 164

25 SPECIAL RELATIVITY: SYNCHRONIZING CLOCKS ... 165

26 TIME DILATION: A WORKED EXAMPLE ... 168

26.1 "Moving Clocks Run Slow" plus "Moving Clocks Lose Synchronization" plus "Length Contraction" leads to consistency! ... 168

27 MORE RELATIVITY: THE TRAIN AND THE TWINS 173

27.1 Einstein's Definition of Common Sense 173

27.2 Trapping a Train in a Tunnel 173

27.3 The Tunnel Doors are Closed Simultaneously 174

27.4 Or are They? 174

27.5 Does the Fitzgerald Contraction Work Sideways? 175

27.6 How to Give Twins Very Different Birthdays 175

27.7 The Twins Stay in Touch 176

27.8 Figuring the Observed Time between Flashes 176

27.9 What does she see? 177

27.10 What does he see? 178

27.11 The Doppler Effect 178

28 MOMENTUM, WORK AND ENERGY 179

28.1 Momentum 179

28.2 Momentum Conservation and Newton's Laws 180

28.3 Work 181

28.4 Energy 183

28.5 Kinetic Energy 184

29 ADDING VELOCITIES: A WALK ON THE TRAIN 185

29.1 The Formula 185

29.2 Testing the Addition of Velocities Formula 187

30 CONSERVING MOMENTUM: THE RELATIVISTIC MASS INCREASE 188

30.1 Momentum has Direction 188

30.2 Momentum Conservation on the Pool Table 189

30.3 A Symmetrical Spaceship Collision 190

30.4 Just How Symmetrical Is It? 190

30.5 Einstein Rescues Momentum Conservation 190

30.6 Mass Really *Does* Increase with Speed 191

30.7 Kinetic Energy and Mass for Very Fast Particles 191

30.8 Kinetic Energy and Mass for Slow Particles 193

30.9 $E = mc^2$ 194

31 GENERAL RELATIVITY 195

31.1 Einstein's Parable 195

31.2 Some Consequences of the Equivalence Principle 196

31.3 General Relativity and the Global Positioning System 197

1 Introduction: What is this course about?

1.1 Some Basic Ideas

This course traces the historical development of some key scientific ideas: space, time, motion, mass and force. Philosophers, and more practical people, have struggled with these concepts since the earliest recorded times. Their combined efforts have been fruitful: real progress in understanding has evolved over the centuries.

In fact, the first real understanding of the mathematics and physics of motion, by Galileo and Newton in the seventeenth century, began the series of developments that underlie our modern technology. Just think for a moment how much the way we live now depends on this technology: a hundred years ago there were no computers, few phones and few automobiles. Actually, though, a lot *had* been accomplished in the century prior to that: in 1800, the fastest transportation on land was a horse, at sea a well-designed sailing ship, and messages took weeks to cross the Atlantic. By 1900, steamships were crossing the Atlantic in a week or so, railroads spanned the continents, there were thousands of miles of telegraph cables under the oceans. In 1904, a steam train in England reached one hundred miles an hour. The pace of change has been picking up ever since it began with Galileo and Newton using Greek methods to analyze motions of projectiles and planets, and thereby discovering the underlying laws.

1.2 Babylonians and Greeks

Many civilizations have made significant contributions to the understanding of these key ideas: in a one semester course, we can only look at a few. After a brief look at the Babylonians who were carrying out sophisticated mathematical accounting and keeping precise astronomical records four thousand years ago, we'll spend some time reviewing the astonishing contribution of the ancient Greeks, who not only invented science, but also developed the modern approach to mathematics. Greek geometry played a central role in the work of Galileo and Newton almost two thousand years later, at the beginning of the modern scientific era.

1.3 Greek Classics Come to Baghdad

Despite these achievements, Greek interest in science gradually waned during the first centuries of the Christian era. However, much of the earlier work was preserved, thanks, remarkably, to the Arabs. As Europe fell into the chaos of the Dark Ages, Greek classics were being translated into Arabic in Baghdad, and over the next five hundred years Arab scholars clarified and extended Greek science and mathematics.

About a thousand years ago, parts of Europe began to settle down, and the classical learning preserved and strengthened by the Arabs was rediscovered. The centers of learning in the West at that time were the monasteries. The Arabic texts were translated into Latin (the common language of the European monks), and later some of the surviving original Greek texts were uncovered and translated.

1.4 Monasteries and Universities

This influx of classical learning certainly broadened the intellectual horizons in the monasteries, and also in the universities, new religious teaching foundations that began to appear over the next two or three centuries. But there was a catch—the monks were, naturally, of a theological disposition, and they were so overwhelmed by the excellence of the rediscovered classics, in particular the all-encompassing works of Aristotle, that they treated them in effect as holy writ. Aristotle was right about most things, but unfortunately not about the motion of falling stones, projectiles, forces, or indeed the motion of the earth: central topics in this course. Consequently, there was little progress in understanding the science of motion—called *dynamics*—during this period.

1.5 Galileo

The first real advance beyond Aristotle in understanding of motion came with Galileo. Others had in fact sensed some of the truth, but Galileo was the first to bring together experimental data and mathematical analysis in the modern fashion. In this sense, Galileo was arguably the first *scientist* (although that word wasn't invented until the 1830's). An essential part of science is that it is not faith-based: it's O.K. to have a theory, which you might even term a belief, but it is only provisionally true until firmly established by unambiguous experiment or observation. This was Galileo's revolutionary contribution: he refused to accept assertions from Aristotle, or for that matter what he considered naïve literal interpretations of the bible, if they contradicted what could be demonstrated or observed. But, like every other great scientist, he wasn't right about everything, as we shall see.

1.6 Newton

Galileo made many contributions to physics and astronomy, his telescopic observations made clear that the Moon was a big rock, not some mysterious ethereal substance. Together with some results of Kepler, this led Isaac Newton in the late 1600's to his Equations of Motion and theory of a Universal Law of Gravitation, establishing the unity of familiar earthly dynamics, like throwing a ball, and the motion of the planets through the night sky. This unification of earth and the heavens had a tremendous impact: suddenly the natural world seemed more rational, more capable of systematic human explanation. It was the Age of Enlightenment. Newton's methods were applied in many fields, proving brilliantly successful in analyzing natural phenomena, but much less convincing in the social sciences.

1.7 From Newton to Einstein

After Newton, the mathematical and physical sciences developed rapidly, especially in France and England. Over the next two hundred years, astronomical observations became orders of magnitude more exact, and Newton's theory of gravitation, based on much cruder observations, continued to predict the planetary motions successfully, as more demanding precision was needed. But finally it stumbled: the planet Mercury goes around the Sun in an elliptical orbit, and the axis of this ellipse itself moves slowly around in a circle. The rate of this axis turning, or *precession*, as it's called, is predicted by Newton's theory: it's caused by attraction of other planets, and the sun not being a perfect sphere. But the Newtonian prediction was a little bit wrong: about 1% from what was observed. And the precession is very small anyway: the discrepancy amounted to a difference of about a one degree turn in ten thousand years. But this tiny effect was a key to Einstein's General Theory of Relativity, which turns out to have far more dramatic manifestations, such as black holes.

Perhaps the most important development in the physical sciences in the two centuries following Newton was progress in understanding electricity and magnetism, and the realization that light was a wave of electric and magnetic fields. This made possible much of the basic technology underpinning our civilization: electric power was distributed in the late 1800's, radio waves were first transmitted around 1900. Quite unexpectedly, imagining someone in motion measuring the electric and magnetic fields in a light wave led to Einstein's epiphany, that time was not as absolute as everyone had always taken for granted, but flowed at different rates for people moving at different speeds. This in turn led to the Special Theory of Relativity, to $E = mc^2$, and nuclear power.

1.8 What about Other Civilizations?

Why is this course all about Europe and the Middle East? What about China, India, or Mayans, for example? Good questions—but we are not attempting here a history of all scientific development, or even of all basic physics. We're focusing instead on one important aspect: the understanding of dynamics, meaning how motion relates to forces. Analyzing motion quantitatively, which is necessary for getting anywhere in this subject, turns out to be very tricky. The ancient Greeks, for all their brilliant discoveries and writings, didn't manage it, nor did the Chinese, Indians or Mayans. Galileo was the first to give a satisfactory mathematical description of acceleration, and link it firmly with observation, although partial success had been achieved by others earlier. His advances are the basis for this course: but he would have gotten nowhere without tools provided by Greek geometry, itself dependent on earlier work by Babylonians and Egyptians. So we begin at the beginning, or at least at the recorded beginning. A crucial contribution from India was the so-called Arabic number system, which is actually Hindu. The Arabs in Baghdad developed classical learning thanks to cheap paper, a discovery imported from China, and the revival of learning in Europe after the Dark Ages owed much to newly efficient farming, using horses with Chinese harnesses in place of less efficient oxen—this gave some people time to think. The Chinese also contributed printing, the magnetic compass and gunpowder, all of which had a great impact on the West, but, as we've said, here we're just focusing on some basic ideas, and there was no Chinese Galileo.

1.9 Plan of the Course

The course falls rather naturally into three parts, which will take approximately equal times. There will be in-class midterms at the ends of the first two parts.

Part I covers the needed developments in mathematics and science from the earliest times until Galileo: the Babylonians and Egyptians, the Greeks, the Arabs, and early Western Europe. Obviously, we cannot give a complete picture of all these developments: we pick and choose those most relevant to our later story. (We won't even cover all the material in these notes—see the syllabus—those sections we skip contain background material you might find interesting.)

Part II will be devoted to the works of Galileo and Newton, culminating in Newton's Laws of Motion and his Law of Universal Gravitation: the revolution revealing that the heavens obeyed the same laws as earthly phenomena.

Part III begins with some puzzles about the nature and speed of light, accurately measured in the nineteenth century thanks to new developments in technology. These puzzles were only resolved decades later by Einstein, and led to a new and different understanding of time, space, mass and energy.

2 Counting in Babylon

2.1 The Earliest Written Language

Sumer and Babylonia, located in present-day Iraq, were probably the first peoples to have a written language, beginning in Sumer in about 3100 BC. The language continued to be written until the time of Christ, but then it was completely forgotten, even the name Sumer became unknown until the nineteenth century.

From the earliest times, the language was used for business and administrative documents. Later, it was used for writing down epics, myths, etc., which had earlier probably been handed down by oral tradition, such as the Epic of Gilgamesh.

2.2 Weights and Measures: 60s everywhere!

In about 2500 BC, by Royal Edict, weights and measures were standardized in Babylon. This was a practical business decision, which without doubt eliminated much tension in the marketplace.

The smallest unit of **weight** was the grain (about 45 milligrams). What use was that? At first, the currency was in fact barleycorn! (They later moved to silver and gold ingots.) The shekel was 180 grains (about ¼ ounce), the mina 60 shekels, and the talent 3600 shekels (about 67 pounds). More details here.

1 talent =	**60 minas =**	**3600 shekels =**		**approx 60 lbs**
	1 mina =	**60 shekels =**		**approx 1 lb**
		1 shekel =	**180 grains =**	**approx ¼ oz**
			1 grain =	**approx 45 mg**

The smallest unit of **length** was—surprise—the **barleycorn**, called **she**, about **1/10 inch**.

Next came the **finger**, or **shu-si**, equal to 6 she, about **2/3 of an inch**.

The **cubit** (or **kush**) was 30 fingers, about **20 inches**.

The **nindan** (or **GAR**, or rod) was 12 cubits, **20 feet** or 6 meters.

The **cord** or rope (used in surveying) was 120 cubits, **200 feet**, that is, 3600 fingers.

The **league** (also called stage and beru) was 180 cords, about **seven miles**.

The basic unit of **area** was the **sar**, **one square nindan**, 400 sq ft, a garden plot.

The **gin** was 1/60 sar.

By 2000 BC, there was a calendar with a year of 360 days, 12 months of 30 days each, with an extra month thrown in every six years or so to keep synchronized with astronomical observations. (According to Dampier, *A History of Science*, Cambridge, page 3, the day was divided into hours, minutes and seconds, and the sundial invented. He implies this is about 2000 BC. He doesn't say how many hours in a day, and Neugebauer (*The Exact Sciences in Antiquity*, Dover, page 86) claims the Egyptians were the first to come up with twenty-four.)

The circle was divided into 360 degrees.

Notice that all these standards of measurement include *multiples of 60* frequently—obviously, 60 was the Babylonians' favorite number.

2.3 Number Systems: Ours, the Roman and the Babylonian

To appreciate what constitutes a good counting system, it is worthwhile reviewing briefly our own system and that of the Romans. The Roman system is in a way more primitive than ours: X always means 10, C means 100 and I means 1. (You might be thinking: this isn't *quite* true—they reversed numbers to indicate subtraction, such as IV for 4. In fact it appears they didn't, they used IIII, and IV is more recent. There's an article on all this in Wikipedia, which is interesting but currently unreliable.)

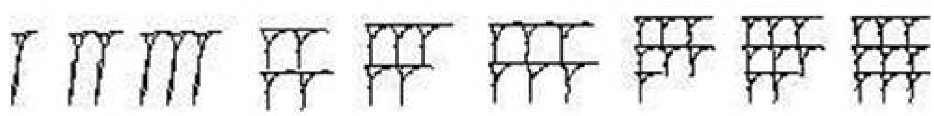
By contrast, in our system 1 can mean 1 or 10 or 100 depending on where it appears in the expression—the 1 in 41 means a different quantity from the 1 in 145, for example. We say the value of a symbol has "*positional dependence*"—its actual value depends on where in the expression it appears. Our convention, as you well know, is that the number to the far right in our system is the number of 1's, the number to its immediate left is the number of 10's, to the left of that comes the number of 10×10's, then of 10×10×10's and so on. We use the same set of symbols, 1,2,3,4,5,6,7,8,9,0 in each of these positions, so the value of such a symbol in a number depends on its position in that number.

To express quantities less than 1, we use the decimal notation. We put a dot (in some countries a comma is used) and it is understood that the number to the immediate left of the dot is the number of 1's, that to the immediate right the number of tenths (10^{-1}'s in mathematical notation), the next number is the number of hundredths (10^{-2}'s) and so on. With this convention, ½ is written .5 or 0.5 and 1/5 is .2. Unfortunately, 1/3 becomes .33333..., rather inconveniently, and 1/6 and 1/7 similarly go on forever. (Actually, this decimal system with the dot is, historically speaking, rather a recent invention—it was created by a Scotsman called Napier about 400 years ago.)

To get back to comparing the Roman system with our own, notice that the Romans did not have a 0, zero. This is why it is important to have a different symbol for ten and one, X and I are easily distinguished. If we didn't have a zero, one and ten would both be represented by 1, although we might be able to distinguish them in a column of figures by placing them in different columns.

After those preliminary remarks, we are ready to look at the Babylonian system. It's written on clay tablets – that's why we still have original copies around!

Their number system has only two basic elements, the first of which is clear on examining the first nine numbers:



Evidently, these nine numbers are all constructed of a single element, a mark easily gouged with one twist of a stick in the soft clay, and the number of times this element is repeated is the number represented. The sticks used to make the marks were wedge shaped,

The numbers 10, 20, 30, 40, 50, are represented by the symbols:

It is clear that again we have simple repetition of a basic element, which we will conveniently represent by <, and again it's a mark not difficult to make in the soft clay. Thus, any number between 1 and 59 is represented by a symbol from the second diagram followed in the usual case by one from the first diagram, so 32 would be written <<<11, approximately.

When they get to 60, the Babylonians start again in a similar way to our starting again at 10. Thus, 82 is written as 1<<11, where the first 1 represents 60.

So the Babylonian system is based on the number 60 the same way ours is based on 10. Ours is called a "**decimal**" system, theirs a "**sexagesimal**" system.

There are some real problems with the Babylonian number system, the main one being that nobody thought of having a zero, so both *sixty and one look exactly the same*, that is both are represented by 1! Actually, it's even worse—since there is no decimal point, the way to write 1/2, which we write 0.5, for five tenths, they would write <<<, for thirty sixtieths—but with no zero, of course, and no dot either. So if we see <<< on a clay tablet, we don't know if it means 1/2, 30 or for that matter 30×60, that is, 1800.

This is in fact not as bad as it sounds—sixty is a very big factor, and it will usually be clear from the context if <<< should be interpreted as 1/2 or 30. Also, in columns of figures, a <<< representing 30 was often put to the left of <<< representing 1/2.

2.4 Fractions

In real life commercial transactions, simple addition and even multiplication are not that difficult in most number systems. The hard part is division, in other words, working with *fractions*, and this comes up all the time when resources must be divided among several individuals. The Babylonian system is really wonderful for fractions!

The most common fractions, 1/2, 1/3, 1/4, 1/5, 1/6 all are represented by a single number (1/2= <<< , 1/3= << , 1/5= <11, etc.). That is, these fractions are exact numbers of sixtieths—sixty is the lowest number which exactly divides by 2, 3, 4, 5, and 6. This is a vast improvement on the decimal system, which has infinite recurrences for 1/3 and 1/6, and even ¼ needs two figures: .25.

(Of course, even in Babylonian, eventually we are forced to go to the second "sexagesimal" number, which would be the number of sixtieths of sixtieths, that is, of three-thousand-six-hundredths. For example, 1/8 is seven-and-a-half sixtieths, so would be written as seven followed by thirty—for seven sixtieths plus thirty sixtieths of a sixtieth. And, 1/7 is as much of a headache as it is in our own system.)

2.5 Ancient Math Tables

In order to make their bookkeeping as painless as possible, the Babylonians had math tables: clay tablets with whole lists of *reciprocals*. The reciprocal of a number is what you have to multiply it by to get 1, so the reciprocal of 2 is 1/2 written 0.5 in our system, the reciprocal of 5 is 1/5 written 0.2 and so on.

The point of having reciprocal tables is that dividing by something is the same as multiplying by the reciprocal, so using the tables you can replace division by multiplication, which is a lot easier.

Surviving clay tablet examples of Babylonian reciprocal tablets look like this:

```
11          <<<

111         <<

1111        <11111

11111       <11

111111      <

11111111    1111111 <<<
```

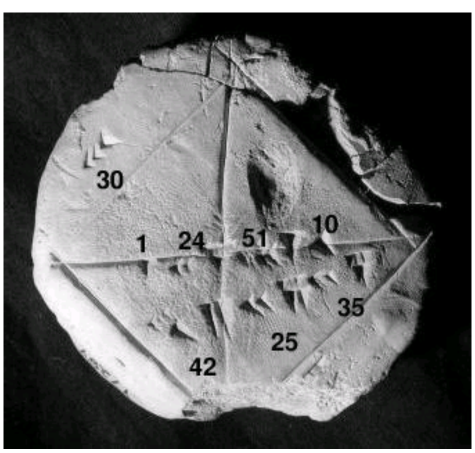
We have cheated a bit here—the numbers 4, 5, 6, etc. in both columns should really have their 1's stacked as in the first figure above.

2.6 How Practical are Babylonian Weights and Measures?

Let's take as an example how much food a family needs. If they consume 120 *shekels* of grain each day, for example, that's 12 *talents* of grain per year. (One talent = 3600 shekels). Just imagine the parallel calculation now: if the family consumes 30 ounces of grain a day, what is that in tons per year? If you were transported to the Babylon of four thousand years ago, you would hardly miss your calculator! (Admittedly, the Babylonian calculation is a bit more difficult every six years when they throw in an extra month.)

2.7 Pythagoras' Theorem a Thousand Years before Pythagoras

Some of the clay tablets discovered contain lists of **triplets** of numbers, starting with (3, 4, 5) and (5, 12, 13) which are the lengths of sides of right angled triangles, obeying Pythagoras'



"sums of squares" formula. In particular, one tablet, now in the Yale Babylonian Collection, this photograph by Bill Casselman, shows a picture of a square with the diagonals marked, and the lengths of the lines are marked on the figure: the side is marked <<< meaning thirty (fingers?) long, the diagonal is marked: <<<<11 <<11111 <<<11111. This translates to 42, 25, 35, meaning 42 + 25/60 + 35/3600. Using these figures, the ratio of the length of the diagonal to the length of the side of the square works out to be 1.414213…

Now, if we use Pythagoras' theorem, the diagonal of a square forms with two of the sides a right angled triangle, and if we take the sides to have length one, the length of the diagonal squared equals 1 + 1, so the length of the diagonal is the square root of 2. The figure on the clay tablet is incredibly accurate—the true value is 1.414214… Of course, this Babylonian value is far too accurate to have been found by measurement from an accurate drawing—it was clearly checked by arithmetic multiplication by itself, giving a number very close to two.

2.8 Babylonian Pythagorean Triplets

Question: the Babylonians catalogued many Pythagorean triplets of numbers (centuries before Pythagoras!) including the enormous 3,367 : 3,456 : 4,825. Obviously, they didn't check every triplet of integers, even plausible looking ones, up to that value. How could they possibly have come up with that set?

Suppose they *did* discover a few sets of integers by trial and error, say 3 : 4: 5, 5 : 12 :13, 7: 24 : 25, 8 : 15 : 17. We'll assume they didn't count 6 : 8 : 10, and other triplets where all three numbers have a common factor, since that's not really anything new.

Now they contemplate their collection of triplets. Remember, they're focused on *sums of squares* here. So, they probably noticed that all their triplets had a remarkable common property: the largest member of each triplet (whose square is of course the sum of the squares of the other two members) is in fact *itself* a sum of two squares! Check it out: $5 = 2^2 + 1^2$, $13 = 3^2 + 2^2$, $25 = 4^2 + 3^2$, $17 = 4^2 + 1^2$.

Staring at the triplets a little longer they might have seen that once you express the largest member as a *sum* of two squares, one of the other two members of the triplet is the *difference* of the same two squares! That is, $3 = 2^2 - 1^2$, $5 = 3^2 - 2^2$, $7 = 4^2 - 3^2$, $15 = 4^2 - 1^2$.

How does the third member of the triplet relate to the numbers we squared and added to get the largest member? It's just twice their product! That is, $4 = 2\cdot2\cdot1$, $12 = 2\cdot3\cdot2$, $24 = 2\cdot4\cdot3$, $8 = 2\cdot4\cdot1$.

This at least suggests a way to manufacture larger triplets, which can then be checked by multiplication. We need to take the squares of two numbers that don't have a common factor (otherwise, all members of the triplet will have that factor). *The simplest possible way to do that is to have the first square an even power of 2, the second an even power of 3.* (The order in the difference term could of course be reversed, depending on which one's bigger).

In fact, notice that 5 : 12 : 13 and 7 : 24 : 25 are already of this form, with 2^2, 3^2 and 2^4, 3^2. What about 2^6, 3^2? That gives 48 : 55 : 73. Try another: 2^6, 3^4 gives 17 : 144 : 145. But why stop there? Let's try something bigger: 2^{12}, 3^6. *That* gives the triplet 3,367 : 3,456 : 4,825. Not so mysterious after all.

3 Early Greek Science: Thales to Plato

3.1 The Milesians

The first recorded important contributions to Greek science are from the city of **Miletus**, near the coast of what is now Turkey, beginning with **Thales** in about **585 B.C.**, followed by **Anaximander** about **555 B.C.**, then **Anaximenes** in **535 B.C.** We shall argue below that these Milesians were the first to do real science, immediately recognizable as such to a modern scientist, as opposed to developing new technologies.

The crucial contribution of Thales to scientific thought was the **discovery of nature**. By this, we mean the idea that the natural phenomena we see around us are explicable in terms of matter interacting by natural laws, and are not the results of arbitrary acts by gods.

An example is Thales' theory of earthquakes, which was that the (presumed flat) earth is actually floating on a vast ocean, and disturbances in that ocean occasionally cause the earth to shake or even crack, just as they would a large boat. (Recall the Greeks were a seafaring nation.) The common Greek belief at the time was that the earthquakes were caused by the anger of Poseidon, god of the sea. Lightning was similarly the anger of Zeus. Later, Anaximander suggested lightning was caused by clouds being split up by the wind, which in fact is not far from the truth.

The main point here is that *the gods are just not mentioned* in analyzing these phenomena. The Milesians' view is that nature is a dynamic entity evolving in accordance with some admittedly not fully understood laws, but not being micromanaged by a bunch of gods using it to vent their anger or whatever on hapless humanity.

An essential part of the Milesians' success in developing a picture of nature was that they engaged in open, rational, critical debate about each others ideas. It was tacitly assumed that all the theories and explanations were directly competitive with one another, and all should be open to public scrutiny, so that they could be debated and judged. This is still the way scientists work. Each contribution, even that of an Einstein, depends heavily on what has gone before.

The theories of the Milesians fall into two groups:

(1) theories regarding particular phenomena or problems, of the type discussed above,

(2) speculations about the nature of the universe, and human life.

Concerning the universe, *Anaximander suggested that the earth was a cylinder*, and the sun, moon and stars were located on concentric rotating cylinders: the first recorded attempt at a *mechanical model*. He further postulated that the stars themselves were rings of fire. Again, a very bold conjecture—all heavenly bodies had previously been regarded as *living gods*.

He also considered the problem of the origin of life, which is of course more difficult to explain if you don't believe in gods! He suggested that the lower forms of life might be generated by the action of sunlight on moist earth. He also realized that a human baby is not self-sufficient for quite a long time, so postulated that the first humans were born from a certain type of fish.

All three of these Milesians struggled with the puzzle of the origin of the universe, what was here at the beginning, and what things are made of. Thales suggested that in the beginning there was only water, so somehow everything was made of it. Anaximander supposed that initially there was a boundless chaos, and the universe grew from this as from a seed. Anaximenes had a more sophisticated approach, to modern eyes. His suggestion was that originally there was only air (really meaning a gas) and the liquids and solids we see around us were formed by condensation. Notice that this means a simple initial state develops into our world using physical processes which were already familiar. Of course this leaves a lot to explain, but it's quite similar to the modern view.

3.2 Early Geometry

One of the most important contributions of the Greeks was their development of geometry, culminating in Euclid's *Elements*, a giant textbook containing all the known geometric theorems at that time (about 300 BC), presented in an elegant logical fashion.

Notice first that the word "geometry" is made up of "geo", meaning the earth, and "metry" meaning measurement of, in Greek. (The same literal translations from the Greek give geo*graphy* as *picturing* the earth (as in *graphic*) and geo*logy* as *knowledge* about the earth. Of course, the precise meanings of all these words have changed somewhat since they were first introduced.)

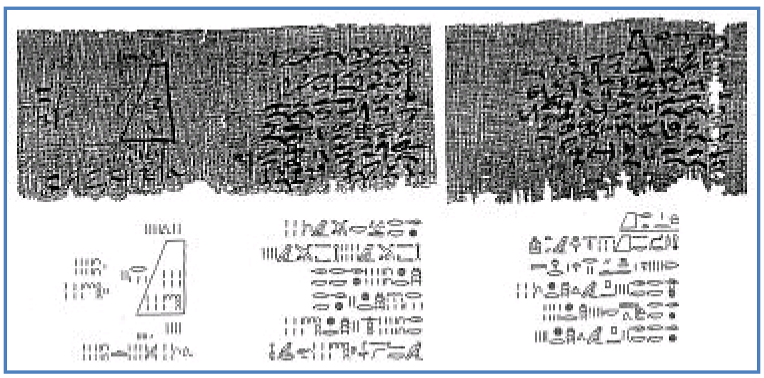
The first account we have of the beginnings of geometry is from the Greek historian **Herodotus**, writing (in **440 B.C.** or so) about the Egyptian king Sesotris (**1300 B.C.**):

"*This king moreover (so they said) divided the country among all the Egyptians by giving each an equal square parcel of land, and made this the source of his revenue, appointing the payment of a yearly tax. And any man who was robbed by the river of a part of his land would come to Sesotris and declare what had befallen him; then the king would send men to look into it and measure the space by which the land was diminished, so that thereafter it should pay the* *appointed tax in proportion to the loss. From this, to my thinking, the Greeks learnt the art of measuring land...*"

On the other hand Aristotle, writing a century later, had a more academic, and perhaps less plausible, theory of the rise of geometry:

"*..the sciences which do not aim at giving pleasure or at the necessities of life were discovered, and first in the places where men first began to have leisure. That is why the mathematical arts were founded in Egypt, for there the priestly class was allowed to be at leisure*."

However, as Thomas Heath points out in *A History of Greek Mathematics*, page 122, one might imagine that if this (that is, Aristotle's theory) were true, Egyptian geometry "would have advanced beyond the purely practical stage to something more like a theory or science of geometry. But the documents which have survived do not give any grounds for this supposition; the art of geometry in the hands of the priests never seems to have advanced beyond mere routine. The most important available source of information about Egyptian mathematics is the Papyrus Rhind written probably about 1700 BC, but copied from an original of the time of King Amenemhat III (Twelfth Dynasty), say 2200 BC."



Heath goes on to give details of what appears in this document: areas of rectangles, trapezia and triangles, areas of circles given as $(8d/9)^2$, where d is the diameter, corresponding to pi equal to 3.16 or so, about 1% off. There are approximate volume measures for hemispherical containers, and volumes for pyramids.

Another important Egyptian source is the Moscow Papyrus, which includes the very practical problem of calculating the volume of a pyramid! (Actually with a flat top: look at the figure, from Wikipedia.)

3.3 Early Geometry According to Proclus

Here's a brief overview of the early history of geometry, up to Euclid, by the Greek author Proclus Diadochus, AD 410-485. He asserts that geometry was first brought to Greece by Thales, after he spent some years in Egypt.

From his book: *Commentary on Euclid's Elements I:*

We must next speak of the origin of geometry in the present world cycle. For, as the remarkable Aristotle tells us, the same ideas have repeatedly come to men at various periods of the universe. It is not, he goes on to say, in our time or in the time of those known to us that the sciences have first arisen, but they have appeared and again disappeared, and will continue to appear and disappear, in various cycles, of which the number both past and future is countless. But since we must speak of the origin of the arts and sciences with reference to the present world cycle, it was, we say, among the Egyptians that geometry is generally held to have been discovered. It owed its discovery to the practice of land measurement. For the Egyptians had to perform such measurements because the overflow of the Nile would cause the boundary of each person's land to disappear. Furthermore, it should occasion no surprise that the discovery both of this science and of the other sciences proceeded from utility, since everything that is in the process of becoming advances from the imperfect to the perfect. The progress, then, from sense perception to reason and from reason to understanding is a natural one. And so, just as the accurate knowledge of numbers originated with the Phoenicians through their commerce and their business transactions, so geometry was discovered by the Egyptians for the reason we have indicated.

It was Thales, who, after a visit to Egypt, first brought this study to Greece. Not only did he make numerous discoveries himself, but laid the foundation for many other discoveries on the part of his successors, attacking some problems with greater generality and others more empirically. After him Mamercus the brother of the poet Stesichorus, is said to have embraced the study of geometry, and in fact Hippias of Elis writes that he achieved fame in that study.

After these Pythagoras changed the study of geometry, giving it the form of a liberal discipline, seeking its first principles in ultimate ideas, and investigating its theorems abstractly and in a purely intellectual way.

[*He then mentions several who developed this abstract approach further: Anaxagoras, Hippocrates, Theodorus, etc.]*

Plato, who lived after Hippocrates and Theodorus, stimulated to a very high degree the study of mathematics and of geometry in particular because of his zealous interest in these subjects. For he filled his works with mathematical discussions, as is well known, and everywhere sought to awaken admiration for mathematics in students of philosophy.

[*He then lists several mathematicians, including Eudoxus and Theatetus, who discovered many new geometric theorems, and began to arrange them in logical sequences-this process culminated in the work of Euclid, called his* Elements *(of geometry) about 300 BC*.]

Euclid composed *Elements*, putting in order many of the theorems of Eudoxus, perfecting many that had been worked out by Theatetus, and furnishing with rigorous proofs propositions that had been demonstrated less rigorously by his predecessors … the *Elements* contain the complete and irrefutable guide to the scientific study of the subject of geometry.

3.4 The Pythagoreans: a Cult with a Theorem, and an Irrational Discovery

Pythagoras was born about **570 B.C.** on the island of **Samos** (on the map above), less than a hundred miles from Miletus, and was thus a contemporary of Anaximenes. However, the island of Samos was ruled by a tyrant named Polycrates, and to escape an unpleasant regime, Pythagoras moved to **Croton**, a Greek town in southern Italy (at 39 05N, 17 7 30E), about 530 B.C.

Pythagoras founded what we would nowadays call a cult, a religious group with strict rules about behavior, including diet (no beans), and a belief in the immortality of the soul and reincarnation in different creatures. This of course contrasts with the Milesians' approach to life.

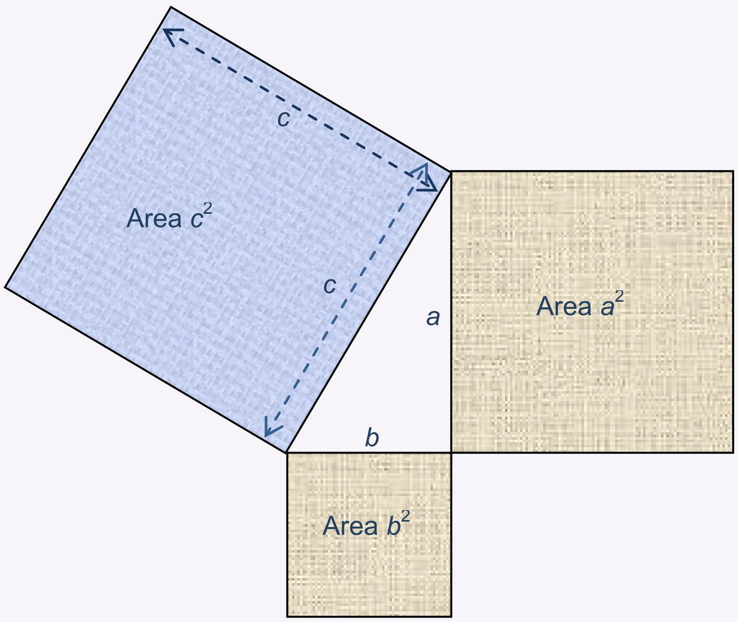
The Pythagoreans believed strongly that numbers, by which they meant the positive integers 1,2,3, ..., had a fundamental, mystical significance. The numbers were a kind of eternal truth, perceived by the soul, and not subject to the uncertainties of perception by the ordinary senses. In fact, they thought that the numbers had a physical existence, and that the universe was somehow constructed from them. In support of this, they pointed out that different musical notes differing by an octave or a fifth, could be produced by pipes (like a flute), whose lengths were in the ratios of whole numbers, 1:2 and 2:3 respectively. Note that this is an *experimental* verification of an hypothesis.

They felt that the motion of the heavenly bodies must somehow be a perfect harmony, giving out a music we could not hear since it had been with us since birth. Interestingly, they did not consider the earth to be at rest at the center of the universe. They thought it was round, and orbited about a central point daily, to account for the motion of the stars. Much was wrong with their picture of the universe, but it was not geocentric, for religious reasons. They felt the earth was not noble enough to be the center of everything, where they supposed there was a central fire. (Actually there is some debate about precisely what their picture was, but there is no doubt they saw the earth as round, and accounted for the stars' motion by the earth's rotation.)

To return to their preoccupation with numbers, they coined the term "square" number, for 4,9, etc., drawing square patterns of evenly spaced dots to illustrate this idea. The first square number, 4, they equated with justice. 5 represented marriage, of man (3) and woman (2). 7 was a mystical number. Later Greeks, like Aristotle, made fun of all this.

3.5 The Square on the Hypotenuse

Pythagoras is of course most famous for the theorem about right angled triangles, that the sum of the squares of the two sides enclosing the right angle is equal to the square of the long side, called the hypotenuse.



Pythagoras' Theorem: $c^2 = a^2 + b^2$. This means the total amount of blue material (the biggest square) is equal to the total amount of brown material (the other two squares).

This is easily proved by drawing *two* diagrams, one having four copies of the triangle arranged so that their hypotenuses form a square, and their right angles are all pointing outward, forming a larger overall square, in the other this larger square is divided differently - the four triangles are formed into two rectangles, set into corners of the square, leaving over two other square areas which are seen to be the squares on the other two sides.

You can prove it yourself by clicking here!

Actually, it seems very probable that this result was known to the Babylonians a thousand years earlier (see the discussion in the lecture on Babylon), and to the Egyptians, who, for example, used lengths of rope 3, 4 and 5 units long to set up a large right-angle for building and surveying purposes.

3.6 Rational and Irrational Numbers

As we discussed above, the Pythagoreans greatly revered the integers, the whole numbers 1, 2, 3, …, and felt that somehow they were the key to the universe. One property of the integers we'll need is the distinction between prime numbers and the rest: prime numbers have no divisors. So, no even number is prime, because all even numbers divide exactly by 2. You can map out the primes by writing down all the integers, say up to 100, cross out all those divisible by 2 (not counting 2 itself), then cross out those divisible by 3, then 5, etc. The numbers surviving this process have no divisors, they are the primes: 1, 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, … Now, any integer can be written as a product of primes: just divide it systematically first by 2, then if it divides, by 2 again, until you get something that doesn't divide by 2 (and give a whole number). Then redo the process with 3, then 5, until you're done. You can then write, for example, 12 = 2×2×3, 70 = 2×5×7 and so on.

Notice now that if you express a number as a product of its prime factors in this way, then the square of that number is the product of the same factors, but each factor appears twice as often: $(70)^2 = 2\times2\times5\times5\times7\times7$. And, in particular, note that the square of an even number has 2 appearing at least twice in its list of factors, but the square of an odd number must still be odd: if 2 wasn't on the list of factors of the number, then it won't be on the list for its square, since this is the same list with the factors just appearing twice as often.

Of course, from the earliest times, from Babylon and Egypt, people had been dealing with numbers that were not whole numbers---fractions, for example, or numbers which were integers plus fractions, such as one-and-a-half. This didn't bother the Pythagoreans too much, because after all fractions are simply ratios of two whole numbers, so they fit nicely into a slightly extended scheme.

Let's think about all possible numbers between one and ten, say, including all those with fractional parts, such as 3/2 or 4567/891, to choose a number at random. Suppose we take a piece of paper, mark on it points for the whole numbers 1, 2, 3, ...,10. Then we put marks for the halves, then the quarters and three quarters. Next we put marks at the thirds, 4/3, 5/3, 7/3, up to 29/3. Then we do the fifths, then the sevenths, ... Then we buy a supercomputer with a great graphics program to put in the higher fractions one after the other at lightning speed!

The question is: is this list of fractions *all the numbers there are* between one and ten?

In other words, can we prove that there's a number you could never ever reach by this method, no matter how fast your computer?

Two thousand five hundred years ago, the Pythagoreans figured out the answer to this question.

The answer is ***yes***: there *are* numbers which are not fractions—that is, they cannot be expressed as ratios of integers.

This discovery greatly upset the Pythagoreans, since they revered the integers as the mystical foundation of the universe, and now apparently they were not even sufficient foundation for the numbers. Ironically, this unnerving discovery followed from applying their very own theorem—Pythagoras' theorem—to the simplest possible right-angled triangle: half a square, a triangle with its two shorter sides both equal to one.

This means its long side—the hypotenuse—has a length whose square is two.

We shall now go through their argument showing that the length of this longest side cannot be written as a ratio of two integers, no matter how large you choose the integers to be.

The basic strategy of the proof is to assume it can be written as a ratio of integers, then prove this leads to a contradiction.

So, we assume we *can* write this number—the length of the longest side—as a ratio of two whole numbers, in other words a fraction m/n. This is the length whose square is 2, so $m^2/n^2 = 2$, from which $m^2 = 2n^2$.

Now all we have to do is to find two whole numbers such that the square of one is exactly twice the square of the other. How difficult can this be? To get some idea, let's write down the squares of some numbers and look:

$1^2 = 1$, $2^2 = 4$, $3^2 = 9$, $4^2 = 16$, $5^2 = 25$, $6^2 = 36$, $7^2 = 49$, $8^2 = 64$, $9^2 = 81$, $10^2 = 100$, $11^2 = 121$, $12^2 = 144$, $13^2 = 169$, $14^2 = 196$, $15^2 = 225$, $16^2 = 256$, $17^2 = 289$, … .

On perusing this table, you will see we have some near misses: 3^2 is only one more than twice 2^2, 7^2 is only one less than twice 5^2, and 17^2 is only one more than twice 12^2. It's difficult to believe that if we keep at it, we're not going to find a direct hit eventually.

In fact, though, it turns out *this never happens*, and that's what the Pythagoreans proved. Here's how they did it.

First, assume we *canceled* any common factors between numerator and denominator.

This means that m and n can't both be even.

Next, notice that ***the square of an even number is even***. This is easy to check: if a is an even number, it can be written $a = 2b$, where b is another whole number. Therefore, $a^2 = 2 \text{x} 2 \text{x} b^2$, so on fact a^2 is not only even, *it has 4 as a factor*.

On the other hand, ***the square of an odd number is always odd***. If a number doesn't have 2 as a factor, multiplying it by itself won't give a number that has 2 as a factor.

Now, back to the length of the square's diagonal, m/n, with $\mathbf{m^2 = 2n^2}$.

Evidently, ***m^2 must be even***, because it equals $2n^2$, which has a factor 2.

Therefore, from what we have just said above about squares of even and odd numbers, ***m must itself be even***.

This means, though, that ***m^2 must be divisible by 4***.

This means that $2n^2$ must be divisible by 4, since $m^2 = 2n^2$ -- but in this case, ***n^2 must be divisible by 2!***

It follows that ***n must itself be even***—BUT we stated at the beginning that we had *canceled* any common factors between m and n. This would include any factor of 2, so they *can't* both be even!

Thus a watertight logical argument has led to a contradiction.

The only possible conclusion is: the original assumption is incorrect.

This means that the diagonal length of a square of side 1 **cannot** be written as the ratio of two integers, no matter how large we are willing to let them be.

This was the first example of an *irrational* number—one that is not a *ratio* of integers.

Legend has it that the Pythagoreans who made this discovery public died in a shipwreck.

3.7 What's so Important about Irrational Numbers?

The historical significance of the above proof is that it establishes something new in mathematics, which couldn't have been guessed, and, in fact, *something the discoverers didn't want to be true*. Although fractions very close to the square root of 2 had been found by the Babylonians and Egyptians, there is no hint that they considered the possibility that *no fraction would ever be found* representing the square root of 2 exactly.

The kind of abstract argument here is far removed from practical considerations where geometry is used for measurement. In fact, it is irrelevant to measurement - one can easily find approximations better than any possible measuring apparatus. The reason the Pythagoreans worked on this problem is because they thought they were investigating the fundamental structure of the universe.

Abstract arguments of this type, and the beautiful geometric arguments the Greeks constructed during this period and slightly later, seemed at the time to be merely mental games, valuable for developing the mind, as Plato emphasized. In fact, these arguments have turned out, rather surprisingly, to be on the right track to modern science, as we shall see.

3.8 Change and Constancy in the Physical World

Over the next century or so, 500 B.C.- 400 B.C., the main preoccupation of philosophers in the Greek world was that when we look around us, we see things changing all the time. How is this to be reconciled with the feeling that the universe must have some constant, eternal qualities? **Heraclitus**, from Ephesus, claimed that **"everything flows"**, and even objects which appeared static had some inner tension or dynamism. **Parminedes**, an Italian Greek, came to the opposite conclusion, that **nothing ever changes**, and apparent change is just an illusion, a result of our poor perception of the world.

This may not sound like a very promising debate, but in fact it is, because, as we shall see, trying to analyze **what is changing and what isn't** in the physical world leads to the ideas of **elements**, **atoms** and **conservation laws**, like the conservation of matter.

The first physicist to give a clear formulation of a possible resolution of the problem of change was **Empedocles** around **450 B.C**., who stated that everything was made up of f**our elements**: **earth, water, air and fire**. He asserted that **the elements themselves were eternal and unchanging**. Different substances were made up of the elements in different proportions, just as all colors can be created by mixing three primary colors in appropriate proportions. Forces of attraction and repulsion (referred to as love and strife) between these elements cause coming together and separation, and thus apparent change in substances. Another physicist, **Anaxogoras**, argued that no natural substance can be more elementary than any other, so there were an infinite number of elements, and everything had a little bit of everything else in it. He was particularly interested in nutrition, and argued that food contained small amounts of hair, teeth, etc., which our bodies are able to extract and use.

The most famous and influential of the fifth century B.C. physicists, though, were the **atomists**, **Leucippus** of Miletus and **Democritus** of Abdera. They claimed that the physical world consisted of atoms in constant motion in a void, rebounding or cohering as they collide with each other. Change of all sorts is thus accounted for on a basic level by the atoms separating and recombining to form different materials. The atoms themselves do not change. This sounds amazingly like our modern picture, but of course it was all conjecture, and when they got down to relating the atoms to physical properties, Democritus suggested, for example, that things made of sharp, pointed atoms tasted acidic, those of large round atoms tasted sweet. There was also some confusion between the idea of physical indivisibility and that of mathematical indivisibility, meaning something that only exists at a point. The atoms of Democritus had shapes, but it is not clear if he realized this implied they could, at least conceptually, be divided. This caused real problems later on, especially since at that time there was no experimental backing for an atomic theory, and it was totally rejected by Aristotle and others.

3.9 Hippocrates and his Followers

It is also worth mentioning that at this same time, on the island of **Kos** (see map) just a few miles from Miletus, lived the first great doctor, **Hippocrates**. He and his followers adopted the Milesian point of view, applied to disease, that it was not caused by the gods, even epilepsy, which was called the sacred disease, but there was some rational explanation, such as infection, which could perhaps be treated.

Here's a quote from one of Hippocrates' followers, writing about epilepsy in about 400 B.C.:

"*It seems to me that the disease called sacred … has a natural cause, just as other diseases have. Men think it divine merely because they do not understand it. But if they called everything divine that they did not understand, there would be no end of divine things! … If you watch these fellows treating the disease, you see them use all kinds of incantations and magic—but they are also very careful in regulating diet. Now if food makes the disease better or worse, how can they say it is the gods who do this? … It does not really matter whether you call such things divine or not. In Nature, all things are alike in this, in that they can be traced to preceding causes*."

The Hippocratic doctors criticized the philosophers for being too ready with postulates and hypotheses, and not putting enough effort into careful observation. These doctors insisted on careful, systematic observation in diagnosing disease, and a careful sorting out of what was relevant and what was merely coincidental. Of course, this approach is the right one in all sciences.

3.10 Plato

In the fourth century B.C., Greek intellectual life centered increasingly in Athens, where first Plato and then Aristotle established schools, the Academy and the Lyceum respectively, which were really the first universities, and attracted philosophers and scientists from all over Greece.

Actually, this all began somewhat earlier with Socrates, Plato's teacher, who, however, was not a scientist, and so not central to our discussion here. One of Socrates' main concerns was how to get the best people to run the state, and what were the ideal qualities to be looked for in such leaders. He believed in free and open discussion of this and other political questions, and managed to make very clear to everybody that he thought the current leaders of Athens were a poor lot. In fact, he managed to make an enemy of almost everyone in a position of power, and he was eventually brought to trial for corrupting the young with his teachings. He was found guilty, and put to death.

This had a profound effect on his pupil Plato, a Greek aristocrat, who had originally intended to involve himself in politics. Instead, he became an academic-in fact, he invented the term! He, too, pondered the question of what is the ideal society, and his famous book *The Republic* is his suggested answer. He was disillusioned with Athenian democracy after what had happened to Socrates, and impressed with Sparta, an authoritarian state which won a war, the Peloponnesian war, against Athens. Hence his Republic has rather a right wing, antidemocratic flavor. However, he tries to ensure that the very best people in each generation are running the state, and he considers, being a philosopher, that the best possible training for these future leaders is a strong grounding in logic, ethics and dealing with abstract ideas. This is made particularly clear on p 67,8 of Lloyd, where a quote from the Republic is given, in which Socrates is emphasizing how important it is for future leaders to study astronomy. Glaucon agrees that astronomy is useful in navigation, military matters and accurately determining seasons for planting, etc., to which Socrates responds emphatically that these reasons are not nearly as important as the training in abstract reasoning it provides.

Plato, then, had a rather abstract view of science, reminiscent of the Pythagoreans. In particular, he felt that the world we apprehend with our senses is less important than the underlying world of pure eternal forms we perceive with our reason or intellect, as opposed to our physical senses. This naturally led him to downgrade the importance of careful observation, for instance in astronomy, and to emphasize the analytical, mathematical approach.

Plato believed the universe was created by a rational god, who took chaotic matter and ordered it, but he also believed that because of the inherent properties of the matter itself, his god was not omnipotent, in the sense that there were limits as to how good the universe could be: one of his examples was that smart people have large brains (he thought), but if you make the brain too large by having a very thin skull, they won't last long! He felt this need to compromise was the explanation of the presence of evil in a universe created by a beneficent god.

Plato's concentration on perfect underlying forms did in fact lead to a major contribution to astronomy, despite his own lack of interest in observation. He stated that the main problem in astronomy was to account for the observed rather irregular motion of the planets by some combination of perfect motions, that is, circular motions. This turned out to be a very fruitful way of formulating the problem.

Plato's theory of matter was based on Empedocles' four elements, fire, air, water and earth. However, he did not stop there. He identified each of these elements with a perfect form, one of the regular solids, fire with the tetrahedron, air with the octahedron, water with the icosahedron and earth with the cube. He divided each face of these solids into elementary triangles (45 45 90 and 30 60 90) which he regarded as the basic units of matter. He suggested that water could be decomposed into fire and air by the icosahedron breaking down to two octahedra and a tetrahedron. This looks like a kind of atomic or molecular theory, but his strong conviction that all properties of matter could eventually be deduced by pure thought, without resort to experiment, proved counterproductive to the further development of scientific understanding for centuries. It should perhaps be mentioned, though, that the latest theory in elementary particle physics, string theory, known modestly as the theory of everything, also claims that all physical phenomena should be deducible from a very basic mathematical model having in its formulation no adjustable parameters—a perfect form.

3.11 References

Lloyd, G. E. R. (1970). Early Greek Science: Thales to Aristotle, Norton.

Heath, Sir Thomas (1921, 1981). *A History of Greek Mathematics*, *Volume I*. Dover.

4 Motion in the Heavens: Stars, Sun, Moon, Planets

4.1 Introduction

The purpose of this lecture is just to review the various motions observed in the heavens in the simplest, most straightforward way. We shall ignore for the moment refinements like tiny deviations from simple motion, but return to them later.

It is illuminating to see how these observed motions were understood in early times, and how we see them now. Of course, *you* know the Earth rotates and orbits around the Sun. However, I want you to be *bilingual* for this session: to be able to visualize *also* the ancient view of a fixed Earth, and rotating heavens, and be able to think from *both* points of view.

This is really largely an exercise in three-dimensional visualization—that's the hard part! But without some effort to see the big picture, you will not be able to appreciate some really nice things, like the phases of the moon, eclipses, and even just the seasons. You really need to have a clear picture of the Earth orbiting around the Sun and at the same time rotating about an axis tilted relative to the plane the orbit lies in, with the axis of rotation always pointing at the same star, and not changing its direction as the Earth goes around the Sun. Then you must add to your picture the Moon orbiting around the Earth once a month, the plane of its orbit tilted five degrees from the plane of the Earth's orbit around the Sun. Then we add in the planets … .

Some of these topics are treated nicely in *Theories of the World from Antiquity to the Copernican Revolution*, by Michael J. Crowe, Dover.

4.2 Looking at the Stars

There is one star that always stays in the same place in the sky, as seen from Charlottesville (or anywhere else in the northern hemisphere). This is Polaris, the North Star. All the other stars move in circular paths around Polaris, with a period of 24 hours. This was understood in ancient times by taking the stars to be fixed to the inside surface of a large sphere, the "starry vault", which was the outer boundary of the Universe, and contained everything else.

Of course, we only see the stars move around *part* of their circular path, because when the Sun comes up, the bright blue scattered Sunlight—the blue sky—drowns out the starlight. If there were no atmosphere, we would see the stars all the time, and see the complete circles for those that stayed above the horizon.

Try to picture yourself inside this large, spherical rotating starry vault with stars attached, and visualize the paths of the stars as they wheel overhead. Think about the paths the stars would take as seen from the North Pole, from the Equator, and from Charlottesville.

4.3 Motion of the Sun

Every day the Sun rises in the east, moves through the southern part of the sky and sets in the west. If there were no atmosphere so that we could see Polaris all the time, would the Sun also be going in a circular path centered on Polaris?

The answer is yes. (Well, almost).

If you were at the North Pole in the middle of summer, lying on your back, you would see the Sun go around in a circle in the sky, anticlockwise. The circle would be centered on Polaris, which is directly overhead, except for the fact that you wouldn't see Polaris all summer, since it wouldn't be dark. Here of course we see the Sun circling part of the time, and see Polaris the other part of the time, so it isn't completely obvious that the Sun's circling Polaris. Does the Sun circle clockwise or anticlockwise for us? It depends on how you look at it—in winter, when it's low in the sky, we tend to look "from above", see the Sun rise in the east, move in a low path via the south towards the west, and that looks clockwise—unless you're lying on your back.

Actually the Sun moves very slightly each day relative to the starry vault. This would be obvious if there were no atmosphere, so we could just watch it, but this can also be figured out, as the Greeks and before them the Babylonians did, by looking closely at the stars in the west just after sunset and seeing where the Sun fits into the pattern.

It turns out that the Sun moves almost exactly one degree per day against the starry vault, so that after one year it's back where it started. This is no coincidence—no doubt this is why the Babylonians chose their angular unit as the degree (they also liked 60).

Anyway, the Sun goes around in the circular path along with the starry vault, and at the same time slowly progresses along a path in the starry vault. This path is called the *ecliptic*.

If we visualize Polaris as the "North Pole" of the starry vault, and then imagine the vault's "Equator", the ecliptic is a great circle tilted at 23 ½ degrees to the "equator". The Sun moves along the ecliptic from west to east. (Imagine the Earth were not rotating at all relative to the stars. How would the Sun appear to move through the year?)

