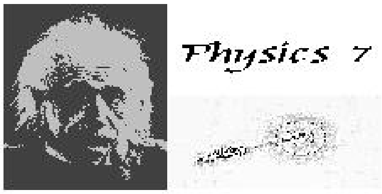
Relativity, Space-Time and Cosmology

Jose Wudka

Contents



Chapter 1

Introduction

1.1 Overview

These notes cover the development of the current scientific concepts of space and time through history, emphasizing the newest developments and ideas. The presentation will be non-mathematical: the concepts will be introduced and explained, but no real calculations will be performed. The various concepts will be introduced in a historical order (whenever possible), this provides a measure of understanding as to how the ideas on which the modern theory of space and time is based were developed. In a real sense this has been an adventure for humanity, very similar to what a child undergoes from the moment he or she first looks at the world to the point he or she understands some of its rules. Part of this adventure will be told here.

Every single culture has had a theory of the formation of the universe and the laws that rule it. Such a system is called a cosmology (from the Greek *kosmos*: world, and *logia* from *legein*: to speak). The first coherent non-religious cosmology was developed during ancient Greece, and much attention will be paid to it after a brief overview of Egyptian and Babyonian comologies [1] The system of the world devised by the Greeks described correctly all phenomena known at the time, and was able to predict most astronomical phenomena with great accuracy. Its most refined version, the Ptolemaic system, survived for more than one thousand years.

[1] A few other comologies will be only summarily described. This is for lack of erudition, Indian, Chinese and American comologies are equally fascinating.

These promising developments came to a stop during the Middle Ages, but took off with a vengeance during the Renaissance; the next landmark in this saga. During this time Copernicus developed his system of the world, where the center of the Universe was the Sun and not the Earth. In the same era Galileo defined and developed the science of mechanics with all its basic postulates; he was also the creator of the idea of relativity, later used by Einstein to construct his Special and General theories.

The next great player was Isaac Newton, who provided a framework for understanding all the phenomena known at the time. In fact most of our daily experience is perfectly well described by Newton's mathematical formulae.

The cosmology based on the ideas of Galileo and Newton reigned supreme up until the end of the 19th century: by this time it became clear that Newton's laws were unable to describe correctly electric and magnetic phenomena. It is here that Einstein enters the field, he showed that the Newtonian approach does not describe correctly situations in which bodies move at speeds close to that of light (in particular it does not describe light accurately). Einstein also provided the generalization of Newton's equations to the realm of such high speeds: the Special Theory of Relativity. Perhaps more importantly, he also demonstrated that certain properties of space and time taken for granted are, in fact, incorrect. We will see, for example, that the concept of two events occurring at the same time in different places is not absolute, but depends on the state of motion of the observer.

Not content with this momentous achievements, Einstein argued that the Special Theory of Relativity itself was inapplicable under certain conditions, for example, near very heavy bodies. He then provided the generalization which encompasses these situations as well: the General Theory of Relativity. This is perhaps the most amazing development in theoretical physics in 300 years: without any experimental motivation, Einstein single handedly developed this modern theory of gravitation and used it to predict some of the most surprising phenomena observed to date. These include the bending of light near heavy bodies and the existence of black holes, massive objects whose gravitational force is so strong it traps all objects, including light.

These notes provide an overview of this saga. From the Greeks and their measuring of the Earth, to Einstein and his description of the universe. But before plunging into this, it is natural to ask how do scientific theories are born, and why are they discarded. Why is it that we believe Einstein is right and Aristotle is wrong? Why is it that we claim that our current understating of the universe is deeper than the one achieved by the early Greeks? The answer to these questions lies in the way in which scientists evaluate the information derived from observations and experiments, and is the subject of the next section.

1.2 The scientific method

> *Science is best defined as a careful, disciplined, logical search for knowledge about any and all aspects of the universe, obtained by examination of the best available evidence and always subject to correction and improvement upon discovery of better evidence. What's left is magic. And it doesn't work.*
>
> *James Randi*

It took a long while to determine how is the world better investigated. One way is to just talk about it (for example Aristotle, the Greek philosopher, stated that males and females have different number of teeth, without bothering to check; he then provided long arguments as to why this is the way things ought to be). This method is unreliable: arguments cannot determine whether a statement is correct, this requires *proofs*.

A better approach is to do experiments and perform careful observations. The results of this approach are universal in the sense that they can be reproduced by any skeptic. It is from these ideas that the *scientific method* was developed. Most of science is based on this procedure for studying Nature.

1.2.1 What is the "scientific method"?

The scientific method is the best way yet discovered for winnowing the truth from lies and delusion. The simple version looks something like this:

1. Observe some aspect of the universe.

2. Invent a tentative description, called a *hypothesis*, that is consistent with what you have observed.

3. Use the hypothesis to make predictions.

4. Test those predictions by experiments or further observations and modify the hypothesis in the light of your results.

5. Repeat steps 3 and 4 until there are no discrepancies between theory and experiment and/or observation.

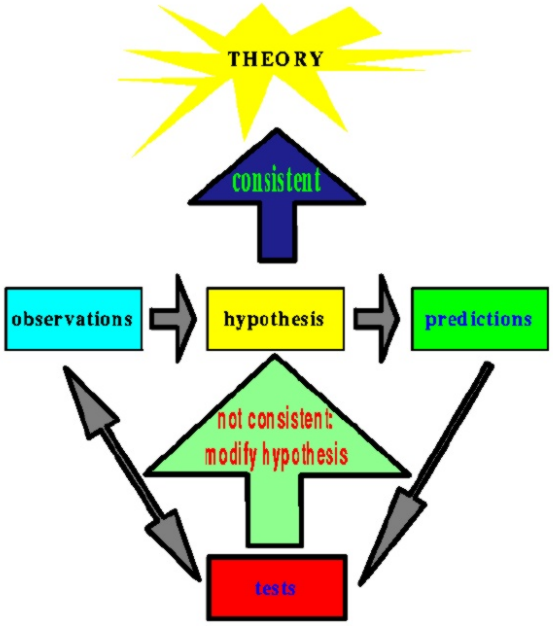


Figure 1.1: Flow diagram describing the scientific method.

When consistency is obtained the hypothesis becomes a *theory* and provides a coherent set of propositions which explain a class of phenomena. A theory is then a framework within which observations are explained and predictions are made.

The scientific method is unprejudiced

The great advantage of the scientific method is that it is unprejudiced: one does not have to believe a given researcher, one can redo the experiment and determine whether his/her results are true or false. The conclusions will hold irrespective of the state of mind, or the religious persuasion, or the state of consciousness of the investigator and/or the subject of the investigation. Faith, defined as [2] *belief that does not rest on logical proof or material evidence*, does not determine whether a scientific theory is adopted or discarded.

A theory is accepted not based on the prestige or convincing powers of the proponent, but on the results obtained through observations and/or ex-

[2] The American Heritage Dictionary (second college edition)

periments which *anyone* can reproduce: the results obtained using the scientific method are repeatable. In fact, most experiments and observations *are* repeated many times (certain experiments are not repeated independently but are repeated as parts of other experiments). If the original claims are not verified the origin of such discrepancies is hunted down and exhaustively studied.

The results obtained using the scientific method are repeatable

When studying the cosmos we cannot perform experiments; all information is obtained from observations and measurements. Theories are then devised by extracting some regularity in the observations and coding this into physical laws.

There is a very important characteristic of a scientific theory or hypothesis which differentiates it from, for example, an act of faith: a theory must be "falsifiable". This means that there must be some experiment or possible discovery that could prove the theory untrue. For example, Einstein's theory of Relativity made predictions about the results of experiments. These experiments could have produced results that contradicted Einstein, so the theory was (and still is) falsifiable.

Every scientific theory must be "falsifiable"

In contrast, the theory that "the moon is populated by little green men who can read our minds and will hide whenever anyone on Earth looks for them, and will flee into deep space whenever a spacecraft comes near" is not falsifiable: these green men are designed so that no one can ever see them. On the other hand, the theory that there are no little green men on the moon is scientific: you can disprove it by catching one. Similar arguments apply to abominable snow-persons, UFOs and the Loch Ness Monster(s?).