The motion of the Sun across the starry vault has been known at least since the Babylonians, and interpreted in many colorful ways. Compare our present view of the stars, thermonuclear reactions in the sky, with the ancient view (see *Hemisphaerium Boreale*, Appendix to Heath's *Greek Astronomy*).

Many of the ancients believed, to varying degrees, that there were spirits in the heavens, and the arrangements of stars suggested animals, and some people.

The Sun's path through all this, the ecliptic, endlessly repeated year after year, and the set of *constellations* (the word just means "group of stars") and the animals they represented became known as the Zodiac. ("zo" being the same Greek word for animal that appears in "Zoo".) So this is your sign: where in its path through this zoo was the Sun on the day you were born?

Notice that the print shows the Sun's path through the northern hemisphere, that is, for our summer. The furthest north (closest to Polaris) it gets is on June 21, when it is in Cancer, it is then overhead on the Tropic of Cancer, 23½ degrees north of the Equator.

In other words, the spherical Earth's surface is visualized as having the same center as the larger sphere of the starry vault, so when in its journey across this vault the Sun reaches the tropic of the vault, it will naturally be overhead at the corresponding point on the Earth's tropic which lies directly below the tropic on the vault.

Here's a more spectacular demonstration of the same thing: notice, for example, the Plough (also known as Ursa Major, the great bear) in the tail and body of the bear, and the familiar astrological collection of animals around the zodiac (from http://www.atlascoelestis.com/5.htm)

4.4 Motion of the Moon against the Starry Vault

The Sun goes around the starry vault once a year, the Moon goes completely around every month.

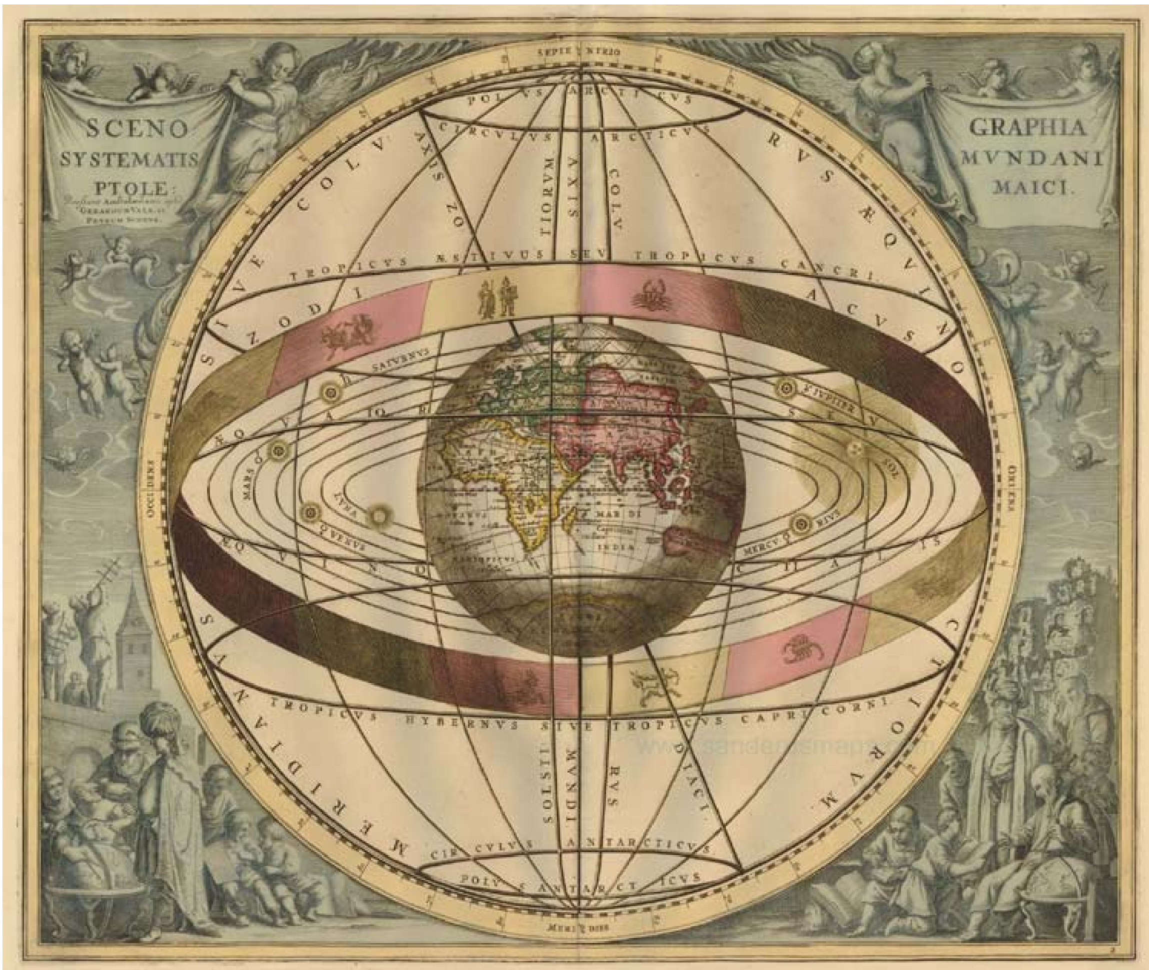
Does it follow the same path as the Sun?

The answer is no, but it's close. It always stays within 5 degrees of the ecliptic, so it goes through the same set of constellations, "the Moon is in the Seventh House" and all that. In fact, the "houses"—the signs of the Zodiac—are defined to occupy a band of the stars that stretches eight degrees either way from the ecliptic, because that turns out to be wide enough that the Sun, Moon and all the planets lie within it.

How can we understand the Moon's motion from our present perspective? If the Earth, the Moon and the Sun were all in the same plane, in other words, if the moon's orbit was in the same plane as the Earth's orbit around the Sun, the Moon would follow the ecliptic. In fact, the Moon's orbit is tilted at 5 degrees to the Earth's orbit around the Sun.

This also explains why eclipses of the Moon (and Sun) don't happen every month, which they would if everything was in the same plane. In fact, they only occur when the moon's path crosses the ecliptic, hence the name.

A nice three-dimensional representation, published by Cellario in 1627, can be found at http://www.atlascoelestis.com/Cell%2009.htm : here it is:



Notice the band representing the zodiac.

4.5 Motion of the Planets

Since ancient times it has been known that five of the "stars" moved across the sky: Mercury, Venus, Mars, Jupiter and Saturn. They were termed "planets" which simply means wanderers.

Are their paths in the starry vault also related to the ecliptic?

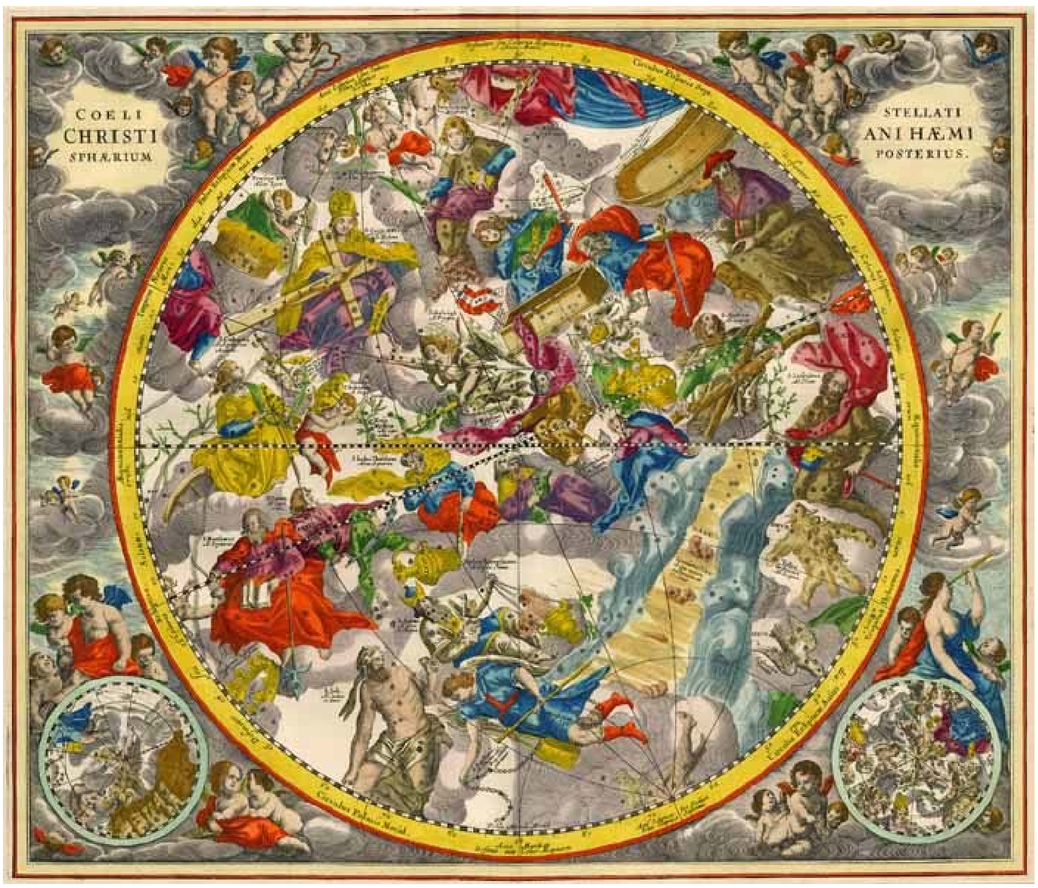
The answer is yes. They all stay within 8 degrees of the ecliptic, and in fact this is the *definition* of the Zodiac: the band of sky within eight degrees of the ecliptic, and for this reason.

Do they go all the way round?

Yes they do, but Mercury never gets more than 28 degrees away from the Sun, and Venus never more than 46 degrees. Thus as the Sun travels around the ecliptic, these two swing backwards and forwards across the Sun.

The other planets are not tethered to the Sun in the same way, but they also have some notable behavior—in particular, they occasionally loop backwards for a few weeks before resuming their steady motion.

Cultural note: an attempt was made about the same time by Julius Schiller to replace the barbaric twelve signs of the zodiac with the twelve apostles: http://www.atlascoelestis.com/epi%20schiller%20cellario.htm



It didn't catch on.

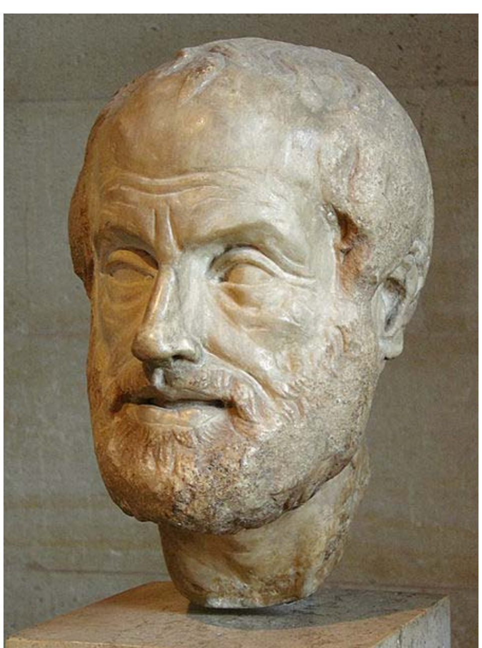
5 Aristotle

5.1 Beginnings of Science and Philosophy in Athens

Let us first recap briefly the emergence of philosophy and science in Athens after around 450 B.C. It all began with **Socrates**, who was born in 470 B.C. Socrates was a true philosopher, a lover of wisdom, who tried to elicit the truth by what has become known as the Socratic method, in which by a series of probing questions he forced successive further clarification of thought. Of course, such clarity often reveals that the other person's ideas don't in fact make much sense, so that although Socrates made a lot of things much clearer, he wasn't a favorite of many establishment politicians. For example, he could argue very convincingly that traditional morality had no logical basis. He mostly lectured to the sons of well-to-do aristocrats, one of whom was **Plato**, born in 428 B.C. Plato was a young man when Athens was humiliated by Sparta in the Peloponnesian War, and Plato probably attributed the loss to Athens' being a democracy, as opposed to the kind of fascist war-based state Sparta was. Plato founded an Academy. The name came (at least in legend) from one *Academus*, a landowner on whose estate Plato and other philosophers met regularly. The important point is that this was the first university. All the people involved were probably aristocrats, and they discussed everything: politics, economics, morality, philosophy, mathematics and science. One of their main concerns was to find what constituted an ideal city-state. Democracy didn't seem to have worked very well in their recent past. Plato's ideas are set out in the *Republic*.

5.2 Plato's Idea of a Good Education

What is interesting about the *Republic* from our point of view is the emphasis on a good education for the elite group in charge of Plato's ideal society. In particular, Plato considered education in mathematics and astronomy to be excellent ways of sharpening the mind. He believed that intense mental exercise of this kind had the same effect on the mind that a rigorous physical regimen did on the body. Students at the Academy covered a vast range of subjects, but there was a sign over the door stating that some knowledge of *mathematics* was needed to enter—nothing else was mentioned! Plato in particular loved geometry, and felt that the beauty of the five regular solids he was the first to categorize meant they must be fundamental to nature, they must somehow be the shapes of the atoms. Notice that this approach to physics is not heavily dependent on observation and experiment.



5.3 Aristotle and Alexander

We turn now to the third member of this trio, **Aristotle**, born in 384 B.C. in Stagira, in Thrace, at the northern end of the Aegean, near Macedonia.

Aristotle's father was the family physician of King Philip of Macedonia. At the age of eighteen, Aristotle came to Athens to study at Plato's Academy, and stayed there twenty years until Plato's death in 348 B.C. (Statue is a Roman copy of a Greek original, in the Louvre, photographer Eric Gaba (User:Sting), July 2005.)

Five years after Plato's death, Aristotle took a position as tutor to King Philip of Macedonia's thirteen year old son Alexander. He stayed for three years. It is not clear what impact, if any, Aristotle's lessons had, but Alexander, like his father, was a great admirer of Greek civilization, even though the Athenians considered Macedonia the boondocks. In fact, when his father Philip died in 336 B.C., Alexander did his best to spread Greek civilization as far as he could. Macedonia had an excellent army, and over the next thirteen years Alexander organized Greece as a federation of city states, conquered Persia, the Middle East, Egypt, southern Afghanistan, some of Central Asia and the Punjab in India.

The picture below is a fortress built by Alexander's army in Herat, Afghanistan, and still standing. (Picture from http://flickr.com/photos/koldo/67606119/ , author koldo / Koldo Hormaza .)

He founded Greek cities in many places, the greatest being Alexandria in Egypt, which in fact became the most important center of Greek science later on, and without which all of Greek learning might have been lost. The Greek cities became restless, predictably but rather ungratefully, when he demanded to be treated as a god. He died of a fever at age 33.

5.4 Aristotle Founds the Lyceum

Aristotle came back to Athens in 335 B.C., and spent the next twelve years running his own version of an academy, which was called the Lyceum, named after the place in Athens where it was located, an old temple of Apollo. (French high schools are named *lycee* after Aristotle's establishment.) Aristotle's preferred mode of operation was to spend a lot of time walking around talking with his colleagues, then write down his arguments. The Aristotelians are often called the *Peripatetics*: people who walk around.

Aristotle wrote extensively on all subjects: politics, metaphysics, ethics, logic and science. He didn't care for Plato's rather communal Utopia, in which the women were shared by the men, and the children raised by everybody, because for one thing he feared the children would be raised by nobody. His ideal society was one run by cultured gentlemen. He saw nothing wrong with slavery, provided the slave was naturally inferior to the master, so slaves should not be Greeks. This all sounds uncomfortably similar to Jefferson's Virginia, perhaps not too surprising since Greek was a central part of a gentleman's education in Jefferson's day.

5.5 Aristotle's Science

Aristotle's approach to science differed from Plato's. He agreed that the highest human faculty was reason, and its supreme activity was contemplation. However, in addition to studying what he called "first philosophy" - metaphysics and mathematics, the things Plato had worked on, Aristotle thought it also very important to study "second philosophy": the world around us, from physics and mechanics to biology. Perhaps being raised in the house of a physician had given him an interest in living things.

What he achieved in those years in Athens was to begin a school of organized scientific inquiry on a scale far exceeding anything that had gone before. He first clearly defined what was scientific knowledge, and why it should be sought. In other words, he single-handedly invented science as the collective, organized enterprise it is today. Plato's Academy had the equivalent of a university mathematics department, Aristotle had the first science department, truly excellent in biology, but, as we shall see, a little weak in physics. After Aristotle, there was no comparable professional science enterprise for over 2,000 years, and his work was of such quality that it was accepted by all, and had long been a part of the official orthodoxy of the Christian Church 2,000 years later. This was unfortunate, because when Galileo questioned some of the assertions concerning simple physics, he quickly found himself in serious trouble with the Church.

5.6 Aristotle's Method

Aristotle's method of investigation varied from one natural science to another, depending on the problems encountered, but it usually included:

1. defining the subject matter
2. considering the difficulties involved by reviewing the generally accepted views on the subject, and suggestions of earlier writers
3. presenting his own arguments and solutions.

Again, this is the pattern modern research papers follow, Aristotle was laying down the standard professional approach to scientific research. The arguments he used were of two types: *dialectical*, that is, based on logical deduction; and *empirical*, based on practical considerations.

Aristotle often refuted an opposing argument by showing that it led to an absurd conclusion, this is called *reductio ad absurdum* (reducing something to absurdity). As we shall see later, Galileo used exactly this kind of argument against Aristotle himself, to the great annoyance of Aristotelians 2,000 years after Aristotle.

Another possibility was that an argument led to a *dilemma*: an apparent contradiction. However, dilemmas could sometimes be resolved by realizing that there was some ambiguity in a definition, say, so *precision of definitions* and usage of terms is *essential* to productive discussion in any discipline.

5.7 "Causes"

In contrast to Plato, who felt the only worthwhile science to be the contemplation of abstract forms, Aristotle practiced detailed observation and dissection of plants and animals, to try to understand how each fitted into the grand scheme of nature, and the importance of the different organs of animals. His motivation is made clear by the following quote from him (in Lloyd, p105):

For even in those kinds [of animals] that are not attractive to the senses, yet to the intellect the craftsmanship of nature provides extraordinary pleasures for those who can recognize the causes in things and who are naturally inclined to philosophy.

His study of nature was a search for "causes." What, exactly are these "causes"? He gave some examples (we follow Lloyd's discussion here). He stated that any object (animal, plant, inanimate, whatever) had four *attributes*:

- matter
- form
- moving cause
- final cause

For a table, the matter is wood, the form is the shape, the moving cause is the carpenter and the final cause is the reason the table was made in the first place, for a family to eat at, for example. For man, he thought the matter was provided by the mother, the form was a rational two-legged animal, the moving cause was the father and the final cause was to become a fully grown human being. He did not believe nature to be conscious, he believed this final cause to be somehow innate in a human being, and similarly in other organisms. Of course, fulfilling this final cause is not inevitable, some accident may intervene, but apart from such exceptional circumstances, nature is regular and orderly.

To give another example of this central concept, he thought the "final cause" of an acorn was to be an oak tree. This has also been translated by Bertrand Russell (*History of Western Philosophy*) as the "nature" of an acorn is to become an oak tree. It is certainly very natural on viewing the living world, especially the maturing of complex organisms, to view them as having innately the express purpose of developing into their final form.

It is interesting to note that this whole approach to studying nature fits very well with Christianity. The idea that every organism is beautifully crafted for a particular function - its "final cause" - in the grand scheme of nature certainly leads naturally to the thought that all this has been designed by somebody.

5.8 Biology

Aristotle's really great contribution to natural science was in biology. Living creatures and their parts provide far richer evidence of form, and of "final cause" in the sense of design for a particular purpose, than do inanimate objects. He wrote in detail about five hundred different animals in his works, including a hundred and twenty kinds of fish and sixty kinds of insect. He was the first to use dissection extensively. In one famous example, he gave a precise description of a kind of dog-fish that was not seen again by scientists until the nineteenth century, and in fact his work on this point was disbelieved for centuries.

Thus both Aristotle and Plato saw in the living creatures around them overwhelming evidence for "final causes", that is to say, evidence for design in nature, a different design for each species to fit it for its place in the grand scheme of things. Empedocles, on the other hand, suggested that maybe creatures of different types could come together and produce mixed offspring, and those well adapted to their surroundings would survive. This would seem like an early hint of Darwinism, but it was not accepted, because as Aristotle pointed out, men begat men and oxen begat oxen, and there was no evidence of the mixed creatures Empedocles suggested.

Although this idea of the "nature" of things accords well with growth of animals and plants, it leads us astray when applied to the motion of inanimate objects, as we shall see.

5.9 Elements

Aristotle's theory of the basic constituents of matter looks to a modern scientist perhaps something of a backward step from the work of the atomists and Plato. Aristotle assumed all substances to be compounds of four *elements*: earth, water, air and fire, and each of these to be a combination of two of four *opposites*, hot and cold, and wet and dry. (Actually, the words he used for wet and dry also have the connotation of softness and hardness).

Aristotle's whole approach is more in touch with the way things present themselves to the senses, the way things really seem to be, as opposed to abstract geometric considerations. Hot and cold, wet and dry are qualities immediately apparent to anyone, this seems a very natural way to describe phenomena. He probably thought that the Platonic approach in terms of abstract concepts, which do not seem to relate to our physical senses but to our reason, was a completely wrongheaded way to go about the problem. It has turned out, centuries later, that the atomic and mathematical approach was on the right track after all, but at the time, and in fact until relatively recently, Aristotle seemed a lot closer to reality. He discussed the properties of real substances in terms of their "elemental" composition at great length, how they reacted to fire or water, how, for example, water evaporates on heating because it goes from cold and wet to hot and wet, becoming air, in his view. Innumerable analyses along these lines of commonly observed phenomena must have made this seem a coherent approach to understanding the natural world.

5.10 Dynamics: Motion, And Why Things Move

It is first essential to realize that the world Aristotle saw around him in everyday life was very different indeed from that we see today. Every modern child has since birth seen cars and planes moving around, and soon finds out that these things are not alive, like people and animals. In contrast, most of the motion seen in fourth century Greece *was* people, animals and birds, all very much alive. This motion all had a purpose, the animal was moving to someplace it would rather be, for some reason, so the motion was directed by the animal's *will*. For Aristotle, this motion was therefore fulfilling the "nature" of the animal, just as its natural growth fulfilled the nature of the animal.

To account for motion of things obviously *not* alive, such as a stone dropped from the hand, he extended the concept of the "nature" of something to inanimate matter. He suggested that the motion of such inanimate objects could be understood by postulating that *elements tend to seek their natural place* in the order of things, so earth moves downwards most strongly, water flows downwards too, but not so strongly, since a stone will fall through water. In contrast, air moves up (bubbles in water) and fire goes upwards most strongly of all, since it shoots upward through air. This general theory of how elements move has to be elaborated, of course, when applied to real materials, which are mixtures of elements. He would conclude that wood, say, has both earth and air in it, since it does not sink in water.

5.11 Natural Motion and Violent Motion

Of course, things also sometimes move because they are pushed. A stone's natural tendency, if left alone and unsupported, is to fall, but we can lift it, or even throw it through the air. Aristotle termed such forced motion "violent" motion as opposed to natural motion. The term "violent" here connotes that some external force is applied to the body to cause the motion. (Of course, from the modern point of view, gravity is an external force that causes a stone to fall, but even Galileo did not realize that. Before Newton, the falling of a stone was considered natural motion that did not require any outside help.)

(*Question*: I am walking steadily upstairs carrying a large stone when I stumble and both I and the stone go clattering down the stairs. Is the motion of the stone before the stumble natural or violent? What about the motion of the stone (and myself) after the stumble?)

5.12 Aristotle's Laws of Motion

Aristotle was the first to think *quantitatively* about the speeds involved in these movements. He made two quantitative assertions about how things fall (natural motion):

1. Heavier things fall faster, the speed being proportional to the weight.
2. The speed of fall of a given object depends *inversely* on the density of the medium it is falling through, so, for example, the same body will fall twice as fast through a medium of half the density.

Notice that these rules have a certain elegance, an appealing quantitative simplicity. And, if you drop a stone and a piece of paper, it's clear that the heavier thing does fall faster, and a stone falling through water is definitely slowed down by the water, so the rules at first appear plausible. The surprising thing is, in view of Aristotle's painstaking observations of so many things, he didn't check out these rules in any serious way. It would not have taken long to find out if half a brick fell at half the speed of a whole brick, for example. Obviously, this was not something he considered important.

From the second assertion above, he concluded that *a vacuum cannot exist*, because if it did, since it has zero density, all bodies would fall through it at infinite speed which is clearly nonsense.

For *violent* motion, Aristotle stated that the *speed* of the moving object was *in direct proportion to* the applied *force*.

This means first that if you stop pushing, the object stops moving. This certainly sounds like a reasonable rule for, say, pushing a box of books across a carpet, or a Grecian ox dragging a plough through a field. (This intuitively appealing picture, however, fails to take account of the large frictional force between the box and the carpet. If you put the box on a sled and pushed it across ice, it wouldn't stop when you stop pushing. Galileo realized the importance of friction in these situations.)

5.13 Planetary Dynamics

The idea that motion (of inanimate objects) can be accounted for in terms of them seeking their natural place clearly cannot be applied to the planets, whose motion is apparently composed of circles. Aristotle therefore postulated that the heavenly bodies were not made up of the four elements earth, water, air and fire, but of a fifth, different, element called *aither*, whose natural motion was circular. This was not very satisfying for various reasons. Somewhere between here and the moon a change must take place, but where? Recall that Aristotle did not believe that there was a void anywhere. If the sun has no heat component, why does sunlight seem so warm? He thought it somehow generated heat by friction from the sun's motion, but this wasn't very convincing, either.

5.14 Aristotle's Achievements

To summarize: Aristotle's philosophy laid out an approach to the investigation of all natural phenomena, to determine form by detailed, systematic work, and thus arrive at final causes. His logical method of argument gave a framework for putting knowledge together, and deducing new results. He created what amounted to a fully-fledged professional scientific enterprise, on a scale comparable to a modern university science department. It must be admitted that some of his work - unfortunately, some of the physics - was not up to his usual high standards. He evidently found falling stones a lot less interesting than living creatures. Yet the sheer scale of his enterprise, unmatched in antiquity and for centuries to come, gave an authority to all his writings.

It is perhaps worth reiterating the difference between Plato and Aristotle, who agreed with each other that the world is the product of rational design, that the philosopher investigates the form and the universal, and that the only true knowledge is that which is irrefutable. The essential difference between them was that Plato felt *mathematical reasoning* could arrive at the truth with little outside help, but Aristotle believed *detailed empirical investigations* of nature were essential if progress was to be made in understanding the natural world.

Books I used to prepare this lecture:

Early Greek Science: Thales to Aristotle, G. E. R. Lloyd, Norton, N.Y., 1970. An excellent inexpensive paperback giving a more detailed presentation of many of the subjects we have discussed. My sections on Method and Causes, in particular, follow Lloyd's treatment.

History of Western Philosophy, Bertrand Russell. An opinionated but very entertaining book, mainly on philosophy but with a fair amount of science and social analysis.

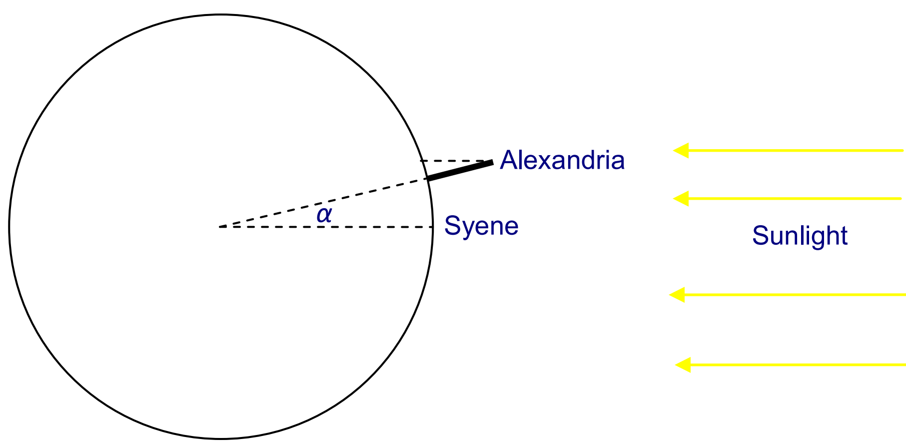
6 Measuring the Solar System

In this lecture, we shall show how the Greeks made the first real measurements of astronomical distances: the size of the earth and the distance to the moon, both determined quite accurately, and the distance to the sun, where their best estimate fell short by a factor of two.

6.1 How Big is the Earth?

The first reasonably good measurement of the earth's size was done by Eratosthenes, a Greek who lived in Alexandria, Egypt, in the third century B.C. He knew that far to the south, in the town of Syene (present-day Aswan, where there is now a huge dam on the Nile) there was a deep well and at midday on June 21, the sunlight reflected off the water far down in this well, something that happened on no other day of the year. The point was that the sun was exactly vertically overhead at that time, and at no other time in the year. Eratosthenes also knew that the sun was never vertically overhead in Alexandria, the closest it got was on June 21, when it was off by an angle he found to be about 7.2 degrees, by measuring the shadow of a vertical stick.

The distance from Alexandria to Syene was measured at 5,000 stades (a stade being 500 feet), almost exactly due south. From this, and the difference in the angle of sunlight at midday on June 21, Eratosthenes was able to figure out how far it would be to go completely around the earth.



If the Sun is directly overhead at Syene, and the angle between a vertical stick and its shadow is α at Alexandria, then α is also the angle between lines from Alexandria and Syene to the center of the Earth

Of course, Eratosthenes fully recognized that the Earth is spherical in shape, and that "vertically downwards" anywhere on the surface just means the direction towards the center from that point. Thus two vertical sticks, one at Alexandria and one at Syene, were not really parallel. On the other hand, the rays of sunlight falling at the two places *were* parallel. Therefore, if the sun's rays were parallel to a vertical stick at Syene (so it had no shadow) the angle they made with the stick at Alexandria was the same as how far around the Earth, in degrees, Alexandria was from Syene.

According to the Greek historian Cleomedes, Eratosthenes measured the angle between the sunlight and the stick at midday in midsummer in Alexandria to be 7.2 degrees, or one-fiftieth of a complete circle. It is evident on drawing a picture of this that this is the same angle as that between Alexandria and Syene as seen from the center of the earth, so the distance between them, the 5,000 stades, must be one-fiftieth of the distance around the earth, which is therefore equal to 250,000 stades, about 23,300 miles. The correct answer is about 25,000 miles, and in fact Eratosthenes may have been closer than we have stated here---we're not quite sure how far a stade was, and some scholars claim it was about 520 feet, which would put him even closer.

6.2 How High is the Moon?

How do we begin to measure the distance from the earth to the moon? One obvious thought is to measure the angle to the moon from two cities far apart at the same time, and construct a similar triangle, like Thales measuring the distance of the ship at sea. Unfortunately, the angle difference from two points a few hundred miles apart was too small to be measurable by the techniques in use at the time, so that method wouldn't work.

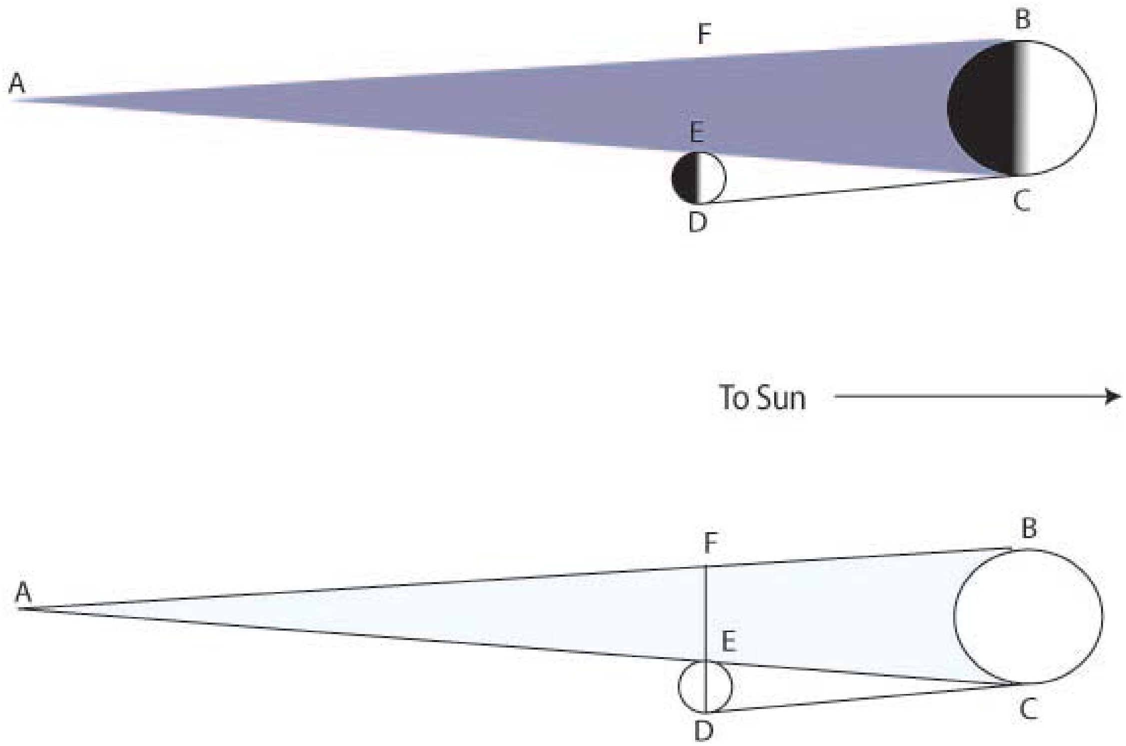
Nevertheless, Greek astronomers, beginning with Aristarchus of Samos (310-230 B.C., approximately) came up with a clever method of finding the moon's distance, by careful observation of a lunar eclipse, which happens when the earth shields the moon from the sun's light.

For a Flash movie of a lunar eclipse, click here!

To better visualize a lunar eclipse, just imagine holding up a quarter (diameter one inch approximately) at the distance where it just blocks out the sun's rays from one eye. Of course you shouldn't try this---you'll damage your eye! You *can* try it with the full moon, which happens to be the same apparent size in the sky as the sun. It turns out that the right distance is about nine feet away, or 108 inches. If the quarter is further away than that, it is not big enough to block out all the sunlight. If it is closer than 108 inches, it will totally block the sunlight from some small circular area, which gradually increases in size moving towards the quarter. Thus the part of space where the sunlight is *totally* blocked is conical, like a long slowly tapering icecream cone, with the point 108 inches behind the quarter. Of course, this is surrounded by a fuzzier area, called the "penumbra", where the sunlight is partially blocked. The fully shaded area is called the "umbra". (This is Latin for shadow. Umbrella means little shadow in Italian.) If you tape a quarter to the end of a thin stick, and hold it in the sun appropriately, you can see these different shadow areas.

Question: If you used a dime instead of a quarter, how far from your eye would you have to hold it to just block the full moonlight from that eye? How do the different distances relate to the relative sizes of the dime and the quarter? Draw a diagram showing the two conical shadows.

Now imagine you're out in space, some distance from the earth, looking at the earth's shadow. (Of course, you could only really *see* it if you shot out a cloud of tiny particles and watched which of them glistened in the sunlight, and which were in the dark.) Clearly, the earth's shadow must be conical, just like that from the quarter. And it must also be *similar* to the quarter's in the technical sense---it must be 108 earth diameters long! That is because the point of the cone is the furthest point at which the earth can block all the sunlight, and the ratio of that distance to the diameter is determined by the angular size of the sun being blocked. This means the cone is 108 earth diameters long, the far point 864,000 miles from earth.



Now, during a total lunar eclipse the moon moves into this cone of darkness. Even when the moon is completely inside the shadow, it can still be dimly seen, because of light scattered by the earth's atmosphere. By observing the moon carefully during the eclipse, and seeing how the earth's shadow fell on it, the Greeks found that *the diameter of the earth's conical shadow at the distance of the moon was about two-and-a-half times the moon's own diameter*.

Note: It is possible to check this estimate either from a photograph of the moon entering the earth's shadow, or, better, by actual observation of a lunar eclipse.

Question: At this point the Greeks knew the size of the earth (approximately a sphere 8,000 miles in diameter) and therefore the size of the earth's conical shadow (length 108 times 8,000 miles). They knew that when the moon passed through the shadow, the shadow diameter at that distance was two and a half times the moon's diameter. Was that enough information to figure out how far away the moon was?

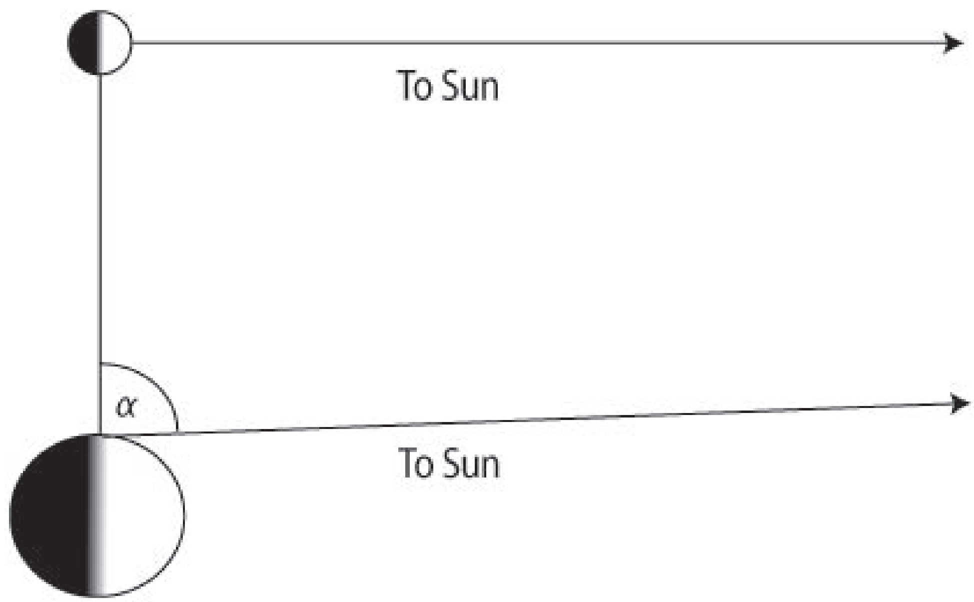
Well, it did tell them the moon was no further away than 108x8,000 = 864,000 miles, otherwise the moon wouldn't pass through the earth's shadow at all! But from what we've said so far, it could be a tiny moon almost 864,000 miles away, passing through that last bit of shadow near the point. However, such a tiny moon could never cause a *solar* eclipse. In fact, as the Greeks well knew, *the moon is the same apparent size in the sky as the sun*. This is the crucial extra fact they used to nail down the moon's distance from earth.

They solved the problem using geometry, constructing the figure below. In this figure, the fact that the moon and the sun have the same apparent size in the sky means that the angle ECD is the same as the angle EAF. Notice now that the length FE is the diameter of the earth's shadow at the distance of the moon, and the length ED is the diameter of the moon. The Greeks found by observation of the lunar eclipse that the ratio of FE to ED was 2.5 to 1, so looking at the similar isosceles triangles FAE and DCE, we deduce that AE is 2.5 times as long as EC, from which AC is 3.5 times as long as EC. But they knew that AC must be 108 earth diameters in length, and taking the earth's diameter to be 8,000 miles, the furthest point of the conical shadow, A, is 864,000 miles from earth. From the above argument, this is 3.5 times further away than the moon is, so the distance to the moon is 864,000/3.5 miles, about 240,000 miles. This is within a few percent of the right figure. The biggest source of error is likely the estimate of the ratio of the moon's size to that of the earth's shadow as it passes through.

6.3 How Far Away is the Sun?

This was an even more difficult question the Greek astronomers asked themselves, and they didn't do so well. They did come up with a very ingenious method to measure the sun's distance, but it proved too demanding in that they could not measure the important angle accurately enough. Still, they did learn from this approach that the sun was much further away than the moon, and consequently, since it has the same apparent size, it must be much bigger than either the moon or the earth.

Their idea for measuring the sun's distance was very simple in principle. They knew, of course, that the moon shone by reflecting the sun's light. Therefore, they reasoned, when the moon appears to be exactly half full, the line from the moon to the sun must be exactly perpendicular to the line from the moon to the observer (see the figure to convince yourself of this). So, if an observer on earth, on observing a half moon in daylight, measures carefully the angle between the direction of the moon and the direction of the sun, the angle α in the figure, he should be able to construct a long thin triangle, with its baseline the earth-moon line, having an angle of 90 degrees at one end and α at the other, and so find the ratio of the sun's distance to the moon's distance.



The problem with this approach is that the angle α turns out to differ from 90 degrees by about a sixth of a degree, too small to measure accurately. The first attempt was by Aristarchus, who estimated the angle to be 3 degrees. This would put the sun only five million miles away. However, it would already suggest the sun to be *much larger than the earth*. It was probably this realization that led Aristarchus to suggest that the sun, rather than the earth, was at the center of the universe. The best later Greek attempts found the sun's distance to be about half the correct value (92 million miles).

The presentation here is similar to that in Eric Rogers, *Physics for the Inquiring Mind*, Princeton, 1960.

7 Greek Science after Aristotle

7.1 Strato

As we mentioned before, Aristotle's analysis of motion was criticized by **Strato** (who died around 268 B.C., he is sometimes called Straton), known as "the Physicist" who was the third director of the Lyceum after Aristotle (the founder) and Theophrastus, who was mainly a botanist.

Strato's career was curiously parallel to Aristotle's. Recall Aristotle spent twenty years at Plato's academy before going to Macedonia to be tutor to Alexander, after which Aristotle came back to Athens to found his own "university", the Lyceum. A few years later, Alexander conquered most of the known world, dividing it into regions with his old friends in charge. In particular, he had his boyhood friend Ptolemy in charge of Egypt, where Alexander founded the new city of Alexandria. Now Strato, after a period of study at the Lyceum, was hired by Ptolemy to tutor his son Ptolemy II Philadelphus (as he became known) in Alexandria. Subsequently Strato returned to Athens where he was in charge of the Lyceum for almost twenty years, until his death.

Strato, like Aristotle, believed in close observation of natural phenomena, but in our particular field of interest here, the study of motion, he observed much more carefully than Aristotle, and realized that falling bodies usually accelerate. He made two important points: rainwater pouring off a corner of a roof is clearly moving faster when it hits the ground than it was when it left the roof, because a continuous stream can be seen to break into drops which then become spread further apart as they fall towards the ground. His second point was that if you drop something to the ground, it lands with a bigger thud if you drop it from a greater height: compare, say, a three foot drop with a one inch drop. One is forced to conclude that falling objects do *not* usually reach some final speed in a very short time and then fall steadily, which was Aristotle's picture. Had this line of investigation been pursued further at the Lyceum, we might have saved a thousand years or more, but after Strato the Lyceum concentrated its efforts on literary criticism.

7.2 Aristarchus

Strato did, however, have one very famous pupil, **Aristarchus** of Samos (310 - 230 B.C.). Aristarchus claimed that the earth rotated on its axis every twenty-four hours and also went round the sun once a year, and that the other planets all move in orbits around the sun. In other words*, he anticipated Copernicus in all essentials*. In fact, Copernicus at first acknowledged Aristarchus, but later didn't mention him (see *Penguin Dictionary of Ancient History*). Aristarchus' claims were not generally accepted, and in fact some thought he should be indicted on a charge of impiety for suggesting that the earth, thought to be the fixed center of the universe, was in motion (Bertrand Russell, quoting Plutarch about Cleanthes). The other astronomers didn't believe Aristarchus' theory for different reasons. It was known that the distance to the sun was in excess of one million miles (Aristarchus himself estimated one and a half million miles, which is far too low) and they thought that if the earth is going around in a circle that big, the pattern of stars in the sky would vary noticeably throughout the year, because the closer ones would appear to move to some extent against the background of the ones further away. Aristarchus responded that they are all so far away that a million miles or two difference in the point of observation is negligible. This implied, though, the universe was *really* huge—at least *billions* of miles across—which few were ready to believe.

7.3 Euclid

Although the Ptolemies were not exactly nice people, they did a great deal of good for Greek civilization, especially the sciences and mathematics. In their anxiety to prove how cultured and powerful they were, they had constructed a massive museum and a library at Alexandria, a city which grew to half a million people by 200 B.C. It was here that Erastosthenes (275 - 195 B.C.) was librarian, but somewhat earlier Euclid taught mathematics there, about 295 B.C. during the reign of Ptolemy I. His great work is his *Elements*, setting out all of Greek geometry as a logical development from basic axioms in twelve volumes. This is certainly one of the greatest books ever written, but not an easy read.

In fact, Ptolemy I, realizing that geometry was an important part of Greek thought, suggested to Euclid that he would like to get up to speed in the subject, but, being a king, could not put in a great deal of effort. Euclid responded: “There is no Royal Road to geometry.”

Euclid shared Plato’s contempt for the practical. When one of his pupils asked what was in it for him to learn geometry, Euclid called a slave and said “Give this young man fifty cents, since he must needs make a gain out of what he learns.”

The Romans, who took over later on didn’t appreciate Euclid. There is no record of a translation of the *Elements* into Latin until 480 A.D. But the Arabs were more perceptive. A copy was given to the Caliph by the Byzantine emperor in A.D. 760, and the first Latin translation that still survives was actually made from the Arabic in Bath, England, in 1120. From that point on, the study of geometry grew again in the West, thanks to the Arabs.

7.4 Plato, Aristotle and Christianity

It is interesting to note that it was in Alexandria that the first crucial connection between classical Greek philosophy and Christian thought was made. As we have just seen, Alexandria was a major center of Greek thought, and also had a very large Jewish community, which had self-governing privileges. Many Jews never returned to Palestine after the Babylonian captivity, but became traders in the cities around the eastern Mediterranean, and Alexandria was a center of this trade. Thus Alexandria was a melting-pot of ideas and philosophies from these different sources. In particular, St. Clement (A.D. 150-215) and Origen were Greek Christians living in

Alexandria who helped develop Christian theology and incorporated many of the ideas of Plato and Aristotle.

(Actually, this St. Clement was demoted from the Roman martyrology in the ninth century for supposed hereticism (but Isaac Newton admired him!). There is a St. Clement of Rome, who lived in the first century. See the *Columbia Encyclopedia*.) Recall that St. Paul himself was a Greek speaking Jew, and his epistles were written in Greek to Greek cities, like Ephesus near Miletus, Phillipi and Thessalonica on the Aegean, and Corinth between Athens and Sparta. After St. Paul, then, many of the early Christian fathers were Greek, and it is hardly surprising that as the faith developed in Alexandria and elsewhere it included Greek ideas. This Greek influence had of course been long forgotten in the middle ages. Consequently, when monks began to look at the works of Plato and Aristotle at the dawn of the Renaissance, they were amazed to find how these pre-Christian heathens had anticipated so many of the ideas found in Christian theology. (*A History of Science*, W. C. Dampier, end of Chapter 1.)

The most famous Alexandrian astronomer, Ptolemy, lived from about 100 AD to 170 AD. He is *not to be confused* with all the Ptolemies who were the rulers! We will discuss Ptolemy later, in comparing his scheme for the solar system with that of Copernicus.

There were two other great mathematicians of this period that we must mention: Archimedes and Apollonius.

7.5 Archimedes

Archimedes, 287 - 212 B.C., lived at Syracuse in Sicily, but also studied in Alexandria. He contributed many new results to mathematics, including successfully computing areas and volumes of two and three dimensional figures with techniques that amounted to calculus for the cases he studied. He calculated pi by finding the perimeter of a sequence of regular polygons inscribed and escribed about a circle.

Two of his major contributions to physics are his understanding of the principle of buoyancy, and his analysis of the lever. He also invented many ingenious technological devices, many for war, but also the Archimedean screw, a pumping device for irrigation systems.

7.6 Archimedes' Principle

We turn now to Syracuse, Sicily, 2200 years ago, with Archimedes and his friend king Heiro. The following is quoted from Vitruvius, a Roman historian writing just before the time of Christ:

Heiro, after gaining the royal power in Syracuse, resolved, as a consequence of his successful exploits, to place in a certain temple a golden crown which he had vowed to the immortal gods. He contracted for its making at a fixed price and weighed out a precise amount of gold to the *contractor. At the appointed time the latter delivered to the king's satisfaction an exquisitely finished piece of handiwork, and it appeared that in weight the crown corresponded precisely to what the gold had weighed.*

But afterwards a charge was made that gold had been abstracted and an equivalent weight of silver had been added in the manufacture of the crown. Heiro, thinking it an outrage that he had been tricked, and yet not knowing how to detect the theft, requested Archimedes to consider the matter. The latter, while the case was still on his mind, happened to go to the bath, and on getting into a tub observed that the more his body sank into it the more water ran out over the tub. As this pointed out the way to explain the case in question, without a moments delay and transported with joy, he jumped out of the tub and rushed home naked, crying in a loud voice that he had found what he was seeking; for as he ran he shouted repeatedly in Greek, "Eureka, Eureka."

Taking this as the beginning of his discovery, it is said that he made two masses of the same weight as the crown, one of gold and the other of silver. After making them, he filled a large vessel with water to the very brim and dropped the mass of silver into it. As much water ran out as was equal in bulk to that of the silver sunk in the vessel. Then, taking out the mass, he poured back the lost quantity of water, using a pint measure, until it was level with the brim as it had been before. Thus he found the weight of silver corresponding to a definite quantity of water.

After this experiment, he likewise dropped the mass of gold into the full vessel and, on taking it out and measuring as before, found that not so much water was lost, but a smaller quantity: namely, as much less as a mass of gold lacks in bulk compared to a mass of silver of the same weight. Finally, filling the vessel again and dropping the crown itself into the same quantity of water, he found that more water ran over for the crown than for the mass of gold of the same weight. Hence, reasoning from the fact that more water was lost in the case of the crown than in that of the mass, he detected the mixing of silver with the gold and made the theft of the contractor perfectly clear.

What is going on here is simply a measurement of the density—the mass per unit volume—of silver, gold and the crown. To measure the masses some kind of scale is used, note that at the beginning a precise amount of gold is weighed out to the contractor. Of course, if you had a nice rectangular brick of gold, and knew its weight, you wouldn't need to mess with water to determine its density, you could just figure out its volume by multiplying together length, breadth and height, and divide the mass, or weight, by the volume to find the density in, say, pounds per cubic foot or whatever units are convenient. (Actually, the units most often used are the metric ones, grams per cubic centimeter. These have the nice feature that water has a density of 1, because that's how the gram was defined. In these units, silver has a density of 10.5, and gold of 19.3. To go from these units to pounds per cubic foot, we would multiply by the weight in pounds of a cubic foot of water, which is 62.)

The problem with just trying to find the density by figuring out the volume of the crown is that it is a very complicated shape, and although one could no doubt find its volume by measuring each tiny piece and calculating a lot of small volumes which are then added together, it would take a long time and be hard to be sure of the accuracy, whereas lowering the crown into a filled bucket of water and measuring how much water overflows is obviously a pretty simple procedure. (You do have to allow for the volume of the string!). Anyway, the bottom line is that if the crown displaces more water than a block of gold of the same weight, the crown isn't pure gold.

Actually, there is one slightly surprising aspect of the story as recounted above by Vitruvius. Note that they had a weighing scale available, and a bucket suitable for immersing the crown. Given these, there was really no need to measure the amount of water slopping over. All that was necessary was first, to weigh the crown when it was fully immersed in the water, then, second, to dry it off and weigh it out of the water. The difference in these two weighings is just the buoyancy support force from the water. ***Archimedes' Principle*** *states that the buoyancy support force is exactly equal to the weight of the water displaced* by the crown, that is, it is equal to the weight of a volume of water equal to the volume of the crown.

This is definitely a less messy procedure—there is no need to fill the bucket to the brim in the first place, all that is necessary is to be sure that the crown is fully immersed, and not resting on the bottom or caught on the side of the bucket, during the weighing.

Of course, maybe Archimedes had not figured out his Principle when the king began to worry about the crown, perhaps the above experiment led him to it. There seems to be some confusion on this point of history.

7.7 Archimedes and Leverage

Although we know that leverage had been used to move heavy objects since prehistoric times, it appears that Archimedes was the first person to appreciate just how much weight could be shifted by one person using appropriate leverage.

Archimedes illustrated the principle of the lever very graphically to his friend the king, by declaring that if there were another world, and he could go to it, he could move this one. To quote from Plutarch,

Heiro was astonished, and begged him to put his proposition into execution, and show him some great weight moved by a slight force. Archimedes therefore fixed upon a three-masted merchantman of the royal fleet, which had been dragged ashore by the great labours of many men, and after putting on board many passengers and the customary freight, he seated himself at some distance from her, and without any great effort, but quietly setting in motion a system *of compound pulleys, drew her towards him smoothly and evenly, as though she were gliding through the water.*

Just in case you thought kings might have been different 2200 years ago, read on:

Amazed at this, then, and comprehending the power of his art, the king persuaded Archimedes to prepare for him offensive and defensive weapons to be used in every kind of siege warfare.

This turned out to be a very smart move on the king's part, since some time later, in 215 B.C., the Romans attacked Syracuse. To quote from Plutarch's Life of Marcellus (the Roman general):

When, therefore, the Romans assaulted them by sea and land, the Syracusans were stricken dumb with terror; they thought that nothing could withstand so furious an onslaught by such forces. But Archimedes began to ply his engines, and shot against the land forces of the assailants all sorts of missiles and immense masses of stones, which came down with incredible din and speed; nothing whatever could ward off their weight, but they knocked down in heaps those who stood in their way, and threw their ranks into confusion. At the same time huge beams were suddenly projected over the ships from the walls, which sank some of them with great weights plunging down from on high; others were seized at the prow by iron claws, or beaks like the beaks of cranes, drawn straight up into the air, and then plunged stern foremost into the depths, or were turned round and round by means of enginery within the city, and dashed upon the steep cliffs that jutted out beneath the wall of the city, with great destruction of the fighting men on board, who perished in the wrecks. Frequently, too, a ship would be lifted out of the water into mid-air, whirled hither and thither as it hung there, a dreadful spectacle, until its crew had been thrown out and hurled in all directions, when it would fall empty upon the walls, or slip away from the clutch that had held it... .

Then, in a council of war, it was decided to come up under the walls while it was still night, if they could; for the ropes which Archimedes used in his engines, since they imported great impetus to the missiles cast, would, they thought, send them flying over their heads, but would be ineffective at close quarters, since there was no space for the cast. Archimedes, however, as it seemed, had long before prepared for such an emergency engines with a range adapted to any interval and missiles of short flight, and, through many small and contiguous openings in the wall, short-range engines called "scorpions" could be brought to bear on objects close at hand without being seen by the enemy.

When, therefore, the Romans came up under the walls, thinking themselves unnoticed, once more they encountered a great storm of missiles; huge stones came tumbling down upon them almost perpendicularly, and the wall shot out arrows at them from every point; they therefore retired.... . At last, the Romans became so fearful that, whenever they saw a bit of rope or a stick of timber projecting a little over the wall, "There it is," they cried, "Archimedes is training some *engine upon us," and turned their backs and fled. Seeing this, Marcellus desisted from all fighting and assault, and thenceforth depended on a long siege.*

It is sad to report that the long siege was successful and a Roman soldier killed Archimedes as he was drawing geometric figures in the sand, in 212 B.C. Marcellus had given orders that Archimedes was not to be killed, but somehow the orders didn't get through.

7.8 Apollonius

Apollonius probably did most of his work at Alexandria, and lived around 220 B.C., but his exact dates have been lost. He greatly extended the study of conic sections, the ellipse, parabola and hyperbola.

As we shall find later in the course, the conic sections play a central role in our understanding of everything from projectiles to planets, and both Galileo and Newton, among many others, acknowledge the importance of Apollonius' work. This is not, however, a geometry course, so we will not survey his results here, but, following Galileo, rederive the few we need when we need them.

7.9 Hypatia

The last really good astronomer and mathematician in Greek Alexandria was a woman, Hypatia, born in 370 AD the daughter of an astronomer and mathematician Theon, who worked at the museum. She wrote a popularization of Apollonius' work on conics. She became enmeshed in politics, and, as a pagan who lectured on neoplatonism to pagans, Jews and Christians (who by now had separate schools) she was well known. In 412 Cyril became patriarch. He was a fanatical Christian, and became hostile to Orestes, the Roman prefect of Egypt, a former student and a friend of Hypatia. In March 415, Hypatia was killed by a mob of fanatical Christian monks in particularly horrible fashion. The details can be found in the book *Hypatia's Heritage* (see below).

Books I used in preparing this lecture:

Greek Science after Aristotle, G. E. R. Lloyd, Norton, N.Y., 1973

A Source Book in Greek Science, M. R. Cohen and I. E. Drabkin, Harvard, 1966

Hypatia's Heritage: A History of Women in Science, Margaret Alic, The Women's Press, London 1986

A History of Science, W. C. Dampier, Cambridge, 1929

8 Basic Ideas in Greek Mathematics

8.1 Closing in on the Square Root of 2

In our earlier discussion of the irrationality of the square root of 2, we presented a list of squares of the first 17 integers, and remarked that there were several "near misses" to solutions of the equation $m^2 = 2n^2$. Specifically, $3^2 = 2\times2^2 + 1$, $7^2 = 2\times5^2 - 1$, $17^2 = 2\times12^2 + 1$. These results were also noted by the Greeks, and set down in tabular form as follows:

3	2
7	5
17	12

After staring at this pattern of numbers for a while, the pattern emerges: 3 + 2 = 5 and 7 + 5 = 12, so the number in the right-hand column, after the first row, is the sum of the two numbers in the row above. Furthermore, 2 + 5 = 7 and 5 + 12 = 17, so the number in the left-hand column is the sum of the number to its right and the number immediately above that one.

The question is: does this pattern continue? To find out, we use it to find the next pair. The right hand number should be 17 + 12 = 29, the left-hand 29 + 12 = 41. Now $41^2 = 1681$, and $29^2 = 841$, so $41^2 = 2\times 29^2 - 1$. Repeating the process gives 41 + 29 = 70 and 70 + 29 = 99. It is easy to check that $99^2 = 2\times70^2 + 1$. So $99^2/70^2 = 2 + 1/70^2$. In other words, the difference between the square root of 2 and the rational number 99/70 is approximately of the magnitude $1/70^2$. (You can check this with your calculator).

The complete pattern is now evident. The recipe for the numbers is given above, and the +1's and -1's alternate on the right hand side. In fact, the Greeks managed to prove (it can be done with elementary algebra) that pairs of numbers can be added indefinitely, and their ratio gives a better and better approximation to the square root of 2.

The essential discovery here is that, although it is established that the square root of 2 is not a rational number, we can by the recipe find a rational number as close as you like to the square root of two. This is sometimes expressed as "there are rational numbers infinitely close to the square root of 2" but that's not really a helpful way of putting it. It's better to think of a sort of game - you name a small number, say, one millionth, and I can find a rational number (using the table above and finding the next few sets of numbers) which is within one millionth of the square root of 2. However small a number you name, I can use the recipe above to find a rational that close to the square root of 2. Of course, it may take a lifetime, but the method is clear!

8.2 Zeno's Paradoxes

Zeno of Elea (495-435 BC) is said to have been a self-taught country boy. He was a friend of a well-known philosopher, Parmenides, and visited Athens with him, where he perplexed Socrates, among others, with a set of paradoxes. (Plato gives an account of this in *Parmenides*.) We shall look at two of them here.

8.3 Achilles and the Tortoise.

A two hundred yard race is set up between Achilles, who can run at 10 yards per second, and the tortoise, who can run at one yard per second (perhaps rather fast for a tortoise, but I'm trying to keep the numbers simple).

To give the tortoise a chance, he is given a one-hundred yard start.

Now, when Achilles has covered that first 100 yards, to get to where the tortoise was, the tortoise is 10 yards ahead.

When Achilles has covered that 10 yards, the tortoise is 1 yard ahead.

When Achilles has covered that 1 yard, the tortoise is 1/10 yard ahead.

Now, Zeno says, there is no end to this sequence! We can go on forever dividing by 10! So, Zeno concludes, Achilles has to cover an infinite number of smaller and smaller intervals before he catches the tortoise. But to do an infinite number of things takes an infinitely long time - so he'll never catch up.

What is wrong with this argument? Try to think it through before you read on!

The essential point becomes clearer if you figure out how long it takes Achilles to cover the sequence of smaller and smaller intervals. He takes 10 seconds to cover the first 100 yards, 1 second to cover the next 10 yards, 1/10 second for the next yard, 1/100 second for the next 1/10 of a yard, and so on. If we write down running totals of time elapsed to each of these points we get 10 seconds, 11 seconds, 11.1 seconds, 11.11 seconds and so on. It is apparent that the *total* time elapsed for *all* the infinite number of smaller and smaller intervals is going to be 11.1111111…, with the 1's going on forever. But this recurring decimal, 0.111111… is just 1/9, as you can easily check.

*The essential point is that it is possible to add together an **infinite** number of time intervals and still get a **finite** result*. That means there *is* a definite time-11 1/9 seconds-at which Achilles catches up with the tortoise, and after that instant, he's passed the tortoise.

8.4 The Arrow

Consider the flight of an arrow through the air. Let us divide time up into instants, where an instant is an indivisibly small time. Now, during an instant, the arrow can't move at all, because if it did, we could divide up the instant using the changing position of the arrow to indicate which bit of the instant we are in.

However, a finite length of time-like a second-is made up of instants. Therefore, if the arrow doesn't move at all during an instant, it doesn't move in a sum of instants. Hence, it can't move in one second!

What's wrong with this argument?

Now there certainly *is* such a thing as an instant of time: for example, if the arrow is in the air from time zero to time two seconds, say, then there is one instant at which it has been in the air for exactly one second.

The catch is, *there is no way to divide time up into such instants*. Imagine the time from zero to two seconds represented by a geometric line two inches long on a piece of paper. By geometric, I mean an ideal line, not one that's really a collection of microscopic bits of pencil lead, but a true continuous line of the kind the Greeks imagined. Time has that kind of continuity-there aren't little gaps in time (at least, none we've found so far). Now try to imagine the line made up of instants. You could start by putting dots every millionth of a second, say. But then you could imagine putting a million dots between each of those pairs of dots, to have a dot every trillionth of a second. And why stop there? You could keep on indefinitely with this division. But if you spend the rest of your life on this mental exercise, you will never put a dot at the instant corresponding to the time being the square root of two! And it has been proved by the mathematicians that there are infinitely more irrational numbers than there are rational numbers.

So there really is no way to divide time up into instants. If you're still not sure, think about the following problem: what's the next instant after the instant at time equals one second?

8.5 Instants and Intervals

On the other hand, it is obviously useful in analyzing the motion of the arrow to look at the motion one bit at a time-in other words, to divide the time up somehow, to get a grip on how the arrow's speed may be varying throughout the flight. So how should we proceed? Zeno's dividing of time into instants was not very easy to understand, as we've seen. It's much easier to visualize dividing time into *intervals*. For example, the two seconds the arrow is in the air could be divided into two hundred intervals, each of length one-hundredth of a second. Then we could find its average speed in each of those intervals by measuring how far it went in the one-hundredth of a second, and multiplying by one hundred. That is, if it went two feet in the one-hundredth of a second interval, it was traveling at two hundred feet per second during that interval. (Of course, it might not be going at that speed for the whole flight-that's why we've divided it into intervals, so that we can monitor the speed the whole time). Of course, if the arrow hits something, it will slow down very rapidly-there will be a big change in speed in one hundredth of a second. If we want to describe the motion of the arrow in this situation, we must divide time up into smaller intervals, say thousandths of a second, or even ten-thousandths of a second, depending on how precisely we want to follow the change in speed.

8.6 Speed at an Instant

There is still a problem here we haven't quite faced. All this dividing time up into small intervals and finding the average speed in each interval gives a pretty good idea of the arrow's progress, but it's still a reasonable question to ask: just what is the arrow's speed *at the instant* one second after the flight began?

How do we answer that question? Think about it before you read on.

The essential point about speed is that it is a rate of change of position-this is obvious when you think about measuring speed, it's in units like miles per hour, feet per second, etc. This implies that to make any statement about speed we have to say how far the arrow moved between two specified times. Therefore, to find the speed at the time one second after takeoff, we would need to find where the arrow is at, say, 0.995 seconds after takeoff, then at 1.005 seconds after takeoff. I've chosen here two times that are one-hundredth of a second apart. If the arrow moves one and a half feet during that period, it's going at 150 feet per second.

You might object, though, that this is still not very precise. Probably 150 feet per second is pretty close to the arrow's speed at one second after takeoff, but it's really an average over a time interval of one-hundredth of a second, so may not be *exactly* the speed in the middle of the time. This is true-it may not be. What we must do, at least in principle, is to take a smaller time interval, say one-millionth of a second, again centered at time one second, as before. We now measure how far the arrow moves in the one-millionth of a second, and multiply that distance by one million to get the arrow's average speed over that very short time.

Of course, you could say you're still not satisfied. You want to know the *precise* speed at the one second mark, not some approximation based on the average over a time interval. But, as we've just said, all speed measurements necessarily involve some time interval, which, however, can be as short as we like. This suggests how we should define what we mean by the speed at one instant of time-we take a sequence of shorter and shorter time intervals, each one centered at the time in question, and find the average speed in each. This series of speed measurements will close in on the exact speed at the time one second.

This should remind you of the discussion of the square root of two. There we had a sequence of rational numbers such that if you come up with some small number such as a millionth of a trillionth, we could always find a rational within that distance of root two. Here we are saying that if you want the speed to some preassigned accuracy, we can find it by taking a sufficiently small time interval around the time in question, and computing the average speed in that interval.

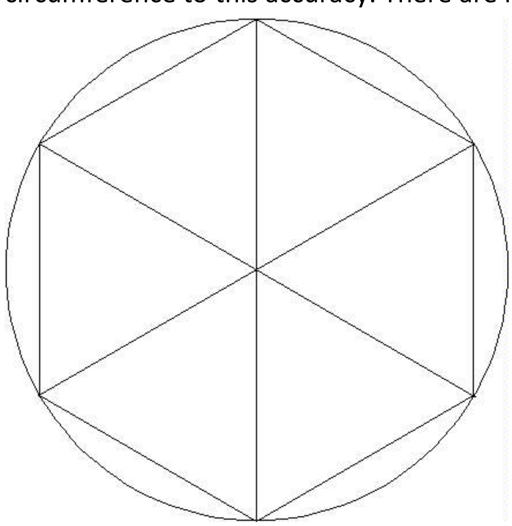
Actually, this may not be as difficult as it sounds. For example, imagine an arrow moving far out in space at a steady speed, with no air resistance or gravity to contend with. Then it will go at a steady speed, and the average speed over all time intervals will be the same. This means we can find (in principle) the *exact* speed at any given time without having to worry about indefinitely small time intervals. Another fairly simple case is an arrow *gaining speed at a steady rate*. Its speed in the middle of a time interval turns out to be *exactly* equal to its average speed in the interval. We shall be discussing this case further when we get to Galileo.

8.7 The Beginning of Calculus

We should emphasize that the above discussion of intervals, instants and so on was not the response of the Athenians to Zeno. Only with later work by Eudoxus, Euclid and Archimedes did the way to deal with these small quantities gradually become apparent. Zeno's contribution was that he initiated the discussion that ultimately led to the calculus. In fact, according to Bertrand Russell (*History of Western Philosophy*) Zeno taught Socrates the Socratic method-the method of seeking knowledge by systematic question and answer. Unhappily, Zeno's approach did not win him powerful friends, and "he finally lost his head for treason or something of the sort" (Bell, *Men of Mathematics*).

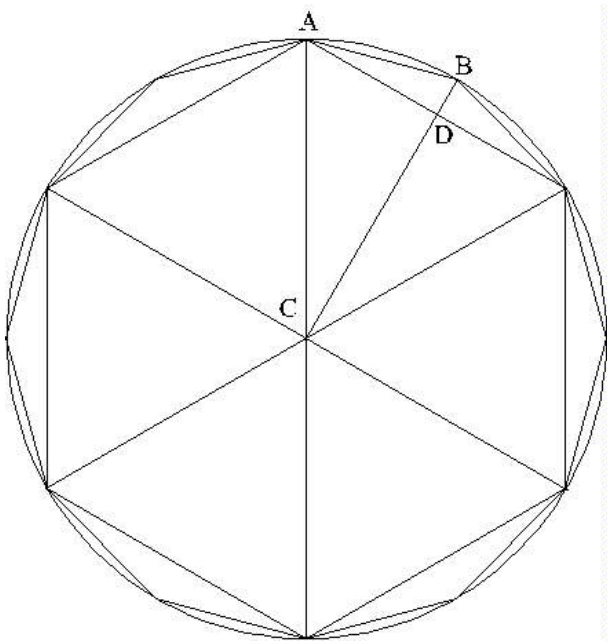
8.8 Archimedes Begins Calculating Pi

Both the Babylonians and the Egyptians used approximations to pi, the ratio of the circumference of a circle to its diameter. The Egyptians used a value 3.16, within one per cent of the true value. (Further details can be found in Neugebauer, *The Exact Sciences in Antiquity*, Dover, page 78.) Actually, this value follows from their rule for the *area* of a circle, $(8/9.d)^2$, but it is reasonable to suppose they could have constructed a circle and measured the circumference to this accuracy. There are no indications that they tried to *calculate* pi, using geometric arguments as Archimedes did.



Following Archimedes, we first draw a circle of radius equal to one (so the *diameter* is 2), and inscribe in it a regular (that is, all sides of equal length) hexagon. It is evident that the hexagon is made up of six equilateral triangles, since the 360 degree angle at the center of the circle is equally divided into six, and the angles of a triangle add to 180 degrees. Therefore, each side of each triangle is equal to the radius of the circle, that is, equal to one. Thus the perimeter of the hexagon is exactly 6. It is clear from the figure that the circumference of the circle, the total distance around, is greater than the perimeter of the hexagon, because the hexagon can be seen as a series of shortcuts on going around the circle. We conclude that pi, the ratio of the circumference of the circle to its diameter, is greater than 3, but not much-the hexagon looks quite close. (For example, much closer than, going around a square boxed around the circle, which would be a distance of 8 radii. If we approximated the circumference of the circle by this square, we would guess pi = 4.)

So the first step-comparing the circle with the hexagon-tells us that pi is greater than three. Archimedes' next move was to find a polygon inscribed in the circle that was closer to the circle than the hexagon, so that its perimeter would be closer to the circumference of the circle. His strategy was to double the number of sides of the polygon, that is, to replace the hexagon by a twelve-sided regular polygon, a dodecagon. Obviously, from the figure, the perimeter of the dodecagon is much closer to that of the circle than the hexagon was (but it's still obviously less, since, like the hexagon, it is a series of shortcuts on going around the circle).



Calculating the perimeter of the dodecagon is not as simple as it was for the hexagon, but all it require is Pythagoras' theorem. Look at the figure. We need to find the length of one side, like AB, and multiply it by 12 to get the total perimeter. AB is the hypotenuse of the right-angled triangle ABD. We know the length AD is just ½ (recall the radius of the circle = 1). We don't know the other length, BD, but we do know that BC must equal 1, because it's just the radius of the circle again. Switching our attention to the right-angled triangle ACD, we see its hypotenuse equals 1, and one side (AD) equals ½. So from Pythagoras, the square of CD must be ¾. We will write CD = ½×sqrt3.

Having found CD, we can find DB since CD + DB = CB = 1, that is, DB = 1 - ½×sqrt3. So we know the two shorter sides of the right-angled triangle ADB, and we can find the hypotenuse using Pythagoras again.

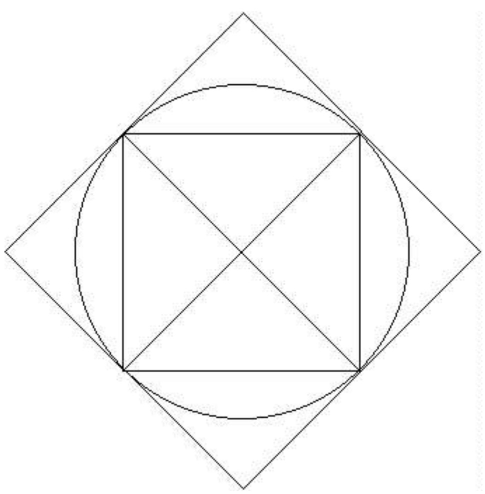
The dodecagon turns out to have a perimeter 6.21, giving pi greater than 3.1. This is not quite as close as the Egyptians, but Archimedes didn't stop here. He next went to a 24-sided regular polygon inscribed in the circle. Again, he just needed to apply Pythagoras' theorem twice, exactly as in the preceding step. The perimeter of the 24-sided regular polygon turns out to be 6.26, giving pi greater than 3.13. (We are giving a slightly sloppy version of his work: he always worked with rationals, and where the square root of 3 came in, he used 265/153 < sqrt3 < 1351/780. These limits came from an algorithm originating with the Babylonians.)

In fact, Archimedes went on as far as the 96-sided regular polygon inscribed in the circle. He then started all over again with regular polygons circumscribed *about* the circle, so that the circle is touching the middle of each side of the polygon, and is completely contained by it. Such a polygon clearly has a perimeter *greater* than that of the circle, but getting closer to it as we consider polygons with more and more sides. Archimedes considered such a polygon with 96 sides.

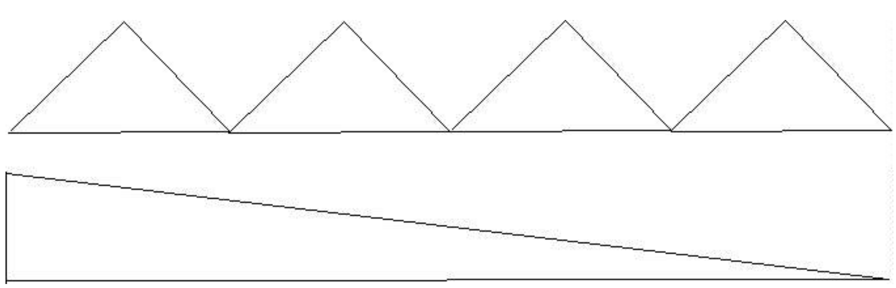
So, with a series of polygons inside the circle, and another series outside it, he managed to bracket the length of the circumference between two sets of numbers which gradually approached each other. This is again reminiscent of the Greek strategy in approximating the square root of 2. The result of all his efforts was the inequality: 3 10/71 < pi < 3 1/7. If we take the average of these two numbers, we find 3.14185. The correct value is 3.14159… .

8.9 Squaring the Circle

This phrase refers to the famous problem of finding an area with straight-line boundaries equal in area to a circle of given diameter. Archimedes proved that the area of a circle is equal to that of a right-angled triangle having the two shorter sides equal to the radius of the circle and its circumference respectively.



The idea of his proof is as follows. Consider first a square inscribed in the circle. The square is made up of four triangles, each of height h, say, and base length b. (Actually, $b = 2h$, but we'll keep them separate.) The total area of the square is equal to the total area of the 4 triangles, which is 4 times $½×h×b$, or $½×h×4b$. Notice that *this is the area of a long thin triangle*, with *height* equal to the distance h from the middle of the side of the square to the center of the circle, and *base* equal to the perimeter length $4b$ of the square.



The area of the square isn't a very good approximation to that of the circle, but we can improve it by replacing the square by a regular octagon, with all its points on the circle. Now, this octagon can by divided into eight triangles, following the same procedure as for the square. The height of each of these triangles equals the distance from the center of the circle to the middle of one side of the octagon. Just as for the square case, the total area of these eight triangles is equal to that of a long thin triangle of the same height, and with base length equal to the perimeter of the octagon.

It is evident that the height of the octagon's triangles is closer to the radius of the circle than the height of the square's triangles, and the perimeter of the octagon is closer to the circumference of the circle than the perimeter of the square was.

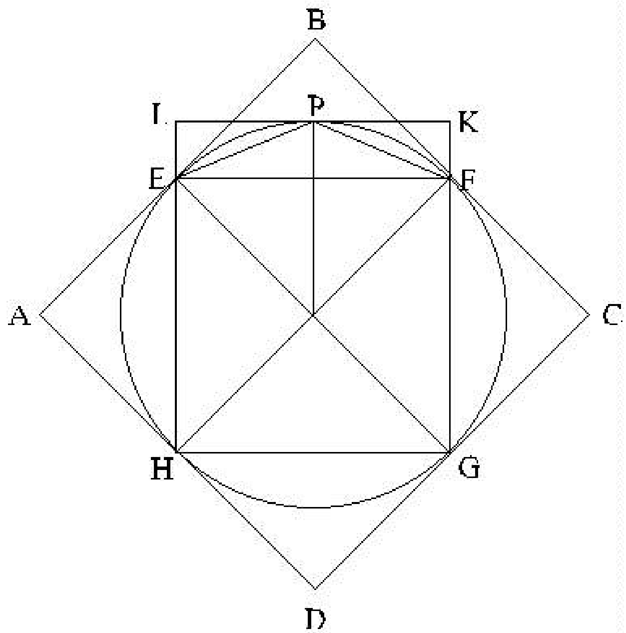
The process is repeated: the octagon is replaced by a regular 16-sided polygon, with all its points on the circle. This polygon is equal in area to the sum of the 16 triangles formed by drawing lines from the center of the circle to its points. These triangles all have the same height, so they have total area the same as a long thin triangle having the same height, and base length equal to the perimeter of the 16-sided polygon.

At this point, the pattern should be clear-as we go to polygons of 32, 64, … sides, the total area of the polygon is the same as that of a right angled triangle with a long side equal to the perimeter of the polygon, which approaches the circumference of the circle as the polygons have more and more sides, and the height of the triangle approaches the radius of the circle. Therefore, the area of the polygons approaches ½×base×height = ½×2×pi×r×r = pi×r^2.

8.10 Eudoxus' Method of Exhaustion

This section and the next are optional—they won't appear on any tests, etc. I just put them in for completeness.

In fact, the account given above doesn't do justice to the tightness of the Greeks' geometric arguments. The approach to the limit of more and more sided polygons approximating the circle better and better is a bit vague. It's not very clear how quickly this is happening.



Eudoxus clarified the situation by giving a procedure putting a lower limit on how much more of the circle's total area was covered by the new polygon created at each step. Let's begin with the square. In the figure, we show the inscribed square EFGH, and also a circumscribed square ABCD. Clearly, the area of square EFGH is exactly half of that of square ABCD. Since the circle lies entirely inside ABCD, it follows that EFGH covers more than half of the area of the circle.

Now consider how much more of the circle's total area is covered when we go from the square to the octagon. We add triangular areas like EPF to each side of the square. Now, notice that the triangle EPF has area exactly half of the rectangular area ELKF. If we had added rectangular areas like that to the four sides of the square, the circle's area would have been completely contained. This implies that by adding just the triangles, that is, going from the square to the octagon, we are covering more than half of the area of the circle that lay outside the square.

This same argument works at each step: so, the inscribed square covers more than half the circle's area, going to the octagon covers more than half the rest, so the octagon covers more than three-quarters of the circle's area, the 16-sided inscribed polygon covers more than seven-eighths of the circle's area, and so on.

Archimedes used Eudoxus' approach to prove that the area of a circle was equal to that of the right-angled triangle with shorter sides equal to the radius and the circumference of the circle. Suppose, he said, that the triangle's area is less than the circle's. Then in the sequence of polygons with 4, 8, 16, 32, … sides, we will get to one with area greater than the triangle's. But that polygon will have an area equal to that of a number of triangles equal to its number of sides, and, as we've argued above, the sum of their areas is equal to that of a triangle having their height and base length equal to the perimeter of the polygon. But their height is less than the radius of the circle, and the perimeter is less than the circumference of the circle. Hence their total area must be less that that of the triangle having height the radius of the circle and base the circumference. This gives a contradiction, so that triangle cannot have area less than the circle's.

Supposing that the triangle's area is greater than the circles leads to another contradiction on considering a sequence of polygons *circumscribed* about the circle-so the two must be exactly equal.

8.11 Archimedes does an Integral

Archimedes realized that in finding the area of a circle, another problem was solved, that of finding the area of the curved surface of a cone (like an old-fashioned ice-cream cone). If such a cone is opened out by cutting a straight line up from its point, it will have the shape of a fan-that is, a segment of a circle. Its area will then be that fraction of the full circle's area that its curved edge is of the full circle's circumference. He also showed how to find the curved area of a "slice" of a cone, such as you'd get by cutting off the top of an ice-cream cone, by which we mean the other end from the point, cutting parallel to the top circle, to get a sort of ring-shaped bit of cone. He then managed to calculate the surface area of a sphere. His approach was as follows: imagine where Charlottesville appears on a globe, on the 38^{th} parallel. This parallel is a ring going all the way around the globe at a constant distance down from the North Pole. Now consider the part of the globe surface between this 38^{th} parallel and the 39^{th} parallel. This is a ribbon of surface going around, and is very close to a slice of a cone, if we choose a cone of the right size and angle. Archimedes' strategy was to divide the whole surface into ribbons like this, and find the area of each ribbon by taking it to be part of a cone. He then summed up the ribbon areas. Lastly, he took thinner and thinner ribbons to get an accurate result, using the method of exhaustion to prove that the area of the sphere was $4\times\text{pi}\times r^2$. This is precisely equivalent to a modern integral calculus solution of the same problem, and just as rigorous (but more difficult!)

8.12 Conclusion

It is clear from the above discussion that the Greeks laid the essential groundwork and even began to build the structure of much of modern mathematics. It should also be emphasized that although some great mathematicians devoted their lives to this work, it nevertheless took three centuries of cumulative effort, each building on the previous work. Evidently, this required a stable, literate culture over many generations. Geometric results are difficult to transmit in an oral tradition! Recall that Archimedes was killed drawing diagrams in the sand for his pupils. This level of mathematical analysis attained by Archimedes, Euclid and others is far in advance of anything recorded by the Babylonians or Egyptians.

In preparing this lecture I used :

A Source Book in Greek Science, M. R. Cohen and I. E. Drabkin, Harvard, 1966

9 How the Greeks Used Geometry to Understand the Stars

9.1 Crystal Spheres: Plato, Eudoxus, Aristotle

Plato, with his belief that the world was constructed with geometric simplicity and elegance, felt certain that the sun, moon and planets, being made of aither, would have a natural circular motion, since that is the simplest uniform motion that repeats itself endlessly, as their motion did. However, although the "fixed stars" did in fact move in simple circles about the North star, the sun, moon and planets traced out much more complicated paths across the sky. These paths had been followed closely and recorded since early Babylonian civilization, so were very well known. Plato suggested that perhaps these complicated paths were actually *combinations* of simple circular motions, and challenged his Athenian colleagues to prove it.

The first real progress on the problem was made by Eudoxus, at Plato's academy. Eudoxus placed all the fixed stars on a huge sphere, the earth itself a much smaller sphere fixed at the center. The huge sphere rotated about the earth once every twenty-four hours. So far, this is the standard "starry vault" picture. Then Eudoxus assumed the sun to be attached to another sphere, concentric with the fixed stars' sphere, that is, it was also centered on the earth. This new sphere, lying entirely inside the sphere carrying the fixed stars, had to be transparent, since the fixed stars are very visible. The new sphere was attached to the fixed stars' sphere so that it, too, went around every twenty-four hours, but *in addition* it rotated slowly about the two axis points where it was attached to the big sphere, and this extra rotation was once a year. This meant that the sun, viewed against the backdrop of the fixed stars, traced out a big circular path which it covered in a year. This path is the *ecliptic*. To get it all right, the ecliptic has to be tilted at 23½ degrees to the "equator" line of the fixed stars, taking the North star as the "north pole".

This gives a pretty accurate representation of the sun's motion, but it didn't quite account for all the known observations at that time. For one thing, if the sun goes around the ecliptic at an exactly uniform rate, the time intervals between the solstices and the equinoxes will all be equal. In fact, they're not-so the sun moves a little faster around some parts of its yearly journey through the ecliptic than other parts. This, and other considerations, led to the introduction of three more spheres to describe the sun's motion. Of course, to actually *show* that the combination of these motions gave an accurate representation of the sun's observed motion required considerable geometric skill! Aristotle wrote a summary of the "state of the art" in accounting for all the observed planetary motions, and also those of the sun and the moon. This required the introduction of *fifty-five* concentric transparent spheres. Still, it *did* account for everything observed in terms of simple circular motion, the only kind of motion thought to be allowed for aether. Aristotle himself believed the crystal spheres existed as physical entities, although Eudoxus may have viewed them as simply a computational device.

It is interesting to note that, despite our earlier claim that the Greeks "discovered nature", Plato believed the planets to be animate beings. He argued that it was not possible that they should accurately describe their orbits year after year if they didn't know what they were doing—that is, if they had no soul attached.

9.2 Measuring the Earth, the Moon and the Sun: Eratosthenes and Aristarchus

A little later, Eratosthenes and Aristarchus between them got some idea of the size of the earth-sun-moon system, as we discussed in an earlier lecture.

And, to quote from Archimedes (see Heath, *Greek Astronomy*),

"Aristarchus of Samos brought out a book consisting of certain hypotheses, in which the premises lead to the conclusion that the universe is many times greater than it is presently thought to be. His hypotheses are that the fixed stars and the sun remain motionless, that the earth revolves about the sun in the circumference of a circle, the sun lying in the middle of the orbit, and that the sphere of the fixed stars, situated about the same center as the sun, is so great that the circular orbit of the earth is as small as a point compared with that sphere."

The tiny size of the earth's orbit is necessary to understand why the fixed stars do not move relative to each other as the earth goes around its orbit.

Aristarchus' model was not accepted, nor even was the suggestion that the earth rotates about its axis every twenty-four hours.

However, the model of the fifty-five crystal spheres was substantially improved on. It did have some obvious defects. For example, the sun, moon and planets necessarily each kept a constant distance from the earth, since each was attached to a sphere centered on the earth. Yet it was well-known that the apparent size of the moon varied about ten per cent or so, and the obvious explanation was that its distance from the earth must be varying. So how *could* it be attached to a sphere centered on the earth? The planets, too, especially Mars, varied considerably in brightness compared with the fixed stars, and again this suggested that the distance from the earth to Mars must vary in time.

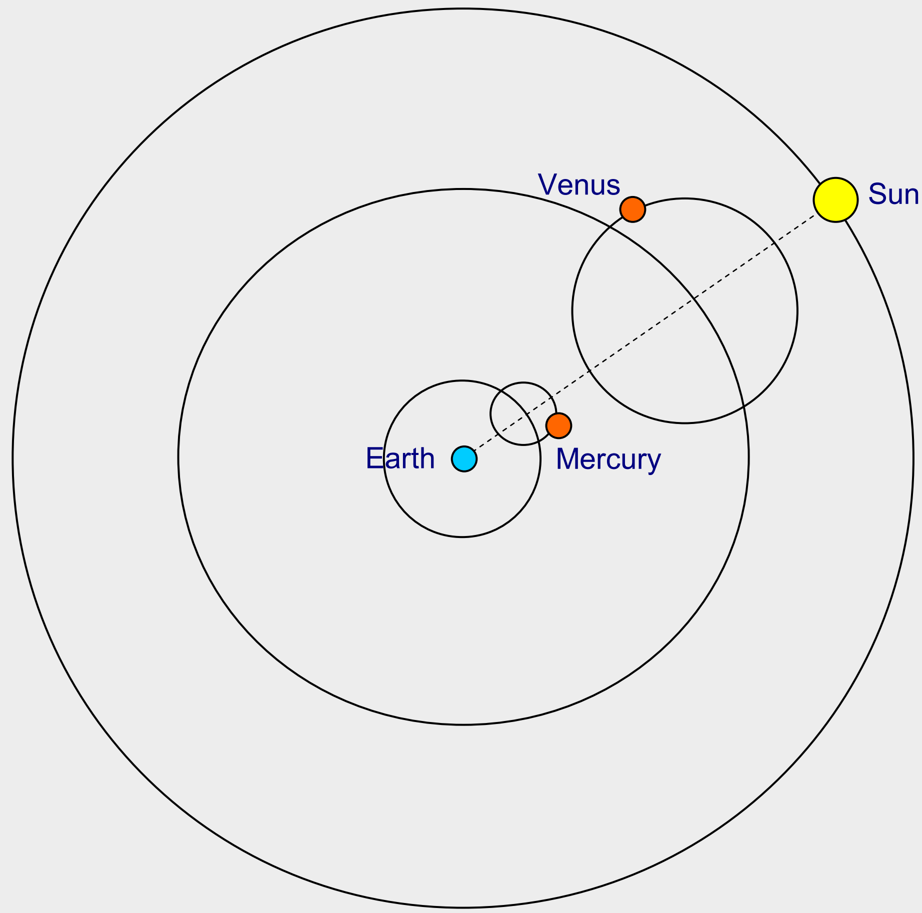
9.3 Cycles and Epicycles: Hipparchus and Ptolemy

A new way of combining circular motions to account for the movements of the sun, moon and planets was introduced by Hipparchus (second century BC) and realized fully by Ptolemy (around AD 150). Hipparchus was aware the seasons weren't quite the same length, so he suggested that the sun went around a circular path at uniform speed, but that the earth wasn't in the center of the circle. Now the solstices and equinoxes are determined by how the tilt of the earth's axis lines up with the sun, so the directions of these places from the earth are at right angles. If the circle is off center, though, some of these seasons will be shorter than others. We know the shortest season is fall (in our hemisphere).

Another way of using circular motions was provided by Hipparchus' theory of the moon. This introduced the idea of the "epicycle", a small circular motion riding around a big circular motion. (See below for pictures of epicycles in the discussion of Ptolemy.) The moon's position in the sky could be well represented by such a model. In fact, so could all the planets. One problem was that to figure out the planet's position in the sky, that is, the line of sight from the earth, given its position on the cycle and on the epicycle, needs trigonometry. Hipparchus developed trigonometry to make these calculations possible.

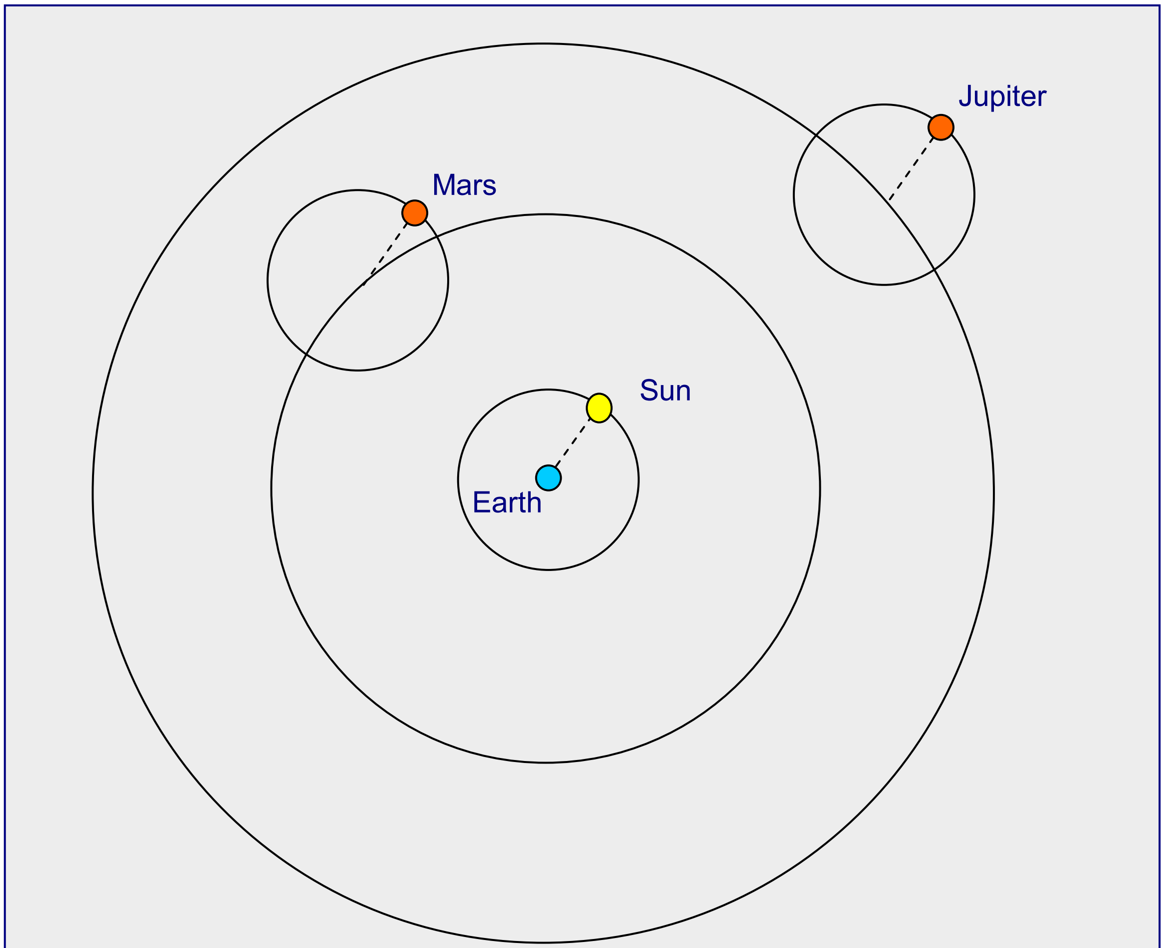
Ptolemy wrote the "bible" of Greek (and other ancient) astronomical observations in his immense book, the "Almagest". This did for astronomy at the time what Euclid's Elements did for geometry. It gave huge numbers of tables by which the positions of planets, sun and moon could be accurately calculated for centuries to come. We cannot here do justice to this magnificent work, but I just want to mention one or two significant points which give the general picture.

To illustrate the mechanism, we present here a slightly simplified version of his account of how the planets moved. The main idea was that each planet (and also, of course, the sun and moon) went around the earth in a cycle, a large circle centered at the center of the earth, but at the same time the planets were describing smaller circles, or epicycles, about the point that was describing the cycle. Mercury and Venus, as shown in the figure, had epicycles centered on the line from the earth to the sun. This picture does indeed represent fairly accurately their apparent motion in the sky—note that they always appear fairly close to the sun, and are not visible in the middle of the night.



Cycles and Epicycles for the Inner Planets: Basic Version

The planets Mars, Jupiter and Saturn, on the other hand, can be seen through the night in some years. Their motion is analyzed in terms of cycles greater than the sun's, but with epicycles exactly equal to the sun's cycle, and with the planets at positions in their epicycles which correspond to the sun's position in its cycle—see the figure below.



Cycles and Epicycles for the Outer Planets: Basic Version

This system of cycles and epicycles was built up to give an accurate account of the observed motion of the planets. Actually, we have significantly simplified Ptolemy's picture. He caused some of the epicycles to be not quite centered on the cycles, they were termed eccentric. This departure from apparent perfection was necessary for full agreement with observations, and we shall return to it later. Ptolemy's book was called the *Almagest* in the Middle Ages, the Arabic prefix *al* with the Greek for "the greatest" the same as our prefix *mega*.

9.4 Ptolemy's View of the Earth

It should perhaps be added that Ptolemy, centuries after Aristarchus, certainly did not think the earth rotated. (Heath, *Greek Astronomy*, page 48). His point was that the aither was lighter than any of the earthly elements, even fire, so it would be easy for it to move rapidly, motion that would be difficult and unnatural for earth, the heaviest material. And if the earth *did* rotate, Athens would be moving at several hundred miles per hour. How could the air keep up? And even if somehow it did, since it was light, what about heavy objects falling through the air? If somehow the air was carrying them along, they must be very firmly attached to the air, making it difficult to see how they could ever move relative to the air at all! Yet they can be, since they can fall, so the whole idea must be wrong.

Ptolemy did, however, know that the earth was spherical. He pointed out that people living to the east saw the sun rise earlier, and how much earlier was proportional to how far east they were located. He also noted that, though all must see a lunar eclipse simultaneously, those to the east will see it as later, e.g. at 1 a.m., say, instead of midnight, local time. He also observed that on traveling to the north, Polaris rises in the sky, so this suggests the earth is curved in that direction too. Finally, on approaching a hilly island from far away on a calm sea, he noted that the island seemed to rise out of the sea. He attributed this phenomenon (correctly) to the curvature of the earth.

10 How Classical Knowledge Reached Baghdad

10.1 The Classical Achievement in Mathematics and Science

With Ptolemy's *Almagest* giving detailed accounts and predictions of the movement of the planets, we reach the end of the great classical period in science. Let's review what was achieved.

First, the Babylonians developed a very efficient system of numbers and measures of all kind, primarily for business purposes. Unfortunately, it did not pass through to the Greeks and Romans, except for measures of time and angle, presumably those are the units relevant for recording astronomical observations. The Babylonians kept meticulous astronomical records over many centuries, mainly for astrological purposes, but also to maintain and adjust the calendar. They had tables of squares they used to aid multiplication, and even recorded solutions to word problems which were a kind of pre-algebra, a technique broadened and developed millennia later in Baghdad, as we shall see.

The Egyptians developed geometry for land measurement (that's what it means!), the land being measured for tax assessment.

The Greeks, beginning with Thales then Pythagoras, later Euclid, Archimedes and Apollonius, greatly extended geometry, building a logical system of theorems and proofs, based on a few axioms. An early result of this very abstract approach was the Pythagoreans' deduction that the square root of 2 could not be expressed as a ratio of whole numbers. This was a result they didn't want to be true, and that no-one would have guessed. Remember, they believed that God constructed the Universe out of pure numbers! Their accepting of this new "irrational" truth was a testimony to their honesty and clear mindedness.

The development of geometry took many generations: it could only happen because people with some leisure were able to record and preserve for the next generation complicated arguments and results. They went far beyond what was of immediate practical value and pursued it as an intellectual discipline. Plato strongly believed such efforts led to clarity of thought, a valuable quality in leaders. In fact, above the door of his academy he apparently wrote: "let no one who cannot think geometrically enter here."

Over this same period, the Greeks began to think *scientifically*, meaning that they began to talk of natural origins for phenomena, such as lightning, thunder and earthquakes, rather than assuming they were messages from angry gods. Similarly, Hippocrates saw epilepsy as a physical disease, possibly treatable by diet or life style, rather than demonic possession, as was widely believed at the time (and much later!).

The geometric and scientific came together in analyzing the motion of the planets in terms of combinations of circular motions, an approach suggested by Plato, and culminating in Ptolemy's *Almagest*. This Greek approach to astronomy strongly contrasted with that of the Babylonians, who had made precise solar, lunar and planetary observations for many hundreds of years, enough data to predict future events, such as eclipses, fairly accurately, yet they never attempted to construct geometric models to analyze those complex motions.

10.2 Why did Mathematics and Science Grind to a Halt?

Why did the development of science on the ancient world pretty much end after 800 years, around 200 AD or so? For one thing, the Romans were now dominant, and although they were excellent engineers, building thousands of miles of roads, hundreds of military garrisons, and so on, they did very little science. And, *the Greeks themselves lost interest*: Plato's Academy began to concentrate on rhetoric, the art of speechmaking. Perhaps this had been found to be more valuable for an aspiring leader than the ability to think geometrically or scientifically—or perhaps better for winning elections and persuading people. Furthermore, with the conversion of the Roman empire to Christianity around 300 AD, saving souls became a top priority in the Catholic church. As St. Augustine put it,

"Nor need we be afraid lest the Christian should be rather ignorant of the force and number of the elements, the motion, order and eclipses of the heavenly bodies, the form of the heavens, the kinds and natures of animals, shrubs and stones ... It is enough for the Christian to believe that the cause of all created things, whether heavenly or earthly, whether visible or invisible, is none other than the goodness of the Creator, who is the one true God."

It's a little puzzling to put this together with Botticelli's picture, showing Augustine looking prayerful but with scientific instruments in plain sight! (Augustine *was* very interested in science and many other unholy things earlier in life.)

St. Augustine by Botticelli (Wikipedia Commons).

10.3 But Some Christians Preserved the Classical Knowledge…

Actually, the story of the treatment of the Greek mathematical and scientific knowledge by the early Christian church is complicated, like the church itself. Recall that mathematics and science effectively ended in Alexandria with the murder of Hypatia in 415 AD, ordered by the Patriarch Cyril. This same Cyril engaged in a violent theological quarrel with the Patriarch of Constantinople, Nestorius. The question was the relative importance of the Virgin Mary. Cyril demanded that she be referred to as the Mother of God, Nestor would only accept Mother of Christ. This was all part of a debate about the nature of Christ: did he have two natures, human and divine, or one nature? Nestor thought two, of which only one, the human, died on the cross. Getting this right was very important: it was believed that salvation depended on it. However, the dispute was also (and perhaps principally) a struggle for power. At the Council of Ephesus in 431, Cyril arrived early with a large group of strong men, handed out bribes, and got the assembled bishops to condemn Nestor as a heretic. (Further complications ensued at later Councils, see for example *The Closing of the Western Mind*, Charles Freeman, Knopf, 2002, page 259 on, but it was all bad news for Nestor and his followers, who became known as Nestorians.)

10.4 How the Nestorians Helped Science Survive

What has this got to do with science? It is a crucial link in the chain. In contrast to most of the rest of the church, the Nestorians preserved and read the works of Aristotle, Plato, etc., and translated many of them into Syriac. They felt that clear thinking was useful in theology. Being declared heretics meant that it was no longer a good idea to stay in the Roman Empire, and, in fact, they were expelled.

Let's briefly review the extent of the Roman Empire to understand what expulsion implied.

The maps below are from:

http://www.roman-empire.net/maps/empire/extent/augustus.html

At its greatest extent, in 116 AD, pictured above, notice that the Empire included almost all of present-day Iraq, including the port of Basra (bottom right, on the Persian Gulf). However, this didn't last long—the Romans' most powerful enemy, the Persians (now known as Iranians), recaptured the territory after a short Roman occupation.

At the time of the death of Constantine, 337 AD, the Empire was officially Christian. The eastern part of the Empire, ruled from Constantinople and Greek speaking, became known as Byzantium. The Empire's total extent is shown below:

The Nestorians found temporary refuge with Syriac speaking sympathizers in Edessa (see Google map below, 37 10 N, 38 47 E. Istanbul (top left) is of course Constantinople):

(Nestor was a pupil of Theodore of Mopsuestia in Antioch, Syria. When Nestor was condemned, these Arab Christians broke with the Byzantine church, forming the Assyrian Church of the East, see Wikipedia.)

10.5 On into Persia

This was all during the time of the second Persian Empire (226-651), the Sassanid Empire.

The Sassanid Persian kings saw an opportunity to handle their own considerable number of Christian subjects better. They granted protection to Nestorians in 462, then in 484, they executed the Bishop of Nisibis (37 04 N, 41 13 E) (who was anti Nestorian, pro Byzantine) and replaced him with a Nestorian. (This is from Wikipedia.) The Nestorians settled in the Persian Empire, moving eventually to Gundishapur (near modern Dezful, at 32 25 N, 48 26 E). These Nestorians sent out many missionaries, for example reaching China in 635, and even Korea, and founding many churches, races still remain today. (However, foreign religions were suppressed in China in the 800's.)

The academy at Gundishapir had Syraic as the working language. Under a Sassanid monarch, Khosrau I, 531 – 579 AD, it became famous for learning. Although Khosrau I was a Zoroastrian, the dominant Persian religion, he was tolerant of all religions, in fact one of his sons became a Christian. He greatly improved the infrastructure, building palaces, strong defenses, and irrigation canals. He encourages science and art, collecting books from all over the known world, and introducing chess from India. (*Trivial Fact*: "Checkmate" is a corruption of the Persian shah mat, meaning the king is dead.) He had Syriac and Greek works translated into Persian. He also sent a famous physician Borzuyeh to India to invite Indian and Chinese scholars to Gundishapur.

10.6 The Advent of Islamic Rule

In 622, the prophet Muhammad left hostile Mecca to found his own theocratic state in Medina (just over two hundred miles to the north, both in western Saudi Arabia). He readily attracted converts, and built an army that captured Mecca eight years later. He died in 632, but his armies continued to conquer. Both Romans and Persians were by this point rather weak militarily, having spent decades fighting each other. The Sassanid dynasty fell to Muslim Arab armies in 638 AD. Alexandria was conquered in 642. These Muslims, although at war with Byzantium, were tolerant of their ethnic brethren, the Arab Christians. The first dynasty, the Umayyad (660 – 750), centered in Damascus, included Hisham ibn Abd as-Malik, who encouraged the arts, education, and "translation of numerous literary and scientific masterpieces into Arabic" (Wikipedia). (The Muslim Empire was now vast: a Hindu rebellion in Sindh was subdued; at the same time Umayyad armies went north from Spain, but were defeated at Tours, France, in 732. It has been argued that if the Arab armies had won at Tours, all Europe would have become Islamic, and still would be.)

In 749, a second dynasty, the Abbasid caliphate, began. In 762 the Abbasid Caliph al-Mansur built a magnificent new capital: Baghdad. Al-Mansur emulated the Persian rulers, building a palace library like the Sassanid Imperial Library, except that now everything was to be translated into Arabic. Harun ar-Rashid, Caliph from 786 to 808, sent agents to buy Greek manuscripts from Constantinople, to be translated into Arabic. At the same time, the *Siddhantas* wrrived from India: a set of Indian astronomical works, including trigonometric tables that likely originated with Hipparchus, and had then found their way to the Greek cities in India and Afghanistan founded by Alexander. (It's worth noting that the first paper mill outside China was built in Baghdad in 794, the secret having been given by prisoners of war from a battle against the Chinese in Central Asia. In fact, the cheap availability of paper made the complex Abbasid bureaucracy reasonably efficient.)

Meanwhile, Gundishapur wasn't far away: generously funded court appointments drew physicians (including al-Mansur's personal physician) and teachers to Baghdad.

Later, under the Abbasid Caliph al-Ma'mun (813 – 833), the *House of Wisdom* was founded (in 828): a large library and translation center into Arabic: first from Persian, then Syriac, then Greek. Many works were translated from Syriac into Arabic, including some Archimedes and all Euclid. Hunayn, a Christian, from Jundishapur, redid many translations to make them more readable.