A frequent criticism made of the scientific method is that it cannot accommodate anything that has not been proved. The argument then points out that many things thought to be impossible in the past are now everyday realities. This criticism is based on a misinterpretation of the scientific method. When a hypothesis passes the test it is adopted as a theory it correctly explains a range of phenomena it *can*, at any time, be falsified by new experimental evidence. When exploring a new set or phenomena scientists do use existing theories but, since this is a new area of investigation, it is always kept in mind that the old theories might fail to explain the new experiments and observations. In this case new hypotheses are devised and tested until a new theory emerges.

There are many types of "pseudo-scientific" theories which wrap themselves in a mantle of apparent experimental evidence but that, when examined closely, are nothing but statements of faith. The argument [3], cited by

[3] From `http://puffin.ptialaska.net/~svend/award.html`

some creationists, that science is just another kind of faith is a philosophic stance which ignores the trans-cultural nature of science. Science's theory of gravity explains why both creationists and scientists don't float off the earth. All you have to do is jump to verify this theory – no leap of faith required.

1.2.2 What is the difference between a fact, a theory and a hypothesis?

A theory is a conceptual framework that explains existing observations and predicts new ones

In popular usage, a theory is just a vague and fuzzy sort of fact and a hypothesis is often used as a fancy synonym to 'guess'. But to a scientist a theory is a conceptual framework that explains existing observations and predicts new ones. For instance, suppose you see the Sun rise. This is an existing observation which is explained by the theory of gravity proposed by Newton. This theory, in addition to explaining why we see the Sun move across the sky, also explains many other phenomena such as the path followed by the Sun as it moves (as seen from Earth) across the sky, the phases of the Moon, the phases of Venus, the tides, just to mention a few. You can today make a calculation and *predict* the position of the Sun, the phases of the Moon and Venus, the hour of maximal tide, all 200 years from now. The *same* theory is used to guide spacecraft all over the Solar System.

A hypothesis is a working assumption

A hypothesis is a working assumption. Typically, a scientist devises a hypothesis and then sees if it "holds water" by testing it against available data (obtained from previous experiments and observations). If the hypothesis does hold water, the scientist declares it to be a theory.

1.2.3 Truth and proof in science.

Experiments sometimes produce results which cannot be explained with existing theories. In this case it is the job of scientists to produce new theories which replace the old ones. The new theories should explain all the observations and experiments the old theory did *and*, in addition, the new set of facts which lead to their development. One can say that new theories devour and assimilate old ones (see Fig, 1.2). Scientists continually test existing theories in order to probe how far can they be applied.

When a new theory cannot explain new observations it will be (eventually) replaced by a new theory. This does *not* mean that the old ones are "wrong" or "untrue", it only means that the old theory had a limited applicability and could not explain all current data. The only certain thing about currently accepted theories is that they explain all available data, which, if

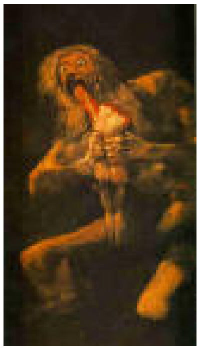


Figure 1.2: Saturn devouring his sons (by F. Goya). A paradigm of how new theories encompass old ones.

course, does not imply that they will explains all future experiments!

In some cases new theories provide not only extensions of old ones, but a completely new insight into the workings of nature. Thus when going from Newton's theory of gravitation to Einstein's our understanding of the nature of space and time was revolutionized. Nonetheless, no matter how beautiful and simple a new theory might be, it must explain the same phenomena the old one did. Even the most beautiful theory can be annihilated by a single ugly fact.

Scientific theories have various degrees of reliability and one can think of them as being on a scale of certainty. Up near the top end we have our theory of gravitation based on a staggering amount of evidence; down at the bottom we have the theory that the Earth is flat. In the middle we have our theory of the origin of the moons of Uranus. Some scientific theories are nearer the top than others, but none of them ever actually reach it.

An extraordinary claim is one that contradicts a fact that is close to the top of the certainty scale and will give rise to a lot of skepticism. So if you are trying to contradict such a fact, you had better have facts available that are even higher up the certainty scale: "extraordinary evidence is needed for an extraordinary claim".