10.7 The House of Wisdom: al-Khwarismi

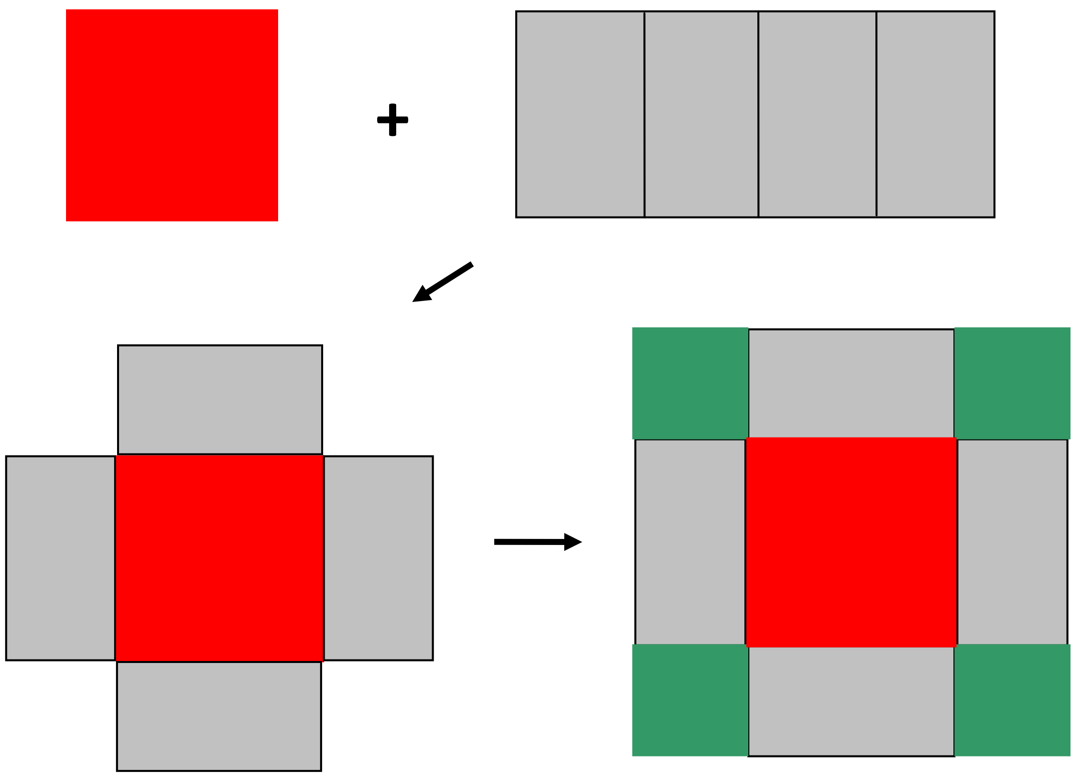
Perhaps the most famous scholar from the House of Wisdom is Al-Khwarismi (780 – 850). The word *algorithm*, meaning some kind of computational procedure, is just a mangling of his name. This is because he wrote the book that introduced the Hindu numbering system (now known as Arabic) to the Western world, and medieval scholars used his name to refer to routines for multiplication using Arabic numbers, far more efficient than anything possible with the previously used Roman numerals!

He also wrote the book on algebra: that word is actually "al-jabr" meaning completion. (We'll see below why this is an appropriate term.) Actually, he didn't use symbols to denote unknown quantities, now the essence of algebra. Ironically, such symbols had been used by the Greek Diophantus, in Alexandria, in the 200's AD, but that work was apparently unknown to the Arabs. Instead, al-Khwarismi stated algebraic problems as word problems, as the Babylonians had over two millennia earlier, but he also gave geometric representations of his solutions.

Let's look at one of his examples: $x^2 + 10x = 39$. (OK, I've cheated by using *x*: he wrote it all out in words, but his thought process was as outlined below.)

This he thought of in terms of equating *areas*: a very natural approach to something beginning with a square! On the left we have a square of side *x* and a rectangle of sides *x* and 10.

His strategy is to add area to this to make it *one big square*—he takes the rectangle and divides it into four equal rectangles each having sides *x* and 10/4 = 5/2. He then glues these to the *x* square:



***al-jabr*: completing the square**

The next step is to extend this to give just one square, by adding the green bits. But to keep the equation valid, the same amount must of course be added to the other side. That is, 5/2×5/2×4 is to be added to each side. We can see that on the left we now have a square of side x + 5. on the right hand side, we have 39 + 25 = 64 = 8×8. Therefore, x + 5 = 8, and x = 3.

So by adding to both sides we have "completed the square", and *al-jabr* is this adding to get completion. Negative numbers were not in use at that time, so quadratics like $x^2 = 10x + 39$, for example, were treated separately, and several distinct cases had to be explained.

It's not clear that al-Khwarismi's own contribution, by which I mean really new mathematics, was great, but his influence was tremendous: his presentation of algebra, and of the Arab numerals, sparked much further mathematical development, both in Baghdad and, later, in the West, as we shall see.

11 Later Islamic Science

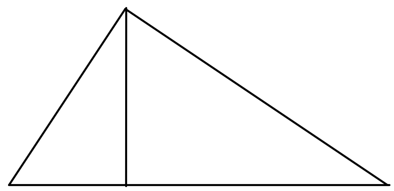
11.1 The Islamic World

Our interest here is in the scientific developments that took place in the Islamic world. We will look at a few of the most famous of the Islamic scientists, and only mention very briefly the political context in each case: the spread of Islam over much of the known world, and the subsequent political changes, were very complex. For example, after Baghdad, Cordoba in Spain became the preeminent center for science, but Spain was under the Umayyads, not the Abbasid caliph. Furthermore, some of the greatest Islamic scientists were Persian, and political developments there included a Shia revival in the tenth century, the Sunni Abbasids thereby losing their eastern empire, followed by a Turkish (Sunni) takeover—the Turks having been brought in as a palace guard.

11.2 Omar Khayyam

Omar Khayyam was born in Nishapur, in present-day northeastern Iran (see map) in 1048, a time when most of Persia (Iran) was under (Seljuk) Turkish rule. Initially, he did not find it a good environment for scholarly work, and in 1070 moved to Samarkand (see map). He did manage to write a famous book on Algebra.

His main contribution to that subject is a serious attack on cubic equations, such as finding *x* given that $2x^3 - 2x^2 + 2x - 1 = 0$. This particular problem has a geometric origin:



Given that for the right-angled triangle shown, the sum of the height and the shortest side is equal to the hypotenuse, find the ratio of the length of the shortest side to that of the other side.

Later Malik Shah, the third Seljuk sultan, and his Persian vizier al-Mulk, invited Khayyam to head up his observatory in Esfahan (his capital city, directly south of Teheran, see map). Khayyam measured the length of the year, getting 365.242198…days. This is correct to within one second: the error is in that last digit only!

Unfortunately for Khayyam, his friend the vizier was murdered by a terrorist group, the Assassins, who specifically targeted important political figures, on the road to Baghdad in 1092, and Malik-Shah died soon after that. His widow discontinued the observatory funding, but later his son Sanjar founded a center in Turkmenistan where Khayyam continued to do mathematics.

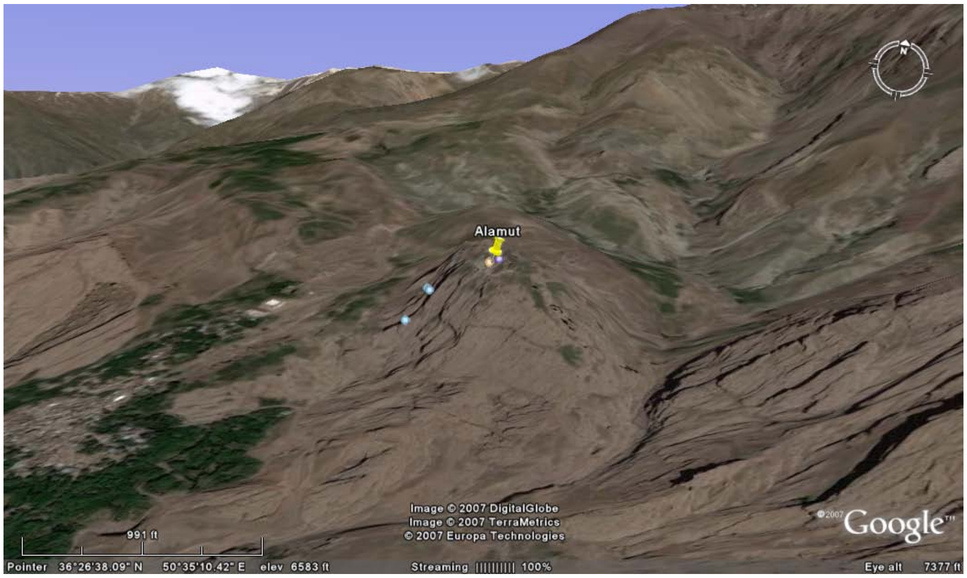
Omak Khayyam is also famous for his writings, such as the Rubaiyat. However, these have a distinctly irreligious flavor, and he had to tread carefully to minimize trouble with the Muslim religious authorities.

Note: many of the above facts are from the St Andrew's website.

11.3 Al-Tusi

Nasir al-Din al-Tusi was born in Tus, in northwest Iran (near Nishapur) in 1201.

He studies as a child at a mainly Shia religious school in Tus, followed by secondary education in Nishapur. However, around 1220, the Mongols invaded the area, causing much destruction. Al-Tusi found refuge with the Assassins in their mountain fortress at Alamut :



It isn't clear whether or not al-Tusi was actually a prisoner, but it is clear that he did some important scholarly work in this relatively quiet environment, writing on astronomy, mathematics, philosophy and ethics.

Nevertheless, when the Mongols, led by Ghengis Khan's grandson Hulagu (pictured below), took Alamut in 1256, al-Tusi switched sides, and the Mongols appointed him their scientific advisor.

Al-Tusi was with the Mongols when they attacked Baghdad in 1258. Apparently if the Caliph (the last of the Abbasids) had surrendered, little damage would have been done, but instead he refused, told Hulagu that God would avenge, but the Caliph did little to prepare defenses. The Mongols attacked (after a brief siege organized by a Chinese general), wrecking the Grand Library and throwing all the books in the river, burning down mosques and other buildings that were the work of generations, slaughtering the citizenry with abandon, breaking up the canal system that had kept the area fertile, and leaving too few survivors to repair the canals. This was the end of Baghdad as a cultural center for many centuries. The Mongols went on to fight with Egypt, but this time they were turned back in what is now the West Bank by superior Egyptian cavalry, in 1260: the same year that the Hulagu's brother Kublai Khan became Emperor of China, with his capital at Beijing. The Mongols in the Far East reached their limit when they attempted to invade Japan in the 1270's and 80's: their fleet was destroyed by a massive typhoon, one the Japanese termed *kamikaze*, meaning divine wind.

After Hulagu destroyed Baghdad, he constructed, at al-Tusi's suggestion, a magnificent observatory at Marageh in northwest Iran (see map above) with al-Tusi in charge. The observatory opened in 1262, and al-Tusi brought together many scholars and scientists. The observatory became, essentially, a university: al-Tusi had several pupils who made important contributions, and in fact his role was central in reviving Islamic science.

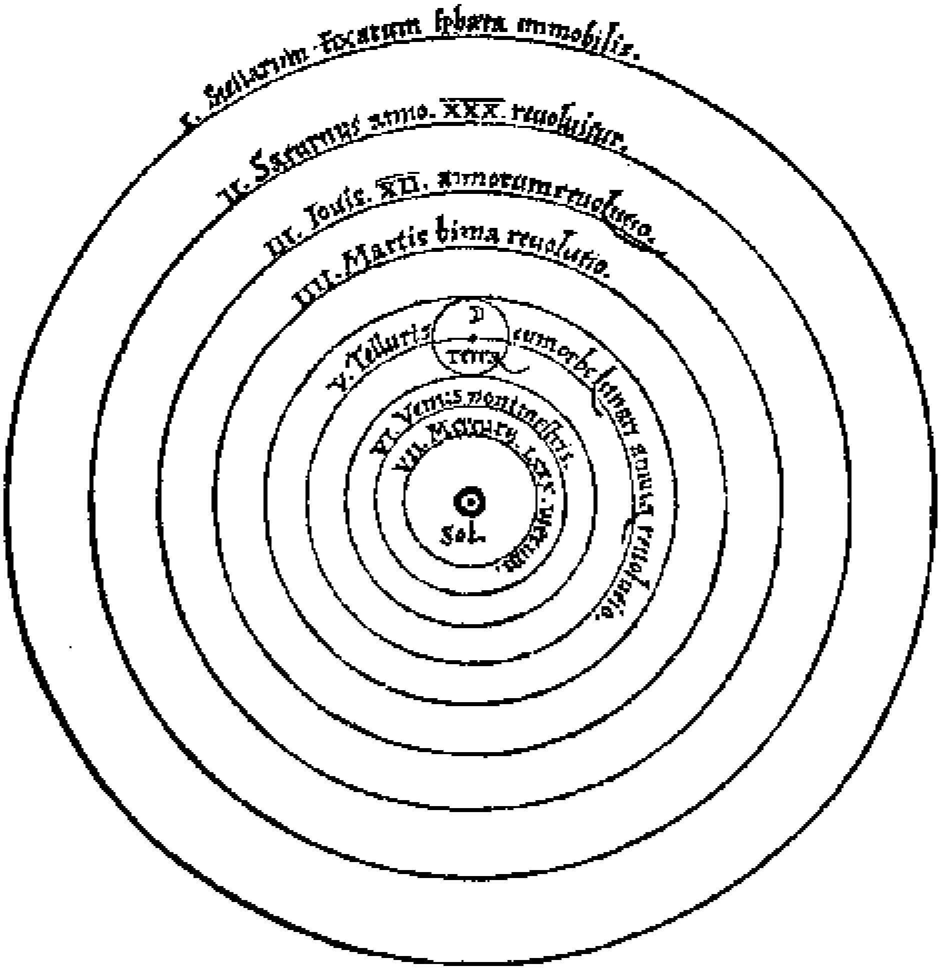
Al-Tusi himself developed plane and spherical trigonometry and wrote the first complete book on the subject. He also made the first really significant advance on Ptolemy's *Almagest*. Although Ptolemy's work described the planetary motions well, it contained some aesthetically unappealing features—it had strayed far from Plato's long ago suggestion that all should be described in terms of combinations of circular motions. In particular, accounting for the lack of coplanarity of planetary orbits required what amounted to an up-and-down linear component in planetary motion. Perhaps al-Tusi's most famous achievement was to demonstrate how such motion could be generated by a combination of two circular motions, see the animation at Tusi couple! Here's his original explanation, from a Vatican exhibit:

12 Galileo and the Telescope

12.1 Copernicus Challenges Ptolemy's Scheme

Ptolemy's picture of the solar system was almost fully accepted for the next fourteen hundred years, to be challenged by Copernicus (real name: Nicolaus Koppernigk) a mathematician and astronomer with a Polish father and a German mother, in 1530.

Copernicus' picture of the solar system had the sun at the center, and the earth went around it, as did the other planets.



We show here the picture from his original publication. Notice that the only exception to the rule that everything goes around the sun is the moon, which continues to go around the earth. One objection to the picture was that if the earth was indeed just another planet, how come it was the only one with a moon?

Other objections were based on the Aristotelian point of view—it was difficult to believe that all the other planets were composed of aither, and the earth of the other four elements, if they were all behaving in so similar a fashion. A further objection, which had long ago been raised by Aristotle to the idea of a rotating earth, was that the stresses would cause it to fly apart, and furthermore, anything thrown in the air would land far to the west.

Despite these problems, Pope Clement VII approved of a summary of Copernicus' work in 1530, and asked for a copy of the full work when it was available. This was not until 1543, the year Copernicus died.

As Copernicus' new picture of the universe became more widely known, misgivings arose. The universe had after all been created for mankind, so why wasn't mankind at the center? An intellectual revolutionary called Giordano Bruno accepted Copernicus' view, and went further, claiming that the stars were spread through an infinite space, not just on an outer sphere, and there were infinitely many inhabited worlds. Bruno was burned at the stake in 1600.

The real breakthrough that ultimately led to the acceptance of Copernicus' theory was due to Galileo, but was actually a *technological* rather than a conceptual breakthrough. It was Galileo's refinement and clever use of the telescope that persuaded people that the moon was a lot like the earth, and in some ways, so were the planets.

12.2 The Evolution of the Telescope

(*This section is mostly just a summary of Van Helden's excellent Introduction to* Sidereus Nuncius*, University of Chicago Press, 1989*)

The first known use of a magnifying glass to aid in reading was in the 1200s, by Roger Bacon at Oxford. It proved a boon to aging scholars, many of whom had been forced to retire while still relatively young. The idea spread throughout Europe, and Italian craftsmen, were making glasses for old men before 1300, (lens means lentil in Italian, so called because of the shape of the pieces of glass used) and for the myopic young not until a hundred and fifty years later. The reading glasses for the old men, who were longsighted, were *convex* lenses, (bulging in the middle like ()), whereas the glasses required by the shortsighted young were *concave* lenses, thinner in the middle than at the edges like)(, and hence more difficult to make and not so robust. The first time, as far as we know, that anyone put two lenses together to make a telescope-like optical instrument was in 1608, in Holland. The inventor of an opera-glass like telescope was called Lipperhey. He was unable to get a patent, however, because his invention was deemed too easy to reproduce. Perhaps the reason it had not been done before was that to get magnification, one needs a concave lens stronger than the convex lens being used with it, and commonly the lenses in wide use were the other way around.

Galileo found out about this invention in the spring of 1609, and immediately set about improving it. He saw it as a possible way out of his financial difficulties. He was an oldest son, and so was responsible for his younger sisters' dowries. He also had three children of his own, by his mistress. At the time, he was a Professor of Mathematics in the University of Padua, in the Venetian Republic. He soon put together a spyglass with a magnification of three, which many other people had already done. Galileo was an excellent experimentalist, and working with different lenses, he realized that the magnification was proportional to the ratio of the power of the concave (eyepiece) lens to the convex (more distant) lens. In other words, to get high magnification he needed a weak convex lens and a strong concave lens. the problem was that the opticians only made glasses in a narrow range of strengths, and three or so was the best magnification available with off the shelf lenses. Galileo therefore learned to grind his own lenses, and by August, he had achieved about ninefold linear magnification. This was an enormous improvement over everything else on the market. Galileo therefore approached the Senate of Venice to demonstrate his instrument. Many senators climbed the highest belltowers in Venice to look through the glass at ships far out at sea, and were impressed by the obvious military potential of the invention.

Galileo then wrote a letter to the Doge:

Galileo Galilei, a most humble servant of Your Serene Highness, being diligently attentive, with all his spirit, not only to discharging the duties pertaining to the lecturing of mathematics at the University of Padua, but also to bringing extraordinary benefit to Your Serene Highness with some useful and remarkable invention, now appear before You with a new contrivance of glasses, drawn from the most recondite speculations of perspective, which render visible objects so close to the eye and represent them so distinctly that those that are distant, for example, nine miles appear as though they were only one mile distant. This is a thing of inestimable benefit for all transactions and undertakings, maritime or terrestrial, allowing us at sea to discover at a much greater distance than usual the hulls and sails of the enemy, so that for two hours or more we can detect him before he detects us...

Galileo concludes the letter by asking for tenure:

....(the telescope is) one of the fruits of the science which he has professed for the past 17 years at the University of Padua, with the hope of carrying on his work in order to present You greater ones, if it shall please the Good Lord and Your Serene Highness that he, according to his desire, will pass the rest of his life in Your service.

It is nice to report that Galileo was granted tenure, and a reasonable salary, but—the bad news—with a proviso that further raises would not be forthcoming.

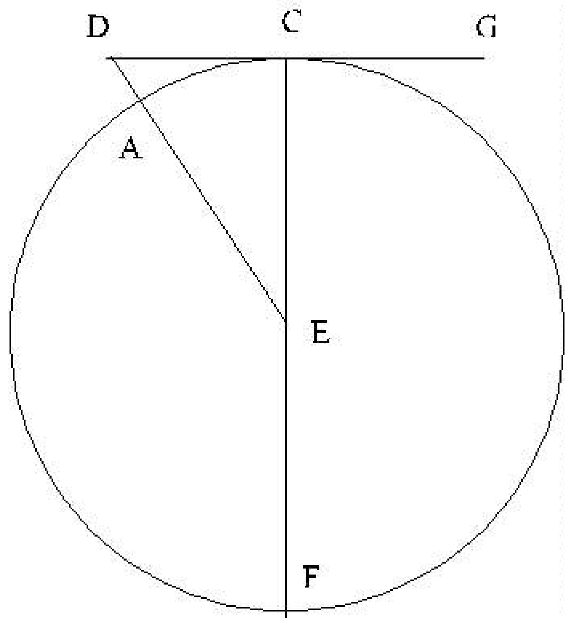
12.3 Mountains on the Moon

Galileo's first major astronomical discovery with the telescope was that the Moon's surface is mountainous, and not a perfect sphere as had always been assumed (see his drawings in *Sidereus Nuncius*). He built a convincing case for the reality of the mountains by sketching the appearance of parts of the Moon's surface at different times of the month, that is, under different angles of lighting, and showing how the light and shadow seen could be simply and naturally accounted for topographically, rendering the prevailing theory at the time, that the variations in light arose from something inside a perfect sphere, a cumbersome and unappealing alternative. This caused an uproar.

1. 

From the National Central Library of Florence (BNCF).

He was able to estimate the *height of the mountains on the moon* by seeing how far into the dark part bright spots could be discerned.



At half moon, a little geometry is enough to calculate the heights! Galileo himself worked an example: suppose a bright spot, presumably an illuminated mountaintop, is visible one-twentieth of a moon diameter into the dark side, at half-moon. Then the picture is as shown here (and is taken from *Sidereus Nuncius*). The light from the sun fully illuminates the right-hand half of the moon, plus, for example, the mountaintop at D. (GCD is a ray from the sun.) If the base of the mountain, vertically below D, is at A, and E is the moon's center, this is exactly the same problem as how far away is the horizon, for a person of given height on a flat beach. It can be solved using Pythagoras' theorem as we did for that problem, with the center of the moon E one of the points in the triangle, that is, the triangle is EDC.

A problem with asserting the existence of mountains is the apparent smooth roundness of the edge of the Moon, for which Galileo had two arguments. First, ranges behind those on the edge would tend to fill in the gaps. This is correct. Second, maybe things were fuzzed out by the Moon's atmosphere. This is wrong.

Galileo's next major discovery began with his observation on January 7, 1610, of what he took to be a rather odd set of three small fixed stars near Jupiter, and, in fact, collinear with the planet. These stars were invisible to the naked eye. He looked again at Jupiter on successive nights, and by the 15^{th} had realized that he was looking at *moons of Jupiter*, which were going around the planet with periods of the order of days. This caused even more consternation than the demystification of the Moon. Seven was a sacred number, and there were seven planets, wanderers, or moving stars. Jupiter's moons spoiled this. Furthermore, they suggested that it was o.k. to go in a circle about something other than the center of the universe, i.e. the Earth.

This made Copernicus' argument, that the Moon goes around the Earth and the Earth around the Sun, more plausible.

Again, Galileo's grantsmanship is admirable. In a masterstroke of public relations, he named the satellites after the Medici family, Dukes of Tuscany, where he applied for the position of mathematician to the court. He sent his most recent 20X telescope to the Duke, so that he could peruse the stars named after him and his brothers, and emphasized its military applicability.

13 Life of Galileo

13.1 Books

NOTE: Many books have been written about Galileo, and, in particular, about his interaction with the Church. An excellent short biography is *Galileo*, Stillman Drake, Oxford. Drake has also written *Galileo at Work: His Scientific Biography*, Dover. An enlightening book on the social context, and Galileo's adaptation to it, is *Galileo Courtier* by Mario Biagioli.

One classic is *The Crime of Galileo*, Giorgio de Santillana, 1955, University of Chicago Press. A fairly recent biography by a journalist is *Galileo, a Life*, James Reston, Jr., HarperCollins.

(I am certainly no expert in this complex field of study, and just present a collection of facts below to try to give the flavor of Galileo's life and times.)

13.2 Like Father, like Son

Galileo was born in Pisa, Tuscany in 1564, the son of Florentine musician Vincenzio Galilei. Actually, Vincenzio was a revolutionary musician—he felt the formal church music that then dominated the scene had become sterile, and that classic Greek poetry and myths had a power the church music lacked, that perhaps could be translated into modern music. He attempted some of this, and his work began the development that culminated in Italian opera.

To understand something of Galileo's early upbringing, here is a quote from his father, Vincenzio Galileo:

"It appears to me that those who rely simply on the weight of authority to prove any assertion, without searching out the arguments to support it, act absurdly. I wish to question freely and to answer freely without any sort of adulation. That well becomes any who are sincere in the search for truth."

Vincenzio wrote this in the introduction to *Dialogue on Ancient and Modern Music* (Google books).

Dialogo della Musica Antica et della Moderna, Florence, 1581. I took the quote from Reston quoting J.J. Fahie, *Galileo*, page 3. (Google books.)

Now Vincenzio had studied music with the leading musical theorist of the day, Gioseffo Zarlino in Venice. Zarlino had a Pythagorean approach: he believed any two notes that sounded right together were related by simple numerical ratios: if a plucked string gives middle C, a string exactly half the length will give the next C up, then a quarter the next, etc: you divide the string by the *same factor* each time. Similarly, this is how appropriate intervals are determined for *adjacent* notes. For example, a semitone is a 9/8 ratio (meaning that for a string held under constant tension, decreasing the length by this ratio will increase the note by a semitone. Notes two semitones apart would have a frequency difference 9/8 X 9/8. But what about notes *half* a semitone apart? Zarlino claimed that was impossible to divide a semitone evenly. His reason was that it would need a ratio of the square root of 9/8. The square root of 9 is 3, that of 8 = 4 x 2 = 2 x √2, so it couldn't be expresses as a ratio. (Remember √2 is *irrational*!)

But Vincenzio Galileo ridiculed this theory—he could *play* a note just half way between! In other words, he took a practical rather than a theoretical approach to music: what sounds right trumps any abstract mathematical discussion of music. In fact, this was a very old argument: Vincenzio Galileo was following Aristoxenus (4^{th} century BC Greek) instead of Pythagoras—see his book, *Dialogue on Ancient and Modern Music* .

It was also widely believed that if the tension in a string was doubled, that would be the same as halving the length at constant tension—another example of the linear/ratio/proportion mentality, probably from Aristotle. And, it wasn't difficult to believe: the tension was varied by tightening the string until it sounded right, there was actual measurement of tension. Vincenzio proved, by hanging weights on strings, that in fact the tension had to be *quadrupled* to have the same effect as halving the length.

So Galileo was brought up to believe that theoretical claims needed to be checked experimentally and, in particular, simple linear rules might not always be right.

At age 17, Galileo went to the University of Pisa. He enrolled as a medical student, following his father's advice, but turned to math, after persuading his father that he didn't want to be a doctor. His father allowed him to be tutored by the Tuscan court mathematician, Ricci, who designed fortifications, which no doubt impressed Galileo (Reston, page 15).

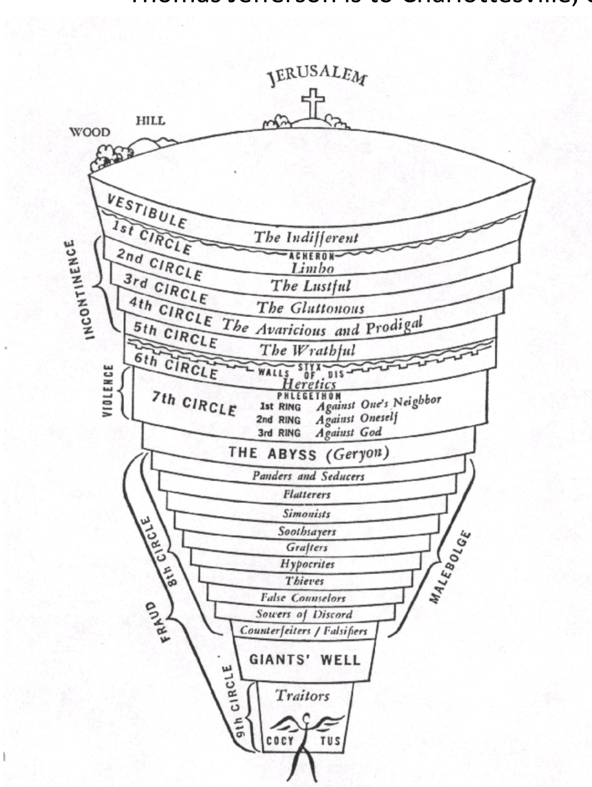
Possenti's Lamp in the Cathedral, Pisa.

13.3 Pendulums and Pulses

When he was eighteen, Galileo watched a lamp being lit in the cathedral at Pisa. (Fahie, page 9; but apparently Fahie is wrong on one (irrelevant!) detail, the Possenti lamp he shows was only hung a few years later.) These cathedral lamps were very ornate affairs, with many candles. The lamplighter would pull the lamp towards himself, light the candles then let it go. Galileo watched the swinging lit lamp for some time. He timed its swinging with his pulse, and realized that the period of one swing stayed the same, or very close to the same, as the swings became smaller and smaller. He realized this constancy of swing time could be used to measure a patient's pulse in hospital. (This was before watches had been invented—even pendulum clocks came a little later, both Galileo and Huygens developed the concept. Huygens was better.) Anyway, Galileo, working with a physician friend Santorio, constructed a simple device, which became known as a pulsilogia, and became widely used. Basically, it was a pendulum of adjustable length, the length being set by the physician so the swings coincided with the patient's pulse, a built-in ruler than read off the pulse rate.

13.4 The Roof of Hell

Galileo proved to be an extremely talented mathematician, and in his early twenties he wrote some tracts extending results of Archimedes on centers of gravity of shapes. At age 25, he was appointed to the Chair of Mathematics at Pisa. His job interview was to give two lectures to the Florentine Academy on mathematical topics. This Academy's main function was to glorify the Medicis, the ruling family, and of course Florence itself. Now Dante was to Florence what Thomas Jefferson is to Charlottesville, only on a bigger scale: perhaps his most famous work was his delineation of hell in Dante's Inferno. Hell started below ground, and continued down to the center of the earth. In contrast to heaven, out there somewhere and doubtless made of aetherial material, Hell was constructed of familiar stuff, so could be mathematically analyzed, like any architectural construction.



Hell

Galileo presented his lecture as a description of two different plans of hell, one by Antonio Manetti, a Florentine, and one by Alessandro Vellutello, from Lucca, a rival city. Manetti's Hell extended 8/9ths of the way from the Earth's center to the surface, Vellutello's only one-tenth of the way, giving a much smaller volume available, as Galileo pointed out. He showed off some of his new mathematical results here: the volume of a cone goes as the cube of its height, so the factor of about 10 difference in height of the two hells meant a difference of about a thousand in volume. People had been skeptical of Manetti's hell, though: they doubted that vast arched roof could hold up against gravity. Galileo argued that this was no problem: the dome in the cathedral in

Florence held up beautifully, and was relatively thinner than the 405 miles that Galileo estimated to be the thickness of Dante's cover.

The lectures went over very well, and he got the job. But then it dawned on him that he'd made a mistake, and his whole analysis, comparing the roof of hell with the cathedral, was seriously flawed. He kept very quiet. No-one else noticed. We'll return to this important point later.

At age 28, in 1592, Galileo moved to a better position at Padua, in the Venetian Republic, where he stayed until the age of 46.

13.5 Venice: Wine, Women and Dialogue

Reston's book certainly paints a vivid picture of the Venetian Republic at the time Galileo moved there! Venice, a city of 150,000 people, apparently consumed 40 million bottles of wine annually. There were more courtesans than in Rome. In 1599, Galileo met one Marina Gamba, 21 years old. He had three children by her, greatly upsetting his mother. Galileo also spent a lot of time with *Sagredo*, a young Venetian nobleman, both in the town and at Sagredo's very fancy house, or palace. Sagredo is featured as one of the disputants in *Dialogue Concerning the Two Chief World Systems*, and *Two New Sciences*. Another close friend during this period was Fra Paolo Sarpi, a Servite friar, and official theologian to the Republic of Venice in 1606, when Pope Paul put Venice under the interdict. Tensions between Venice and Rome were partly generated because Venice wanted to be able to tax churches built in Venice by Rome. Sarpi advised the Venetians to ignore the interdict, and the Jesuits were expelled from Venice. A nearly successful attempt on Sarpi's life was generally blamed on the Jesuits (from Drake, page 28).

13.6 The Telescope: Heaven Abolished?

(*In Bertolt Brecht's play Galileo , on page 65 Sagredo tells Galileo that he's as naïve as his daughter, and the pope is not going to just scribble a note in his diary: "Tenth of January, 1610, Heaven abolished."*)

When Galileo was 46 years old, in 1610, he developed the telescope, secured tenure and a big raise at Padua, then went on to make all the discoveries announced in *Sidereus Nuncius*: mountains on the moon, the moons of Jupiter, phases of Venus, etc. By naming the moons of Jupiter after the Medici family, Galileo landed the job of Mathematician and Philosopher (meaning Physicist) to the Grand Duke of Tuscany, and was able to return to his native land. This move upset his friends in Venice who had worked so hard to secure his promotion at Padua only months before.

Of course, Galileo's belief that his discoveries with the telescope strongly favored the Copernican world view meant he was headed for trouble with the Church. In fact, his Venetian friends warned him that it might be dangerous to leave the protection of the Venetian state.

13.7 Galileo Wins Over Some Jesuit Astronomers…

Nevertheless, in 1611, Galileo went to Rome and met with the Jesuit astronomers. Probably he felt that if he could win them over, he would smooth his path in any future problems with the Church. Father Clavius, author of Gregorian Calendar and undisputed leader of Jesuit astronomy had a hard time believing there were mountains on the moon, but he surrendered with good grace on looking through the telescope (Sant., pages 18, 20)

One archbishop wrote (p 20): "Bellarmine asked the Jesuits for an opinion on Galileo, and the learned fathers sent the most favorable letter you can think of …" Bellarmine was chief theologian of the Church, and a Jesuit himself. Bellarmine wrote in a letter to A. Foscarini, 12 April 1615:

Third, I say that if there were a true demonstration that the sun is at the center of the world and the earth in the third heaven, and that the sun does not circle the earth but the earth circles the sun, then one would have to proceed with great care in explaining the Scriptures that appear contrary, and say rather that we do not understand them than that what is demonstrated is false. But I will not believe that there is such a demonstration, until it is shown me."

(Quote from Feldhay, *Galileo and the Church*, Cambridge, 1995, page 35)

This was far from a mindless rejection of the Copernican picture—it just demanded a more convincing demonstration.

Somewhat earlier—Dec 1613—Galileo had written a letter to Castelli (a Benedictine abbot and former pupil of Galileo's) saying in essence that Scripture cannot contradict what we see in nature, so scripture, written for the business of saving souls and readable by everybody, sometimes is metaphorical in describing nature. It seems that Bellarmine and Galileo might have been able to come to some agreement on a world view.

Incidentally, Galileo was thinking about quite a different series of physics problems at this same time-trying to understand when things will float and when they sink. He believed Archimedes' Principle, that denser objects than water sink in water. (To be precise, the Principle states that the buoyant supporting force from the water on an immersed object is equal to the weight of the water displaced by the object. That is, it is equal to the weight of a volume of water equal to the volume of the object. So if the object is denser than water, its weight is greater than the buoyancy force and it sinks.)

It was pointed out to him that a ball made of ebony sinks in water, but a flat chip of ebony floats. We now understand this in terms of surface tension, but that had not been understood in Galileo's time. Nevertheless, Galileo gave an essentially correct answer: he observed that the chip floated somewhat below the previous level of the surface, dragging the water down slightly around its edges, so one should consider the floating body to be the chip plus the thin sheet of air over it, and putting these together gives an average density equal to that of water. Galileo discussed problems of this kind with a Florentine patrician, Filippo *Salviati*, and a group of his acquaintances. As usual, Galileo's style and ability to pulverize the opposition did not win many friends. (see Drake, pages 49-51). Salviati appears as one of the three disputants in Galileo's *Dialogue*.

13.8 …but Alienates Some Others

One more source of tension between Galileo and the Jesuits arose at this point. Since 1611, Galileo had been observing the motion of sunspots: small dark spots on the surface of the sun, easily visible through a telescope at sunset. They were observed independently at about the same time by Christopher Scheiner, a German Jesuit from Ingoldstadt. (It is possible that Scheiner had somehow heard of Galileo's observations.) Scheiner thought they were small dark objects circling the sun at some distance, Galileo correctly surmised they were actually on the sun's surface, another blow to the perfect incorruptibility of a heavenly body. Galileo published his findings in 1613, with a preface asserting his priority of discovery. This greatly upset Scheiner.

About this time, some members of another order of the Church, the Dominicans, were becoming aware of the Copernican world view, and began to preach against it. In 1613, Father Nicolo Lorini, a professor of ecclesiastical history in Florence, inveighed against the new astronomy, in particular "Ipernicus". (Sant p 25). He wrote a letter of apology after being reproved. In 1614, another Dominican, Father Tommaso Caccini, who had previously been reprimanded for rabble-rousing, preached a sermon with the text "Ye men of Galilee, why stand ye gazing up into the heaven?" He attacked mathematicians, and in particular Copernicus. (In the popular mind, mathematician tended to mean astrologer.) It should be added that these two were by no means representative of the order as a whole. The Dominican Preacher General, Father Luigi Maraffi, wrote Galileo an apology, saying "unfortunately I have to answer for all the idiocies that thirty or forty thousand brothers may or actually do commit".

According to De Santillana (page 45) in 1615 Father Lorini sends an altered copy of Galileo's letter to Castelli (mentioned above) to the Inquisition. He made two changes, one of which was to go from "There are in scripture words which, taken in the strictly literal meaning, look as if they differ from the truth" to "which are false in the literal meaning". Still, the inquisitor who read it thought it passable, although open to being misconstrued.

Nevertheless, in February 1616, the Copernican System was condemned. According to Drake (page 63): "A principal area of contention between Catholics and Protestants was freedom to interpret the Bible, which meant that any new Catholic interpretation could be used by the Protestants as leverage: if one reinterpretation could be made, why not wholesale reinterpretations? A dispute between the Dominicans and the Jesuits over certain issues of free will was still fresh in the pope's mind, as he had to take action in 1607 to stop members of the two great teaching orders from hurling charges of heresy at each other. These things suggest that Paul V, if not temperamentally anti-intellectual, had formed a habit of nipping in the bud any intellectual dispute that might grow into factionalism within the Church and become a source of strength for the contentions of the Protestants."

The pope asked Bellarmine to convey the ruling against the Copernican system to Galileo. Bellarmine had a meeting with Galileo, and apparently there were also some Dominicans present. Just what happened at this meeting is not quite clear, at least to me. Later (in May) Galileo was given an affidavit by Bellarmine stating that he must no longer *hold or defend* the propositions that the earth moves and the sun doesn't. Another document, however, which was unsigned (and therefore perhaps of questionable accuracy), stated that the Commissary of the Inquisition, in the name of the pope, ordered that Galileo could no longer *hold, defend or teach* the two propositions (Drake, page 67). This second document was not given to Galileo. The inclusion of *teach* was a crucial difference: it meant Galileo couldn't even *describe* the Copernican system. A week later (early March) books describing a moving earth were placed on the Index of Prohibited Books, some pending correction.

In the fall of 1618, three comets appeared. A book by a prominent Jesuit argued that the comets followed orbits close to those of planets, although they had short lifetimes. Galileo knew the comets moved in almost straight line motion much of the time. As usual, Galileo could not conceal his contempt of the incorrect views of others:

In Sarsi I seem to discern the belief that in philosophizing one must support oneself on the opinion of some celebrated author, as if our minds ought to remain completely sterile and barren unless wedded to the reasoning of someone else. Possibly he thinks that philosophy is a book of fiction by some author, like the Iliad … . Well, Sarsi, that is not how things are. Philosophy is written in this grand book of the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and to read the alphabet in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles and other geometric figures, without which it is humanly impossible to understand a single word of it; without these, one wanders in a dark labyrinth.

(This is from *The Assayer*, 1623)

Naturally, this further alienated the Jesuits.

In 1623, Galileo's admirer the Florentine Maffeo Barberini becomes Pope Urban VIII. The new pope saw himself as a widely educated man, who appreciated even Galileo's current theories. He had written a poem "In Dangerous Adulation" about Galileo's ideas. He also suggests his own pet theory to Galileo: even though the universe may be most simply understood by thinking of the sun at rest, God could have arranged it that way, but really with the earth at rest.

Galileo felt that with his friend and admirer as pope, and his affidavit from Bellarmine that didn't actually forbid him from describing the Copernican system, it was safe to write further about his world view. His ambition was to prove that the Copernican system must be correct, even though the more cumbersome Ptolemaic system might be fixed up to describe observations. (For example, the Danish Astronomer Tycho Brahe suggested that the sun went around the earth, but all the other planets went around the sun. That would account correctly for the phases of Venus.) Galileo was searching for some real proof that the earth was moving. He thought he found it in the tides. Why should all the water on the surface of the earth slosh around once or twice a day? Galileo decided it was because the earth was both rotating and moving around the sun, so for a given place on earth, its speed varies throughout the day, depending on whether its speed from the daily rotation is in the same direction as its speed from the earth's moving around the sun. This constant speeding up and slowing down is what Galileo thought generated the tides, so the tides were proof the earth was moving! (Actually this is not a good argument-the tides are really caused by the moon's gravity.)

Galileo worked on his new book, which he intended to call "Dialogue on the Tides", from 1624 to 1630. He was warned as he completed the work that that title seemed to imply he really held the view that the earth was moving, so he changed the title to *Dialogue Concerning the Two Chief Systems of the World-Ptolemaic and Copernican*. As usual, Galileo spared no-one in the book. He mocked the pope himself, by putting Urban's suggestion (see above) in the mouth of Simplicio, then dismissing it contemptuously (Reston, page 195).

The book was published in March 1632 in Florence. In August, an order came from the Inquisition in Rome to stop publication, and Galileo was ordered to stand trial. Apparently, someone—probably Scheiner, now living in Rome—had shown the pope the unsigned memo from the 1616 meeting, forbidding Galileo even to describe the Copernican system. Galileo was not too upset at the thought of a trial, because he held a trump card: the affidavit from Bellarmine. At the trial, Galileo said he had no memory of being forbidden to teach, and no signed document could be found to support the unsigned memo.

The trial did not address the scientific merits of the case, it was about whether or not Galileo had disobeyed an official order. It was suggested that he admit to some wrongdoing, and he would get off lightly. He agreed to tone down the *Dialogue*, pleading that he had been carried away by his own arguments. He was condemned to indefinite imprisonment, and, after some negotiation, was confined to his villa until his death in 1642. During this period, he wrote *Two New Sciences*, a book on the strength of materials and on the science of motion.

Galileo wrote in his old age, in his own copy of the Dialogue:

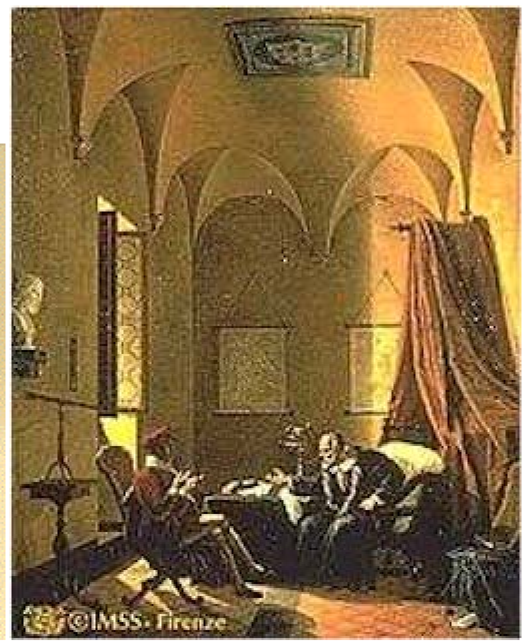
Take note, theologians, that in your desire to make matters of faith out of propositions relating to the fixity of sun and earth, you run the risk of eventually having to condemn as heretics those who would declare the earth to stand still and the sun to change position-eventually, I say, at *such a time as it might be physically or logically proved that the earth moves and the sun stands still.*

(Quoted in Drake, page 62).

It's perhaps worth adding one last word from the Jesuits (Reston, page 273) :

"*If Galileo had only known how to retain the favor of the fathers of this college, he would have stood in renown before the world; he would have been spared all his misfortunes, and could have written about everything, even about the motion of the earth.*"

(and here are some pictures from the museum in Florence: Galileo's preserved forefinger, and his meeting with Milton.)



14 Scaling: Why Giants Don't Exist

Galileo begins "Two New Sciences" with the striking observation that if two ships, one large and one small, have identical proportions and are constructed of the same materials, so that one is purely a scaled up version of the other in every respect, nevertheless the larger one will require proportionately more scaffolding and support on launching to prevent its breaking apart under its own weight. He goes on to point out that similar considerations apply to animals, the larger ones being more vulnerable to stress from their own weight (page 4):

> Who does not know that a horse falling from a height of three or four cubits will break his bones, while a dog falling from the same height or a cat from a height of eight or ten cubits will suffer no injury? ... and just as smaller animals are

> proportionately stronger and more robust than the larger, so also smaller plants are able to stand up better than the larger. I am certain you both know that an oak two hundred cubits high would not be able to sustain its own branches if they were distributed as in a tree of ordinary size; and that nature cannot produce a horse as large as twenty ordinary horses or a giant ten times taller than an ordinary man unless by miracle or by greatly altering the proportions of his limbs and especially his bones, which would have to be considerably enlarged over the ordinary.

For more of the text, click here.

To see what Galileo is driving at here, consider a chandelier lighting fixture, with bulbs and shades on a wooden frame suspended from the middle of the ceiling by a thin rope, just sufficient to take its weight (taking the electrical supply wires to have negligible strength for this purpose). Suppose you like the design of this particular fixture, and would like to make an exactly similar one for a room twice as large in every dimension. The obvious approach is simply to double the dimensions of all components. Assuming essentially all the weight is in the wooden frame, its height, length and breadth will all be doubled, so its volume—and hence its weight—will increase eightfold. Now think about the rope between the chandelier and the ceiling. The new rope will be eight times bigger than the old rope just as the wooden frame was. But the weight-bearing capacity of a uniform rope does *not* depend on its length (unless it is so long that its own weight becomes important, which we take not to be the case here). How much weight a rope of given material will bear depends on the *cross-sectional area* of the rope, which is just a count of the number of rope fibers available to carry the weight. The crucial point is that if the rope has all its dimensions doubled, this cross-sectional area, and hence its weight-carrying capacity, is only increased *fourfold*. Therefore, the doubled rope will not be able to hold up the doubled chandelier, the weight of which increased eightfold. For the chandelier to stay up, it will be necessary to use a new rope which is considerably fatter than that given by just doubling the dimensions of the original rope.

This same problem arises when a weight is supported by a pillar of some kind. If enough weight is piled on to a stone pillar, it begins to crack and crumble. For a uniform material, the weight it can carry is proportional to the cross-sectional area. Thinking about doubling all the dimensions of a stone building supported on stone pillars, we see that the weights are all increased eightfold, but the supporting capacities only go up fourfold. Obviously, there is a definite limit to how many times the dimensions can be doubled and we still have a stable building.

As Galileo points out, this all applies to animals and humans too (page 130): "(large) increase in height can be accomplished only by employing a material which is harder and stronger than usual, or by enlarging the size of the bones, thus changing their shape until the form and appearance of the animals suggests a monstrosity."

He even draws a picture:

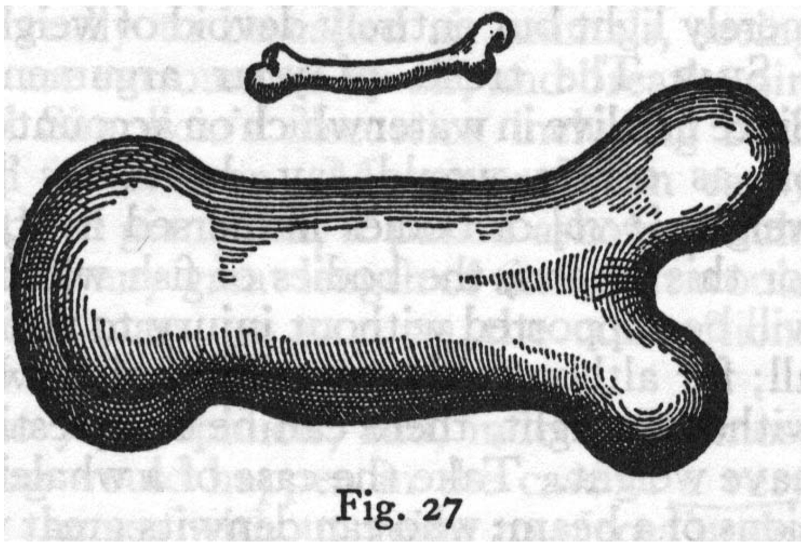


Fig. 27

Galileo understood that you cannot have a creature looking a lot like an ordinary gorilla except that it's sixty feet high. What about Harry Potter's friend Hagrid? Apparently he's twice normal height (according to the book) and three times normal width (although he doesn't look it on this link). But even that's not enough extra width (if the bone width is in proportion).

There is a famous essay on this point by the biologist J. B. S. Haldane, in which he talks of the more venerable giants in Pilgrim's Progress, who were ten times bigger than humans in every dimension, so their weight would have been a thousand times larger, say eighty tons or so. As Haldane says, their thighbones would only have a hundred times the cross section of a human thighbone, which is known to break if stressed by ten times the weight it normally carries. So these giants would break their thighbones on their first step. Or course, big creatures could get around this if they could evolve a stronger skeletal material, but so far this hasn't happened.

Another example of the importance of size used by Galileo comes from considering a round stone falling through water at its terminal speed. What happens if we consider a stone of the same material and shape, but one-tenth the radius? It falls much more slowly. Its weight is down by a factor of one-thousand, but the surface area, which gives rise to the frictional retardation, is only down by a factor of one hundred. Thus a fine powder in water---mud, in other words---may take days to settle, even though a stone of the same material will fall the same distance in a second or two. The point here is that as we look on smaller scales, gravity becomes less and less important compared with viscosity, or air resistance—this is why an insect is not harmed by falling from a tree.

This ratio of surface area to volume has also played a crucial role in evolution, as pointed out by Haldane. Almost all life is made up of cells which have quite similar oxygen requirements. A microscopic creature, such as the tiny worm *rotifer*, absorbs oxygen over its entire surface, and the oxygen rapidly diffuses to all the cells.   As larger creatures evolved, if the shape stayed the same more or less, the surface area went down relative to the volume, so it became more difficult to absorb enough oxygen. Insects, for example, have many tiny blind tubes over the surface of their bodies which air enters and diffuses into finer tubes to reach all parts of the body. The limitations on how well air will diffuse are determined by the properties of air, and diffusion beyond a quarter-inch or so takes a long time, so this limits the size of insects. Giant ants like those in the old movie "Them" wouldn't be able to breathe! The evolutionary breakthrough to larger size animals came with the development of blood circulation as a means of distributing oxygen (and other nutrients). Even so, for animals of our size, there has to be a tremendous surface area available for oxygen absorption. This was achieved by the development of lungs---the lungs of an adult human have a surface area of a hundred square meters approximately. Going back to the microscopic worm rotifer, it has a simple straight tube gut to absorb nutrients from food. Again, if larger creatures have about the same requirements per cell, and the gut surface absorbs nutrients at the same rate, problems arise because the surface area of the gut increases more slowly than the number of cells needing to be fed as the size of the creature is increased. this problem is handled by replacing the straight tube gut by one with many convolutions, in which also the smooth surface is replaced by one with many tiny folds to increase surface area. Thus many of the complications of internal human anatomy can be understood as strategies that have evolved for increasing available surface area per cell for oxygen and nutrient absorption towards what it is for simpler but much smaller creatures.

On the other hand, there is some good news about being big—it makes it feasible to maintain a constant body temperature. This has several advantages. For example, it is easier to evolve efficient muscles if they are only required to function in a narrow range of temperatures than if they must perform well over a wide range of temperatures. However, this temperature control comes at a price. Warm blooded creatures (unlike insects) must devote a substantial part of their food energy simply to keeping warm. For an adult human, this is a pound or two of food per day. For a mouse, which has about one-twentieth the dimensions of a human, and hence twenty times the surface area per unit volume, the required food for maintaining the same body temperature is twenty times as much as a fraction of body weight, and a mouse must consume a quarter of its own body weight daily just to stay warm. This is why, in the arctic land of Spitzbergen, the smallest mammal is the fox.

How high can a giant flea jump? Suppose we know that a regular flea can jump to a height of three feet, and a giant flea is one hundred times larger in all dimensions, so its weight is up by a factor of a million. Its amount of muscle is also up by a factor of a million, and when it jumps it rapidly transforms chemical energy stored in the muscle into kinetic energy, which then goes to gravitational potential energy on the upward flight. But the amount of energy stored in the muscle and the weight to be lifted are up by the same factor, so we conclude that the giant flea can also jump three feet! We can also use this argument in reverse—a shrunken human (as in I shrunk the kids) could jump the same height as a normal human, again about three feet, say. So the tiny housewife trapped in her kitchen sink in the movie could have just jumped out, which she'd better do fast, because she's probably very hungry!

15 Galileo's Acceleration Experiment

15.1 Summarizing Aristotle's View

Aristotle held that there are two kinds of motion for inanimate matter, natural and unnatural. Unnatural (or "violent") motion is when something is being pushed, and in this case the speed of motion is proportional to the force of the push. (This was probably deduced from watching oxcarts and boats.) Natural motion is when something is seeking its natural place in the universe, such as a stone falling, or fire rising. (We are only talking here about substances composed of earth, water, air and fire, the "natural circular motion" of the planets, composed of aether, is considered separately).

For the natural motion of heavy objects falling to earth, Aristotle asserted that the *speed of fall was proportional to the weight*, and *inversely proportional to the density of the medium* the body was falling through. He did also mention that there was some acceleration, as the body approached more closely its own element, its weight increased and it speeded up. However, these remarks in Aristotle are very brief and vague, and certainly not quantitative.

Actually, these views of Aristotle did not go unchallenged even in ancient Athens. Thirty years or so after Aristotle's death, Strato pointed out that a stone dropped from a greater height had a greater impact on the ground, suggesting that the stone picked up more speed as it fell from the greater height.

15.2 Two New Sciences

Galileo set out his ideas about falling bodies, and about projectiles in general, in a book called "*Two New Sciences*". The two were the science of motion, which became the foundation-stone of physics, and the science of materials and construction, an important contribution to engineering.

The ideas are presented in lively fashion as a dialogue involving three characters, Salviati, Sagredo and Simplicio. The official Church point of view, that is, Aristotelianism, is put forward by the character called Simplicio, and usually demolished by the others. Galileo's defense when accused of heresy in a similar book was that he was just setting out all points of view, but this is somewhat disingenuous---Simplicio is almost invariably portrayed as simpleminded.

For example, on TNS page 62, Salviati states:

I greatly doubt that Aristotle ever tested by experiment whether it be true that two stones, one weighing ten times as much as the other, if allowed to fall, at the same instant, from a height of, say, 100 cubits, would so differ in speed that when the heavier had reached the ground, the other would not have fallen more than 10 cubits.

Simplicio's response to this is not to think in terms of doing the experiment himself to respond to Salviati's challenge, but to scrutinize more closely the holy writ:

SIMP: His language would seem to indicate that he had tried the experiment, because he says: We see the heavier*; now the word* see *shows he had made the experiment.*

Sagredo then joins in:

SAGR: But I, Simplicio, who have made the test, can assure you that a cannon ball weighing one or two hundred pounds, or even more, will not reach the ground by as much as a span ahead of a musket ball weighing only half a pound, provided both are dropped from a height of 200 cubits.

This then marks the beginning of the modern era in science---the attitude that assertions about the physical world by authorities, no matter how wise or revered, stand or fall by experimental test. Legend has it that Galileo performed this particular experiment from the leaning tower of Pisa.

Galileo goes on to give a detailed analysis of falling bodies. He realizes that for extremely light objects, such as feathers, the air resistance becomes the dominant effect, whereas it makes only a tiny difference in the experiment outlined above.

15.3 Naturally Accelerated Motion

Having established experimentally that heavy objects fall at practically the same rate, Galileo went on to consider the central question about speed of fall barely touched on by Aristotle---*how does the speed vary during the fall?*

The problem is that it's very difficult to answer this question by just watching something fall---it's all over too fast. To make any kind of measurement of the speed, the motion must somehow be slowed down. Of course, some falling motions are naturally slow, such as a feather, or something not too heavy falling through water. Watching these motions, one sees that after being dropped the body rapidly gains a definite speed, then falls steadily at that speed. The mistake people had been making was in assuming that all falling bodies followed this same pattern, so that most of the fall was at a steady speed. Galileo argued that this point of view was false by echoing the forgotten words of Strato almost two thousand years earlier:

(TNS, page 163) *But tell me, gentlemen, is it not true that if a block be allowed to fall upon a stake from a height of four cubits and drive it into the earth, say, four finger-breadths, that coming from a height of two cubits it will drive the stake a much less distance; and finally if the block be lifted only one finger-breadth how much more will it accomplish than if merely laid on top of the stake without percussion? Certainly very little. If it be lifted only the thickness of a leaf, the effect will be altogether imperceptible. And since the effect of the blow depends upon the* *velocity of this striking body, can any one doubt the motion is very slow .. whenever the effect is imperceptible?*

15.4 Galileo's Acceleration Hypothesis

Having established by the above arguments and experiments that a falling body continues to pick up speed, or accelerate, as it falls, Galileo suggested the simplest possible hypothesis (paraphrasing the discussion on TNS page 161):

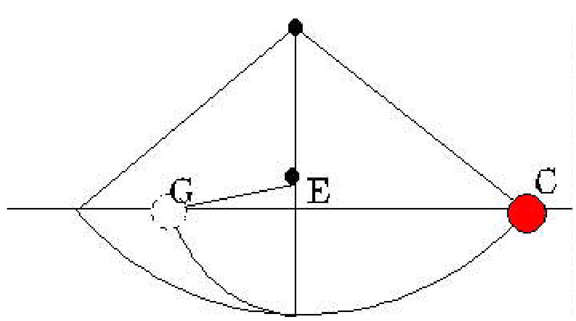
A falling body accelerates uniformly: it picks up equal amounts of speed in equal time intervals, so that, if it falls from rest, it is moving twice as fast after two seconds as it was moving after one second, and moving three times as fast after three seconds as it was after one second.

This is an appealingly simple hypothesis, but not so easy for Galileo to check by experiment---how could he measure the speed of a falling stone twice during the fall and make the comparison?

15.5 Slowing Down the Motion

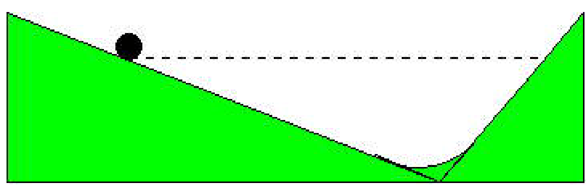
The trick is to *slow down the motion* somehow so that speeds can be measured, *without at the same time altering the character of the motion*. Galileo knew that dropping something through water that fell fairly gently *did* alter the character of the motion, it would land as gently on the bottom dropped from ten feet as it did from two feet, so slowing down the motion by dropping something through water changed things completely.

Galileo's idea for slowing down the motion was to have a ball roll down a ramp rather than to fall vertically. He argued that the speed gained in rolling down a ramp of given height didn't depend on the slope. His argument was based on an experiment with a pendulum and a nail, shown on page 171 of *Two New Sciences*. The pendulum consists of a thread and a lead bullet. It is drawn aside, the string taut, to some point C.



A nail is placed at E directly below the top end of the thread, so that as the pendulum swings through its lowest point, the thread hits the nail and the pendulum is effectively shortened, so that the bullet swings up more steeply, to G with the nail at E. Nevertheless, the pendulum will be seen to swing back up to almost the same *height* it started at, that is, the points G and C are the same height above level ground. Furthermore, when it swings back, it gets up as far as point C again, if we neglect a slight loss caused by air resistance. From this we can conclude that the speed with which the ball passes through the lowest point is the same in both directions. To see this, imagine first the situation *without* the nail at E. The ball would swing backwards and forwards in a *symmetrical* way, an ordinary pendulum, and certainly in this case the speed at the lowest point is the same for both directions (again ignoring gradual slowing down from air resistance). When we do put the nail in, though, we see from the experiment that on the swing back, the ball still manages to get to the beginning point C. We conclude that it must have been going the same speed as it swung back through the lowest point as when the nail wasn't there, because the instant it leaves the nail on the return swing it is just an ordinary pendulum, and how far it swings out from the vertical depends on how fast it's moving at the lowest point.

Galileo argues that a similar pattern will be observed if a ball rolls down a ramp which is smoothly connected to another steeper upward ramp, that is, the ball will roll up the second ramp to a level essentially equal to the level it started at, even though the two ramps have different slopes. It will then continue to roll backwards and forwards between the two ramps, eventually coming to rest because of friction, air resistance, etc.



Thinking about this motion, it is clear that (ignoring the gradual slowing down on successive passes) it must be going the *same speed* coming off one ramp as it does coming off the other. Galileo then suggests we imagine the second ramp steeper and steeper---and we see that if it's steep enough, we can think of the ball as just falling! He concludes that *for a ball rolling down a ramp, the speed at various heights is the same as the speed the ball would have attained (much more quickly!) by just falling vertically from its starting point to that height*. But if we make the ramp gentle enough, the motion will be slow enough to measure. (Actually, there is a difference between a rolling ball and a smoothly sliding or falling ball, but it does not affect the pattern of increase of speed, so we will not dwell on it here.)

15.6 Galileo's Acceleration Experiment

We are now ready to consider Galileo's experiment in which he tested his hypothesis about the way falling bodies gain speed. We quote the account from Two New Sciences, page 178:

A piece of wooden moulding or scantling, about 12 cubits long, half a cubit wide, and three finger-breadths thick, was taken; on its edge was cut a channel a little more than one finger in breadth; having made this groove very straight, smooth, and polished, and having lined it with parchment, also as smooth and polished as possible, we rolled along it a hard, smooth, and very round bronze ball. Having placed this board in a sloping position, by raising one end some one or two cubits above the other, we rolled the ball, as I was just saying, along the channel, noting, in a manner presently to be described, the time required to make the descent. We repeated this experiment more than once in order to measure the time with an accuracy such that the deviation between two observations never exceeded one-tenth of a pulse-beat. Having performed this operation and having assured ourselves of its reliability, we now rolled the ball only one-quarter *the length of the channel; and having measured the time of its descent, we found it precisely* one-half *of the former. Next we tried other distances, compared the time for the whole length with that for the half, or with that for two-thirds, or three-fourths, or indeed for any fraction; in such experiments, repeated a full hundred times, we always found that* the spaces traversed were to each other as the squares of the times*, and this was true for all inclinations of the plane, i.e., of the channel, along which we rolled the ball. We also observed that the times of descent, for various inclinations of the plane, bore to one another precisely that ratio which, as we shall see later, the Author had predicted and demonstrated for them.*

For the measurement of time, we employed a large vessel of water placed in an elevated position; to the bottom of this vessel was soldered a pipe of small diameter giving a thin jet of water which we collected in a small glass during the time of each descent, whether for the whole length of the channel or for part of its length; the water thus collected was weighed, after each descent, on a very accurate balance; the differences and ratios of these weights gave us the differences and ratios of the times, and this with such accuracy that although the operation was repeated many, many times, there was no appreciable discrepancy in the results.

15.7 Actually Doing the Experiment

(We do the experiment in class.) O.K., we don't line the channel with parchment, and we use an ordinary large steel ball (about one inch in diameter). We do use a water clock, with a student letting a jet of water into a polystyrene(!) cup during the interval between another student releasing the ball at some distance up the ramp and it hitting the stop at the bottom. We perform the experiment three times for the full ramp, and three times for a quarter of the distance. We weigh the amount of water in the cup with an ordinary balance. In one run we found (somewhat to our surprise!) that the average amount for the full ramp was 56 grams, for the quarter ramp 28 grams. This was partly luck, there was a scatter of a few grams. However, it does suggest that Galileo was not exaggerating in his claims of accuracy in *Two New Sciences*, since he was far more careful than we were, and repeated the experiment many more times.

16 Naturally Accelerated Motion

16.1 Distance Covered in Uniform Acceleration

In the last lecture, we stated what we called **Galileo's acceleration hypothesis**:

A falling body accelerates uniformly: it picks up equal amounts of speed in equal time intervals, so that, if it falls from rest, it is moving twice as fast after two seconds as it was moving after one second, and moving three times as fast after three seconds as it was after one second.

We also found, from the experiment, that a falling body will fall four times as far in twice the time. That is to say, we found that the time to roll one-quarter of the way down the ramp was one-half the time to roll all the way down.

Galileo asserted that the result of the rolling-down-the-ramp experiment confirmed his claim that the acceleration was uniform. Let us now try to understand why this is so. The simplest way to do this is to put in some numbers. Let us assume, for argument's sake, that the ramp is at a convenient slope such that, after rolling down it for one second, the ball is moving at two meters per second. This means that after two seconds it would be moving at four meters per second, after three seconds at six meters per second and so on until it hits the end of the ramp. (Note: to get an intuitive feel for these speeds, one meter per second is 3.6 km/hr, or 2.25 mph.)

To get a clear idea of what's happening, ***you should sketch a graph of how speed increases with time****. This is a straight line graph, beginning at zero speed at zero time, then going through a point corresponding to two meters per second at time one second, four at two seconds and so on. It sounds trivial, but is surprisingly helpful to have this graph in front of you as you read—so, find a piece of paper or an old envelope (this doesn't have to be too precise) and draw a line along the bottom marked 0, 1, 2, for seconds of time, then a vertical line (or y-axis) indicating speed at a given time—this could be marked 0, 2, 4, … meters per second. Now, put in the points (0,0), (1,2) and so on, and join them with a line.*

From your graph, you can now read off its speed not just at 0, 1, 2 seconds, but at, say 1.5 seconds or 1.9 seconds or any other time within the time interval covered by the graph.

The hard part, though, is figuring out how *far* it moves in a given time. This is the core of Galileo's argument, and it is essential that you understand it before going further, so ***read the next paragraphs slowly and carefully!***

Let us ask a specific question: how far does it get in two seconds? If it were moving *at a steady speed* of four meters per second for two seconds, it would of course move eight meters. But it can't have gotten that far after two seconds, because it just attained the speed of four meters per second when the time reached two seconds, so it was going at slower speeds up to that point. In fact, at the very beginning, it was moving very slowly. Clearly, *to figure out how far it travels during that first two seconds what we must do is to find its **average** speed during that period.*

*This is where the assumption of **uniform** acceleration comes in*. What it means is that the speed starts from zero at the beginning of the period, *increases at a constant rate*, is two meters per second after one second (half way through the period) and four meters per second after two seconds, that is, at the end of the period we are considering. Notice that the speed is one meter per second after half a second, and three meters per second after one-and-a-half seconds. From the graph you should have drawn above of the speed as it varies in time, it should be evident that, for this *uniformly* accelerated motion, the *average speed* over this two second interval *is the speed reached at half-time*, that is, two meters per second.

Now, the distance covered in any time interval is equal to the average speed multiplied by the time taken, so the distance traveled in two seconds is four meters—that is, two meters per second for two seconds.

Now let us use the same argument to figure how far the ball rolls in just one second. At the end of one second, it is moving at two meters per second. At the beginning of the second, it was at rest. At the half-second point, the ball was moving at one meter per second. By the same arguments as used above, then, the average speed during the first second was one meter per second. Therefore, *the total distance rolled during the first second is jus tone meter.*

We can see from the above why, in uniform acceleration, *the ball rolls **four times** as far when the time interval **doubles***. If the average speed were the *same* for the two second period as for the one second period, the distance covered would *double* since the ball rolls for twice as long a time. But since the speed is steadily increasing, *the average speed also doubles.* It is the *combination* of these two factors—moving at twice the average speed for twice the time—that gives the factor of four in distance covered!

It is not too difficult to show using these same arguments that the distances covered in 1, 2, 3, 4, ...seconds are proportional to 1, 4, 9, 16, .., that is, the squares of the times involved. This is left as an exercise for the reader.

16.2 A Video Test of Galileo's Hypothesis

In fact, using a video camera, we can check the hypothesis of uniform acceleration very directly on a falling object. We drop the ball beside a meter stick with black and white stripes each ten centimeters wide, so that on viewing the movie frame by frame, we can estimate where the ball is at each frame. Furthermore, the camera has a built-in clock—it films at thirty frames per second. Therefore, we can constantly monitor the speed by measuring how many centimeters the ball drops from one frame to the next. Since this measures distance traveled in one-thirtieth of a second, we must multiply the distance dropped between frames by thirty to get the (average) speed in that short time interval in centimeters per second.

By systematically going through all the frames showing the ball falling, and finding the (average) speed for each time interval, we were able to draw a graph of speed against time. It was a little rough, a result of our crude measuring of distance, but it was clear that speed was increasing with time at a steady rate, and in fact we could measure the rate by finding the speed reached after, say half a second. We found that, approximately, the rate of increase of speed was ten meters (1,000 cms) per second in each second of fall, so after half a second it was moving at about five meters per second, and after a quarter of a second it was going two and a half meters per second.

This rate of increase of speed is the same for all falling bodies, neglecting the effect of air resistance (and buoyancy for extremely light bodies such as balloons). It is called the *acceleration due to gravity*, written g, and is actually close to 9.8 meters per second per second. However, we shall take it to be 10 for convenience.

16.3 Throwing a Ball Upwards

To clarify ideas on the acceleration due to gravity, it is worth thinking about throwing a ball vertically upwards. If we made another movie, we would find that the motion going upwards is like a mirror image of that on the way down—the distances traveled between frames on the way up get shorter and shorter. In fact, the ball on its way up *loses* speed at a steady rate, and the rate turns out to be ten meters per second per second—the same as the rate of increase on the way down. For example, if we throw the ball straight upwards at 20 meters per second (about 45 mph) after one second it will have slowed to 10 meters per second, and after two seconds it will be at rest momentarily before beginning to come down. After a total of four seconds, it will be back where it started.

An obvious question so: how high did it go? The way to approach this is to find its average speed on the way up and multiply it by the time taken to get up. As before, it is helpful to sketch a graph of how the speed is varying with time. The speed at the initial time is 20 meters per second, at one second it's down to 10, then at two seconds it's zero. It is clear from the graph that the average speed on the way up is 10 meters per second, and since it takes two seconds to get up, the total distance traveled must be 20 meters.

16.4 Speed and Velocity

Let us now try to extend our speed plot to keep a record of the entire fall. The speed drops to zero when the ball reaches the top, then begins to increase again. We could represent this by a V-shaped curve, but it turns out to be more natural to introduce the idea of velocity.

Unfortunately velocity and speed mean the same thing in ordinary usage, but in science velocity means more: it includes speed and direction.

Our convention: velocity upwards is positive, downwards is negative.

This *is* just a convention: it might sometimes be convenient to take the opposite, with downward positive. The important thing is to have opposite signs for the two opposite directions. You can see this is useful if you're trying to calculate distance covered by multiplying together time and average velocity—if you traveled the same speed for equal times in opposite directions, your average *velocity* was zero, and you got nowhere. (In case you're wondering how we deal with velocities in several different directions, we'll come to that soon. For now, we'll stick with up and down motion.)

If we plot the velocity of the ball at successive times, it is +20 initially, +10 after one second, 0 after two seconds, -10 after three seconds, -20 after four seconds. If you plot this on a graph you will see that it is all on the same straight line. Over each one-second interval, the velocity decreases by ten meters per second throughout the flight. In other words, the acceleration due to gravity is -10 meters per second per second, or you could say it is 10 meters per second per second *downwards*.

16.5 What's the Acceleration at the Topmost Point?

Most people on being asked that for the first time say zero. That's wrong. But to see why takes some very clear thinking about just what is meant by velocity and acceleration. Recall Zeno claimed motion was impossible because at each instant of time an object has to be in a particular position, and since an interval of time is made up of instants, it could never move. The catch is that a second of time cannot be built up of instants. It can, however, be built up of *intervals* of time each as short as you wish. Average velocity over an interval of time is defined by dividing the distance moved in that interval by the time taken—the length of the interval. We define *velocity at an instant of time*, such as the velocity of the ball when the time is one second, by taking a small time interval which includes the time one second, finding the average velocity over that time interval, then repeating the process with smaller and smaller time intervals to home in on the answer.

Now, to find acceleration at an instant of time we have to go through the same process. Remember, acceleration is *rate of change of velocity*. This means that acceleration, too, can be positive or negative! You might think negative acceleration is just slowing down, but it could *also* mean speeding up in the direction you've chosen for velocity to be negative—so, be careful!

To find the acceleration at an instant we have to take some short but non-zero time interval that includes the point in question and find how much the velocity changes during that time interval.

Then we divide that velocity change by the time it took to find the acceleration, in, say, meters per second per second.

The point is that at the topmost point of the throw, the ball does come to rest for an instant. Before and after that instant, there is a brief period where the velocity is so small it looks as if the ball is at rest. Also, our eyes tend to lock on the ball, so there is an illusion that the ball has zero velocity for a short but non-zero period of time. But this isn't the case. The ball's velocity is always changing. To find its acceleration at the topmost point, we have to find how its velocity changes in a short time interval which includes that point. If we took, for example, a period of one-thousandth of a second, we would find the velocity to have changed by one centimeter per second. So the ball would fall one two-thousandth of a centimeter during that first thousandth of a second from rest-not too easy to see! The bottom line, though, is that the acceleration of the ball is 10 meters per second per second downwards throughout the flight.

If you still find yourself thinking it's got no acceleration at the top, maybe you're confusing velocity with acceleration. All these words are used rather loosely in everyday life, but we are forced to give them precise meanings to discuss motion unambiguously. In fact, lack of clarity of definitions like this delayed understanding of these things for centuries.

16.6 The Motion of Projectiles

We follow fairly closely here the discussion of Galileo in Two New Sciences, Fourth Day, from page 244 to the middle of page 257.

To analyze how projectiles move, Galileo describes two basic types of motion:

(i) Naturally accelerated vertical motion, which is the motion of a vertically falling body that we have already discussed in detail.

(ii) Uniform horizontal motion, which he defines as straight-line horizontal motion which covers equal distances in equal times.

This uniform horizontal motion, then, is just the familiar one of an automobile going at a steady speed on a straight freeway. Galileo puts it as follows:

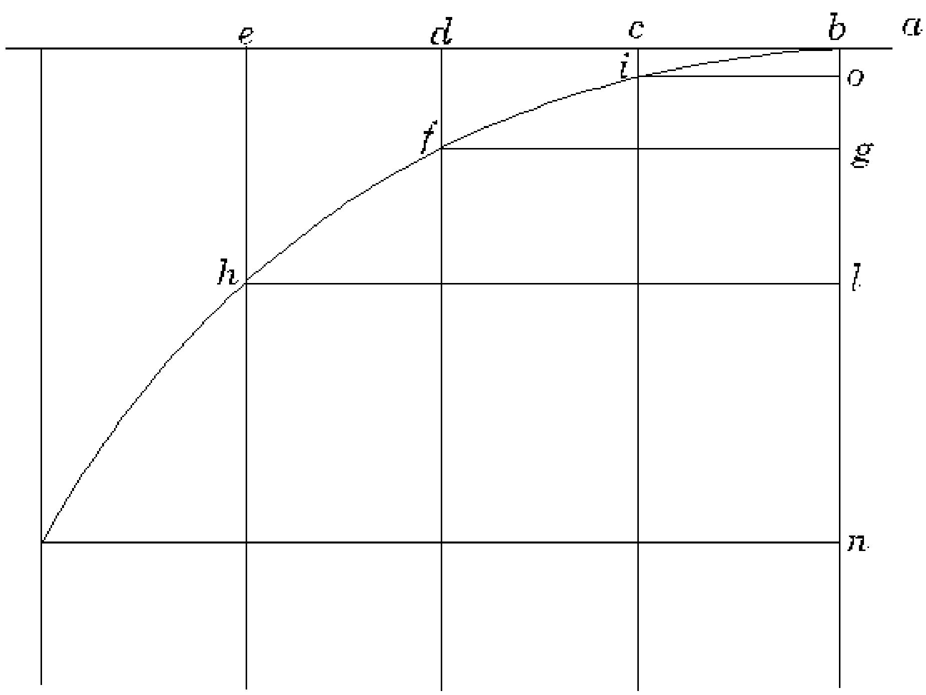
"Imagine any particle projected along a horizontal plane without friction; then we know...that this particle will move along this same plane with a motion that is uniform and perpetual, provided the plane has no limits."

This simple statement is in itself a substantial advance on Aristotle, who thought that an inanimate object could only continue to move as long as it was being pushed. Galileo realized the crucial role played by friction: if there is no friction, he asserted, the motion will continue indefinitely. Aristotle's problem in this was that he observed friction-dominated systems, like oxcarts, where motion stopped almost immediately when the ox stopped pulling. Recall that Galileo, in the rolling a ball down a ramp experiment, went to great pains to get the ramp very smooth, the ball very round, hard and polished. He knew that only in this way could he get reliable, reproducible results. At the same time, it must have been evident to him that if the ramp were to be laid flat, the ball would roll from one end to the other, after an initial push, with very little loss of speed.

16.7 Compound Motion

Galileo introduces projectile motion by imagining that a ball, rolling in uniform horizontal motion across a smooth tabletop, flies off the edge of the table. He asserts that when this happens, the particle's horizontal motion will continue at the same uniform rate, but, in addition, it will acquire a downward vertical motion identical to that of any falling body. He refers to this as a *compound* motion.

The simplest way to see what is going on is to study Galileo's diagram on page 249, which we reproduce here.



Imagine the ball to have been rolling across a tabletop moving to the left, passing the point *a* and then going off the edge at the point *b*. Galileo's figure shows its subsequent position at three equal time intervals, say, 0.1 seconds, 0.2 seconds and 0.3 seconds after leaving the table, when it will be at *i*, *f*, and *h* respectively.

The first point to notice is that the horizontal distance it has travelled from the table increases uniformly with time: *bd* is just twice *bc*, and so on. That is to say, its horizontal motion is just the same as if it had stayed on the table.

The second point is that its vertical motion is identical to that of a vertically falling body. In other words, if another ball had been dropped vertically from *b* at the instant that our ball flew off the edge there, they would always be at the same vertical height, so after 0.1 seconds when the first ball reaches *i*, the dropped ball has fallen to *o*, and so on. It also follows, since we know the falling body falls four times as far if the time is doubled, that *bg* is four times *bo*, so for the projectile *fd* is four times *ic.* This can be stated in a slightly different way, which is the way Galileo formulated it to prove the curve was a parabola:

The ratio of the vertical distances dropped in two different times, for example *bg*/*bo*, is always the square of the ratio of the horizontal distances travelled in those times, in this case *fg*/*io*.

You can easily check that this is always true, from the rule of uniform acceleration of a falling body. For example, *bl* is nine times *bo*, and *hl* is three times *io*.

Galileo proved, with a virtuoso display of Greek geometry, that the fact that the vertical drop was proportional to the square of the horizontal distance meant that the trajectory was a parabola. His definition of a parabola, the classic Greek definition, was that it was the intersection of a cone with a plane parallel to one side of the cone. Starting from *this* definition of a parabola, it takes quite a lot of work to establish that the trajectory is parabolic. However, if we define a parabola as a curve of the form $y = Cx^2$ then of course we've proved it already!

17 Using Vectors to Describe Motion

17.1 Uniform Motion in a Straight Line

Let us consider first the simple case of a car moving at a *steady speed down a straight road*. Once we've agreed on the units we are using to measure speed—such as miles per hour or meters per second, or whatever—a simple number, such as 55 (mph), tells us all there is to say in describing steady speed motion. Well, actually, this is not quite all—it doesn't tell us which way (east or west, say) the car is moving. For some purposes, such as figuring gas consumption, this is irrelevant, but if the aim of the trip is to get somewhere, as opposed to just driving around, it is useful to know the direction as well as the speed.

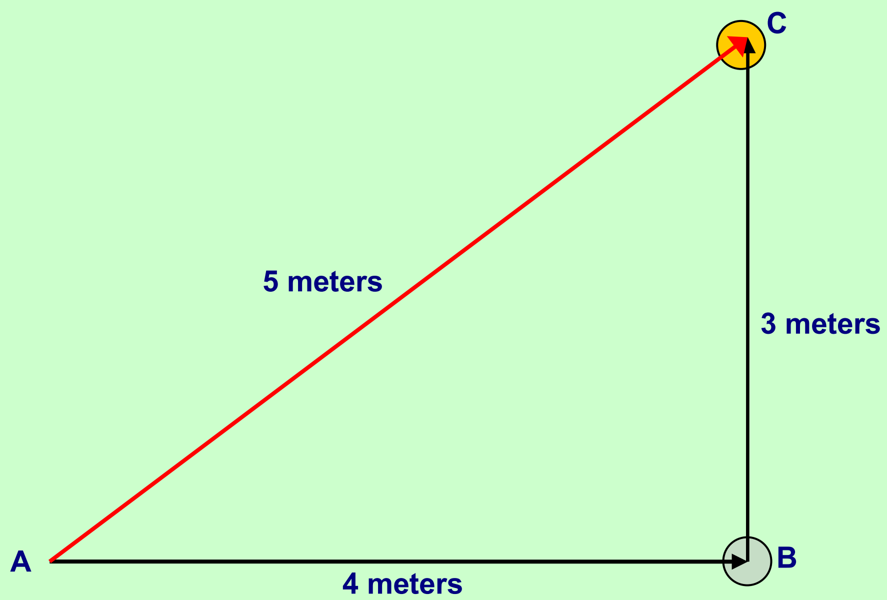
To convey the *direction* as well as the speed, physicists make a distinction between two words that mean the same thing in everyday life: *speed* and *velocity*.

Speed, in physics jargon, keeps its ordinary meaning—it is simply a measure of *how fast something's moving*, and gives *no clue about* which *direction* it's moving in.

Velocity, on the other hand, in physics jargon *includes direction*. For motion along a straight line, velocity can be positive or negative. For a given situation, such as Charlottesville to Richmond, we have to agree beforehand that one particular direction, such as away from Charlottesville, counts as positive, so motion towards Charlottesville would then always be at a *negative* velocity (but, of course, a positive *speed*, since speed is always positive, or zero).

17.2 Uniform Motion in a Plane

Now think about how you would describe quantitatively the motion of a smooth ball rolling steadily on a flat smooth tabletop (so frictional effects are negligible, and we can take the speed to be constant). Obviously, the first thing to specify is the speed—how fast is it moving, say in meters per second? But next, we have to tackle how to give its *direction* of motion, and just positive or negative won't do, since it could be moving at any angle to the table edge.



Vectors add like displacements:

if the ball rolls from A to C along the red line, its displacement from A can be represented by the red arrow.

The *identical* displacement can be achieved by first moving the ball to B, then up to C: *the two black vectors add to the red vector*.

One approach to describing uniform motion in the plane is a sort of simplified version of Galileo's "compound motion" analysis of projectiles. One can think of the motion of the ball rolling steadily across the table as being compounded of two motions, one a steady rolling parallel to the *length* of the table, the other a steady rolling parallel to the *width* of the table. For example, one could say that in its steady motion the ball is proceeding at a steady four meters per second along the length of the table, and, at the same time, it is proceeding at a speed of three meters per second parallel to the width of the table (this is a big table!). To visualize what this means, think about where the ball is at some instant, then where it is one second later. It will have moved four meters along the length of the table, and also three meters along the width. How far did it actually move? And in what direction?

We can see that if the ball's uniform motion is compounded of a steady velocity of 4 meters per second parallel to the length of the table and a steady velocity of 3 meters per second parallel to the width, as shown above, the actual distance the ball moves in one second is 5 meters (remembering Pythagoras' theorem, and in particular that a right angled triangle with the two shorter sides 3 and 4 has the longest side 5—we chose these numbers to make it easy). That is to say, the *speed* of the ball is 5 meters per second.

What, exactly, is its *velocity*? As stated above, the velocity includes both *speed* and *direction* of motion. *The simplest and most natural way to represent direction is with an arrow*. So, we represent velocity by drawing an arrow in the plane indicating the direction the ball is rolling in. We can see on the above representation of a table that this is the direction of the slanting line arrow, which showed from where to where the ball moved in one second, obviously in the direction of its velocity. Hence, we represent the direction of the velocity by drawing an arrow pointing in that direction.

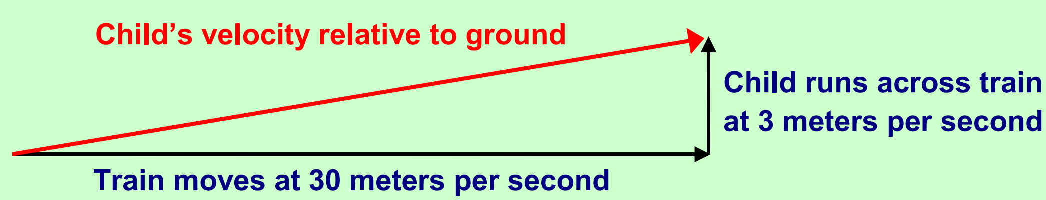
We can make the arrow represent the speed, as well, by agreeing on a rule for its length, such as an arrow 1 cm long corresponds to a speed of 1 meter per second, one 2 cm long represents 2 meters per second, etc. These arrows are usually called *vectors*.

Let us agree that we represent velocities for the moment by arrows pointing in the direction of motion, and an arrow 2 cm long corresponds to a speed of 1 meter per second. Then the velocity of the ball, which is 5 meters per second in the direction of the slanting arrow above, is in fact represented quantitatively by that arrow, since it has the right length—10 cms. Recalling that we began by saying the ball had a velocity 4 meters per second parallel to the length of the table, and 3 meters per second parallel to the width, we notice from the figure that these individual velocities, which have to be added together to give the total velocity, can themselves be represented by arrows (or vectors), and, in fact, are represented by the horizontal and vertical arrows in the figure. All we are saying here is that the arrows showing how far the ball moves in a given direction in one second also represent its velocity in that direction, because for uniform motion velocity just means how far something moves in one second.

The total velocity of 5 meters per second in the direction of the dashed arrow can then be thought of as the sum of the two velocities of 4 meters per second parallel to the length and 3 meters per second parallel to the width. Of course, the *speeds* don't add. Staring at the figure, we see *the way to add these vectors is to place the tail of one of them at the head of the other, then the sum is given by the vector from the other tail to the other head*. In other words, putting the two vectors together to form two sides of a triangle with the arrows pointing around the triangle the same way, the sum of them is represented by the third side of the triangle, but with the arrow pointing the other way.

17.3 Relative Velocities: a Child Running in a Train

As we shall see, relative velocities play an important role in relativity, so it is important to have a clear understanding of this concept. As an example, consider a child running at 3 meters per second (about 6 mph) in a train. The child is running parallel to the length of the train, towards the front, and the train is moving down the track at 30 meters per second. What is the child's velocity relative to the ground? It is 33 meters per second in the direction the train is moving along the track (notice we always specify *direction* for a velocity). To really nail this down, you should think through just how far the child moves relative to the ground in one second—three meters closer to the front of the train, and the train has covered 30 meters of ground. A trickier point arises if the child is running *across* the train, from one side to the other. (This run will only last about one second!) Again, the way to find the child's velocity relative to the ground is to visualize how much ground the child covers in one second—three meters in the direction across the track, from one side to the other, plus thirty meters in the direction along the track.

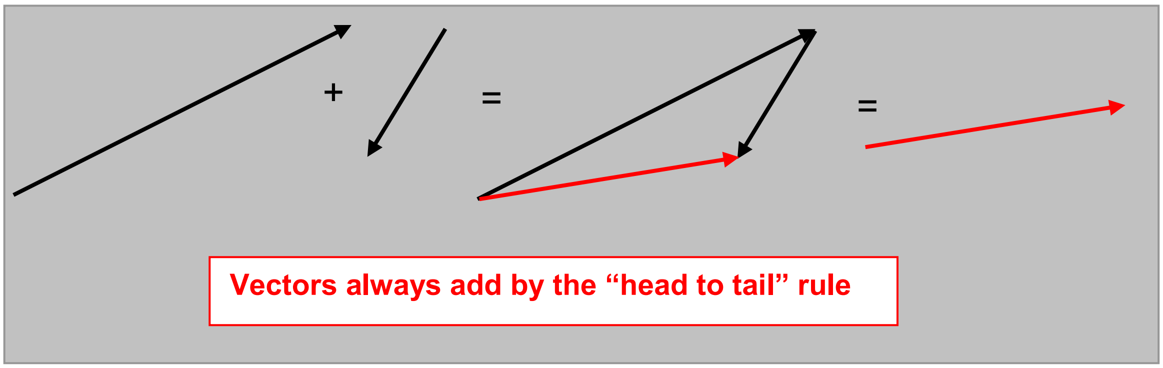


Velocity vectors add just like displacements:

Two vectors are added by putting the tail of the second at the head of the first, the *sum* is then *the vector from the tail of the first to the head of the second*. To see this makes sense for velocities, think through how far the child running across the train has moved *relative to the ground* in one second.

To find the total velocity, we now have to add two velocities at right angles, using the "head to tail" rule for adding vectors. This is just the same problem as the ball rolling across the table at an angle discussed above, and we need to use Pythagoras' theorem to find the child's *speed* relative to the ground.

Here is another example of vector addition, this time the two vectors to be added are not perpendicular to each other, but the same rules apply:



So in the diagram above, the two vectors on the left add to give the vector on the right. To get a bit less abstract, this could represent relative velocity in the following way: the big arrow on the left might be the speed at which a person is swimming relative to water in a river, the little arrow is the velocity at which the river water is moving over the river bed. then the vector sum of these two represents the velocity of our swimmer relative to the river bed, which is what counts for actually getting somewhere!

Exercise: Suppose you are swimming upstream at a speed relative to the water exactly equal to the rate the water is flowing downstream, so you're staying over the same spot on the river bed. Draw vectors representing your velocity relative to the water, the water's velocity relative to the river bed, and your velocity relative to the river bed. From this trivial example, if I draw a vector **A**, you can immediately draw **-A**, the vector which when added to **A** (using the rule for vector addition stated above) gives zero.

17.4 Aristotle's Law of Horizontal Motion

We restrict our considerations here to an object, such as an oxcart, moving in a horizontal plane. Aristotle would say (with some justification) that it moves in the direction it's being pushed (or pulled), and with a speed proportional to the force being applied. Let us think about that in terms of vectors. He is saying that the magnitude of the velocity of the object is proportional to the applied force, and the direction of the velocity is the direction of the applied force. It seems natural to conclude that not only is the velocity a vector, but so is the applied force! The applied force certainly has magnitude (how hard are we pushing?) and direction, and can be represented by an arrow (we would have to figure out some units of force if we want the length to represent force quantitatively—we will come back to this later). But that isn't quite the whole story—an essential property of vectors is that you can add them to each other, head to tail, as described above. But if you have two forces acting on a body, is their total effect equivalent to that of a force represented by adding together two arrows representing the individual forces head to tail? It turns out that if the two forces act at the same point, the answer is yes, but this is a fact about the physical world, and needs to be established experimentally. (It is not true in the subnuclear world, where the forces of attraction between protons and neutrons in a nucleus are affected by the presence of the other particles.)

So **Aristotle's rule for horizontal motion** is: *velocity is proportional to applied force*.

This rule seems to work well for oxcarts, but doesn't make much sense for our ball rolling across a smooth table, where, after the initial shove, there is *no* applied force in the direction of motion.

17.5 Galileo's Law of Horizontal Motion

Galileo's Law of Horizontal Motion can be deduced from his statement near the beginning of Fourth day in Two New Sciences,

Imagine any particle projected along a horizontal plane without friction; then we know ... that this particle will move along this same plane with a motion which is uniform and perpetual, provided the plane has no limits.

So **Galileo's rule for horizontal motion** is: *velocity = constant*, provided *no force*, including friction, acts on the body.

The big advance from Aristotle here is Galileo's realization that *friction is an important part of what's going on*. He knows that if there were no friction, the ball would keep at a steady velocity. The reason Aristotle thought it was necessary to apply a force to maintain constant velocity was that he failed to identify the role of friction, and to realize that the force applied to maintain constant velocity was just balancing the frictional loss. In contrast, Galileo realized the friction acted as a drag force on the ball, and the external force necessary to maintain constant motion just balanced this frictional drag force, so there was no *total* horizontal force on the ball.

17.6 Galileo's Law of Vertical Motion

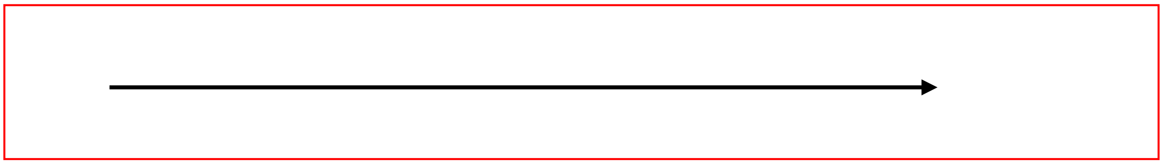
As we have already discussed at length, Galileo's Law of Vertical Motion is:

For **vertical motion**: acceleration = constant (neglecting air resistance, etc.)

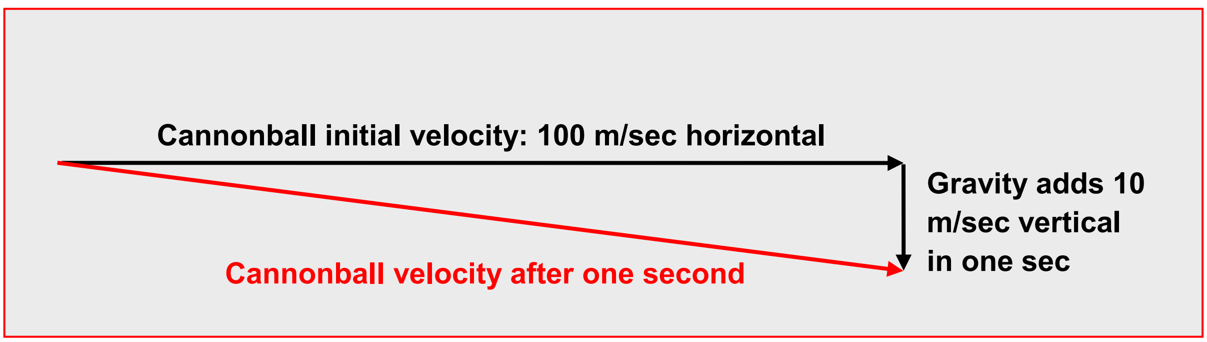
17.7 Describing Projectile Motion with Vectors

As an exercise in using vectors to represent velocities, consider the velocity of a cannonball shot horizontally at 100 meters per second from the top of a cliff: what is the velocity after 1, 2, 3 seconds? As usual, neglect air resistance.

The initial velocity is represented by a horizontal arrow, which we take to be 10 cms long, for convenience:



After one second, the downward velocity will have increased from zero to 10 meters per second, as usual for a falling body. Thus, to find the total velocity after one second, we need to add to the initial velocity, the vector above, a vertically downward vector of length 1 cm, to give the right scale:



It is worth noting that although the velocity has visibly changed in this first second, the speed has hardly changed at all—remember the speed is represented by the length of the slanting vector, which from Pythagoras' theorem is the square root of 101 cms long, or about 10.05 cms, a very tiny change. The velocity after two seconds would be given by adding two of the dashed downward arrows head-to-tail to the initial horizontal arrow, and so on, so that after ten seconds, if the cliff were high enough, the velocity would be pointing downwards at an angle of 45 degrees, and the speed by this point would have increased substantially.

17.8 Acceleration

Galileo defined naturally accelerated motion as downward motion in which speed increased at a steady rate, giving rise to units for acceleration that look like a misprint, such as 10 meters per second per second.

In everyday life, this is just what acceleration means—how fast something's picking up speed.

However, in physics jargon, acceleration (like velocity) has a more subtle meaning: *the acceleration of an object is its rate of change of velocity*. From now on, this is what we mean when we say acceleration.

At first this might seem to you a nitpicking change of definition—but it isn't. Remember velocity is a *vector*. *It can change without its length changing*—it could just swing around and point in a different direction. This means a body can accelerate *without* changing speed!

Why would we want to define acceleration in such a nonintuitive way? It almost seems as if we are trying to make things difficult! It turns out that our new definition is what Galileo might call the *natural* definition of acceleration. In the true laws of motion that describe things that happen in the universe, as we shall discuss below, if a body has a net force acting on it, it accelerates. But it doesn't necessarily change speed—it might just swing its velocity around, in other words veer off in a different direction. Therefore, as we shall see, this new definition of acceleration is what we need to describe the real world.

For motion in a straight line, our definition is the same as Galileo's—we agree, for example, that the acceleration of a falling body is 10 meters per second per second downwards.

NOTE: the next topics covered in the course are the contributions of two very colorful characters, Tycho Brahe and Johannes Kepler. I gave a more complete account of these two and their works in an earlier version of this course. If you would like to read the more complete (and more interesting) version, click on Tycho Brahe.

18 Tycho Brahe and Johannes Kepler

These two colorful characters made crucial contributions to our understanding of the universe: Tycho's observations were accurate enough for Kepler to discover that the planets moved in elliptic orbits, and his other laws, which gave Newton the clues he needed to establish universal inverse-square gravitation.

Tycho Brahe (1546-1601), from a rich Danish noble family, was fascinated by astronomy, but disappointed with the accuracy of tables of planetary motion at the time. He decided to dedicate his life and considerable resources to recording planetary positions ten times more accurately than the best previous work. After some early successes, and in gratitude for having his life saved by Tycho's uncle, the king of Denmark gave Tycho tremendous resources: an island with many families on it, and money to build an observatory. (One estimate is that this was 10% of the gross national product at the time!) Tycho built vast instruments to set accurate sights on the stars, and used multiple clocks and timekeepers.

He achieved his goal of measuring to one minute of arc. This was a tremendous feat before the invention of the telescope. His aim was to confirm his own picture of the universe, which was that the earth was at rest, the sun went around the earth and the planets all went around the sun - an intermediate picture between Ptolemy and Copernicus.

Johannes Kepler (1571-1630) believed in Copernicus' picture. Having been raised in the Greek geometric tradition, he believed God must have had some geometric reason for placing the six planets at the particular distances from the sun that they occupied. He thought of their orbits as being on spheres, one inside the other. One day, he suddenly remembered that there were just five perfect Platonic solids, and this gave a reason for there being six planets - the orbit spheres were maybe just such that between two successive ones a perfect solid would just fit. He convinced himself that, given the uncertainties of observation at the time, this picture might be the right one. However, that was before Tycho's results were used. Kepler realized that Tycho's work could settle the question one way or the other, so he went to work with Tycho in 1600. Tycho died the next year, Kepler stole the data, and worked with it for nine years.

He reluctantly concluded that his geometric scheme was wrong. In its place, he found his three laws of planetary motion:

I *The planets move in elliptical orbits with the sun at a focus.*

II *In their orbits around the sun, the planets sweep out equal areas in equal times.*

III *The squares of the times to complete one orbit are proportional to the cubes of the average distances from the sun.*

These are the laws that Newton was able to use to establish universal gravitation.

Kepler was the first to state clearly that the way to understand the motion of the planets was in terms of some kind of force from the sun. However, in contrast to Galileo, Kepler thought that a continuous force was necessary to maintain motion, so he visualized the force from the sun like a rotating spoke pushing the planet around its orbit.

On the other hand, Kepler did get right that the tides were caused by the moon's gravity. Galileo mocked him for this suggestion.

A much ***fuller treatment*** of Tycho Brahe and Johannes Kepler can be found in my earlier notes:

Links to: ***Tycho Brahe*** ***Kepler*** ***More Kepler***

19 Isaac Newton

19.1 Newton's Life

In 1642, the year Galileo died, Isaac Newton was born in Woolsthorpe, Lincolnshire, England on Christmas Day. His father had died three months earlier, and baby Isaac, very premature, was also not expected to survive. It was said he could be fitted into a quart pot. When Isaac was three, his mother married a wealthy elderly clergyman from the next village, and went to live there, leaving Isaac behind with his grandmother. The clergyman died, and Isaac's mother came back, after eight years, bringing with her three small children. Two years later, Newton went away to the Grammar School in Grantham, where he lodged with the local apothecary, and was fascinated by the chemicals. The plan was that at age seventeen he would come home and look after the farm. He turned out to be a total failure as a farmer.

His mother's brother, a clergyman who had been an undergraduate at Cambridge, persuaded his mother that it would be better for Isaac to go to university, so in 1661 he went up to Trinity College, Cambridge. Isaac paid his way through college for the first three years by waiting tables and cleaning rooms for the fellows (faculty) and the wealthier students. In 1664, he was elected a scholar, guaranteeing four years of financial support. Unfortunately, at that time the plague was spreading across Europe, and reached Cambridge in the summer of 1665. The university closed, and Newton returned home, where he spent two years concentrating on problems in mathematics and physics. He wrote later that during this time he first understood the theory of gravitation, which we shall discuss below, and the theory of optics (he was the first to realize that white light is made up of the colors of the rainbow), and much mathematics, both integral and differential calculus and infinite series. However, he was always reluctant to publish anything, at least until it appeared someone else might get credit for what he had found earlier.

On returning to Cambridge in 1667, he began to work on alchemy, but then in 1668 Nicolas Mercator published a book containing some methods for dealing with infinite series. Newton immediately wrote a treatise, *De Analysi*, expounding his own wider ranging results. His friend and mentor Isaac Barrow communicated these discoveries to a London mathematician, but only after some weeks would Newton allow his name to be given. This brought his work to the attention of the mathematics community for the first time. Shortly afterwards, Barrow resigned his Lucasian Professorship (which had been established only in 1663, with Barrow the first incumbent) at Cambridge so that Newton could have the Chair.

Newton's first major *public* scientific achievement was the invention, design and construction of a reflecting telescope. He ground the mirror, built the tube, and even made his own tools for the job. This was a real advance in telescope technology, and ensured his election to membership in the Royal Society. The mirror gave a sharper image than was possible with a large lens because a lens focusses different colors at slightly different distances, an effect called *chromatic aberration*. This problem is minimized nowadays by using compound lenses, two lenses of different kinds of glass stuck together, that err in opposite directions, and thus tend to cancel each other's shortcomings, but mirrors are still used in large telescopes.

Later in the 1670's, Newton became very interested in theology. He studied Hebrew scholarship and ancient and modern theologians at great length, and became convinced that Christianity had departed from the original teachings of Christ. He felt unable to accept the current beliefs of the Church of England, which was unfortunate because he was required as a Fellow of Trinity College to take holy orders. Happily, the Church of England was more flexible than Galileo had found the Catholic Church in these matters, and King Charles II issued a royal decree excusing Newton from the necessity of taking holy orders! Actually, to prevent this being a wide precedent, the decree specified that, in perpetuity, the Lucasian professor need not take holy orders. (The current Lucasian professor is Stephen Hawking.)

In 1684, three members of the Royal Society, Sir Christopher Wren, Robert Hooke and Edmond Halley, argued as to whether the elliptical orbits of the planets could result from a gravitational force towards the sun proportional to the inverse square of the distance. Halley writes:

Mr. Hook said he had had it, but that he would conceal it for some time so that others, triing and failing might know how to value it, when he should make it publick.

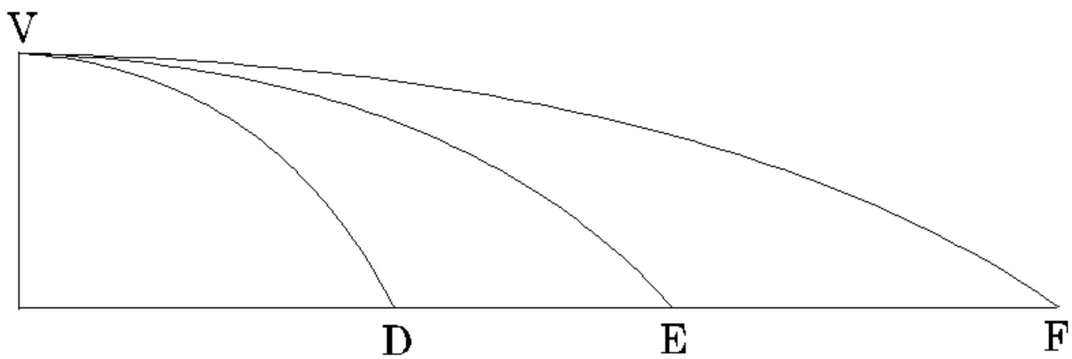
Halley went up to Cambridge, and put the problem to Newton, who said he had solved it four years earlier, but couldn't find the proof among his papers. Three months later, he sent an improved version of the proof to Halley, and devoted himself full time to developing these ideas, culminating in the publication of the *Principia* in 1686. This was the book that really did change man's view of the universe, as we shall shortly discuss, and its importance was fully appreciated very quickly. Newton became a public figure. He left Cambridge for London, where he was appointed Master of the Mint, a role he pursued energetically, as always, including prosecuting counterfeiters. He was knighted by Queen Anne. He argued with Hooke about who deserved credit for discovering the connection between elliptical orbits and the inverse square law until Hooke died in 1703, and he argued with a German mathematician and philosopher, Leibniz, about which of them invented calculus. Newton died in 1727, and was buried with much pomp and circumstance in Westminster Abbey—despite his well-known reservations about the Anglican faith.

An excellent, readable book is *The Life of Isaac Newton*, by Richard Westfall, Cambridge 1993, which I used in writing the above summary of Newton's life.

A fascinating collection of articles, profusely illustrated, on Newton's life, work and impact on the general culture is *Let Newton Be!*, edited by John Fauvel and others, Oxford 1988, which I also consulted.

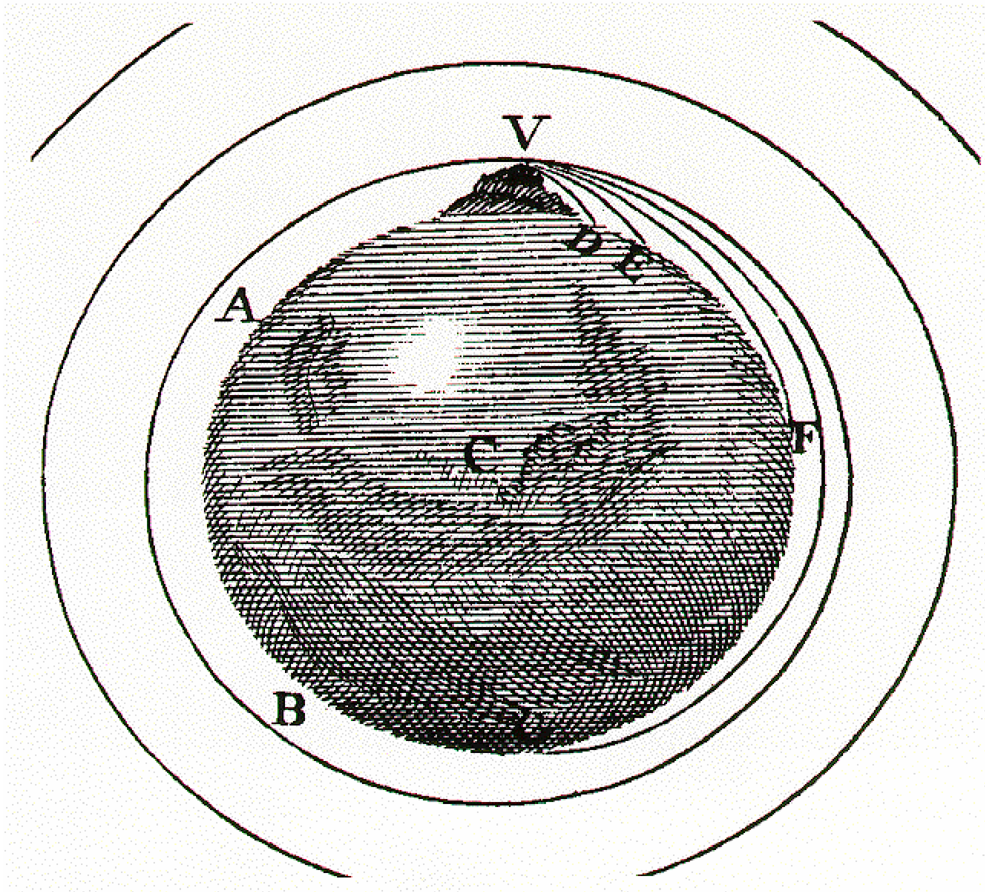
19.2 Projectiles and Planets

Let us now turn to the central topic of the *Principia*, the universality of the gravitational force. The legend is that Newton saw an apple fall in his garden in Lincolnshire, thought of it in terms of an attractive gravitational force towards the earth, and realized the same force might extend as far as the moon. He was familiar with Galileo's work on projectiles, and suggested that the moon's motion in orbit could be understood as a natural extension of that theory. To see what is meant by this, consider a gun shooting a projectile horizontally from a very high mountain, and imagine using more and more powder in successive shots to drive the projectile faster and faster.