1.2.4 If scientific theories keep changing, where is the Truth?

In 1666 Isaac Newton proposed his theory of gravitation. This was one of the greatest intellectual feats of all time. The theory explained all the observed facts, and made predictions that were later tested and found to be correct within the accuracy of the instruments being used. As far as anyone could see, Newton's theory was "the Truth".

During the nineteenth century, more accurate instruments were used to test Newton's theory, these observations uncovered some slight discrepancies. Albert Einstein proposed his theories of Relativity, which explained the newly observed facts and made more predictions. Those predictions have now been tested and found to be correct within the accuracy of the instruments being used. As far as anyone can see, Einstein's theory is "the Truth".

So how can the Truth change? Well the answer is that it hasn't. The Universe is still the same as it ever was. When a theory is said to be "true" it means that it agrees with all known experimental evidence. But even the best of theories have, time and again, been shown to be incomplete: though they might explain a lot of phenomena using a few basic principles, and even predict many new and exciting results, eventually new experiments (or more precise ones) show a discrepancy between the workings of nature and the predictions of the theory. In the strict sense this means that the theory was not "true" after all; but the fact remains that it is a very good approximation to the truth, at lest where a certain type of phenomena is concerned.

When a theory is said to be "true" it means that it agrees with all known experimental evidence

When an accepted theory cannot explain some new data (which has been confirmed), the researchers working in that field strive to construct a new theory. This task gets increasingly more difficult as our knowledge increases, for the new theory should not only explain the new data, but also all the old one: a new theory has, as its first duty, to devour and assimilate its predecessors.

One other note about truth: science does not make moral judgments. Anyone who tries to draw moral lessons from the laws of nature is on very dangerous ground. Evolution in particular seems to suffer from this. At one time or another it seems to have been used to justify Nazism, Communism, and every other -ism in between. These justifications are all completely bogus. Similarly, anyone who says "evolution theory is evil because it is used to support Communism" (or any other -ism) has also strayed from the path of Logic (and will not live live long nor prosper).

1.2.5 What is Ockham's Razor?

When a new set of facts requires the creation of a new theory the process is far from the orderly picture often presented in books. Many hypothses are proposed, studied, rejected. Researchers discuss their validity (sometimes quite heatedly) proposing experiments which will determine the validity of one or the other, exposing flaws in their least favorite ones, etc. Yet, even when the unfit hypotheses are discarded, several options may remain, in some cases making the exact same predictions, but having very different underlying assumptions. In order to choose among these possible theories a very useful tool is what is called *Ockham's razor*.

Ockham's Razor is the principle proposed by William of Ockham in the fourteenth century: "Pluralitas non est ponenda sine neccesitate", which translates as "entities should not be multiplied unnecessarily".

In many cases this is interpreted as "keep it simple", but in reality the Razor has a more subtle and interesting meaning. Suppose that you have two competing theories which describe the same system, if these theories have different predictions than it is a relatively simple matter to find which one is better: one does experiments with the required sensitivity and determines which one give the most accurate predictions. For example, in Copernicus' theory of the solar system the planets move in circles around the sun, in Kepler's theory they move in ellipses. By measuring carefully the path of the planets it was determined that they move on ellipses, and Copernicus' theory was then replaced by Kepler's.

But there are are theories which have the very same predictions and it is here that the Razor is useful. Consider form example the following two theories aimed at describing the motions of the planets around the sun

- The planets move around the sun in ellipses because there is a force between any of them and the sun which decreases as the square of the distance.

- The planets move around the sun in ellipses because there is a force between any of them and the sun which decreases as the square of the distance. This force is generated by the will of some powerful aliens.

Since the force between the planets and the sun determines the motion of the former and both theories posit the same type of force, the predicted motion of the planets will be identical for both theories. the second theory, however, has additional baggage (the will of the aliens) which is unnecessary for the description of the system.