The parabolic paths would become flatter and flatter, and, if we imagine that the mountain is so high that air resistance can be ignored, and the gun is sufficiently powerful, *eventually the point of landing is so far away that we must consider the curvature of the earth in finding where it lands.*

In fact, the real situation is more dramatic—the earth's curvature may mean the projectile *never lands at all.* This was envisioned by Newton in the *Principia*. The following diagram is from his later popularization, *A Treatise of the System of the World*, written in the 1680's:



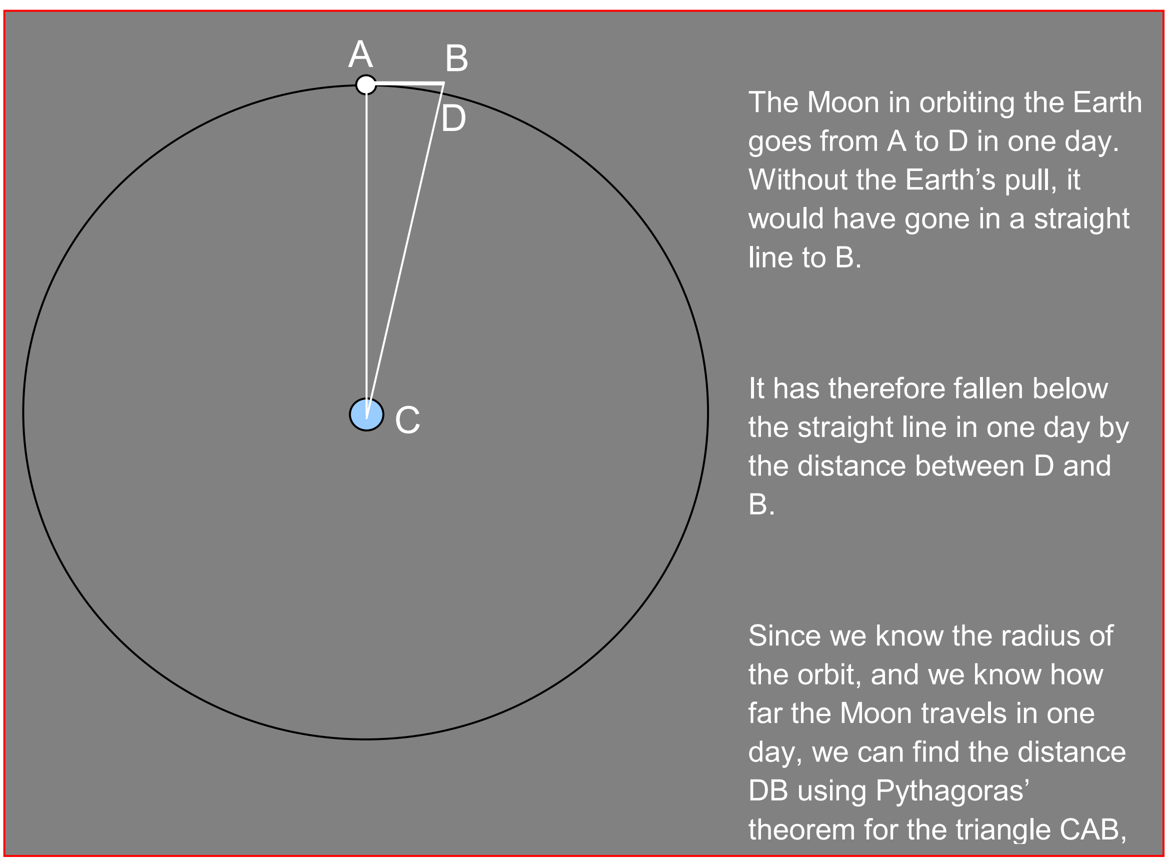
The mountaintop at V is supposed to be above the earth's atmosphere, and for a suitable initial speed, the projectile orbits the earth in a circular path. In fact, the earth's curvature is such that the surface falls away below a truly flat horizontal line by about five meters in 8,000 meters (five miles). Recall that five meters is just the vertical distance an initially horizontally moving projectile will fall in the first second of motion. But this implies that if the (horizontal) muzzle velocity were 8,000 meters per second, the downward fall of the cannonball would be just matched by the earth's surface falling away, and it would never hit the ground! This is just the motion, familiar to us now, of a satellite in a low orbit, which travels at about 8,000 meters (five miles) a second, or 18,000 miles per hour. (Actually, Newton drew this mountain impossibly high, no doubt for clarity of illustration. A satellite launched horizontally from the top would be far above the usual shuttle orbit, and go considerably more slowly than 18,000 miles per hour.) For an animated version of Newton's cannon on a mountain, click here!

19.3 The Moon is Falling

Newton realized that the moon's circular path around the earth could be caused in this way by the same gravitational force that would hold such a cannonball in low orbit, in other words, the same force that causes bodies to fall.

To think about this idea, let us consider the moon's motion, beginning at some particular instant, as deviating downwards—falling—from some initial "horizontal" line, just as for the cannonball shot horizontally from a high mountain. The first obvious question is: does the moon fall five meters below the horizontal line, that is, towards the earth, in the first second? This was not difficult for Newton to check, because the path of the moon was precisely known by this time. The moon's orbit is approximately a circle of radius about 384,000 kilometers (240,000 miles), which it goes around in a month (to be precise, in 27.3 days), so the distance covered in one second is, conveniently, very close to one kilometer. It is then a matter of geometry to figure out how far the curved path falls below a "horizontal" line in one second of flight, and the answer turns out to be not five meters, but only a little over one *millimeter*! (Actually around 1.37 millimeters.)

It's completely impossible to draw a diagram showing how far it falls in one second, but the geometry is the same if we look how far it falls in one *day*, so here it is:



For one *second*, AB would be only one kilometer, so since AC is 384,000 km., the triangle ABC is *really* thin, but we can still use Pythagoras' theorem!

Thus the "natural acceleration" of the moon towards the earth, measured by how far it falls below straight line motion in one second, is less than that of an apple here on earth by the ratio of five meters to 1.37 millimeters, which works out to be about 3,600.

What can be the significance of this much smaller rate of fall? Newton's answer was that the natural acceleration of the moon was much smaller than that of the cannonball because they were both caused by a *force*—a *gravitational attraction* towards the earth, and that *the gravitational force became weaker on going away from the earth*.

In fact, the figures we have given about the moon's orbit enable us to compute how fast the gravitational attraction dies away with distance. The distance from the center of the earth to the earth's surface is about 6,350 kilometers (4,000 miles), so the moon is about 60 times further from the center of the earth than we and the cannonball are.

From our discussion of how fast the moon falls below a straight line in one second in its orbit, we found that the gravitational acceleration for the moon is down by a factor of 3,600 from the cannonball's (or the apple's).

Putting these two facts together, and noting that 3,600 = 60 x 60, led Newton to his famous *inverse square law*: *the force of gravitational attraction between two bodies decreases with increasing distance between them as* ***the inverse of the square of that distance****, so if the distance is doubled, the force is down by a factor of four.*

20 How Newton built on Galileo's Ideas

20.1 Newton's Laws

We are now ready to move on to Newton's Laws of Motion, which for the first time presented a completely coherent analysis of motion, making clear that the motion in the heavens could be understood in the same terms as motion of ordinary objects here on earth.

20.2 Acceleration Again

The crucial Second Law, as we shall see below, links the *acceleration* of a body with the force acting on the body. To understand what it says, it is necessary to be completely clear what is meant by acceleration, so let us briefly review.

Speed is just how fast something's moving, so is fully specified by a positive number and suitable units, such as 55 mph or 10 meters per second.

Velocity, on the other hand, means to a scientist more than speed---it also *includes* a specification of the *direction* of the motion, so 55 mph *to the northwest* is a velocity. Usually wind *velocities* are given in a weather forecast, since the direction of the wind affects future temperature changes in a direct way. The standard way of representing a velocity in physics is with an arrow pointing in the appropriate direction, its length representing the speed in suitable units. These arrows are called "*vectors*".

(WARNING: Notice, though, that for a moving object such as a projectile, *both* its position at a given time (compared with where it started) *and* its velocity at that time can be represented by vectors, so you must be clear what your arrow represents!)

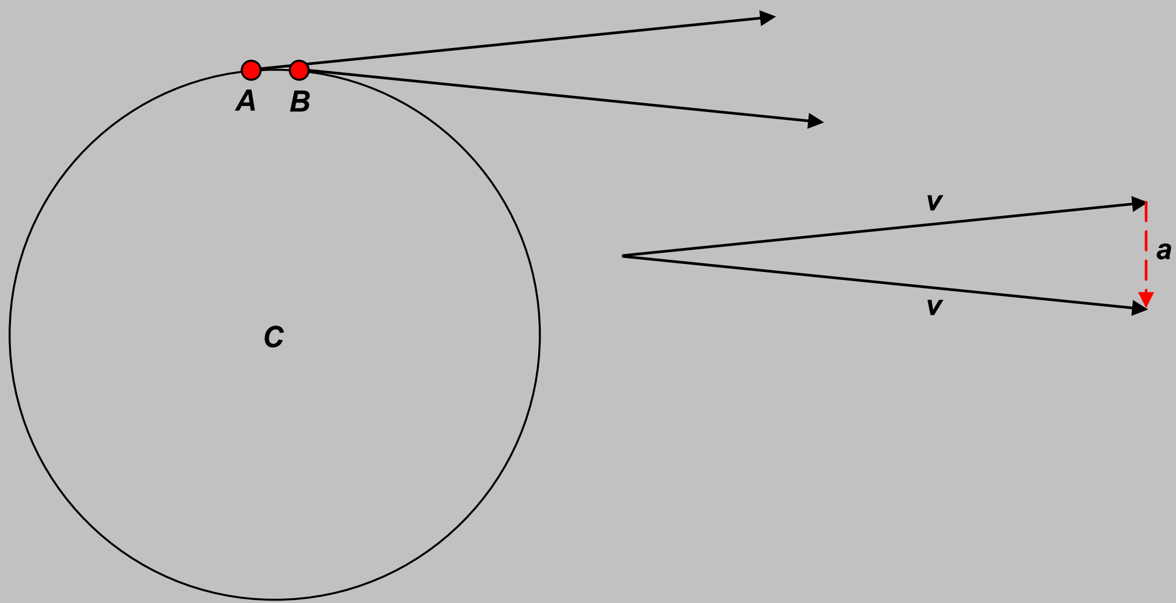
Acceleration: as we have stated, acceleration is defined as *rate of change of velocity*.

It is *not* defined as rate of change of speed. *A body can have nonzero acceleration while moving at constant speed!*

20.3 An Accelerating Body that isn't Changing Speed

Consider Newton's cannon on an imaginary high mountain above the atmosphere, that shoots a ball so fast it circles the earth at a steady *speed. Of course, its velocity is changing constantly, because velocity includes direction.*

Let us look at how its *velocity* changes over a period of one second. (Actually, in the diagram below we exaggerate how far it would move in one second, the distance would in fact be one-five thousandth of the distance around the circle, impossible to draw.)



The velocity vectors (length *v*) for a ball moving at steady speed in a circle are shown one second apart. To find the acceleration—the difference between these two vectors—we must put their tails together and draw a new vector (the red dotted one) from the head of the first to the head of the second. This new vector (length *a*) is the difference: what must be added to the first to give the second.

Here we show the cannonball (greatly exaggerated in size!) at two points in its orbit, and the velocity vectors at those points. On the right, we show the two velocity vectors again, but we put their ends together so that we can see the difference between them, which is the small dashed vector.

In other words, the small dashed vector is the velocity that has to be added to the first velocity to get the second velocity: it is the *change* in velocity on going around that bit of the orbit.

Now, if we think of the two points in the orbit as corresponding to positions of the cannonball one second apart, the small dashed vector will represent the change in velocity in one second, and that is—by definition—the acceleration. The acceleration is the rate of change of velocity, and that is how much the velocity changes in one second (for motions that change reasonably smoothly over the one-second period, which is certainly the case here. To find the rate of change of velocity of a fly's wing at some instant, we obviously would have to measure its velocity change over some shorter interval, maybe a thousandth of a second).

So we see that, with our definition of acceleration as the rate of change of velocity, which is a vector, a body moving at a steady speed around a circle is accelerating towards the center all the time, although it never gets any closer to it. If this thought makes you uncomfortable, it is because you are still thinking that acceleration must mean a change of speed, and just changing direction doesn't count.

20.4 Finding the Acceleration in Circular Motion

It is possible to find an explicit expression for the magnitude of the acceleration towards the center (sometimes called the *centripetal* acceleration) for a body moving on a circular path at speed *v*. Look again at the diagram above showing two values of the velocity of the cannonball one second apart. As is explained above, the magnitude *a* of the acceleration is the length of the small dashed vector on the right, where the other two sides of this long narrow triangle have lengths equal to the speed *v* of the cannonball. We'll call this the "*vav*" triangle, because those are the lengths of its sides. What about the *angle* between the two long sides? That is just the angle the velocity vector turns through in one second as the cannonball moves around its orbit. Now look over at the circle diagram on the left showing the cannonball's path. Label the cannonball's position at the beginning of the second *A*, and at the end of the second *B*, so the length *AB* is how far the cannonball travels in one second, that is, *v*. (It's true that the part of the path *AB* is slightly curved, but we can safely ignore that very tiny effect.) Call the center of the circle *C*. Draw the triangle *ACB*. (The reader should sketch the figure and actually draw these triangles!) The two long sides *AC* and *BC* have lengths equal to the radius of the circular orbit. We could call this long thin triangle an "*rvr*" triangle, since those are the lengths of its sides.

The important point to realize now is that the "*vav*" triangle and the "*rvr*" triangle are *similar*, because since the velocity vector is always perpendicular to the radius line from the center of the circle to the point where the cannonball is in orbit, *the angle the velocity vector rotates by in one second is the same as the angle the radius line turns through in one second*. Therefore, the two triangles are similar, and their corresponding sides are in the same ratios, that is, $a/v = v/r$. It follows immediately that *the magnitude of the acceleration a for an object moving at steady speed v in a circle of radius r is* v^2/r *directed towards the center of the circle.*

This result is true for all circular motions, even those where the moving body goes round a large part of the circle in one second. To establish it in a case like that, recall that the acceleration is the rate of change of velocity, and we would have to pick a smaller time interval than one second, so that the body didn't move far around the circle in the time chosen. If, for example, we looked at two velocity vectors one-hundredth of a second apart, and they were pretty close, then the acceleration would be given by the difference vector between them *multiplied by one-hundred*, since acceleration is defined as what the velocity change in one second would be if it continued to change at that rate. (In the circular motion situation, the acceleration is of course changing all the time. To see why it is sometimes necessary to pick small time intervals, consider what would happen if the body goes around the circle completely in one second. Then, if you pick two times one second apart, you would conclude the velocity isn't changing at all, so there is no acceleration.)

20.5 An Accelerating Body that isn't Moving

We've stated before that a ball thrown vertically upwards has constant *downward* acceleration of 10 meters per second in each second, even when it's at the very top and isn't moving at all. The key point here is that acceleration is rate of change of velocity. You can't tell what the rate of change of something is unless you know its value at more than one time. For example, speed on a straight road is rate of change of distance from some given point. You can't get a speeding ticket just for being at a particular point at a certain time—the cop has to prove that a short time later you were at a point well removed from the first point, say, three meters away after one-tenth of a second. That would establish that your speed was thirty meters per second, which is illegal in a 55 m.p.h. zone. In just the same way that speed is rate of change of position, acceleration is rate of change of velocity. Thus to find acceleration, you need to know velocity at two different times. The ball thrown vertically upwards does have zero velocity at the top of its path, but that is only at a *single instant* of time. One second later it is dropping at ten meters per second. One millionth of a second after it reached the top, it is falling at one hundred-thousandth of a meter per second. Both of these facts correspond to a downward acceleration, or rate of change of velocity, of 10 meters per second in each second. *It would only have zero acceleration if it stayed at rest at the top for some finite period of time*, so that you could say that its velocity remained the same—zero—for, say, a thousandth of a second, and during that period the rate of change of velocity, the acceleration, would then of course be zero. Part of the problem is that the speed is very small near the top, and also that our eyes tend to lock on to a moving object to see it better, so there is the illusion that it comes to rest and stays there, even if not for long.

20.6 Galileo's Analysis of Motion: Two Kinds

Galileo's analysis of projectile motion was based on two concepts:

1. *Naturally accelerated motion*, describing the *vertical* component of motion, in which the body picks up speed at a uniform rate.

2. *Natural horizontal motion*, which is motion at a *steady speed in a straight line*, and happens to a ball rolling across a smooth table, for example, when frictional forces from surface or air can be ignored.

20.7 Newton Puts Them Together

Newton's major breakthrough was to show that these two different kinds of motion can be thought of as *different aspects of the same thing*. He did this by introducing the idea of motion being affected by a *force*, then expressing this idea in a quantitative way. Galileo, of course, had been well aware that motion is affected by external forces. Indeed, his definition of natural horizontal motion explicitly states that it applies to the situation where such forces can be neglected. He knew that friction would ultimately slow the ball down, and—very important—a force pushing it from behind would cause it to accelerate. What he didn't say, though, and Newton did, was that just as a force would cause acceleration in horizontal motion, the natural acceleration actually observed in vertical motion *must* be the result of a vertical force on the body, without which the natural vertical motion would *also* be at a constant speed, just like natural horizontal motion. This vertical force is of course just the force of gravity.

20.8 Force is the Key

Therefore the point Newton is making is that the essential difference between Galileo's natural steady speed horizontal motion and the natural accelerated vertical motion is that vertically, there is always the force of gravity acting, and without that—for example far into space—the natural motion (that is, with no forces acting) in *any* direction would be at a steady speed in a straight line.

(Actually, it took Newton some time to clarify the concept of force, which had previously been unclear. This is discussed at length in *Never at Rest*, by Richard Westfall, and I have summarized some of the points here.)

20.9 Newton's First Law: no Force, no Change in Motion

To put it in his own words (although actually he wrote it in Latin, this is from an 1803 translation):

Law 1

Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed thereon.

He immediately adds, tying this in precisely with Galileo's work:

Projectiles persevere in their motions, so far as they are not retarded by the resistance of the air, or impelled downwards by the force of gravity.

Notice that here “persevere in their motions” must mean in *steady speed straight line* motions, because he is adding the gravitational acceleration on to this.

This is sometimes called “The Law of Inertia”: in the absence of an external force, a body in motion will continue to move at constant speed and direction, that is, at constant velocity.

So *any* acceleration, or change in speed (or direction of motion), of a body *signals that it is being acted on by some force*.

20.10 Newton’s Second Law: Acceleration of a Body is Proportional to Force

Newton’s next assertion, based on much experiment and observation, is that, *for a given body*, the acceleration produced is proportional to the strength of the external force, so doubling the external force will cause the body to pick up speed twice as fast.

Law 2

The alteration of motion is ever proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.

20.11 What About Same Force, Different Bodies?

Another rather obvious point he doesn’t bother to make is that for a given *force*, such as, for example, the hardest you can push, applied to two different objects, say a wooden ball and a lead ball of the same size, with the lead ball weighing seven times as much as the wooden ball, then the lead ball will only pick up speed at one-seventh the rate the wooden one will.

20.12 Falling Bodies One More Time: What is Mass?

Now let us consider the significance of this law for falling bodies. Neglecting air resistance, bodies of all masses accelerate downwards at the same rate. This was Galileo’s discovery.

Let us put this well established fact together with Newton’s Second Law: the acceleration is *proportional* to the external force, but *inversely proportional* to the *mass* of the body the force acts on.

Consider two falling bodies, one having *twice* the mass of the other. Since their acceleration is the same, the body having twice the mass must be experiencing a gravitational force which is twice as strong. Of course, we are well aware of this, all it’s saying is that two bricks weigh twice as much as one brick. Any weight measuring device, such as a bathroom scales, is just measuring the force of gravity. However, this proportionality of mass and weight is not a completely trivial point. Masses can be measured against each other *without using gravity at all*, for example far into space, by comparing their relative accelerations when subject to a standard force, a push. If one object accelerates at half the rate of another when subject to our standard push, we conclude it has twice the mass. Thinking of the mass in this way as a measure of resistance to having velocity changed by an outside force, Newton called it *inertia*. (Note that this is a bit different from everyday speech, where we think of inertia as being displayed by something that stays at rest. For Newton, steady motion in a straight line is the same as being at rest. That seems perhaps counterintuitive, but that's because in ordinary life, steady motion in a straight line usually causes some frictional or resistive forces to come into play).

20.13 Mass and Weight

To return to the concept of mass, it is really just a measure of the *amount of stuff*. For a uniform material, such as water, or a uniform solid, the mass is the volume multiplied by the density—the density being defined as the mass of a unit of volume, so water, for example, has a density of one gram per cubic centimeter, or sixty-two pounds per cubic foot.

Hence, from Galileo's discovery of the uniform acceleration of all falling bodies, we conclude that the *weight* of a body, which is the gravitational attraction it feels towards the earth, is directly proportional to its mass, the amount of stuff it's made of.

20.14 The Unit of Force

All the statements above about force, mass and acceleration are statements about proportionality. We have said that for a body being accelerated by a force acting on it the acceleration is proportional to the (total) external force acting on the body, and, for a given force, inversely proportional to the mass of the body.

If we denote the force, mass and acceleration by F, m and a respectively (bearing in mind that really F and a are vectors pointing in the same direction) we could write this:

F is proportional to ma

To make any progress in applying Newton's Laws in a real situation, we need to choose some unit for measuring forces. We have already chosen units for mass (the kilogram) and acceleration (meters per second per second). The most natural way to define our unit of force is:

The unit of force is that force which causes a unit mass (one kilogram) to accelerate with unit acceleration (one meter per second per second).

This unit of force is named, appropriately, the ***newton***.

If we now agree to measure forces in newtons, the statement of proportionality above can be written as a simple equation:

$$F = ma$$

which is the usual statement of Newton's Second Law.

If a mass is now observed to accelerate, it is a trivial matter to find the total force acting on it. The force will be in the direction of the acceleration, and its magnitude will be the product of the mass and acceleration, measured in newtons. For example, a 3 kilogram falling body, accelerating downwards at 10 meters per second per second, is being acted on by a force *ma* equal to 30 newtons, which is, of course, its weight.

20.15 Newton's Third Law: Action and Reaction

Having established that a force—the action of another body—was necessary to cause a body to change its state of motion, Newton made one further crucial observation: such forces *always* arise as a *mutual interaction* of two bodies, and the other body also feels the force, but in the opposite direction.

Law 3

To every action there is always opposed an equal and opposite reaction*: or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.*

Newton goes on:

Whatever draws or presses another is as much drawn or pressed by that other. If you press a stone with your finger, the finger is also pressed by the stone. If a horse draws a stone tied to a rope, the horse (if I may so say) will be equally drawn back towards the stone: for the distended rope, by the same endeavour to relax or unbend itself, will draw the horse as much towards the stone, as it does the stone towards the horse, and will obstruct the progress of the one as much as it advances that of the other. If a body impinge upon another, and by its force change the motion of the other, that body also (because of the equality of the mutual pressure) will undergo an equal change, in its own motion, towards the contrary part. The changes made by these actions are equal, not in the velocities but in the motions of bodies; that is to say, if the bodies are not hindered by any other impediments. For, because the motions are equally changed, the changes of the velocities made towards contrary parts are reciprocally proportional to the bodies. This law takes place also in attractions.

All this maybe sounds kind of obvious. Anyone who's had a dog on a leash, especially a big dog, is well aware that tension in a rope pulls both ways. If you push against a wall, the wall is pushing you back. If that's difficult to visualize, imagine what would happen if the wall suddenly evaporated. Newton's insight here, his realization that every acting force has a reacting force, and that acceleration of a body only occurs when an *external* force acts on it, was one of the big forward steps in our understanding of how the Universe works.

20.16 Newton's Second Law in Everyday Life

The Second Law states that if a body is accelerating, there must be an external force acting on it. It's not always obvious what this external force is even in the most trivial everyday occurrences. Suppose you're standing still, then begin to walk. What was the external force that caused you to accelerate? Think about that for a while. Here's a clue: it's very hard to start walking if you're wearing smooth-bottomed shoes and standing on smooth ice. You tend to skid around in the same place. If you understand that, you also know what external force operates when a car accelerates.

The reason the external force causing the acceleration may not be immediately evident is that it may not be what's doing the work. Consider the following scenario: you are standing on level ground, on rollerskates, facing a wall with your palms pressed against it. You push against the wall, and roll away backwards. You accelerated. Clearly, you did the work that caused the acceleration. But from Newton's second law, your acceleration was, in fact, caused by the reactive external force of the wall pushing your hands, and hence the rest of you. That is to say, the force causing the acceleration may not be generated directly by what—or who—is doing the work! In this example, it's generated indirectly, as a reaction force to that of the hands pushing on the wall. But if the wall were on wheels, and it accelerated away when you pushed (having taken off your roller skates) the force causing the acceleration of the wall *would* be generated directly by the agent doing the work, you.

Now imagine two people on roller skates, standing close facing each other, palms raised and pushing the other person away. According to Newton's discussion above following his Third Law, the two bodies involved will undergo equal changes of motion, but to contrary parts, that is, in opposite directions. That sounds reasonable. They obviously both move off backwards. Notice, however, that Newton makes a special point of the fact that these equal (but opposite) "motions" do not imply equal (but opposite) velocities—this becomes obvious when you imagine the experiment with a 100 pound person and a 200 pound person. Newton tells us that in that situation the heavier person will roll backwards at half the speed—notice he says the velocities are "*reciprocally proportional to the bodies*".

Roller skates actually provide a pretty good example of the necessity of generating an external force if you want to accelerate. If you keep the skates pointing strictly forwards, and only the wheels are in contact with the ground, it's difficult to get going. The way you start is to turn the skates some, so that there is some sideways push on the wheels. Since the wheels can't turn sideways, you are thus able to push against the ground, and therefore it is pushing you—you've managed to generate the necessary external force to accelerate you. Note that if the wheels were to be replaced by ball bearings somehow, you wouldn't get anywhere, unless you provided some other way for the ground to push you, such as a ski pole, or maybe twisting your foot so that some fixed part of the skate contacted the ground.

20.17 Gravity

We have now reached the last sentence in Newton's discussion of his Third Law: "*This law also takes place in attractions*". This of course is central to Newton's (and our) view of the Universe. If the Earth is attracting the Moon gravitationally with a certain force holding it in its orbit, then the Moon is attracting the Earth with an equal force. So why isn't the Earth going around the Moon? The answer is that the masses are so different. The Earth's mass is more than one hundred times that of the Moon. Consequently, the Earth's acceleration, "falling" towards the Moon, is very small. What actually happens is that they both circle around a balance point between them, which in fact lies within the Earth. This motion of the Earth is easily detectable with instruments, but tiny compared with the daily rotation. Of course, it also follows from the above considerations that since the Earth is attracting you downwards with a force equal to your weight, you are attracting the Earth upwards—towards you—with a force of exactly the same strength.

20.18 The Law of Gravity

Let us now put together what we know about the gravitational force:

1. The gravitational force on a body (its weight, at the Earth's surface) is proportional to its mass.

2. If a body *A* attracts a body *B* with a gravitational force of a given strength, then *B* attracts *A* with a force of equal strength in the opposite direction.

3. The gravitational attraction between two bodies decreases with distance, being proportional to the inverse square of the distance between them. That is, if the distance is doubled, the gravitational attraction falls to a quarter of what it was.

One interesting point here—think about how the earth is gravitationally attracting you. Actually, all the different parts of the earth are attracting you! Mount Everest is pulling you one way, the Antarctic ice mass a different way, and the earth's core is pulling you downwards. Newton managed to prove, after thinking about it for years, that if the earth is a sphere (which is a very good approximation) then all these different attractions add up to what you would feel if *all* the earth's mass were concentrated in *one point at the center*. So, when we're talking about the gravitational attraction between you and the earth, and we talk about the distance of separation, we mean the distance between you and the *center* of the earth, which is just less than four thousand miles (6300 kilometers).

Let's denote the gravitational attractive force between two bodies A and B (as mentioned in item 2 above) by F. The forces on the two bodies are really equal and opposite vectors, each pointing to the other body, so our letter F means the *length* of these vectors, the strength of the force of attraction.

Now, item 1 tells us that the gravitational attraction between the earth and a mass m is proportional to m. This is an immediate consequence of the experimental fact that falling bodies accelerate at the same rate, usually written g (approximately 10 meters per second per second), and the definition of force from Newton's Second Law above. Thus we have

$$F \text{ is proportional to mass } m$$

for the earth's gravitational attraction on a body (often written weight $W = mg$), and Newton generalized this finding to assert that this proportionality to mass would be true for *any* gravitational attraction on the body.

From the symmetry of the force (item 2 above) and the proportionality to the mass (item 1), it follows that the gravitational force between two bodies must be proportional to *both* masses. So, if we double both masses, say, the gravitational attraction between them increases by a factor of four. We see that if the force is proportional to both masses, let's call them M and m, it is actually proportional to the *product* Mm of the masses. From item 3 above, the force is also proportional to $1/r^2$, where r is the distance between the bodies, so for the gravitational attractive force between two bodies

$$F \text{ is proportional to } Mm/r^2$$

This must mean that by measuring the gravitational force on something, we should be able to figure out the mass of the Earth! But there's a catch—all we know is that the force is *proportional* to the Earth's mass. From that we could find, for instance, the ratio of the mass of the Earth to the mass of Jupiter, by comparing how fast the Moon is "falling" around the Earth to how fast Jupiter's moons are falling around Jupiter. For that matter, we could find the ratio of the Earth's mass to the Sun's mass by seeing how fast the planets swing around the Sun. Still, knowing all these ratios doesn't tell us the Earth's mass in tons. It does tell us that if we find that out, we can then find the masses of the other planets, at least those that have moons, and the mass of the Sun.

20.19 Weighing the Earth

So how do we measure the mass of the Earth? The only way is to compare the Earth's gravitational attraction with that of something we already know the mass of. We don't know the masses of any of the heavenly bodies. What this really means is that we have to take a known mass, such as a lead ball, and measure how strongly it attracts a smaller lead ball, say, and compare that force with the earth's attraction for the smaller lead ball. This is very difficult to accomplish because the forces are so small, but it was done successfully in 1798, just over a century after Newton's work, by Cavendish.

In other words, Cavendish took two lead weights *M* and *m*, a few kilograms each, and actually *detected the tiny gravitational attraction between them* (of order of magnitude millionths of a newton)! This was a sufficiently tough experiment that even now, two hundred years later, it's not easy to give a lecture demonstration of the effect.

Making this measurement amounts to finding the constant of proportionality in the statement about *F* above, so that we can sharpen it up from a statement about proportionality to an actual useable equation,

$$F = GMm/r^2$$

where the constant *G* is what Cavendish measured, and found to be 6.67×10^{-11} in the appropriate units, where the masses are in kilograms, the distance in meters and the force in newtons. (Notice here that we can't get rid of the constant of proportionality *G*, as we did in the equation *F* = *ma*, Newton's Second Law, above. We succeeded there by defining the unit of force appropriately. In the present case, we have already defined our units of mass, distance and force, so we have no further room to maneuver.)

From Newton's theory of *universal* gravitational attraction, the *same* constant *G* determines the gravitational attraction between *any two masses* in the universe. ***This means we can now find the mass of the earth***. We just consider a one kilogram mass at the earth's surface. We know it feels a force of approximately 10 newtons, and is a distance of about 6300 km, or 6,300,000 meters, from the center of the earth. So we know *every* term in the above equation *except* the mass of the earth, and therefore can find it. This is left as an exercise.

21 The Speed of Light

21.1 Early Ideas about Light Propagation

As we shall soon see, attempts to measure the speed of light played an important part in the development of the theory of special relativity, and, indeed, the speed of light is central to the theory.

The first recorded discussion of the speed of light (I think) is in Aristotle, where he quotes Empedocles as saying the light from the sun must take some time to reach the earth, but Aristotle himself apparently disagrees, and even Descartes thought that light traveled instantaneously. Galileo, unfairly as usual, in *Two New Sciences* (page 42) has Simplicio stating the Aristotelian position,

SIMP. Everyday experience shows that the propagation of light is instantaneous; for when we see a piece of artillery fired at great distance, the flash reaches our eyes without lapse of time; but the sound reaches the ear only after a noticeable interval.

Of course, Galileo points out that in fact nothing about the speed of light can be deduced from this observation, except that light moves faster than sound. He then goes on to suggest a possible way to measure the speed of light. The idea is to have two people far away from each other, with covered lanterns. One uncovers his lantern, then the other immediately uncovers his on seeing the light from the first. This routine is to be practised with the two close together, so they will get used to the reaction times involved, then they are to do it two or three miles apart, or even further using telescopes, to see if the time interval is perceptibly lengthened. Galileo claims he actually tried the experiment at distances less than a mile, and couldn't detect a time lag. From this one can certainly deduce that light travels at least ten times faster than sound.

21.2 Measuring the Speed of Light with Jupiter's Moons

The first real measurement of the speed of light came about half a century later, in 1676, by a Danish astronomer, Ole Römer, working at the Paris Observatory. He had made a systematic study of Io, one of the moons of Jupiter, which was eclipsed by Jupiter at regular intervals, as Io went around Jupiter in a circular orbit at a steady rate. Actually, Römer found, for several months the eclipses lagged more and more behind the expected time, but then they began to pick up again. In September 1676,he correctly predicted that an eclipse on November 9 would be 10 minutes behind schedule. This was indeed the case, to the surprise of his skeptical colleagues at the Royal Observatory in Paris. Two weeks later, he told them what was happening: as the Earth and Jupiter moved in their orbits, the distance between them varied. The light from Io (actually reflected sunlight, of course) took time to reach the earth, and took the longest time when the earth was furthest away. When the Earth was furthest from Jupiter, there was an extra distance for light to travel equal to the diameter of the Earth's orbit compared with the point of closest approach. The observed eclipses were furthest behind the predicted times when the earth was furthest from Jupiter.

From his observations, Römer concluded that light took about twenty-two minutes to cross the earth's orbit. This was something of an overestimate, and a few years later Newton wrote in the *Principia* (Book I, section XIV): "For it is now certain from the phenomena of Jupiter's satellites, confirmed by the observations of different astronomers, that light is propagated in succession (*note*: I think this means at finite speed) and requires about seven or eight minutes to travel from the sun to the earth." This is essentially the correct value.

Of course, to find the speed of light it was also necessary to know the distance from the earth to the sun. During the 1670's, attempts were made to measure the parallax of Mars, that is, how far it shifted against the background of distant stars when viewed simultaneously from two different places on earth at the same time. This (very slight) shift could be used to find the distance of Mars from earth, and hence the distance to the sun, since all *relative* distances in the solar system had been established by observation and geometrical analysis. According to Crowe (*Modern Theories of the Universe*, Dover, 1994, page 30), they concluded that the distance to the sun was between 40 and 90 million miles. Measurements presumably converged on the correct value of about 93 million miles soon after that, because it appears Römer (or perhaps Huygens, using Römer's data a short time later) used the correct value for the distance, since the speed of light was calculated to be 125,000 miles per second, about three-quarters of the correct value of 186,300 miles per second. This error is fully accounted for by taking the time light needs to cross the earth's orbit to be twenty-two minutes (as Römer did) instead of the correct value of sixteen minutes.

21.3 Starlight and Rain

The next substantial improvement in measuring the speed of light took place in 1728, in England. An astronomer James Bradley, sailing on the Thames with some friends, noticed that the little pennant on top of the mast changed position each time the boat put about, even though the wind was steady. He thought of the boat as the earth in orbit, the wind as starlight coming from some distant star, and reasoned that the apparent direction the starlight was "blowing" in would depend on the way the earth was moving. Another possible analogy is to imagine the starlight as a steady downpour of rain on a windless day, and to think of yourself as walking around a circular path at a steady pace. The apparent direction of the incoming rain will not be vertically downwards—more will hit your front than your back. In fact, if the rain is falling at, say, 15 mph, and you are walking at 3 mph, to you as observer the rain will be coming down at a slant so that it has a vertical speed of 15 mph, and a horizontal speed towards you of 3 mph. Whether it is slanting down from the north or east or whatever at any given time depends on where you are on the circular path at that moment. Bradley reasoned that the apparent direction of incoming starlight must vary in just this way, but the angular change would be a lot less dramatic. The earth's speed in orbit is about 18 miles per second, he knew from Römer's work that light went at about 10,000 times that speed. That meant that the angular variation in apparent incoming direction of starlight was about the magnitude of the small angle in a right-angled triangle with one side 10,000 times longer than the other, about one two-hundredth of a degree. Notice this would have been just at the limits of Tycho's measurements, but the advent of the telescope, and general improvements in engineering, meant this small angle was quite accurately measurable by Bradley's time, and he found the velocity of light to be 185,000 miles per second, with an accuracy of about one percent.

21.4 Fast Flickering Lanterns

The problem is, all these astronomical techniques do not have the appeal of Galileo's idea of two guys with lanterns. It would be reassuring to measure the speed of a beam of light between two points on the ground, rather than making somewhat indirect deductions based on apparent slight variations in the positions of stars. We can see, though, that if the two lanterns are ten miles apart, the time lag is of order one-ten thousandth of a second, and it is difficult to see how to arrange that. This technical problem was solved in France about 1850 by two rivals, Fizeau and Foucault, using slightly different techniques. In Fizeau's apparatus, a beam of light shone between the teeth of a rapidly rotating toothed wheel, so the "lantern" was constantly being covered and uncovered. Instead of a second lantern far away, Fizeau simply had a mirror, reflecting the beam back, where it passed a second time between the teeth of the wheel. The idea was, the blip of light that went out through one gap between teeth would only make it back through the same gap if the teeth had not had time to move over significantly during the round trip time to the far away mirror. It was not difficult to make a wheel with a hundred teeth, and to rotate it hundreds of times a second, so the time for a tooth to move over could be arranged to be a fraction of one ten thousandth of a second. The method worked. Foucault's method was based on the same general idea, but instead of a toothed wheel, he shone the beam on to a rotating mirror. At one point in the mirror's rotation, the reflected beam fell on a distant mirror, which reflected it right back to the rotating mirror, which meanwhile had turned through a small angle. After this second reflection from the rotating mirror, the position of the beam was carefully measured. This made it possible to figure out how far the mirror had turned during the time it took the light to make the round trip to the distant mirror, and since the rate of rotation of the mirror was known, the speed of light could be figured out. These techniques gave the speed of light with an accuracy of about 1,000 miles per second.

21.5 Albert Abraham Michelson

Albert Michelson was born in 1852 in Strzelno, Poland. His father Samuel was a Jewish merchant, not a very safe thing to be at the time. Purges of Jews were frequent in the neighboring towns and villages. They decided to leave town. Albert's fourth birthday was celebrated in Murphy's Camp, Calaveras County, about fifty miles south east of Sacramento, a place where five million dollars worth of gold dust was taken from one four acre lot. Samuel prospered selling supplies to the miners. When the gold ran out, the Michelsons moved to Virginia City, Nevada, on the Comstock lode, a silver mining town. Albert went to high school in San Francisco. In 1869, his father spotted an announcement in the local paper that Congressman Fitch would be appointing a candidate to the Naval Academy in Annapolis, and inviting applications. Albert applied but did not get the appointment, which went instead to the son of a civil war veteran. However, Albert knew that President Grant would also be appointing ten candidates himself, so he went east on the just opened continental railroad to try his luck. Unknown to Michelson, Congressman Fitch wrote directly to Grant on his behalf, saying this would really help get the Nevada Jews into the Republican party. This argument proved persuasive. In fact, by the time Michelson met with Grant, all ten scholarships had been awarded, but the President somehow came up with another one. Of the incoming class of ninety-two, four years later twenty-nine graduated. Michelson placed first in optics, but twenty-fifth in seamanship. The Superintendent of the Academy, Rear Admiral Worden, who had commanded the *Monitor* in its victory over the *Merrimac*, told Michelson: "If in the future you'd give less attention to those scientific things and more to your naval gunnery, there might come a time when you would know enough to be of some service to your country."

21.6 Sailing the Silent Seas: Galilean Relativity

Shortly after graduation, Michelson was ordered aboard the USS *Monongahela*, a sailing ship, for a voyage through the Carribean and down to Rio. According to the biography of Michelson written by his daughter (*The Master of Light*, by Dorothy Michelson Livingston, Chicago, 1973) he thought a lot as the ship glided across the quiet Caribbean about whether one could decide in a closed room inside the ship whether or not the vessel was moving. In fact, his daughter quotes a famous passage from Galileo on just this point:

[SALV.] Shut yourself up with some friend in the largest room below decks of some large ship and there procure gnats, flies, and other such small winged creatures. Also get a great tub full of water and within it put certain fishes; let also a certain bottle be hung up, which drop by drop lets forth its water into another narrow-necked bottle placed underneath. Then, the ship lying still, observe how those small winged animals fly with like velocity towards all parts of the room; how the fish swim indifferently towards all sides; and how the distilling drops all fall into the bottle placed underneath. And casting anything toward your friend, you need not throw it with more force one way than another, provided the distances be equal; and leaping with your legs together, you will reach as far one way as another. Having observed all these particulars, though no man doubts that, so long as the vessel stands still, they ought to take place in this manner, make the ship move with what velocity you please, so long as the motion is uniform and not fluctuating this way and that. You will not be able to discern the least alteration in all the forenamed effects, nor can you gather by any of them whether the ship moves or stands still. ...in throwing something to your friend you do not need to throw harder if he is towards the front of the ship from you... the drops from the upper bottle still fall into the lower bottle even though the ship may have moved many feet while the drop is in the air ... Of this correspondence of

effects the cause is that the ship's motion is common to all the things contained in it and to the air also; I mean if those things be shut up in the room; but in case those things were above the deck in the open air, and not obliged to follow the course of the ship, differences would be observed, ... smoke would stay behind... .

[SAGR.] Though it did not occur to me to try any of this out when I was at sea, I am sure you are right. I remember being in my cabin wondering a hundred times whether the ship was moving or not, and sometimes I imagined it to be moving one way when in fact it was moving the other way. I am therefore satisfied that no experiment that can be done in a closed cabin can determine the speed or direction of motion of a ship in steady motion.

I have paraphrased this last remark somewhat to clarify it. This conclusion of Galileo's, that everything looks the same in a closed room moving at a steady speed as it does in a closed room at rest, is called *The Principle of Galilean Relativity*. We shall be coming back to it.

21.7 Michelson Measures the Speed of Light

On returning to Annapolis from the cruise, Michelson was commissioned Ensign, and in 1875 became an instructor in physics and chemistry at the Naval Academy, under Lieutenant Commander William Sampson. Michelson met Mrs. Sampson's niece, Margaret Heminway, daughter of a very successful Wall Street tycoon, who had built himself a granite castle in New Rochelle, NY. Michelson married Margaret in an Episcopal service in New Rochelle in 1877.

At work, lecture demonstrations had just been introduced at Annapolis. Sampson suggested that it would be a good demonstration to measure the speed of light by Foucault's method. Michelson soon realized, on putting together the apparatus, that he could redesign it for much greater accuracy, but that would need money well beyond that available in the teaching demonstration budget. He went and talked with his father in law, who agreed to put up $2,000. Instead of Foucault's 60 feet to the far mirror, Michelson had about 2,000 feet along the bank of the Severn, a distance he measured to one tenth of an inch. He invested in very high quality lenses and mirrors to focus and reflect the beam. His final result was 186,355 miles per second, with possible error of 30 miles per second or so. This was twenty times more accurate than Foucault, made the New York Times, and Michelson was famous while still in his twenties. In fact, this was accepted as the most accurate measurement of the speed of light for the next forty years, at which point Michelson measured it again.

The next lecture is on the Michelson-Morley experiment to detect the aether.

22 The Michelson-Morley Experiment

Flashlet of the Experiment!

22.1 The Nature of Light

As a result of Michelson's efforts in 1879, the speed of light was known to be 186,350 miles per second with a likely error of around 30 miles per second. This measurement, made by timing a flash of light travelling between mirrors in Annapolis, agreed well with less direct measurements based on astronomical observations. Still, this did not really clarify the *nature* of light. Two hundred years earlier, Newton had suggested that light consists of tiny *particles* generated in a hot object, which spray out at very high speed, bounce off other objects, and are detected by our eyes. Newton's arch-enemy Robert Hooke, on the other hand, thought that light must be a kind of *wave motion*, like sound. To appreciate his point of view, let us briefly review the nature of sound.

22.2 The Wavelike Nature of Sound

Actually, *sound* was already quite well understood by the ancient Greeks. The essential point they had realized is that sound is generated by a vibrating material object, such as a bell, a string or a drumhead. Their explanation was that the vibrating drumhead, for example, alternately pushes and pulls on the air directly above it, sending out waves of compression and decompression (known as rarefaction), like the expanding circles of ripples from a disturbance on the surface of a pond. On reaching the ear, these waves push and pull on the eardrum with the same frequency (that is to say, the same number of pushes per second) as the original source was vibrating at, and nerves transmit from the ear to the brain both the intensity (loudness) and frequency (pitch) of the sound.

There are a couple of special properties of sound waves (actually any waves) worth mentioning at this point. The first is called *interference*. This is most simply demonstrated with water waves. If you put two fingers in a tub of water, just touching the surface a foot or so apart, and vibrate them at the same rate to get two expanding circles of ripples, you will notice that where the ripples overlap there are quite complicated patterns of waves formed. The essential point is that at those places where the wave-crests from the two sources arrive at the same time, the waves will work together and the water will be very disturbed, but at points where the crest from one source arrives at the same time as the wave trough from the other source, the waves will cancel each other out, and the water will hardly move. You can hear this effect for sound waves by playing a constant note through stereo speakers. As you move around a room, you will hear quite large variations in the intensity of sound. Of course, reflections from walls complicate the pattern. This large variation in volume is *not* very noticeable when the stereo is playing music, because music is made up of many frequencies, and they change all the time. The different frequencies, or notes, have their quiet spots in the room in different places. The other point that should be mentioned is that high frequency tweeter-like sound is much more *directional* than low frequency woofer-like sound. It really doesn't matter where in the room you put a low-frequency woofer—the sound seems to be all around you anyway. On the other hand, it is quite difficult to get a speaker to spread the high notes in all directions. If you listen to a cheap speaker, the high notes are loudest if the speaker is pointing right at you. A lot of effort has gone into designing tweeters, which are small speakers especially designed to broadcast high notes over a wide angle of directions.

22.3 Is Light a Wave?

Bearing in mind the above minireview of the properties of waves, let us now reconsider the question of whether light consists of a stream of particles or is some kind of wave. The strongest argument for a particle picture is that light travels in straight lines. You can hear around a corner, at least to some extent, but you certainly can't see. Furthermore, no wave-like interference effects are very evident for light. Finally, it was long known, as we have mentioned, that sound waves were compressional waves in air. If light is a wave, just what is waving? It clearly isn't just air, because light reaches us from the sun, and indeed from stars, and we know the air doesn't stretch that far, or the planets would long ago have been slowed down by air resistance.

Despite all these objections, it was established around 1800 that light *is* in fact some kind of wave. The reason this fact had gone undetected for so long was that the wavelength is *really* short, about one fifty-thousandth of an inch. In contrast, the shortest wavelength sound detectable by humans has a wavelength of about half an inch. The fact that light travels in straight lines is in accord with observations on sound that the higher the frequency (and shorter the wavelength) the greater the tendency to go in straight lines. Similarly, the interference patterns mentioned above for sound waves or ripples on a pond vary over distances of the same sort of size as the wavelengths involved. Patterns like that would not normally be noticeable for light because they would be on such a tiny scale. In fact, it turns out, there *are* ways to see interference effects with light. A familiar example is the many colors often visible in a soap bubble. These come about because looking at a soap bubble you see light reflected from both sides of a very thin film of water—a thickness that turns out to be comparable to the wavelength of light. The light reflected from the lower layer has to go a little further to reach your eye, so that light wave must wave an extra time or two before getting to your eye compared with the light reflected from the top layer. What you actually *see* is the *sum* of the light reflected from the top layer and that reflected from the bottom layer. Thinking of this now as the sum of two sets of waves, the light will be bright if the crests of the two waves arrive together, dim if the *crests* of waves reflected from the top layer arrive simultaneously with the *troughs* of waves reflected from the bottom layer. Which of these two possibilities actually occurs for reflection from a particular bit of the soap film depends on just how much further the light reflected from the lower surface has to travel to reach your eye compared with light from the upper surface, and that depends on the angle of reflection and the thickness of the film.

Suppose now we shine *white* light on the bubble. White light is made up of all the colors of the rainbow, and these different colors have different wavelengths, so we see colors reflected, because for a particular film, at a particular angle, some colors will be reflected brightly (the crests will arrive together), some dimly, and we will see the ones that win.

22.4 If Light *is* a Wave, What is Waving?

Having established that light is a wave, though, we still haven't answered one of the major objections raised above. Just what is waving? We discussed sound waves as waves of compression in air. Actually, that is only one case—sound will also travel through liquids, like water, and solids, like a steel bar. It is found experimentally that, other things being equal, sound travels faster through a medium that is harder to compress: the material just springs back faster and the wave moves through more rapidly. For media of equal springiness, the sound goes faster through the less heavy medium, essentially because the same amount of springiness can push things along faster in a lighter material. So when a sound wave passes, the material—air, water or solid—waves as it goes through. Taking this as a hint, it was natural to suppose that light must be just waves in some mysterious material, which was called the *aether*, surrounding and permeating everything. This aether must also fill all of space, out to the stars, because we can see them, so the medium must be there to carry the light. (We could never *hear* an explosion on the moon, however loud, because there is no air to carry the sound to us.) Let us think a bit about what properties this aether must have. Since light travels so fast, it must be very light, and very hard to compress. Yet, as mentioned above, it must allow solid bodies to pass through it freely, without aether resistance, or the planets would be slowing down. Thus we can picture it as a kind of ghostly wind blowing through the earth. But how can we prove any of this? Can we detect it?

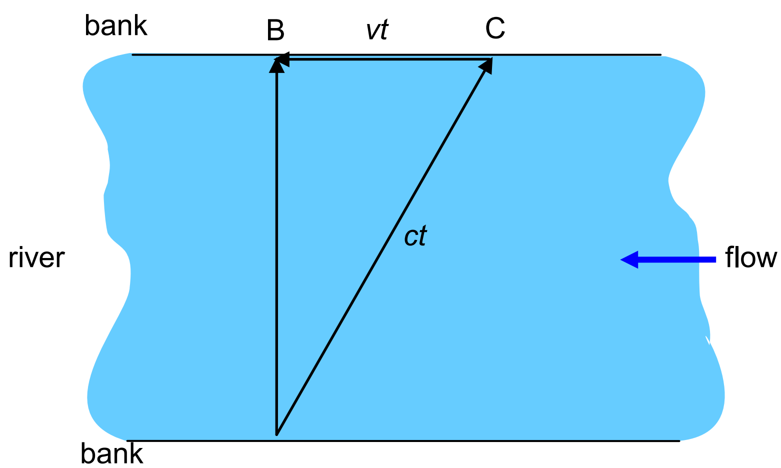
22.5 Detecting the Aether Wind: the Michelson-Morley Experiment

Detecting the aether wind was the next challenge Michelson set himself after his triumph in measuring the speed of light so accurately. Naturally, something that allows solid bodies to pass through it freely is a little hard to get a grip on. But Michelson realized that, just as the speed of sound is relative to the air, so the speed of light must be relative to the aether. This must mean, if you could measure the speed of light accurately enough, you could measure the speed of light travelling upwind, and compare it with the speed of light travelling downwind, and the difference of the two measurements should be twice the windspeed. Unfortunately, it wasn't that easy. All the recent accurate measurements had used light travelling to a distant mirror and coming back, so if there was an aether wind along the direction between the mirrors, it would have opposite effects on the two parts of the measurement, leaving a very small overall effect. There was no technically feasible way to do a one-way determination of the speed of light.

At this point, Michelson had a very clever idea for detecting the aether wind. As he explained to his children (according to his daughter), it was based on the following puzzle:

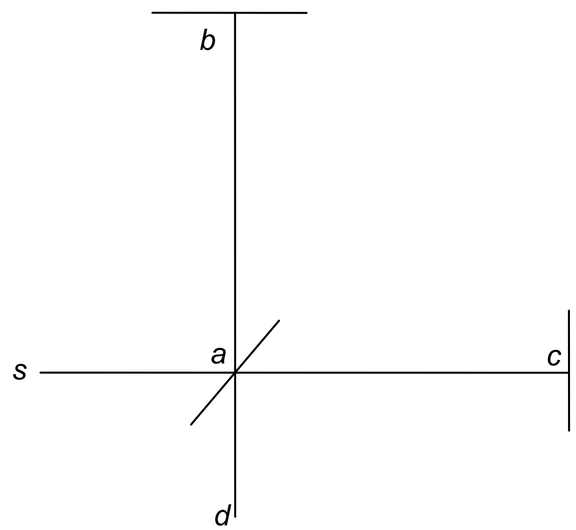
Suppose we have a river of width w (say, 100 feet), and two swimmers who both swim at the same speed v feet per second (say, 5 feet per second). The river is flowing at a steady rate, say 3 feet per second. The swimmers race in the following way: they both start at the same point on one bank. One swims directly across the river to the closest point on the opposite bank, then turns around and swims back. The other stays on one side of the river, swimming upstream a distance (measured along the bank) exactly equal to the width of the river, then swims back to the start. Who wins?

Let's consider first the swimmer going upstream and back. Going 100 feet upstream, the speed relative to the bank is only 2 feet per second, so that takes 50 seconds. Coming back, the speed is 8 feet per second, so it takes 12.5 seconds, for a total time of 62.5 seconds.



In time *t*, the swimmer has moved *ct* relative to the water, and been carried downstream a distance *vt*.

The swimmer going across the flow is trickier. It won't do simply to aim directly for the opposite bank-the flow will carry the swimmer downstream. To succeed in going directly across, the swimmer must actually aim upstream at the correct angle (of course, a real swimmer would do this automatically). Thus, the swimmer is going at 5 feet per second, at an angle, relative to the river, and being carried downstream at a rate of 3 feet per second. If the angle is correctly chosen so that the net movement is directly across, in one second the swimmer must have moved *four feet* across: the distances covered in one second will form a 3,4,5 triangle. So, at a crossing rate of 4 feet per second, the swimmer gets across in 25 seconds, and back in the same time, for a total time of 50 seconds. The cross-stream swimmer wins. This turns out to true whatever their swimming speed. (Of course, the race is only possible if they can swim faster than the current!)



This diagram is from the original paper. The source of light is at *s*, the 45 degree line is the half-silvered mirror, *b* and *c* are mirrors and *d* the observer.

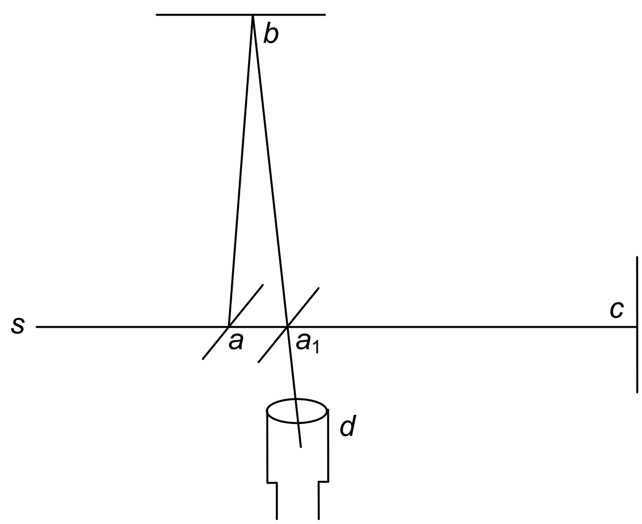
Michelson's great idea was to construct an exactly similar race for pulses of light, with the aether wind playing the part of the river. The scheme of the experiment is as follows: a pulse of light is directed at an angle of 45 degrees at a half-silvered, half transparent mirror, so that half the pulse goes on through the glass, half is reflected. These two half-pulses are the two swimmers. They both go on to distant mirrors which reflect them back to the half-silvered mirror. At this point, they are again half reflected and half transmitted, but a telescope is placed behind the half-silvered mirror as shown in the figure so that half of each half-pulse will arrive in this telescope. Now, if there is an aether wind blowing, someone looking through the telescope should see the halves of the two half-pulses to arrive at slightly different times, since one would have gone more upstream and back, one more across stream in general. To maximize the effect, the whole apparatus, including the distant mirrors, was placed on a large turntable so it could be swung around.

An animated flashlet of the experiment is available here–it makes the account above a lot clearer!

Let us think about what kind of time delay we expect to find between the arrival of the two half-pulses of light. Taking the speed of light to be c miles per second relative to the aether, and the aether to be flowing at v miles per second through the laboratory, to go a distance w miles upstream will take $w/(c-v)$ seconds, then to come back will take $w/(c+v)$ seconds. The total roundtrip time upstream and downstream is the sum of these, which works out to be $2wc/(c^2-v^2)$, which can also be written $(2w/c)\times 1/(1-v^2/c^2)$. Now, we can safely assume the speed of the aether is much less than the speed of light, otherwise it would have been noticed long ago, for example in timing of eclipses of Jupiter's satellites. This means v^2/c^2 is a very small number, and we can use some handy mathematical facts to make the algebra a bit easier. First, if x is very small compared to 1, $1/(1-x)$ is very close to $1+x$. (You can check it with your calculator.) Another fact we shall need in a minute is that for small x, the square root of $1+x$ is very close to $1+x/2$.

Putting all this together,

$$\text{upstream-downstream roundtrip time} \cong \frac{2w}{c}\times\left(1+\frac{v^2}{c^2}\right).$$



This is also from the original paper, and shows the expected path of light relative to the aether with an aether wind blowing.

Now, what about the cross-stream time? The actual cross-stream speed must be figured out as in the example above using a right-angled triangle, with the hypoteneuse equal to the speed c, the shortest side the aether flow speed v, and the other side the cross-stream speed we need to find the time to get across. From Pythagoras' theorem, then, the cross-stream speed is the square root of (c^2-v^2).

Since this will be the same both ways, the roundtrip cross-stream time will be

$$2w/\sqrt{c^2-v^2}.$$

This can be written in the form

$$\frac{2w}{c}\frac{1}{\sqrt{1-v^2/c^2}} \cong \frac{2w}{c}\frac{1}{1-\left(v^2/2c^2\right)} \cong \frac{2w}{c}\left(1+\frac{v^2}{2c^2}\right)$$

where the two successive approximations, valid for $v/c = x \ll 1$, are $\sqrt{1-x} \cong 1-\left(x/2\right)$ and $1/\left(1-x\right) \cong 1+x$.

Therefore the

$$\text{cross-stream roundtrip time } \cong \frac{2w}{c}\times\left(1+\frac{v^2}{2c^2}\right).$$

Looking at the two roundtrip times at the ends of the two paragraphs above, we see that they differ by an amount $(2w/c) \times v^2/2c^2$. Now, $2w/c$ is just the time the light would take if there were no aether wind at all, say, a few millionths of a second. If we take the aether windspeed to be equal to the earth's speed in orbit, for example, v/c is about 1/10,000, so v^2/c^2 is about 1/100,000,000. This means the time delay between the pulses reflected from the different mirrors reaching the telescope is about one-hundred-millionth of a few millionths of a second. It seems completely hopeless that such a short time delay could be detected. However, this turns out *not* to be the case, and Michelson was the first to figure out how to do it. The trick is to use the *interference* properties of the lightwaves. Instead of sending pulses of light, as we discussed above, Michelson sent in a steady beam of light of a single color. This can be visualized as a sequence of ingoing waves, with a wavelength one fifty-thousandth of an inch or so. Now this sequence of waves is split into two, and reflected as previously described. One set of waves goes upstream and downstream, the other goes across stream and back. Finally, they come together into the telescope and the eye. If the one that took longer is half a wavelength behind, its troughs will be on top of the crests of the first wave, they will cancel, and nothing will be seen. If the delay is less than that, there will still be some dimming. However, slight errors in the placement of the mirrors would have the same effect. This is one reason why the apparatus is built to be rotated. On turning it through 90 degrees, the upstream-downstream and the cross-stream waves change places. Now the other one should be behind. Thus, if there is an aether wind, if you watch through the telescope while you rotate the turntable, you should expect to see variations in the brightness of the incoming light.

To magnify the time difference between the two paths, in the actual experiment the light was reflected backwards and forwards several times, like a several lap race. For a diagram, click here. For an actual photograph of the real apparatus, click here.

Michelson calculated that an aether windspeed of only one or two miles a second would have observable effects in this experiment, so if the aether windspeed was comparable to the earth's speed in orbit around the sun, it would be easy to see. In fact, *nothing* was observed. The light intensity did not vary at all. Some time later, the experiment was redesigned so that an aether wind caused by the earth's daily rotation could be detected. Again, nothing was seen. Finally, Michelson wondered if the aether was somehow getting stuck to the earth, like the air in a below-decks cabin on a ship, so he redid the experiment on top of a high mountain in California. Again, no aether wind was observed. It was difficult to believe that the aether in the immediate vicinity of the earth was stuck to it and moving with it, because light rays from stars would deflect as they went from the moving faraway aether to the local stuck aether.

The only possible conclusion from this series of very difficult experiments was that the whole concept of an all-pervading aether was wrong from the start. Michelson was very reluctant to think along these lines. In fact, new theoretical insight into the nature of light had arisen in the 1860's from the brilliant theoretical work of Maxwell, who had written down a set of equations describing how electric and magnetic fields can give rise to each other. He had discovered that his equations predicted there could be waves made up of electric and magnetic fields, and the speed of these waves, deduced from experiments on how these fields link together, would be 186,300 miles per second. This is, of course, the speed of light, so it is natural to assume that light is made up of fast-varying electric and magnetic fields. But this leads to a big problem: Maxwell's equations predict a definite speed for light, and it *is* the speed found by measurements. But what is the speed to be measured relative to? The whole point of bringing in the aether was to give a picture for light resembling the one we understand for sound, compressional waves in a medium. The speed of sound through air is measured relative to air. If the wind is blowing towards you from the source of sound, you will hear the sound sooner. If there isn't an aether, though, this analogy doesn't hold up. So what does light travel at 186,300 miles per second relative to?

There is another obvious possibility, which is called the emitter theory: the light travels at 186,300 miles per second relative to the source of the light. The analogy here is between light emitted by a source and bullets emitted by a machine gun. The bullets come out at a definite speed (called the muzzle velocity) relative to the barrel of the gun. If the gun is mounted on the front of a tank, which is moving forward, and the gun is pointing forward, then relative to the ground the bullets are moving faster than they would if shot from a tank at rest. The simplest way to test the emitter theory of light, then, is to measure the speed of light emitted in the forward direction by a flashlight moving in the forward direction, and see if it exceeds the known speed of light by an amount equal to the speed of the flashlight. Actually, this kind of direct test of the emitter theory only became experimentally feasible in the nineteen-sixties. It is now possible to produce particles, called neutral pions, which decay each one in a little explosion, emitting a flash of light. It is also possible to have these pions moving forward at 185,000 miles per second when they self destruct, and to catch the light emitted in the forward direction, and clock its speed. It is found that, despite the expected boost from being emitted by a very fast source, the light from the little explosions is going forward at the usual speed of 186,300 miles per second. In the last century, the emitter theory was rejected because it was thought the appearance of certain astronomical phenomena, such as double stars, where two stars rotate around each other, would be affected. Those arguments have since been criticized, but the pion test is unambiguous. The definitive experiment was carried out by Alvager et al., *Physics Letters* **12**, 260 (1964).

22.6 Einstein's Answer

The results of the various experiments discussed above seem to leave us really stuck. Apparently light is not like sound, with a definite speed relative to some underlying medium. However, it is also not like bullets, with a definite speed relative to the source of the light. Yet when we measure its speed we always get the same result. How can all these facts be interpreted in a simple consistent way? We shall show how Einstein answered this question in the next lecture.

23 Special Relativity

23.1 Galilean Relativity again

At this point in the course, we finally enter the twentieth century—Albert Einstein wrote his first paper on relativity in 1905. To put his work in context, let us first review just what is meant by "relativity" in physics. The first example, mentioned in a previous lecture, is what is called "Galilean relativity" and is nothing but Galileo's perception that by observing the motion of objects, alive or dead, in a closed room there is no way to tell if the room is at rest or is in fact in a boat moving at a steady speed in a fixed direction. (You *can* tell if the room is accelerating or turning around.) Everything looks the same in a room in steady motion as it does in a room at rest. After Newton formulated his Laws of Motion, describing how bodies move in response to forces and so on, physicists reformulated Galileo's observation in a slightly more technical, but equivalent, way: they said *the laws of physics are the same in a uniformly moving room as they are in a room at rest*. In other words, the same force produces the same acceleration, and an object experiencing no force moves at a steady speed in a straight line in either case. Of course, talking in these terms implies that we have clocks and rulers available so that we can actually time the motion of a body over a measured distance, so the physicist envisions the room in question to have calibrations along all the walls, so the position of anything can be measured, and a good clock to time motion. Such a suitably equipped room is called a "*frame of reference*"—the calibrations on the walls are seen as a frame which you can use to specify the precise position of an object at a given time. (This is the same as a set of "coordinates".) Anyway, the bottom line is that no amount of measuring of motions of objects in the "frame of reference" will tell you whether this is a frame at rest or one moving at a steady velocity.

What exactly do we mean by a frame "at rest" anyway? This seems obvious from our perspective as creatures who live on the surface of the earth—we mean, of course, at rest relative to fixed objects on the earth's surface. Actually, the earth's rotation means this isn't quite a fixed frame, and also the earth is moving in orbit at 18 miles per second. From an astronaut's point of view, then, a frame fixed relative to the sun might seem more reasonable. But why stop there? We believe the laws of physics are good throughout the universe. Let us consider somewhere in space far from the sun, even far from our galaxy. We would see galaxies in all directions, all moving in different ways. Suppose we now set up a frame of reference and check that Newton's laws still work. In particular, we check that the First Law holds—that a body experiencing no force moves at a steady speed in a straight line. *This First law is often referred to as The Principle of Inertia, and a frame in which it holds is called an Inertial Frame*. Then we set up another frame of reference, moving at a steady velocity relative to the first one, and find that Newton's laws are o.k. in this frame too. The point to notice here is that it is not at all obvious which—if either—of these frames is "at rest". We *can*, however, assert that they are both *inertial* frames, after we've checked that in both of them, a body with no forces acting on it moves at a steady speed in a straight line (the speed could be zero). In this situation, Michelson would have said that a frame "at rest" is one at rest relative to the aether. However, his own experiment found motion through the aether to be undetectable, so how would we ever know we were in the right frame?

As we mentioned in the last lecture, in the middle of the nineteenth century there was a substantial advance in the understanding of electric and magnetic fields. (In fact, this advance is in large part responsible for the improvement in living standards since that time.) The new understanding was summarized in a set of equations called Maxwell's equations describing how electric and magnetic fields interact and give rise to each other, just as, two centuries earlier, the new understanding of dynamics was summarized in the set of equations called Newton's laws. The important thing about Maxwell's equations for our present purposes is that they predicted waves made up of electric and magnetic fields that moved at 3×10^8 meters per second, and it was immediately realized that this was no coincidence—light waves must be nothing but waving electric and magnetic fields. (This is now fully established to be the case.)

It is worth emphasizing that Maxwell's work predicted the speed of light from the results of experiments that were not thought at the time they were done to have anything to do with light—experiments on, for example, the strength of electric field produced by waving a magnet. Maxwell was able to deduce a speed for waves like this using methods analogous to those by which earlier scientists had figured out the speed of sound from a knowledge of the density and the springiness of air.

23.2 Generalizing Galilean Relativity to Include Light: Special Relativity

We now come to Einstein's major insight: the Theory of Special Relativity. It is deceptively simple. Einstein first dusted off Galileo's discussion of experiments below decks on a uniformly moving ship, and restated it as :

The Laws of Physics are the same in all Inertial Frames.

Einstein then simply brought this up to date, by pointing out that the Laws of Physics must now include Maxwell's equations describing electric and magnetic fields as well as Newton's laws describing motion of masses under gravity and other forces. (*Note for experts and the curious*: we shall find that Maxwell's equations are completely unaltered by special relativity, but, as will become clear later, Newton's Laws do need a bit of readjustment to include special relativistic phenomena. The First Law is still O.K., the Second Law in the form $F = ma$ is not, because we shall find mass varies; we need to equate force to rate of change of momentum (Newton understood that, of course—that's the way he stated the law!). The Third Law, stated as action equals reaction, no longer holds because if a body moves, its electric field, say, does not readjust instantaneously—a ripple travels outwards at the speed of light. Before the ripple reaches another charged body, the electric forces between the two will be unbalanced. However, the crucial consequence of the Third Law—the conservation of momentum when two bodies interact, still holds. It turns out that the rippling field itself carries momentum, and everything balances.)

Demanding that Maxwell's equations be satisfied in all inertial frames has one major consequence as far as we are concerned. As we stated above, Maxwell's equations give the speed of light to be 3×10^8 meters per second. Therefore, *demanding that the laws of physics are the same in all inertial frames implies that the speed of any light wave, measured in any inertial frame, must be* 3×10^8 *meters per second*.

This then is the entire content of the Theory of Special Relativity: the Laws of Physics are the same in any inertial frame, and, in particular, any measurement of the speed of light in any inertial frame will always give 3×10^8 meters per second.

23.3 You *Really* Can't Tell You're Moving!

Just as Galileo had asserted that observing gnats, fish and dripping bottles, throwing things and generally jumping around would not help you to find out if you were in a room at rest or moving at a steady velocity, Einstein added that no kind of observation at all*, even measuring the speed of light across your room* to any accuracy you like, would help find out if your room was "really at rest". This implies, of course, that the concept of being "at rest" is meaningless. If Einstein is right, there is *no* natural rest-frame in the universe. Naturally, there can be no "aether", no thin transparent jelly filling space and vibrating with light waves, because if there were, *it* would provide the natural rest frame, and affect the speed of light as measured in other moving inertial frames as discussed above.

So we see the Michelson-Morley experiment was doomed from the start. There never was an aether wind. The light was not slowed down by going "upstream"—light *always* travels at the same speed, which we shall now call c,

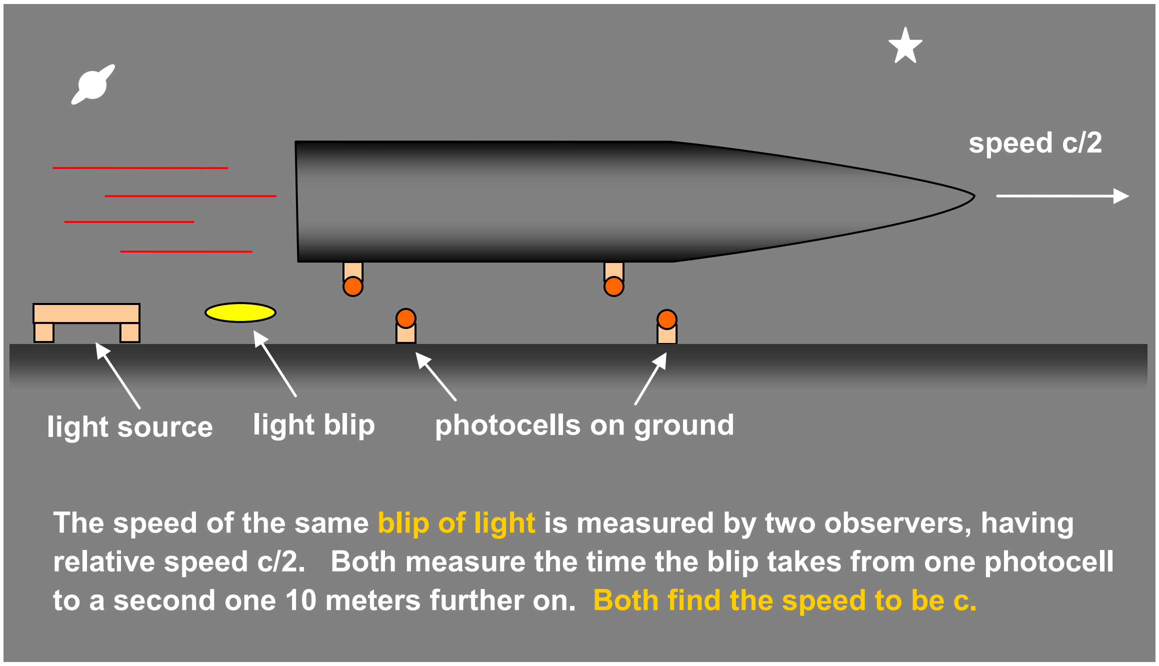
$$c = 3\times10^{8} \text{ meters per second}$$

to save writing it out every time. *This now answers the question of what the speed of light, c, is relative to*. We already found that it is not like sound, relative to some underlying medium. It is also not like bullets, relative to the source of the light (the discredited emitter theory). *Light travels at c relative to the observer*, since if the observer sets up an inertial frame (clocks, rulers, etc.) to measure the speed of light he will find it to be c. (We always assume our observers are very competent experimentalists!)

23.4 Truth and Consequences

The Truth we are referring to here is the seemingly innocuous and plausible sounding statement that all inertial frames are as good as each other—the laws of physics are the same in all of them—and so the *speed of light* is the same in all of them. As we shall soon see, this Special Theory of Relativity has some surprising consequences, which reveal themselves most dramatically when things are moving at relative speeds comparable to the speed of light. Einstein liked to explain his theory using what he called "thought experiments" involving trains and other kinds of transportation moving at these speeds (technically unachievable so far!), and we shall follow his general approach.

To begin with, let us consider a simple measurement of the speed of light carried out at the same time in two inertial frames moving at half the speed of light relative to each other. The setup is as follows: on a flat piece of ground, we have a flashlight which emits a blip of light, like a strobe. We have two photocells, devices which click and send a message down a wire when light falls on them. The photocells are placed 10 meters apart in the path of the blip of light, they are somehow wired into a clock so that the time taken by the blip of light to travel from the first photocell to the second, in other words, the time between clicks, can be measured. From this time and the known distance between them, we can easily find the speed of the blip of light.



Meanwhile, there is another observer, passing overhead in a spaceship traveling at half the speed of light. She is also equipped with a couple of photocells, placed 10 meters apart on the bottom of her spaceship as shown, and she is able to measure the speed of the same blip of light, relative to her frame of reference (the spaceship). *The observer on the spaceship will measure the blip of light to be traveling at c relative to the spaceship, the observer on the ground will measure the same blip to be traveling at c relative to the ground*. That is the unavoidable consequence of the Theory of Relativity.

(*Note*: actually the picture above is not quite the way it would really look. As we shall find, objects moving at relativistic speeds are contracted, and this combined with the different times light takes to reach the eye from different parts of the ship would change the ship's appearance. But this does not affect the validity of the statements above.)

24 Special Relativity: What Time is it?

24.1 Special Relativity in a Nutshell

Einstein's Theory of Special Relativity, discussed in the last lecture, may be summarized as follows:

The Laws of Physics are the same in any Inertial Frame of Reference. (Such frames move at steady velocities with respect to each other.)

These Laws include in particular Maxwell's Equations describing electric and magnetic fields, which predict that light always travels at a particular speed c, equal to about 3×10^8 meters per second, that is,186,300 miles per second.

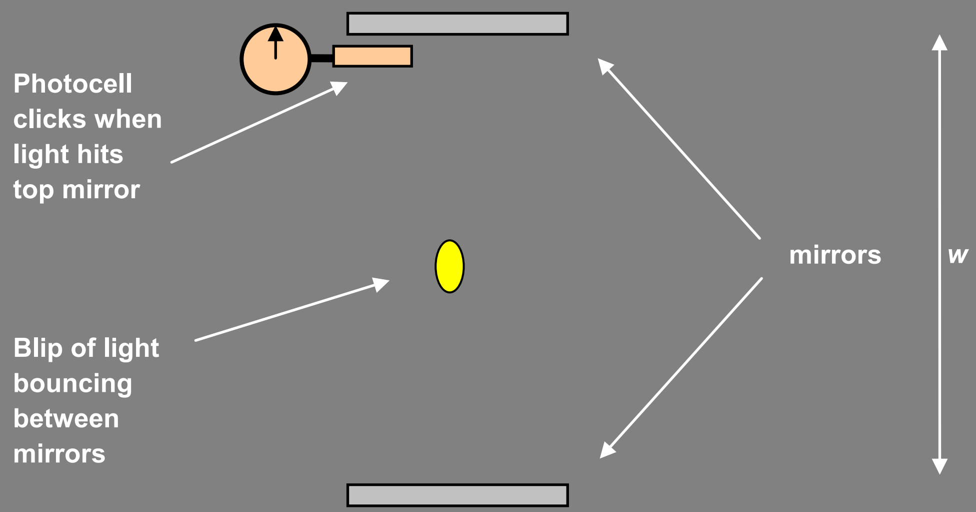
It follows that any measurement of the speed of any flash of light by any observer in any inertial frame will give the same answer c.

We have already noted one counter-intuitive consequence of this, that two different observers moving relative to each other, each measuring the speed of the *same* blob of light relative to himself, will *both* get c, even if their relative motion is in the same direction as the motion of the blob of light.

We shall now explore how this simple assumption changes everything we thought we understood about time and space.

24.2 A Simple but Reliable Clock

We mentioned earlier that each of our (inertial) frames of reference is calibrated (had marks at regular intervals along the walls) to measure distances, and has a clock to measure time. Let us now get more specific about the clock—we want one that is easy to understand in any frame of reference. Instead of a pendulum swinging back and forth, which wouldn't work away from the earth's surface anyway, we have a blip of light bouncing back and forth between two mirrors facing each other. We call this device a *light clock*. To really use it as a timing device we need some way to count the bounces, so we position a photocell at the upper mirror, so that it catches the edge of the blip of light. The photocell clicks when the light hits it, and this regular series of clicks drives the clock hand around, just as for an ordinary clock. Of course, driving the photocell will eventually use up the blip of light, so we also need some provision to reinforce the blip occasionally, such as a strobe light set to flash just as it passes and thus add to the intensity of the light. Admittedly, this may not be an easy way to build a clock, but the basic idea is simple.

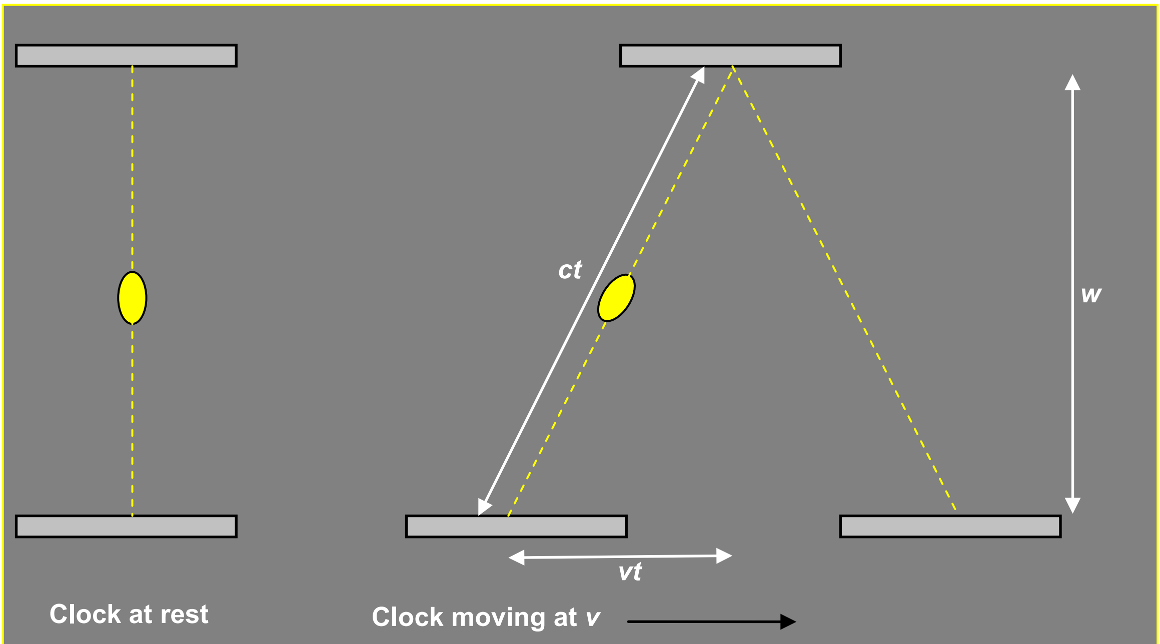


Einstein's Light Clock: a blip of light bounces between two parallel mirrors *w* meters apart; each time it hits the top mirror, the photocell moves the clock on one click.

It's easy to figure out how frequently our light clock clicks. If the two mirrors are a distance w apart, the round trip distance for the blip from the photocell mirror to the other mirror and back is $2w$. Since we know the blip always travels at c, we find the round trip time to be $2w/c$, so this is the time between clicks. This isn't a very long time for a reasonable sized clock! The crystal in a quartz watch "clicks " of the order of 10,000 times a second. That would correspond to mirrors about nine miles apart, so we need our clock to click about 1,000 times faster than that to get to a reasonable size. Anyway, let us assume that such purely technical problems have been solved.

24.3 Looking at Somebody Else's Clock

Let us now consider two observers, Jack and Jill, each equipped with a calibrated inertial frame of reference, and a light clock. To be specific, imagine Jack standing on the ground with his light clock next to a straight railroad line, while Jill and her clock are on a large flatbed railroad wagon which is moving down the track at a constant speed v. Jack now decides to check Jill's light clock against his own. He knows the time for his clock is $2w/c$ between clicks. Imagine it to be a slightly misty day, so with binoculars he can actually see the blip of light bouncing between the mirrors of Jill's clock. How long does he think that blip takes to make a round trip? The one thing he's sure of is that it must be moving at c = 186,300 miles per second, relative to him—that's what Einstein tells him. So to find the round trip time, all he needs is the round trip distance. This will *not* be $2w$, because the mirrors are on the flatbed wagon moving down the track, so, relative to Jack on the ground, when the blip gets back to the top mirror, that mirror has moved down the track some since the blip left, so the blip actually follows a zigzag path as seen from the ground.



Two identical light clocks: **one at rest, one moving relative to us. The light blips in both travel at the same speed *relative to us*, the one in the moving clock goes further, so must take longer between clicks.**

Suppose now the blip in Jill's clock on the moving flatbed wagon takes time t to get from the bottom mirror to the top mirror as measured by Jack standing by the track. Then the length of the "zig" from the bottom mirror to the top mirror is necessarily ct, since that is the distance covered by any blip of light in time t. Meanwhile, the wagon has moved down the track a distance vt, where v is the speed of the wagon. This should begin to look familiar—it is precisely the same as the problem of the swimmer who swims at speed c relative to the water crossing a river flowing at v! We have again a right-angled triangle with hypotenuse ct, and shorter sides vt and w.

From Pythagoras, then,

$$c^2t^2 = v^2t^2 + w^2$$

so

$$t^2(c^2 - v^2) = w^2$$

or

$$t^2(1 - v^2/c^2) = w^2/c^2$$

and, taking the square root of each side, then doubling to get the round trip time, we conclude that Jack sees the time between clicks for Jill's clock to be:

$$\text{time between clicks for moving clock } = \frac{2w}{c}\frac{1}{\sqrt{1-v^2/c^2}}.$$

Of course, this gives the right answer $2w/c$ for a clock at rest, that is, $v = 0$.

This means that Jack sees Jill's light clock to be going slow—a longer time between clicks—compared to his own identical clock. Obviously, the effect is not dramatic at real railroad speeds. The correction factor is $\sqrt{1-v^2/c^2}$, which differs from 1 by about one part in a trillion even for a bullet train! Nevertheless, the effect is real and can be measured, as we shall discuss later.

It is important to realize that the only reason we chose a *light* clock, as opposed to some other kind of clock, is that its motion is very easy to analyze from a different frame. Jill could have a collection of clocks on the wagon, and would synchronize them all. For example, she could hang her wristwatch right next to the face of the light clock, and observe them together to be sure they always showed the same time. Remember, in her frame her light clock clicks every $2w/c$ seconds, as it is designed to do. Observing this scene from his position beside the track, Jack will see the synchronized light clock and wristwatch next to each other, and, of course, note that the wristwatch is *also* running slow by the factor $\sqrt{1-v^2/c^2}$. In fact, *all* her clocks, including her pulse, are slowed down by this factor according to Jack. Jill is aging more slowly because she's moving!

But this isn't the whole story—we must now turn everything around and look at it from Jill's point of view. *Her inertial frame of reference is just as good as Jack's.* She sees his light clock to be moving at speed v (backwards) so from her point of view *his* light blip takes the longer zigzag path, which means *his clock runs slower than hers.* That is to say, each of them will see the other to have slower clocks, and be aging more slowly. This phenomenon is called *time dilation*. It has been verified in recent years by flying very accurate clocks around the world on jetliners and finding they register less time, by the predicted amount, than identical clocks left on the ground. Time dilation is also very easy to observe in elementary particle physics, as we shall discuss in the next section.

24.4 Fitzgerald Contraction

Consider now the following puzzle: suppose Jill's clock is equipped with a device that stamps a notch on the track once a second. How far apart are the notches? From Jill's point of view, this is pretty easy to answer. She sees the track passing under the wagon at *v* meters per second, so the notches will of course be *v* meters apart. But Jack sees things differently. He sees Jill's clocks to be running slow, so he will see the notches to be stamped on the track at intervals of $1/\sqrt{1-v^2/c^2}$ seconds (so for a relativistic train going at *v* = 0.8*c*, the notches are stamped at intervals of 5/3 = 1.67 seconds). Since Jack agrees with Jill that the relative speed of the wagon and the track is *v*, he will assert the notches are not *v* meters apart, but $v/\sqrt{1-v^2/c^2}$ meters apart, a greater distance. Who is right? It turns out that Jack is right, because the notches are in his frame of reference, so he can wander over to them with a tape measure or whatever, and check the distance. This implies that as a result of her motion, Jill observes the notches to be closer together by a factor $\sqrt{1-v^2/c^2}$ than they would be at rest. This is called the *Fitzgerald contraction*, and applies not just to the notches, but also to the track and to Jack—everything looks somewhat squashed in the direction of motion!

24.5 Experimental Evidence for Time Dilation: Dying Muons

The first clear example of time dilation was provided over fifty years ago by an experiment detecting *muons*. (David H. Frisch and James A. Smith, Measurement of the Relativistic Time Dilation Using Muons, *American Journal of Physics*, **31**, 342, 1963). These particles are produced at the outer edge of our atmosphere by incoming cosmic rays hitting the first traces of air. They are unstable particles, with a "half-life" of 1.5 microseconds (1.5 millionths of a second), which means that if at a given time you have 100 of them, 1.5 microseconds later you will have about 50, 1.5 microseconds after that 25, and so on. Anyway, they are constantly being produced many miles up, and there is a constant rain of them towards the surface of the earth, moving at very close to the speed of light. In 1941, a detector placed near the top of Mount Washington (at 6000 feet above sea level) measured about 570 muons per hour coming in. Now these muons are raining down from above, but dying as they fall, so if we move the detector to a lower altitude we expect it to detect fewer muons because a fraction of those that came down past the 6000 foot level will die before they get to a lower altitude detector. Approximating their speed by that of light, they are raining down at 186,300 miles per second, which turns out to be, conveniently, about 1,000 feet per microsecond. Thus they should reach the 4500 foot level 1.5 microseconds after passing the 6000 foot level, so, if half of them die off in 1.5 microseconds, as claimed above, we should only expect to register about 570/2 = 285 per hour with the same detector at this level. Dropping another 1500 feet, to the 3000 foot level, we expect about 280/2 = 140 per hour, at 1500 feet about 70 per hour, and at ground level about 35 per hour. (We have rounded off some figures a bit, but this is reasonably close to the expected value.)

To summarize: given the known rate at which these raining-down unstable muons decay, and given that 570 per hour hit a detector near the top of Mount Washington, we only expect about 35 per hour to survive down to sea level. In fact, when the detector was brought down to sea level, it detected about 400 per hour! How did they survive? The reason they didn't decay is that *in their frame of reference, much less time had passed*. Their actual speed is about 0.994*c*, corresponding to a time dilation factor of about 9, so in the 6 microsecond trip from the top of Mount Washington to sea level, their clocks register only 6/9 = 0.67 microseconds. In this period of time, only about one-quarter of them decay.

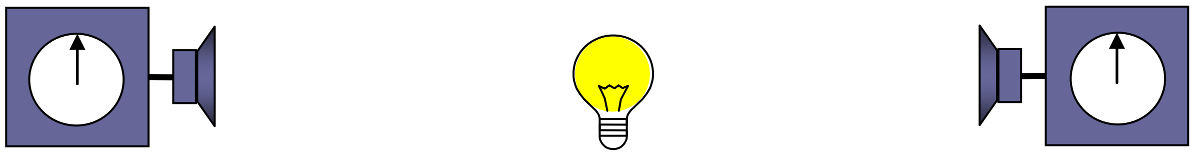
What does this look like from the muon's point of view? How do they manage to get so far in so little time? To them, Mount Washington and the earth's surface are approaching at 0.994c, or about 1,000 feet per microsecond. But in the 0.67 microseconds it takes them to get to sea level, it would seem that to them sea level could only get 670 feet closer, so how could they travel the whole 6000 feet from the top of Mount Washington? The answer is the Fitzgerald contraction. To them, Mount Washington is squashed in a vertical direction (the direction of motion) by a factor of $\sqrt{1-v^2/c^2}$, the same as the time dilation factor, which for the muons is about 9. So, to the muons, Mount Washington is only 670 feet high—this is why they can get down it so fast!

25 Special Relativity: Synchronizing Clocks

Suppose we want to synchronize two clocks that are some distance apart.

We could stand beside one of them and look at the other through a telescope, but we'd have to remember in that case that we are seeing the clock *as it was when the light left it*, and correct accordingly.

Another way to be sure the clocks are synchronized, assuming they are both accurate, is to start them together. How can we do that? We could, for example, attach a photocell to each clock, so when a flash of light reaches the clock, it begins running.



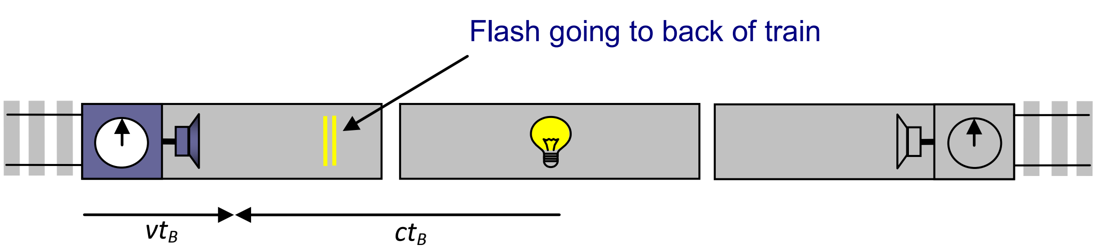
The clocks are triggered when the flash of light from the central bulb reaches the attached photocells.

If, then, we place a flashbulb at the midpoint of the line joining the two clocks, and flash it, the light flash will take the same time to reach the two clocks, so they will start at the same time, and therefore be synchronized.

Let us now put this whole arrangement—the two clocks and the midpoint flashbulb—on a train, and we suppose the train is moving at some speed v to the right, say half the speed of light or so.

Let's look carefully at the clock-synchronizing operation as seen from the ground. In fact, an observer on the ground would say the clocks are *not* synchronized by this operation! The basic reason is that he would see the flash of light from the middle of the train traveling at c *relative to the ground* in each direction, but he would also observe the back of the train coming at v to meet the flash, whereas the front is moving at v away from the bulb, so the light flash must go further to catch up.

In fact, it is not difficult to figure out how much later the flash reaches the front of the train compared with the back of the train, as viewed from the ground. First recall that as viewed from the ground the train has length $L\sqrt{1-v^2/c^2}$.



The train is moving to the right: the central bulb emits a flash of light. Seen from the ground, the part of the flash moving towards the rear travels at c. the rear travels at v to meet it.

Letting t_B be the time it takes the flash to reach the back of the train, it is clear from the figure that

$$vt_B + ct_B = \frac{L}{2}\sqrt{1-\frac{v^2}{c^2}}$$

from which t_B is given by

$$t_B = \frac{1}{c+v}\frac{L}{2}\sqrt{1-\frac{v^2}{c^2}}.$$

In a similar way, the time for the flash of light to reach the front of the train is (as measured by a ground observer)

$$t_F = \frac{1}{c-v}\frac{L}{2}\sqrt{1-\frac{v^2}{c^2}}.$$

Therefore the time difference between the starting of the two clocks, as seen from the ground, is

$$\begin{aligned} t_F - t_B &= \left(\frac{1}{c-v} - \frac{1}{c+v}\right)\frac{L}{2}\sqrt{1-(v/c)^2} \\ &= \left(\frac{2v}{c^2-v^2}\right)\frac{L}{2}\sqrt{1-(v/c)^2} \\ &= \frac{2v}{c^2}\frac{1}{1-(v/c)^2}\frac{L}{2}\sqrt{1-(v/c)^2} \\ &= \frac{vL}{c^2}\frac{1}{\sqrt{1-(v/c)^2}}. \end{aligned}$$

Remember, this is the time difference between the starting of the train's back clock and its front clock as measured by an observer on the ground with clocks on the ground. However, to this observer the clocks on the train appear to tick more slowly, by the factor $\sqrt{1-(v/c)^2}$, so that although the ground observer measures the time interval between the starting of the clock at the back of the train and the clock at the front as $(vL/c^2)\left(1/\sqrt{1-(v/c)^2}\right)$ seconds, he also sees the slow running clock at the back actually reading vL/c^2 seconds at the instant he sees the front clock to start.

To summarize: as seen from the ground, the two clocks on the train (which is moving at *v* in the *x*-direction) are running slowly, registering only $\sqrt{1-(v/c)^2}$ seconds for each second that passes. Equally important, the clocks—which are synchronized by an observer on the train—appear unsynchronized when viewed from the ground, the one at the back of the train reading vL/c^2 seconds ahead of the clock at the front of the train, where *L* is the rest length of the train (the length as measured by an observer on the train).

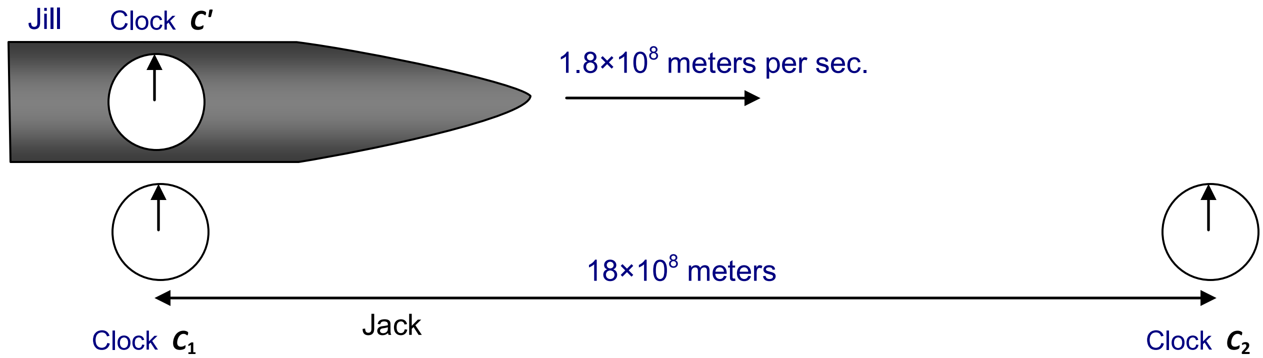
Note that if $L = 0$, that is, if the clocks are together, both the observers on the train and those on the ground will agree that they are synchronized. We need a *distance* between the clocks, as well as relative motion, to get a disagreement about synchronization.

26 Time Dilation: A Worked Example

26.1 "Moving Clocks Run Slow" plus "Moving Clocks Lose Synchronization" plus "Length Contraction" leads to consistency!

The object of this exercise is to show explicitly how it is possible for two observers in inertial frames moving relative to each other at a relativistic speed to each see the other's clocks as running slow and as being unsynchronized, and yet if they both look at the same clock at the same time from the same place (which may be far from the clock), they will *agree* on what time it shows!

Suppose that in Jack's frame we have two synchronized clocks C_1 and C_2 set 18×10^8 meters apart (that's about a million miles, or 6 light-seconds). Jill's spaceship, carrying a clock C', is traveling at 0.6*c*, that is 1.8×10^8 meters per second, parallel to the line C_1C_2, passing close by each clock.



Jill in her relativistic rocket passes Jack's first clock at an instant when both their clocks read zero.

Suppose C' is synchronized with C_1 as they pass, so both read zero.

As measured by Jack the spaceship will take just 10 seconds to reach C_2, since the distance is 6 light seconds, and the ship is traveling at 0.6*c*.

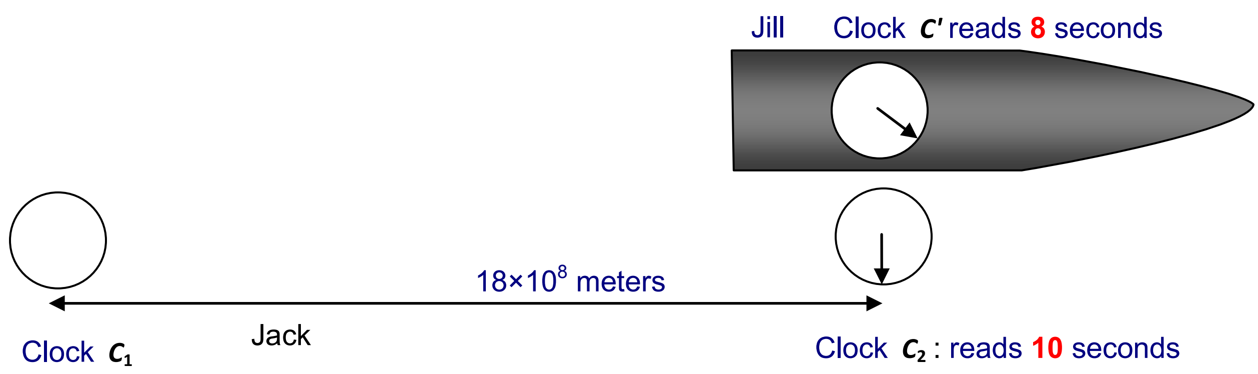
What does clock C' (the clock on the ship) read as it passes C_2?

The time dilation factor

$$\sqrt{1-\left(v^2/c^2\right)}=4/5$$

so C', Jill's clock, will read 8 seconds.

Thus if both Jack and Jill are at C_2 as Jill and her clock C' pass C_2, both will agree that the clocks look like:



As Jill passes Jack's second clock, both see that his clock reads 10 seconds, hers reads 8 seconds.

How, then, can Jill claim that Jack's clocks C_1, C_2 are the ones that are running slow?

To Jill, C_1, C_2 *are* running slow, but remember they are *not synchronized*. To Jill, C_1 is *behind* C_2 by $Lv/c^2 = (L/c)\times(v/c) = 6\times 0.6 = 3.6$ seconds.

Therefore, Jill will conclude that since C_2 reads 10 seconds as she passes it, at that instant C_1 must be registering 6.4 seconds. Jill's own clock reads 8 seconds at that instant, *so she concludes that C_1 is running slow by the appropriate time dilation factor of 4/5*. This is how the change in synchronization makes it possible for both Jack and Jill to see the other's clocks as running slow.

Of course, Jill's assertion that as she passes Jack's second "ground" clock C_2 the first "ground" clock C_1 must be registering 6.4 seconds is not completely trivial to check! After all, that clock is now a million miles away!

Let us imagine, though, that both observers are equipped with Hubble-style telescopes attached to fast acting cameras, so reading a clock a million miles away is no trick.

To settle the argument, the two of them agree that as she passes the second clock, Jack will be stationed at the second clock, and at the instant of her passing they will both take telephoto digital snapshots of the faraway clock C_1, to see what time it reads.

Jack, of course, knows that C_1 is 6 light seconds away, and is synchronized with C_2 which at that instant is reading 10 seconds, so his snapshot must show C_1 to read 4 seconds. That is, looking at C_1 he sees it as it was six seconds ago.

What does Jill's digital snapshot show? It must be identical—two snapshots taken from the same place at the same time must show the same thing! So, Jill *must also* gets a picture of C_1 reading 4 seconds.

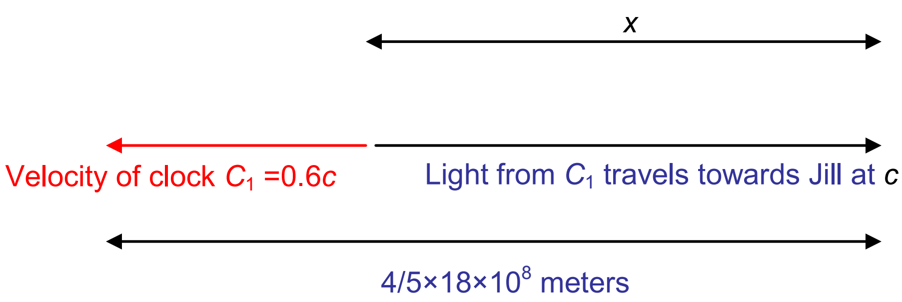
How can she reconcile a picture of the clock reading 4 seconds with her assertion that at the instant she took the photograph the clock was registering 6.4 seconds?

The answer is that she can if she knows her relativity!

First point: length contraction. To Jill, the clock C_1 is actually only 4/5 x 18 x 10^8 meters away (she sees the distance C_1C_2 to be Lorentz contracted!).

Second point: The light didn't even have to go that far! In her frame, the clock C_1 is *moving away*, so the light arriving when she's at C_2 must have left C_1 when it was closer—at distance x in the figure below. The figure shows the light in her frame moving from the clock towards her at speed c, while at the same time the clock itself is moving to the left at 0.6c.

It might be helpful to imagine yourself in her frame of reference, so you are at rest, and to think of clocks C_1 and C_2 as being at the front end and back end respectively of a train that is going past you at speed 0.6c. Then, at the moment the back of the train passes you, you take a picture (through your telescope, of course) of the clock at the front of the train. Obviously, the light from the front clock that enters your camera at that instant left the front clock some time ago. During the time that light traveled towards you at speed c, the front of the train itself was going in the opposite direction at speed 0.6c. But you know the length of the train in your frame is 4/5 x 18 x 10^8 meters, so since at the instant you take the picture the back of the train is passing you, the front of the train must be 4/5 x 18 x 10^8 meters away. Now that distance, 4/5 x 18 x 10^8, is the sum of the distance the light entering your camera traveled plus the distance the train traveled in the same time, that is, (1 + 0.6)/1 times the distance the light traveled.



As Jill passes C_2, she photographs C_1: at that instant, she knows C_1 is $4/5\times18\times10^8$ meters away in her frame, but the light reaching her camera at that moment left C_1 when it was at a distance *x*, not so far away. As the light traveled towards her at speed *c*, C_1 was receding at a speed of 0.6*c*, so the distance $4/5\times18\times10^8$ meters is the *sum* of how far the light traveled towards her and how far the clock traveled away from her, both starting at *x*.

So the image of the first ground clock she sees and records as she passes the second ground clock must have been emitted when the first clock was a distance *x* from her in her frame, where

$$x(1+3/5) = 4/5\times18\times10^8 \text{ meters, so } x = 9\times10^8 \text{ meters.}$$

Having established that the clock image she is seeing as she takes the photograph left the clock when it was only 9×10^8 meters away, that is, 3 light seconds, she concludes that she is observing the first ground clock as it was three seconds ago.

Third point: time dilation. The story so far: she has a photograph of the first ground clock that shows it to be reading 4 seconds. She knows that the light took three seconds to reach her. So, what can she conclude the clock must actually be registering at the instant the photo was taken? If you are tempted to say 7 seconds, you have forgotten that in her frame, the clock is moving at 0.6*c* and hence *runs slow* by a factor 4/5.

Including the time dilation factor correctly, she concludes that in the 3 seconds that the light from the clock took to reach her, the clock itself will have ticked away 3 × 4/5 seconds, or 2.4 seconds.

Therefore, since the photograph shows the clock to read 4 seconds, and she finds the clock must have run a further 2.4 seconds, she deduces that at the instant she took the photograph the clock must actually have been registering 6.4 seconds, which is what she had claimed all along!

The key point of this lecture is that at first it seems impossible for two observers moving relative to each other to both maintain that the other one's clocks run slow. However, by bringing in the other necessary consequences of the theory of relativity, the Lorentz contraction of lengths, and that clocks synchronized in one frame are out of synchronization in another by a precise amount that follows necessarily from the constancy of the speed of light, the whole picture becomes completely consistent!

27 More Relativity: The Train and The Twins

27.1 Einstein's Definition of Common Sense

As you can see from the lectures so far, although Einstein's Theory of Special Relativity solves the problem posed by the Michelson-Morley experiment—the nonexistence of an ether—it is at a price. The simple assertion that the speed of a flash of light is always c in any inertial frame leads to consequences that defy common sense. When this was pointed out somewhat forcefully to Einstein, his response was that common sense is the layer of prejudices put down before the age of eighteen. All our intuition about space, time and motion is based on childhood observation of a world in which no objects move at speeds comparable to that of light. Perhaps if we had been raised in a civilization zipping around the universe in spaceships moving at relativistic speeds, Einstein's assertions about space and time would just seem to be common sense. The real question, from a scientific point of view, is not whether Special Relativity defies common sense, but whether it can be shown to lead to a *contradiction*. If that is so, common sense wins. Ever since the theory was published, people have been writing papers claiming it *does* lead to contradictions. The previous lecture, the worked example on time dilation, shows how careful analysis of an apparent contradiction leads to the conclusion that in fact there was no contradiction after all. In this lecture, we shall consider other apparent contradictions and think about how to resolve them. This is the best way to build up an understanding of Relativity.

27.2 Trapping a Train in a Tunnel

One of the first paradoxes to be aired was based on the Fitzgerald contraction. Recall that any object moving relative to an observer will be seen by that observer to be contracted, foreshortened in the direction of motion by the ubiquitous factor $\sqrt{1-\left(v^2/c^2\right)}$. Einstein lived in Switzerland, a very mountainous country where the railroads between towns often go through tunnels deep in the mountains.