If one accepts the second theory *solely on the basis that it predicts correctly the motion of the planets* one has also accepted the existence of aliens whose will affect the behavior of things, despite the fact that the presence or absence of such beings is irrelevant to planetary motion (the only relevant item is the type of force). In this instance Ockham's Razor would unequivocally reject the second theory. By rejecting this type of additional irrelevant hypotheses guards against the use of solid scientific results (such as the prediction of planetary motion) to justify unrelated statements (such as the existence of the aliens) which may have dramatic consequences. In this case the consequence is that the way planets move, the reason we fall to the ground when we trip, etc. is due to some powerful alien intellect, that this intellect permeates our whole solar system, it is with us even now...and from here an infinite number of paranoid derivations.

For all we know the solar system is permeated by an alien intellect, but the motion of the planets, which can be explained by the simple idea that there is a force between them and the sun, provides no evidence of the aliens' presence nor proves their absence.

A more straightforward application of the Razor is when we are face with two theories which have the same predictions and the available data cannot distinguish between them. In this case the Razor directs us to study in depth the simplest of the theories. It does *not* guarantee that the simplest theory will be correct, it merely establishes priorities.

A related rule, which can be used to slice open conspiracy theories, is Hanlon's Razor: "Never attribute to malice that which can be adequately explained by stupidity".

1.2.6 How much fraud is there in science?

The picture of scientists politely discussing theories, prposing new ones in view of new data, etc. appears to be completely devoid of any emotions. In fact this is far from the truth, the discussions are very human, even though the bulk of the scientific community will eventually accept a single theory based on it explaining the data and making a series of verified predictions. But before this is achieved, does it happen that researchers fake results or experiments for prestige and/or money? How frequent is this kind of scientific fraud?

In its simplest form this question is unanswerable, since undetected fraud is by definition unmeasurable. Of course there are many known cases of fraud in science. Some use this to argue that all scientific findings (especially those they dislike) are worthless.

This ignores the replication of results which is routinely undertaken by scientists. Any important result will be replicated many times by many different people. So an assertion that (for instance) scientists are lying about carbon-14 dating requires that a great many scientists are engaging in a conspiracy. In fact the existence of known and documented fraud is a good illustration of the self-correcting nature of science. It does not matter (for the progress of science) if a proportion of scientists are fraudsters because any important work they do will not be taken seriously without independent verification.

Also, most scientists are idealists. They perceive beauty in scientific truth and see its discovery as their vocation. Without this most would have gone into something more lucrative. These arguments suggest that undetected fraud in science is both rare and unimportant.

The above arguments are weaker in medical research, where *companies* frequently suppress or distort data in order to support their own products. Tobacco companies regularly produce reports "proving" that smoking is harmless, and drug companies have both faked and suppressed data related to the safety or effectiveness or major products. This type of fraud does not, of course, reflect on the validity of the scientific method.

1.2.7 Are scientists wearing blinkers?

One of the commonest allegations against mainstream science is that its practitioners only see what they expect to see. Scientists often refuse to test fringe ideas because "science" tells them that this will be a waste of time and effort. Hence they miss ideas which could be very valuable.

This is the "blinkers" argument, by analogy with the leather shields placed over horses eyes so that they only see the road ahead. It is often put forward by proponents of new-age beliefs and alternative health.

It is certainly true that ideas from outside the mainstream of science can have a hard time getting established. But on the other hand the opportunity to create a scientific revolution is a very tempting one: wealth, fame and Nobel prizes tend to follow from such work. So there will always be one or two scientists who are willing to look at anything new.

If you have such an idea, remember that the burden of proof is on you. The new theory should explain the existing data, provide new predictions and should be testable; remember that all scientific theories are falsifiable. Read the articles and improve your theory in the light of your new knowledge. Starting a scientific revolution is a long, hard slog. Don't expect it to be easy. If it was, we would have them every week. People putting forward extraordinary claims often refer to Galileo as an example of a great genius being persecuted by the establishment for heretic theories. They claim that the scientific establishment is afraid of being proved wrong, and hence is trying to suppress the truth. This is a classic conspiracy theory. The Conspirators are all those scientists who have bothered to point out flaws in the claims put forward by the researchers. The usual rejoinder to someone who says “They laughed at Columbus, they laughed at Galileo” is to say “But they also laughed at Bozo the Clown”.

1.2.8 Why should we worry?

I have argued that the scientific method provides an excellent guideline for studying the world around us. It is, of course, conceivable that there are other “planes of thought” but their presence and properties, and what may happen in them is a matter of belief.

Through time “alternative” sciences regularly rise their head and are debunked. One might be bothered about their presence since it does say something less than flattering about human psychology. But even if one defends these beliefs on the basis of free speech, one should be aware that they sometimes represent more than idle talk. For example, there is this recent news article

- *ALTERNATIVE MEDICINE: REPORT SEEKS TO TAKE NIH INTO A NEW AGE! What may rank as the most credulous document in medical history was unveiled yesterday in a Senate conference room. Senator Tom Harkin (D-IA), who fathered the 1991 legislation that created the NIH Office of Alternative Medicine, admitted that the program had “gotten off to a slow start” due to opposition from “traditional” medicine. It should soar now; the 420-page report, “Alternative Medicine: Expanding Medical Horizons,” lays out an OAM agenda for research into everything from Lakota medicine wheels to laying on of hands and homeopathic medicines. Homeopathic medicines employ dilutions far beyond the point at which a single molecule would remain, but the water “remembers.” Where does physics fit in? Well, when really weird things happen, like mental healing at a distance, it must be quantum mechanics (Brian Josephson is cited for authority). Medical ethics are not ignored; the possibility of distant organisms being harmed by non-local mental influence is raised, and board certification of mental healers is proposed “to protect consumers from predatory quacks.” An entire chapter is devoted to “Bioelectromagnetics.” This is tricky stuff: “Weak EMF may, at the proper frequency and site of application, produce large effects that are either clinically beneficial or harmful.”* [4]

It truly is amazing that people will even consider this statement. In fact it is not dismissed because it refers to science, but imagine a similar situation

[4] Extracted from “What’s New”, by Robert L. Park (March 3, 1995) produced by The American Physical Society.

where "really important matters" are involved, such as money. suppose a banker were to empty an account and claim that, even though there is no money left, the owner of the account is just as rich because his bank book still "remembers" the balance and that this miraculous memory of wealth past can be used to "cure" the owner's credit-card balance. Without a doubt this banker would end up in jail or in the loony bin.

Various tests using the scientific method have proven the fallacy of the "water with deep memory" theory. Yet these items are seriously considered and sometimes funded by Congress, diverting monies from important programs such as education. In the OAM has had an interesting and controversial history [5], despite this it has a budget of $12 million; in 1993-1994 it dispersed about 10% of this in grants.

This is not a unique occurrence. There are many many claims which use high-sounding scientific jargon; for example J. Randi [6] mentions that the NIH Office of Alternative Medicine has given credence to such claims as a cure for multiple sclerosis (despite the fact that the staff *must* know there is no such thing). When such startling claims are investigated, they are found to be merely ridiculous statements. If you are curious about these I provide a list of WWW sites for your amusement

- A page of links, ranging from free universal energy claims to antigravity, is found in `http://www.padrak.com/ine/SUBJECTS.html`
- Free energy `http://jabi.com/ucsa/` which is exposed in `http://www.voicenet.com/~eric/dennis.html`
- Perpetual motion machines `http://www.overunity.de/finsrud.htm`
- Products that miraculously improve your car's performance `http://widget.ecn.purdue.edu/~feiereis/magic.html`
- Flat Earth Society links (pro and against) `http://www.town.hanna.ab.ca/hemaruka/hemlinks.htm`.

And yes, in case you are wondering, some of these people *are* serious.

It is important to differentiate between these "pseudo-scientific" creations and true science-based developments. Pseudo-science is either not

[5] See for example, `http://www.nas.org/nassnl/2-11.htm`, `http://cyberwarped.com/~gcahf/ncahf/newslett/nl19-2.html`, `http://washingtonpost.com/wp-srv/WPlate/1997-08/10/0971-081097-idx.html`

[6] `http://www.mindspring.com/~anson/randi-hotline/1995/0046.html`

falsifiable or its results cannot be reproduced in a laboratory. If anything like this were to happen to a scientific hypothesis it would be dismissed forthright independently of the, belief, feelings, etc. of the researchers.