Suppose a train of length L is moving along a straight track at a relativistic speed and enters a tunnel, also of length L. There are bandits inhabiting the mountain above the tunnel. They observe a short train, one of length $L\sqrt{1-\left(v^2/c^2\right)}$, so they wait until this short train is completely inside the tunnel of length *L*, then they close doors at the two ends, and the train is trapped fully inside the mountain. Now look at this same scenario from the point of view of someone on the train. He sees a train of length *L*, approaching a tunnel of length $L\sqrt{1-\left(v^2/c^2\right)}$, so the tunnel is not as long as the train from his viewpoint! What does he think happens when the bandits close both the doors?

27.3 The Tunnel Doors are Closed Simultaneously

The key to understanding what is happening here is that we said the bandits closed the two doors at the ends of the tunnel *at the same time*. How could they arrange to do that, since the doors are far apart? They could use walkie-talkies, which transmit radio waves, or just flash a light down the tunnel, since it's long and straight. Remember, though, that the train is itself going at a speed close to that of light, so they have to be quite precise about this timing! The simplest way to imagine them synchronizing the closings of the two doors is to assume they know the train's timetable, and at a prearranged appropriate time, a light is flashed halfway down the tunnel, and the end doors are closed when the flash of light reaches the ends of the tunnel. Assuming the light was positioned correctly in the middle of the tunnel, that should ensure that the two doors close simultaneously.

27.4 Or are They?

Now consider this door-closing operation from the point of view of someone on the train. Assume he's in an observation car and has incredible eyesight, and there's a little mist, so he actually sees the light flash, and the two flashes traveling down the tunnels towards the two end doors. Of course*, the train is a perfectly good inertial frame*, so he sees these two flashes to be traveling in opposite directions, but *both at c, relative to the train*. Meanwhile, he sees the tunnel itself to be moving rapidly relative to the train. Let us say the train enters the mountain through the "front" door. The observer will see the door at the other end of the tunnel, the "back" door, to be rushing towards him, and rushing to meet the flash of light. Meanwhile, once he's in the tunnel, the front door is receding rapidly behind him, so the flash of light making its way to that door has to travel further to catch it. So the two flashes of light going down the tunnel in opposite directions do not reach the two doors simultaneously as seen from the train.

The concept of simultaneity, events happening at the same time, is not *invariant as we move from one inertial frame to another*. The man on the train sees the back door close first, and, if it is not quickly reopened, the front of the train will pile into it before the front door is closed behind the train.

27.5 Does the Fitzgerald Contraction Work Sideways?

The above discussion is based on Einstein's prediction that objects moving at relativistic speed appear shrunken in their direction of motion. How do we know that they're not shrunken in all three directions, i.e. moving objects maybe keep the same shape, but just get smaller? This can be seen *not* to be the case through a symmetry argument, also due to Einstein. Suppose two trains traveling at equal and opposite relativistic speeds, one north, one south, pass on parallel tracks. Suppose two passengers of equal height, one on each train, are standing leaning slightly out of open windows so that their noses should very lightly touch as they pass each other. Now, if *N* (the northbound passenger) sees *S* as shrunken in height, *N*'s nose will brush against *S*'s forehead, say, and *N* will feel *S*'s nose brush his chin. Afterwards, then, *N* will have a bruised chin (plus nose), *S* a bruised forehead (plus nose). But this is a perfectly symmetric problem, so *S* would say *N* had the bruised forehead, etc. They can both get off their trains at the next stations and get together to check out bruises. They must certainly be symmetrical! The only *consistent symmetrical solution* is given by asserting that *neither* sees the other to shrink in height (i.e. in the direction perpendicular to their relative motion), so that their noses touch each other. Therefore, the Lorentz contraction *only* operates in the direction of motion, objects get squashed but not shrunken.

27.6 How to Give Twins Very Different Birthdays

Perhaps the most famous of the paradoxes of special relativity, which was still being hotly debated in national journals in the fifties, is the twin paradox. The scenario is as follows. One of two twins—the sister—is an astronaut. (Flouting tradition, we will take fraternal rather than identical twins, so that we can use "he" and "she" to make clear which twin we mean). She sets off in a relativistic spaceship to alpha-centauri, four light-years away, at a speed of, say, 0.6*c*. When she gets there, she immediately turns around and comes back. As seen by her brother on earth, her clocks ran slowly by the time dilation factor $\sqrt{1-\left(v^2/c^2\right)}$, so although the round trip took 8/0.6 years = 160 months by earth time, she has only aged by 4/5 of that, or 128 months. So as she steps down out of the spaceship, she is 32 months younger than her twin brother.

But wait a minute—how does this look from *her* point of view? She sees the earth to be moving at 0.6*c*, first away from her then towards her. So she must see her brother's clock on earth to be running slow! So doesn't she expect her brother on earth to be the younger one after this trip?

The key to this paradox is that this situation is not as symmetrical as it looks. The two twins have quite different experiences. The one on the spaceship is *not* in an inertial frame during the initial acceleration *and* the turnaround and braking periods. (To get an idea of the speeds involved, to get to 0.6*c* at the acceleration of a falling stone would take over six months.) Our analysis of how a clock in one inertial frame looks as viewed from another doesn't work during times when one of the frames isn't inertial—in other words, when one is accelerating.

27.7 The Twins Stay in Touch

To try to see just how the difference in ages might develop, let us imagine that the twins stay in touch with each other throughout the trip. Each twin flashes a powerful light once a month, according to their calendars and clocks, so that by counting the flashes, each one can monitor how fast the other one is aging.

The questions we must resolve are:

If the brother, on earth, flashes a light once a month, how frequently, as measured by her clock, does the sister see his light to be flashing as she moves away from earth at speed 0.6*c*?

How frequently does she see the flashes as she is *returning* at 0.6*c*?

How frequently does the brother on earth see the flashes from the spaceship?

Once we have answered these questions, it will be a matter of simple bookkeeping to find how much each twin has aged.

27.8 Figuring the Observed Time between Flashes

To figure out how frequently each twin observes the other's flashes to be, we will use some results from the previous lecture, on time dilation. In some ways, that was a very small scale version of the present problem. Recall that we had two "ground" clocks only one million miles apart. As the astronaut, conveniently moving at 0.6*c*, passed the first ground clock, both that clock and her own clock read zero. As she passed the second ground clock, her own clock read 8 seconds and the *first* ground clock, which she photographed at that instant, she observed to read 4 seconds.

That is to say, after 8 seconds had elapsed on her own clock, constant *observation* of the first ground clock would have revealed it to have registered only 4 seconds. (This effect is compounded of time dilation and the fact that as she moves away, the light from the clock is taking longer and longer to reach her.)

Our twin problem is the same thing, at the same speed, but over a longer time - we conclude that observation of any earth clock from the receding spacecraft will reveal it to be running *at half speed*, so the brother's flashes will be seen at the spacecraft to arrive every two months, by spacecraft time.

Symmetrically, as long as the brother on earth observes his sister's spacecraft to be moving away at 0.6c, he will see light from her flashes to be arriving at the earth every two months by earth time.

To figure the frequency of her brother's flashes observed as she returns towards earth, we have to go back to our previous example and find how the astronaut traveling at 0.6*c* observes time to be registered by the *second* ground clock, the one she's approaching.

We know that as she passes that clock, it reads 10 seconds and her own clock reads 8 seconds. We must figure out what she would have seen that second ground clock to read had she glanced at it through a telescope as she passed the first ground clock, at which point both her own clock and the first ground clock read zero. But at that instant, the reading she would see on the second ground clock must be the same as would be seen by an observer on the ground, standing by the first ground clock and observing the second ground clock through a telescope. Since the ground observer knows both ground clocks are synchronized, and the first ground clock reads zero, and the second is 6 light seconds distant, it must read -6 seconds if observed at that instant.

Hence the astronaut will observe the second ground clock to progress from -6 seconds to +10 seconds during the period that her own clock goes from 0 to 8 seconds. In other words, she sees the clock she is approaching at 0.6*c* to be running *at double speed*.

Finally, back to the twins. During her journey back to earth, the sister will see the brother's light flashing twice a month. (Evidently, the time dilation effect does not fully compensate for the fact that each succeeding flash has less far to go to reach her.)

We are now ready to do the bookkeeping, first, from the sister's point of view.

27.9 What does she see?

At 0.6*c*, she sees the distance to alpha-centauri to be contracted by the familiar $\sqrt{1-\left(v^2/c^2\right)}=0.8$ to a distance of 3.2 light years, which at 0.6c will take her a time 5.333 years, or, more conveniently, 64 months. During the outward trip, then, she will see 32 flashes from home, she will see her brother to age by 32 months.

Her return trip will also take 64 months, during which time she will see 128 flashes, so over the whole trip she will see 128 + 32 = 160 flashes, so she will have seen her brother to age by 160 months or 13 years 4 months.

27.10 What does he see?

As he watches for flashes through his telescope, the stay-at-home brother will see his sister to be aging at half his own rate of aging as long as he sees her to be moving away from him, then aging at twice his rate as he sees her coming back. At first glance, this sounds the same as what she sees—but it isn't! The important question to ask is *when* does he see her turn around? To him, her outward journey of 4 light years' distance at a speed of 0.6*c* takes her 4/0.6 years, or 80 months. BUT he doesn't *see* her turn around until 4 years later, because of the time light takes to get back to earth from alpha-centauri! In other words, he will actually see her aging at half his rate for 80 + 48 = 128 months, during which time he will see 64 flashes.

When he *sees* his sister turn around, she is already more than half way back! Remember, in his frame the whole trip takes 160 months (8 light years at 0.6*c*) so he will only see her aging at twice his rate during the last 160 - 128 = 32 months, during which period he will see all 64 flashes she sent out on her return trip.

Therefore, by counting the flashes of light she transmitted once a month, he will conclude she has aged 128 months on the trip, which by his clock and calendar took 160 months. So when she steps off the spacecraft 32 months younger than her twin brother, neither of them will be surprised!

27.11 The Doppler Effect

The above analysis hinges on the fact that a traveler approaching a flashing light at 0.6*c* will see it flashing at *double* its "natural" rate—the rate observed by someone standing still with the light—and a traveler receding at 0.6*c* from a flashing light will see it to flash at only *half* its natural rate.

This is a particular example of the *Doppler Effect*, first discussed in 1842 by the German physicist Christian Doppler. There is a Doppler Effect for sound waves too. Sound is generated by a vibrating object sending a succession of pressure pulses through the air. These pressure waves are analogous to the flashes of light. If you are approaching a sound source you will encounter the pressure waves more frequently than if you stand still. This means you will hear a higher frequency sound. If the distance between you and the source of sound is increasing, you will hear a lower frequency. This is why the note of a jet plane or a siren goes lower as it passes you. The details of the Doppler Effect for sound are a little different than those for light, because the speed of sound is not the same for all observers—it's 330 meters per second relative to the air.

An important astronomical application of the Doppler Effect is the *red shift*. The light from very distant galaxies is redder than the light from similar galaxies nearer to us. This is because the further away a galaxy is, the faster it is moving away from us, as the Universe expands. The light is redder because red light is low frequency light (blue is high) and we see low frequency light for the same reason that the astronaut receding from earth sees flashes less frequently. In fact, the farthest away galaxies we can see are receding faster than the 0.6*c* of our astronaut!

In the next lecture, we shall brush up on the pre-relativistic concepts of momentum, work and energy to be ready for their relativistic generalizations.

28 Momentum, Work and Energy

28.1 Momentum

At this point, we introduce some further concepts that will prove useful in describing motion. The first of these, *momentum*, was actually introduced by the French scientist and philosopher Descartes before Newton. Descartes' idea is best understood by considering a simple example: think first about someone (weighing say 45 kg) standing motionless on high quality (frictionless) rollerskates on a level smooth floor. A 5 kg medicine ball is thrown directly at her by someone standing in front of her, and only a short distance away, so that we can take the ball's flight to be close to horizontal. She catches and holds it, and because of its impact begins to roll backwards. Notice we've chosen her weight so that, conveniently, she plus the ball weigh just ten times what the ball weighs by itself. What is found on doing this experiment carefully is that after the catch, she plus the ball roll backwards at just one-tenth the speed the ball was moving just before she caught it, so if the ball was thrown at 5 meters per second, she will roll backwards at one-half meter per second after the catch. It is tempting to conclude that the "total amount of motion" is the same before and after her catching the ball, since we end up with ten times the mass moving at one-tenth the speed.

Considerations and experiments like this led Descartes to invent the concept of "momentum", meaning "amount of motion", and to state that for a moving body the momentum was just the product of the mass of the body and its speed. Momentum is traditionally labeled by the letter *p*, so his definition was:

$$\text{momentum} = p = mv$$

for a body having mass *m* and moving at speed *v*. It is then obvious that in the above scenario of the woman catching the medicine ball, total "momentum" is the same before and after the catch. Initially, only the ball had momentum, an amount 5x5 = 25 in suitable units, since its mass is 5kg and its speed is 5 meters per second. After the catch, there is a total mass of 50kg moving at a speed of 0.5 meters per second, so the final momentum is 0.5x50 = 25, the total final amount is equal to the total initial amount. We have just invented these figures, of course, but they reflect what is observed experimentally.

There is however a problem here—obviously one can imagine collisions in which the "total amount of motion", as defined above, is definitely *not* the same before and after. What about two people on roller skates, of equal weight, coming directly towards each other at equal but opposite velocities—and when they meet they put their hands together and come to a complete halt? Clearly in this situation there was plenty of motion before the collision and none afterwards, so the "total amount of motion" definitely doesn't stay the same! In physics language, it is "not conserved". Descartes was hung up on this problem a long time, but was rescued by a Dutchman, Christian Huygens, who pointed out that the problem could be solved in a consistent fashion if one did not insist that the "quantity of motion" be positive.

In other words, *if something moving to the right was taken to have positive momentum, then one should consider something moving to the left to have negative momentum*. With this convention, two people of equal mass coming together from opposite directions at the same speed would have total momentum *zero*, so if they came to a complete halt after meeting, as described above, the total momentum before the collision would be the same as the total after—that is, zero—and momentum *would* be conserved.

Of course, in the discussion above we are restricting ourselves to motions along a single line. It should be apparent that to get a definition of momentum that is conserved in collisions what Huygens really did was to tell Descartes he should replace speed by *velocity* in his definition of momentum. It is a natural extension of this notion to think of momentum as defined by

momentum = mass x *velocity*

in general, so, *since velocity is a vector, momentum is also a vector*, pointing in the same direction as the velocity, of course.

It turns out experimentally that in *any* collision between two objects (where no interaction with third objects, such as surfaces, interferes), the total momentum before the collision is the same as the total momentum after the collision. It doesn't matter if the two objects stick together on colliding or bounce off, or what kind of forces they exert on each other, so conservation of momentum is a very general rule, quite independent of details of the collision.

28.2 Momentum Conservation and Newton's Laws

As we have discussed above, Descartes introduced the concept of momentum, and the general principle of conservation of momentum in collisions, before Newton's time. However, it turns out that conservation of momentum can be deduced from Newton's laws. Newton's laws in principle fully describe all collision-type phenomena, and therefore must contain momentum conservation.

To understand how this comes about, consider first Newton's Second Law relating the acceleration a of a body of mass m with an external force F acting on it:

$F = ma$, or force = mass x acceleration

Recall that acceleration is rate of change of velocity, so we can rewrite the Second Law:

force = mass x rate of change of velocity.

Now, the momentum is mv, mass x velocity. This means for an object having constant mass (which is almost always the case, of course!)

rate of change of momentum = mass x rate of change of velocity.

This means that Newton's Second Law can be rewritten:

force = rate of change of momentum.

Now think of a collision, or any kind of interaction, between two objects *A* and *B*, say. From Newton's Third Law, the force *A* feels from *B* is of equal magnitude to the force *B* feels from *A*, but in the opposite direction. Since (as we have just shown) force = rate of change of momentum, it follows that throughout the interaction process the rate of change of momentum of *A* is exactly opposite to the rate of change of momentum of *B*. In other words, since these are vectors, they are of equal length but pointing in opposite directions. This means that for every bit of momentum *A* gains, *B* gains the negative of that. In other words, *B loses* momentum at exactly the rate *A gains* momentum so their *total* momentum remains the same. But this is true throughout the interaction process, from beginning to end. Therefore, the total momentum at the end must be what it was at the beginning.

You may be thinking at this point: so what? We already know that Newton's laws are obeyed throughout, so why dwell on one special consequence of them? The answer is that although we know Newton's laws are obeyed, this may not be much use to us in an actual case of two complicated objects colliding, because we may not be able to figure out what the forces are. Nevertheless, we *do* know that momentum will be conserved anyway, so if, for example, the two objects stick together, and no bits fly off, we can find their final velocity just from momentum conservation, without knowing any details of the collision.

28.3 Work

The word "work" as used in physics has a narrower meaning than it does in everyday life. First, it only refers to physical work, of course, and second, something has to be accomplished. If you lift up a box of books from the floor and put it on a shelf, you've done work, as defined in physics, if the box is too heavy and you tug at it until you're worn out but it doesn't move, that doesn't count as work.

Technically, work is done when a force pushes something and the object moves some distance in the direction it's being pushed (pulled is ok, too). Consider lifting the box of books to a high shelf. If you lift the box at a steady speed, the force you are exerting is just balancing off gravity, the weight of the box, otherwise the box would be accelerating. (Of course, initially you'd have to exert a little bit more force to get it going, and then at the end a little less, as the box comes to rest at the height of the shelf.) It's obvious that you will have to do twice as much work to raise a box of twice the weight, so the work done is proportional to the force you exert. It's also clear that the work done depends on how high the shelf is. Putting these together, the definition of work is:

work = force x distance

where only distance traveled in the direction the force is pushing counts. With this definition, carrying the box of books across the room from one shelf to another of equal height doesn't count as work, because even though your arms have to exert a force upwards to keep the box from falling to the floor, you do not move the box in the direction of that force, that is, upwards.

To get a more quantitative idea of how much work is being done, we need to have some units to measure work. Defining work as force x distance, as usual we will measure distance in meters, but we haven't so far talked about units for force. The simplest way to think of a unit of force is in terms of Newton's Second Law, force = mass x acceleration. The natural "unit force" would be that force which, pushing a unit mass (one kilogram) with no friction of other forces present, accelerates the mass at one meter per second per second, so after two seconds the mass is moving at two meters per second, etc. *This unit of force is called one **newton*** (as we discussed in an earlier lecture). Note that a one kilogram mass, when dropped, accelerates downwards at ten meters per second per second. This means that its weight, its gravitational attraction towards the earth, must be equal to ten newtons. From this we can figure out that a one newton force equals the weight of 100 grams, just less than a quarter of a pound, a stick of butter.

The downward acceleration of a freely falling object, ten meters per second per second, is often written g for short. (To be precise, g = 9.8 meters per second per second, and in fact varies somewhat over the earth's surface, but this adds complication without illumination, so we shall always take it to be 10.) If we have a mass of m kilograms, say, we know its weight will accelerate it at g if it's dropped, so its weight is a force of magnitude mg, from Newton's Second Law.

Now back to *work*. Since work is force x distance, the natural "unit of work" would be the work done be a force of one newton pushing a distance of one meter. In other words (approximately) lifting a stick of butter three feet. *This unit of work is called one **joule***, in honor of an English brewer.

Finally, it is useful to have a unit for *rate of working*, also called "power". The natural unit of "rate of working" is manifestly one joule per second, and this is called one *watt*. To get some feeling for rate of work, consider walking upstairs. A typical step is eight inches, or one-fifth of a meter, so you will gain altitude at, say, two-fifths of a meter per second. Your weight is, say (put in your own weight here!) 70 kg. (for me) multiplied by 10 to get it in newtons, so it's 700 newtons. The rate of working then is 700 x 2/5, or 280 watts. Most people can't work at that rate for very long. A common English unit of power is the *horsepower*, which is 746 watts.

28.4 Energy

Energy is the ability to do work.

For example, it takes work to drive a nail into a piece of wood—a force has to push the nail a certain distance, against the resistance of the wood. A moving hammer, hitting the nail, can drive it in. A stationary hammer placed on the nail does nothing. The moving hammer has energy—the ability to drive the nail in—because it's moving. This hammer energy is called "*kinetic energy*". Kinetic is just the Greek word for *motion*, it's the root word for cinema, meaning *movies*.

Another way to drive the nail in, if you have a good aim, might be to simply drop the hammer onto the nail from some suitable height. By the time the hammer reaches the nail, it will have kinetic energy. It has this energy, of course, because the force of gravity (its weight) accelerated it as it came down. But this energy didn't come from nowhere. Work had to be done in the first place to lift the hammer to the height from which it was dropped onto the nail. In fact, the work done in the initial lifting, force x distance, is just the weight of the hammer multiplied by the distance it is raised, in joules. But this is exactly the same amount of work as gravity does on the hammer in speeding it up during its fall onto the nail. Therefore, while the hammer is at the top, waiting to be dropped, it can be thought of as storing the work that was done in lifting it, which is ready to be released at any time. This "stored work" is called *potential energy*, since it has the *potential* of being transformed into kinetic energy just by releasing the hammer.

To give an example, suppose we have a hammer of mass 2 kg, and we lift it up through 5 meters. The hammer's weight, the force of gravity, is 20 newtons (recall it would accelerate at 10 meters per second per second under gravity, like anything else) so the work done in lifting it is force x distance = 20 x 5 = 100 joules, since lifting it at a steady speed requires a lifting force that just balances the weight. This 100 joules is now stored ready for use, that is, it is potential energy. Upon releasing the hammer, the potential energy becomes kinetic energy—the force of gravity pulls the hammer downwards through the same distance the hammer was originally raised upwards, so since it's a force of the same size as the original lifting force, the work done on the hammer by gravity in giving it motion is the same as the work done previously in lifting it, so as it hits the nail it has a kinetic energy of 100 joules. We say that the potential energy is transformed into kinetic energy, which is then spent driving in the nail.

We should emphasize that both energy and work are measured in the same units, joules. In the example above, doing work by lifting just adds energy to a body, so-called potential energy, equal to the amount of work done.

From the above discussion, a mass of *m* kilograms has a weight of *mg* newtons. It follows that the work needed to raise it through a height *h* meters is force x distance, that is, weight x height, or *mgh* joules. This is the potential energy.

Historically, this was the way energy was stored to drive clocks. Large weights were raised once a week and as they gradually fell, the released energy turned the wheels and, by a sequence of ingenious devices, kept the pendulum swinging. The problem was that this necessitated rather large clocks to get a sufficient vertical drop to store enough energy, so spring-driven clocks became more popular when they were developed. A compressed spring is just another way of storing energy. It takes work to compress a spring, but (apart from small frictional effects) all that work is released as the spring uncoils or springs back. The stored energy in the compressed spring is often called *elastic potential energy*, as opposed to the *gravitational potential energy* of the raised weight.

28.5 Kinetic Energy

We've given above an explicit way to find the potential energy increase of a mass *m* when it's lifted through a height *h*, it's just the work done by the force that raised it, force x distance = weight x height = *mgh*.

Kinetic energy is created when a force does work accelerating a mass and increases its speed. Just as for potential energy, we can find the kinetic energy created by figuring out how much work the force does in speeding up the body.

Remember that a force only does work if the body the force is acting on moves in the direction of the force. For example, for a satellite going in a circular orbit around the earth, the force of gravity is constantly accelerating the body downwards, but it never gets any closer to sea level, it just swings around. Thus the body does not actually move any distance in the direction gravity's pulling it, and in this case gravity does no work on the body.

Consider, in contrast, the work the force of gravity does on a stone that is simply dropped from a cliff. Let's be specific and suppose it's a one kilogram stone, so the force of gravity is ten newtons downwards. In one second, the stone will be moving at ten meters per second, and will have dropped five meters. The work done at this point by gravity is force x distance = 10 newtons x 5 meters = 50 joules, so this is the kinetic energy of a one kilogram mass going at 10 meters per second. How does the kinetic energy increase with speed? Think about the situation after 2 seconds. The mass has now increased in speed to twenty meters per second. It has fallen a total distance of twenty meters (average speed 10 meters per second x time elapsed of

2 seconds). So the work done by the force of gravity in accelerating the mass over the first two seconds is force x distance = 10 newtons x 20 meters = 200 joules.

So we find that the kinetic energy of a one kilogram mass moving at 10 meters per second is 50 joules, moving at 20 meters per second it's 200 joules. It's not difficult to check that after three seconds, when the mass is moving at 30 meters per second, the kinetic energy is 450 joules. The essential point is that the speed increases linearly with time, but the work done by the constant gravitational force depends on how far the stone has dropped, and that goes as the square of the time. Therefore, the kinetic energy of the falling stone depends on the square of the time, and that's the same as depending on the square of the velocity. For stones of different masses, the kinetic energy at the same speed will be proportional to the mass (since weight is proportional to mass, and the work done by gravity is proportional to the weight), so using the figures we worked out above for a one kilogram mass, we can conclude that for a mass of m kilograms moving at a speed v the kinetic energy must be:

$$\textbf{kinetic energy} = \tfrac{1}{2}mv^2$$

Exercises for the reader: both momentum and kinetic energy are in some sense measures of the amount of motion of a body. How do they differ?

Can a body change in momentum without changing in kinetic energy?

Can a body change in kinetic energy without changing in momentum?

Suppose two lumps of clay of equal mass traveling in opposite directions at the same speed collide head-on and stick to each other. Is momentum conserved? Is kinetic energy conserved?

As a stone drops off a cliff, both its potential energy and its kinetic energy continuously change. How are these changes related to each other?

29 Adding Velocities: A Walk on the Train

29.1 The Formula

If I walk from the back to the front of a train at 3 m.p.h., and the train is traveling at 60 m.p.h., then common sense tells me that my speed relative to the ground is 63 m.p.h. As we have seen, this obvious truth, the simple addition of velocities, follows from the Galilean transformations. Unfortunately, it can't be quite right for high speeds! We know that for a flash of light going from the back of the train to the front, the speed of the light relative to the ground is exactly the same as its speed relative to the train, not 60 m.p.h. different. Hence it is necessary to do a careful analysis of a fairly speedy person moving from the back of the train to the front as viewed from the ground, to see how velocities *really* add.

We consider our standard train of length *L* moving down the track at steady speed *v*, and equipped with synchronized clocks at the back and the front. The walker sets off from the back of the train when that clock reads zero. Assuming a steady walking speed of *u* meters per second (relative to the train, of course), the walker will see the front clock to read *L/u* seconds on arrival there.

How does this look from the ground? Let's assume that at the instant the walker began to walk from the clock at the back of the train, the back of the train was passing the ground observer's clock, and both these clocks (one on the train and one on the ground) read zero. The ground observer sees the walker reach the clock at the front of the train at the instant that clock reads L/u (this is in agreement with what is observed *on* the train—two simultaneous events *at the same place* are simultaneous to all observers), but at this same instant, the ground observer says the train's *back* clock, where the walker began, reads $L/u + Lv/c^2$. (This follows from our previously established result that two clocks synchronized in one frame, in which they are L apart, will be out of synchronization in a frame in which they are moving at v along the line joining them by a time Lv/c^2.)

Now, how much time elapses as measured by the ground observer's clock during the walk? At the instant the walk began, the ground observer saw the clock at the back of the train (which was right next to him) to read zero. At the instant the walk ended, the ground observer would say that clock read $L/u + Lv/c^2$, from the paragraph above. But the ground observer would see that clock to be running slow, by the usual time dilation factor: so he would measure the time of the walk on his own clock to be:

$$\frac{L/u + Lv/c^2}{\sqrt{1-\left(v^2/c^2\right)}}.$$

How *far* does the walker move as viewed from the ground? In the time t_W, the train travels a distance vt_W, so the walker moves this distance plus the length of the train. Remember that the train is contracted as viewed from the ground! It follows that the distance covered *relative to the ground* during the walk is:

$$\begin{aligned} d_W &= vt_W + L\sqrt{1-\left(v^2/c^2\right)} \\ &= v\frac{L/u + Lv/c^2}{\sqrt{1-\left(v^2/c^2\right)}} + L\sqrt{1-\left(v^2/c^2\right)} \\ &= \frac{vL/u + Lv^2/c^2 + L - L\left(v^2/c^2\right)}{\sqrt{1-\left(v^2/c^2\right)}} \\ &= \frac{L\left(1+v/u\right)}{\sqrt{1-\left(v^2/c^2\right)}}. \end{aligned}$$

The walker's *speed* relative to the ground is simply d_W/t_W, easily found from the above expressions:

$$\frac{d_W}{t_W} = \frac{1+v/u}{1/u+v/c^2} = \frac{u+v}{1+uv/c^2}.$$

This is the appropriate formula for adding velocities. Note that it gives the correct answer, *u* + *v*, in the low velocity limit, and also if *u* or *v* equals *c*, the sum of the velocities is *c*.

Exercise: Suppose a spaceship is equipped with a series of one-shot rockets, each of which can accelerate the ship to *c/2* from rest. It uses one rocket to leave the solar system (ignore gravity here) and is then traveling at *c/2* (relative to us) in deep space. It now fires its second rocket, keeping the same direction. Find how fast it is moving relative to us. It now fires the third rocket, keeping the same direction. Find its new speed. Can you draw any general conclusions from your results?

29.2 Testing the Addition of Velocities Formula

Actually, the first test of the addition of velocities formula was carried out in the 1850s! Two French physicists, Fizeau and Foucault, measured the speed of light in water, and found it to be *c*/*n*, where *n* is the refractive index of water, about 1.33. (This was the result predicted by the wave theory of light.)

They then measured the speed of light (relative to the ground) in *moving* water, by sending light down a long pipe with water flowing through it at speed *v*. They discovered that the speed relative to the ground was not just $v + c/n$, but had an extra term, $v + c/n - v/n^2$. Their (incorrect) explanation was that the light was a complicated combination of waves in the water and waves in the aether, and the moving water was only partially dragging the aether along with it, so the light didn't get the full speed *v* of the water added to its original speed *c*/*n*.

The true explanation of the extra term is much simpler: velocities don't simply add. To add the velocity v to the velocity c/n, we must use the addition of velocities formula above, which gives the light velocity relative to the ground to be:

$$(v + c/n)/(1 + v/nc)$$

Now, v is much smaller than c or c/n, so $1/(1 + v/nc)$ can be written as $(1 - v/nc)$, giving:

$$(v + c/n)(1 - v/nc)$$

Multiplying this out gives $v + c/n - v/n^2 - v/n \times v/c$, and the last term is smaller than v by a factor v/c, so is clearly negligible.

Therefore, the 1850 experiment looking for "aether drag" in fact confirms the relativistic addition of velocities formula! Of course, there are many other confirmations. For example, any velocity added to c still gives c. Also, it indicates that the speed of light is a speed limit for all objects, a topic we shall examine more carefully in the next lecture.

30 Conserving Momentum: the Relativistic Mass Increase

30.1 Momentum has Direction

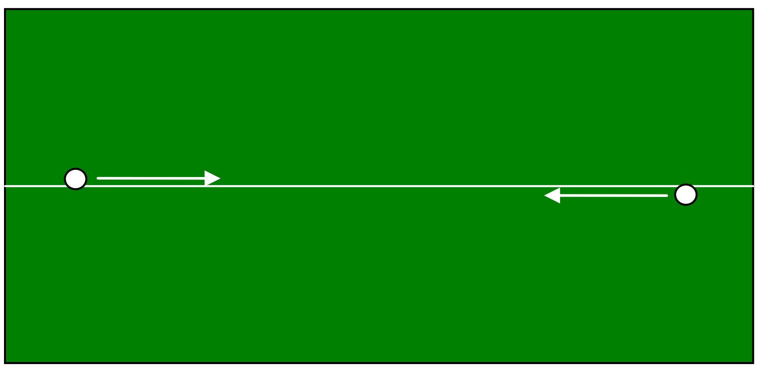
As we discussed in the last lecture, even before Newton formulated his laws, Descartes, with a little help from Huygens, had discovered a deep dynamical truth: in any collision, or in fact in any interaction of any kind, the total amount of "momentum"—a measure of *motion*—always stayed the same. The momentum of a moving object is defined as the product of the mass and the velocity, and so is a *vector*: it has magnitude *and direction*. If you're standing on frictionless skates and you throw a ball, you move backwards: you have momentum equal in magnitude, but *opposite* in direction, to that of the ball, so the total momentum (yours plus the ball's) remains zero. Rockets work the same way, by throwing material out at high speed. They do *not* work by "pushing against the air", they work by pushing against the stuff they're pushing out, just as you push against a ball you're throwing, and it pushes you back, causing your acceleration.

If you still suspect that really rockets push against the air, remember they work just as well in space! In fact, it was widely believed when Goddard, an early American rocketeer (the Goddard Space Flight Center is named after him) talked about rockets in space, he was wasting his time. To quote from a *New York Times* editorial written in 1921: "*Professor Goddard does not know the relation between action and reaction and the need to have something better than a vacuum against which to react. He seems to lack the basic knowledge ladled out daily in our high schools.*" Obviously, *the New York Times* editorial writers of the time lacked the basic knowledge being ladled out in this course!

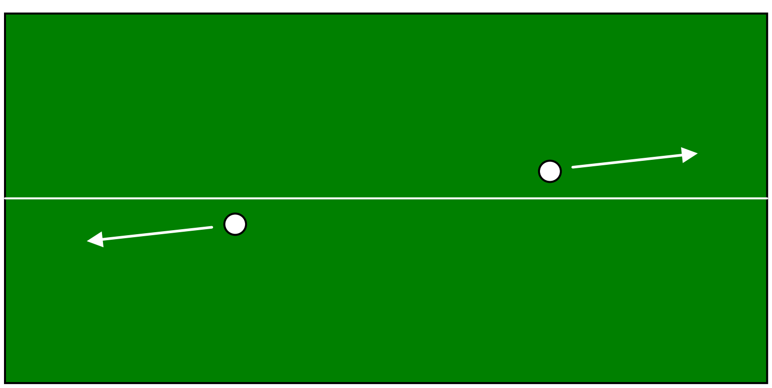
In fact, as we discussed, the conservation of momentum in a collision follows from Newton's laws. However, it is a more general, simpler, concept—it doesn't depend at all on details of the interactions, etc. This simplicity evidently appealed to Einstein, who was convinced that when dynamics was reformulated to include the new ideas about time and space, *conservation of momentum should still hold true in any inertial frame*. This led him to some surprising conclusions, as we shall see.

30.2 Momentum Conservation on the Pool Table

As a warm-up exercise, let us consider conservation of momentum for a collision of two balls on a pool table. We draw a chalk line down the middle of the pool table, and shoot the balls close to, but on opposite sides of, the chalk line from either end, at the same speed, so they will hit in the middle with a glancing blow, which will turn their velocities through a small angle. In other words, if initially we say their (equal magnitude, opposite direction) velocities were parallel to the *x*-direction—the chalk line—then after the collision they will also have equal and opposite small velocities in the *y*-direction. (The *x*-direction velocities will have decreased very slightly).



Balls on pool table moving towards glancing collision



Motion of balls on table after collision

30.3 A Symmetrical Spaceship Collision

Now let us repeat the exercise on a grand scale. Suppose somewhere in space, far from any gravitational fields, we set out a string one million miles long. (It could be between our two clocks in the time dilation experiment). This string corresponds to the chalk line on the pool table. Suppose now we have two identical spaceships approaching each other with equal and opposite velocities parallel to the string from the two ends of the string, aimed so that they suffer a slight glancing collision when they meet in the middle. It is evident from the symmetry of the situation that momentum is conserved in both directions. In particular, the rate at which one spaceship moves away from the string after the collision—its *y*-velocity—is equal and opposite to the rate at which the other one moves away from the string.

But now consider this collision as observed by someone in one of the spaceships, call it *A*. Before the collision, he sees the string moving very fast by the window, say a few meters away. After the collision, he sees the string to be moving away, at, say, 15 meters per second. This is because spaceship *A* has picked up a velocity perpendicular to the string of 15 meters per second. Meanwhile, since this is a completely symmetrical situation, an observer on spaceship *B* would certainly deduce that her spaceship was moving away from the string at 15 meters per second as well.

30.4 Just How Symmetrical Is It?

The crucial question is: *how fast does an observer in spaceship A see spaceship B to be moving away from the string?* Let us suppose that relative to spaceship *A*, spaceship *B* is moving away (in the *x*-direction) at 0.6*c*. First, recall that distances perpendicular to the direction of motion are not Lorentz contracted. Therefore, when the observer in spaceship *B* says she has moved 15 meters further away from the string in a one second interval, the observer watching this movement from spaceship *A* will agree on the 15 meters—but disagree on the one second! He will say her clocks run slow, so as measured by his clocks 1.25 seconds will have elapsed as she moves 15 meters in the *y*-direction.

It follows that, as a result of time dilation, this collision as viewed from spaceship *A* does *not* cause equal and opposite velocities for the two spaceships in the *y*-direction. Initially, both spaceships were moving parallel to the *x*-axis, there was *zero* momentum in the *y*-direction. So how can we argue there is zero total momentum in the *y*-direction *after* the collision, when the identical spaceships do *not* have equal and opposite velocities?

30.5 Einstein Rescues Momentum Conservation

Einstein was so sure that momentum conservation must always hold that he rescued it with a bold hypothesis: the mass of an object must depend on its speed! In fact, the mass must increase with speed in just such a way as to cancel out the lower *y*-direction velocity resulting from time dilation. That is to say, if an object at rest has a mass *M*, moving at a speed *v* it will have a mass $M/\sqrt{1-\left(v^2/c^2\right)}$. Note that this is an undetectably small effect at ordinary speeds, but as an object approaches the speed of light, the mass increases without limit!

30.6 Mass Really *Does* Increase with Speed

Deciding that masses of objects must depend on speed like this seems a heavy price to pay to rescue conservation of momentum! However, it is a prediction that is not difficult to check by experiment. The first confirmation came in 1908, measuring the mass of fast electrons in a vacuum tube. In fact, the electrons in an old style color TV tube are about half a percent heavier than electrons at rest, and this must be allowed for in calculating the magnetic fields used to guide them to the screen.

Much more dramatically, in modern particle accelerators very powerful electric fields are used to accelerate electrons, protons and other particles. It is found in practice that these particles become heavier and heavier as the speed of light is approached, and hence need greater and greater forces for further acceleration. Consequently, the speed of light is a natural absolute speed limit. Particles are accelerated to speeds where their mass is thousands of times greater than their mass measured at rest, usually called the “rest mass”.

30.7 Kinetic Energy and Mass for Very Fast Particles

Let’s think about the kinetic energy of one of these particles traveling close to the speed of light. Recall that in an earlier lecture we found the kinetic energy of an ordinary non-relativistic (i.e. slow moving) mass *m* was $\frac{1}{2}mv^2$. The way we did that was by considering how much work we had to do to raise it through a certain height: we had to exert a force equal to its weight *W* to lift it through height *h*, the total work done, or energy expended, being force x distance, *Wh*. As it fell back down, the force of gravity, *W*, did an exactly equal amount of work *Wh* on the falling object, but this time the work went into accelerating the object, to give it kinetic energy. Since we know how fast falling objects pick up speed, we were able to conclude that the kinetic energy was $\frac{1}{2}mv^2$. (For details, see the previous lecture.)

More generally, we could have accelerated the mass with any constant force *F*, and found the work done by the force (force x distance) to get it to speed *v* from a standing start. The kinetic energy of the mass, $E = \frac{1}{2}mv^2$, is exactly equal to the work done by the force in bringing the mass up to that speed. (It can be shown in a similar way that if a force is applied to a particle already moving at speed *u*, say, and it is accelerated to speed *v*, the work necessary is $\frac{1}{2}mv^2 - \frac{1}{2}mu^2$.)

It is interesting to try to repeat the exercise for a particle moving *very close to the speed of light*, like the particles in the accelerators mentioned in the previous paragraph. Newton's Second Law, in the form

Force = rate of change of momentum

is still true, but *close to the speed of light the speed changes negligibly as the force continues to work*—instead, the *mass* increases! Therefore, we can write to an excellent approximation,

Force = (rate of change of mass) x c

where as usual c is the speed of light. To get more specific, suppose we have a constant force F pushing a particle. At some instant, the particle has mass M, and speed extremely close to c. One second later, since the force is continuing to work on the particle, and thus increase its momentum from Newton's Second Law, the particle will have mass $M+m$ say, where m is the increase in mass as a result of the work done by the force.

What is the increase in the kinetic energy E of the particle during that one second period? By exact analogy with the non-relativistic case reviewed above, it is just the work done by the force during that period. Now, since the mass of the particle changes by m in one second, m is also the *rate of change* of mass. Therefore, from Newton's Second Law in the form

Force = (rate of change of mass) x c,

we can write

Force = mc.

The *increase in kinetic energy E over the one second period is just the work done by the force*,

force x distance.

Since the particle is moving essentially at the speed of light, the *distance* the force acts over in the one-second period is just c meters, $c = 3\times10^8$.

So the total work the force does in that second is force x distance = $mc\times c = mc^2$.

Hence the relationship between the increase in mass of the relativistic particle and its increase in kinetic energy is:

$$E = mc^2$$

30.8 Kinetic Energy and Mass for Slow Particles

Recall that to get Newton's Laws to be true in all inertial frames, we had to assume an increase of mass with speed by the factor $1/\sqrt{1-\left(v^2/c^2\right)}$. This implies that even a slow-moving object has a tiny increase in mass when it moves!

How does that tiny increase relate to the kinetic energy? Consider a mass *M*, moving at speed *v*, much less than the speed of light. Its kinetic energy $E = \frac{1}{2}Mv^2$, as discussed above. Its mass is $M/\sqrt{1-\left(v^2/c^2\right)}$, which we can write as *M* + *m*. What is *m*?

Since we're talking about speeds we are familiar with, like a jet plane, where *v*/*c* is really small, we can use some simple mathematical tricks to make things easier.

The first one is a good approximation for the square root of 1 – *x* when *x* is a lot less than one:

$$\sqrt{1-x} \cong 1-\tfrac{1}{2}x \text{ for } x \ll 1.$$

You can easily check this with your calculator: try $x = \frac{1}{100}$, you find $\sqrt{\frac{99}{100}} = 0.994987...$ which is extremely close to $1-\frac{1}{2}\frac{1}{100} = 0.995$!

The next approximation is

$$\frac{1}{1-x} \cong 1+x \text{ for } x \ll 1.$$

This is also easy to check: again take $x = \frac{1}{100}$: $\dfrac{1}{1-x} = \dfrac{1}{\frac{99}{100}} = \dfrac{100}{99} = 1.01010\ldots$, and $1+x = \frac{101}{100} = 1.01$.

Using these approximations with $x = v/c$, we can approximate $\sqrt{1-\left(v^2/c^2\right)}$ as $1-\frac{1}{2}\left(v^2/c^2\right)$, and then $1/\left(1-\frac{1}{2}\left(v^2/c^2\right)\right)$ as $1+\frac{1}{2}\left(v^2/c^2\right)$.

This means the total mass at speed *v*

$$\frac{M}{\sqrt{1-\left(v^2/c^2\right)}} \cong M\left(1+\tfrac{1}{2}\left(v^2/c^2\right)\right)$$

and writing this as $M + m$, we see the mass increase m equals ½ Mv^2/c^2.

This means that—again—the mass increase m is related to the kinetic energy E by $E = mc^2$.

In fact, it is not difficult to show, using a little calculus, that over the whole range of speed from zero to as close as you like to the speed of light, a moving particle experiences a mass increase related to its kinetic energy by $E = mc^2$. To understand why this isn't noticed in everyday life, try an example, such as a jet airplane weighing 100 tons moving at 2,000mph. 100 tons is 100,000 kilograms, 2,000mph is about 1,000 meters per second. That's a kinetic energy $\frac{1}{2}Mv^2$ of $\frac{1}{2}\times 10^{11}$joules, but the corresponding mass change of the airplane down by the factor c^2, 9×10^{16}, giving an actual mass increase of about half a milligram, not too easy to detect!

30.9 $E = mc^2$

We have seen above that when a force does work accelerating a body to give it kinetic energy, the mass of the body increases by an amount equal to the total work done by the force, the energy E transferred, divided by c^2. What about when a force does work on a body that is *not* speeding it up, so there is no increase in kinetic energy? For example, what if I just lift something at a steady rate, giving it potential energy? It turns out that in this case, too, there is a mass increase given by $E = mc^2$, of course unmeasurably small for everyday objects.

However, this *is* a measurable and important effect in nuclear physics. For example, the helium atom has a nucleus which has two protons and two neutrons bound together very tightly by a strong nuclear attraction force. If sufficient outside force is applied, this can be separated into two "heavy hydrogen" nuclei, each of which has one proton and one neutron. A lot of outside energy has to be spent to achieve this separation, and it is found that the total mass of the two heavy hydrogen nuclei is measurably (about half a percent) *heavier* than the original helium nucleus. This extra mass, multiplied by c^2, is just equal to the energy needed to split the helium nucleus into two. Even more important, this energy can be recovered by letting the two heavy hydrogen nuclei collide and join to form a helium nucleus again. (They are both electrically charged positive, so they repel each other, and must come together fairly fast to overcome this repulsion and get to the closeness where the much stronger nuclear attraction kicks in.) This is the basic power source of the hydrogen bomb, and of the sun.

It turns out that all forms of energy, kinetic and different kinds of potential energy, have associated mass given by $E = mc^2$. For nuclear reactions, the mass change is typically of order one thousandth of the total mass, and readily measurable. For chemical reactions, the change is of order a billionth of the total mass, and not currently measurable.

31 General Relativity

31.1 Einstein's Parable

In Einstein's little book *Relativity: the Special and the General Theory*, he introduces general relativity with a parable. He imagines going into deep space, far away from gravitational fields, where any body moving at steady speed in a straight line will continue in that state for a very long time. He imagines building a space station out there - in his words, "a spacious chest resembling a room with an observer inside who is equipped with apparatus." Einstein points out that there will be no gravity, the observer will tend to float around inside the room.

But now a rope is attached to a hook in the middle of the lid of this "chest" and an unspecified "being" pulls on the rope with a constant force. The chest and its contents, including the observer, accelerate "upwards" at a constant rate. How does all this look to the man in the room? He finds himself moving towards what is now the "floor" and needs to use his leg muscles to stand. If he releases anything, it accelerates towards the floor, and in fact all bodies accelerate at the same rate. If he were a normal human being, he would assume the room to be in a gravitational field, and might wonder why the room itself didn't fall. Just then he would discover the hook and rope, and conclude that the room was suspended by the rope.

Einstein asks: should we just smile at this misguided soul? His answer is no - the observer in the chest's point of view is just as valid as an outsider's. In other words, *being inside* the (from an outside perspective) *uniformly accelerating room is physically equivalent to being in a uniform gravitational field*. This is the basic postulate of general relativity. Special relativity said that all inertial frames were equivalent. General relativity extends this to accelerating frames, and states their equivalence to frames in which there is a gravitational field. This is called the *Equivalence Principle*.

The acceleration could also be used to cancel an existing gravitational field—for example, inside a freely falling elevator passengers are weightless, conditions are equivalent to those in the unaccelerated space station in outer space.

It is important to realize that this equivalence between a gravitational field and acceleration is only possible because the gravitational mass is exactly equal to the inertial mass. There is no way to cancel out electric fields, for example, by going to an accelerated frame, since many different charge to mass ratios are possible.

As physics has developed, the concept of fields has been very valuable in understanding how bodies interact with each other. We visualize the electric field lines coming out from a charge, and know that something is there in the space around the charge which exerts a force on another charge coming into the neighborhood. We can even compute the energy density stored in the electric field, locally proportional to the square of the electric field intensity. It is tempting to think that the gravitational field is quite similar—after all, it's another inverse square field. Evidently, though, this is not the case. If by going to an accelerated frame the gravitational field can be made to vanish, at least locally, it cannot be that it stores energy in a simply defined local way like the electric field.

We should emphasize that going to an accelerating frame can only cancel a *constant* gravitational field, of course, so there is no accelerating frame in which the whole gravitational field of, say, a massive body is zero, since the field necessarily points in different directions in different regions of the space surrounding the body.

31.2 Some Consequences of the Equivalence Principle

Consider a freely falling elevator near the surface of the earth, and suppose a laser fixed in one wall of the elevator sends a pulse of light horizontally across to the corresponding point on the opposite wall of the elevator. Inside the elevator, where there are no fields present, the environment is that of an inertial frame, and the light will certainly be observed to proceed directly across the elevator. Imagine now that the elevator has windows, and an outsider at rest relative to the earth observes the light. As the light crosses the elevator, the elevator is of course accelerating downwards at g, so since the flash of light will hit the opposite elevator wall at precisely the height relative to the elevator at which it began, the outside observer will conclude that the flash of light also accelerates downwards at g. In fact, the light could have been emitted at the instant the elevator was released from rest, so we must conclude that light falls in an initially parabolic path in a constant gravitational field. Of course, the light is traveling very fast, so the curvature of the path is small! Nevertheless, *the Equivalence Principle forces us to the conclusion that the path of a light beam is bent by a gravitational field*.

The curvature of the path of light in a gravitational field was first detected in 1919, by observing stars very near to the sun during a solar eclipse. The deflection for stars observed very close to the sun was 1.7 seconds of arc, which meant measuring image positions on a photograph to an accuracy of hundredths of a millimeter, quite an achievement at the time.

One might conclude from the brief discussion above that a light beam in a gravitational field follows the same path a Newtonian particle would if it moved at the speed of light. This is true in the limit of small deviations from a straight line in a constant field, but is not true even for small deviations for a spatially varying field, such as the field near the sun the starlight travels through in the eclipse experiment mentioned above. We could try to construct the path by having the light pass through a series of freely falling (fireproof!) elevators, all falling towards the center of the sun, but then the elevators are accelerating relative to each other (since they are all falling along *radii*), and matching up the path of the light beam through the series is tricky. If it is done correctly (as Einstein did) it turns out that the angle the light beam is bent through is twice that predicted by a naïve Newtonian theory.

What happens if we shine the pulse of light vertically *down* inside a freely falling elevator, from a laser in the center of the ceiling to a point in the center of the floor? Let us suppose the flash of light leaves the ceiling at the instant the elevator is released into free fall. If the elevator has height h, it takes time h/c to reach the floor. This means the floor is moving downwards at speed gh/c when the light hits.

Question: Will an observer on the floor of the elevator see the light as Doppler shifted?

The answer has to be no, because inside the elevator, by the Equivalence Principle, conditions are identical to those in an inertial frame with no fields present. There is nothing to change the frequency of the light. This implies, however, that to an outside observer, stationary in the earth's gravitational field, the frequency of the light *will* change. This is because he will agree with the elevator observer on what was the initial frequency f of the light as it left the laser in the ceiling (the elevator was at rest relative to the earth at that moment) so if the elevator operator maintains the light had the same frequency f as it hit the elevator floor, which is moving at gh/c relative to the earth at that instant, the earth observer will say the light has frequency $f(1 + v/c) = f(1+gh/c^2)$, using the Doppler formula for very low speeds.

We conclude from this that light shining downwards in a gravitational field is shifted to a higher frequency. Putting the laser in the elevator floor, it is clear that light shining upwards in a gravitational field is red-shifted to lower frequency. Einstein suggested that this prediction could be checked by looking at characteristic spectral lines of atoms near the surfaces of very dense stars, which should be red-shifted compared with the same atoms observed on earth, and this was confirmed. This has since been observed much more accurately. An amusing consequence, since the atomic oscillations which emit the radiation are after all just accurate clocks, is that *time passes at different rates at different altitudes*. The US atomic standard clock, kept at 5400 feet in Boulder, gains 5 microseconds per year over an identical clock almost at sea level in the Royal Observatory at Greenwich, England. Both clocks are accurate to one microsecond per year. This means you would age more slowly if you lived on the surface of a planet with a large gravitational field. Of course, it might not be very comfortable.

31.3 General Relativity and the Global Positioning System

Despite what you might suspect, the fact that time passes at different rates at different altitudes has significant practical consequences. An important *everyday* application of general relativity is the Global Positioning System. A GPS unit finds out where it is by detecting signals sent from orbiting satellites at precisely timed intervals. If all the satellites emit signals simultaneously, and the GPS unit detects signals from four different satellites, there will be three relative time delays between the signals it detects. The signals themselves are encoded to give the GPS unit the precise position of the satellite they came from at the time of transmission. With this information, the GPS unit can use the speed of light to translate the detected time delays into distances, and therefore compute its own position on earth by triangulation.

But this has to be done very precisely! Bearing in mind that the speed of light is about one foot per nanosecond, an error of 100 nanoseconds or so could, for example, put an airplane off the runway in a blind landing. This means the clocks in the satellites timing when the signals are sent out must certainly be accurate to 100 nanoseconds a day. That is one part in 10^{12}. It is easy to check that both the special relativistic time dilation correction from the speed of the satellite, and the general relativistic gravitational potential correction are much greater than that, so the clocks in the satellites must be corrected appropriately. (The satellites go around the earth once every twelve hours, which puts them at a distance of about four earth radii. The calculations of time dilation from the speed of the satellite, and the clock rate change from the gravitational potential, are left as exercises for the student.) For more details, see the lecture by Neil Ashby here.

In fact, Ashby reports that when the first Cesium clock was put in orbit in 1977, those involved were sufficiently skeptical of general relativity that the clock was not corrected for the gravitational redshift effect. But—just in case Einstein turned out to be right—the satellite was equipped with a synthesizer that could be switched on if necessary to add the appropriate relativistic corrections. After letting the clock run for three weeks with the synthesizer turned off, it was found to differ from an identical clock at ground level by precisely the amount predicted by special plus general relativity, limited only by the accuracy of the clock. This simple experiment verified the predicted gravitational redshift to about one percent accuracy! The synthesizer was turned on and left on.