Below I present excerpts from an essay by R. Feynman on this same issue [7].

Cargo Cult Science (excerpts)

by Richard Feynman

During the Middle Ages there were all kinds of crazy ideas, such as that a piece of of rhinoceros horn would increase potency. Then a method was discovered for separating the ideas–which was to try one to see if it worked, and if it didn't work, to eliminate it. This method became organized, of course, into science. And it developed very well, so that we are now in the scientific age. It is such a scientific age, in fact, that we have difficulty in understanding how witch doctors could ever have existed, when nothing that they proposed ever really worked–or very little of it did.

But even today I meet lots of people who sooner or later get me into a conversation about UFO's, or astrology, or some form of mysticism, expanded consciousness, new types of awareness, ESP, and so forth. And I've concluded that it's not a scientific world.

Most people believe so many wonderful things that I decided to investigate why they did. And what has been referred to as my curiosity for investigation has landed me in a difficulty where I found so much junk that I'm overwhelmed. First I started out by investigating various ideas of mysticism and mystic experiences. I went into isolation tanks and got many hours of hallucinations, so I know something about that. Then I went to Esalen, which is a hotbed of this kind of thought (it's a wonderful place; you should go visit there). Then I became overwhelmed. I didn't realize how MUCH there was.

⋮

I also looked into extrasensory perception, and PSI phenomena, and the latest craze there was Uri Geller, a man who is supposed to be able to bend keys by rubbing them with his finger. So I went to his hotel room, on his invitation, to see a demonstration of both mind reading and bending keys. He didn't do any mind reading that succeeded; nobody can read my mind, I guess. And my boy held a key and Geller rubbed it, and nothing happened. Then he told us it works better under water, and so you can picture all of us standing in the bathroom with the water turned on and the key under it, and him rubbing the key with his finger. Nothing happened. So I was unable to investigate that phenomenon.

[7] The complete version can be found in the World-Wide-Web at http://www.pd.infn.it/wwwcdf/science.html

But then I began to think, what else is there that we believe? (And I thought then about the witch doctors, and how easy it would have been to check on them by noticing that nothing really worked.) So I found things that even more people believe, such as that we have some knowledge of how to educate. There are big schools of reading methods and mathematics methods, and so forth, but if you notice, you'll see the reading scores keep going down–or hardly going up–in spite of the fact that we continually use these same people to improve the methods. There's a witch doctor remedy that doesn't work. It ought to be looked into; how do they know that their method should work? Another example is how to treat criminals. We obviously have made no progress–lots of theory, but no progress–in decreasing the amount of crime by the method that we use to handle criminals.

Yet these things are said to be scientific. We study them. And I think ordinary people with common sense ideas are intimidated by this pseudo-science. A teacher who has some good idea of how to teach her children to read is forced by the school system to do it some other way–or is even fooled by the school system into thinking that her method is not necessarily a good one. Or a parent of bad boys, after disciplining them in one way or another, feels guilty for the rest of her life because she didn't do "the right thing," according to the experts.

So we really ought to look into theories that don't work, and science that isn't science.

I think the educational and psychological studies I mentioned are examples of what I would like to call cargo cult science. In the South Seas there is a cargo cult of people. During the war they saw airplanes with lots of good materials, and they want the same thing to happen now. So they've arranged to make things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head to headphones and bars of bamboo sticking out like antennas–he's the controller–and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things *cargo cult science*, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.

Now it behooves me, of course, to tell you what they're missing. But it would be just about as difficult to explain to the South Sea islanders how they have to arrange things so that they get some wealth in their system. It is not something simple like telling them how to improve the shapes of the earphones. But there is one feature I notice that is generally missing in cargo cult science. That is the idea that we all hope you have learned in studying science in school–we never say explicitly what this is, but just hope that you catch on by all the examples of scientific investigation. It is interesting, therefore, to bring it out now and speak of it explicitly. It's a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty–a kind of leaning over backwards. For example, if you're doing an experiment, you should report everything that you think might make it invalid–not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked–to make sure the other fellow can tell they have been eliminated.

Details that could throw doubt on your interpretation must be given, if you know them. You must do the best you can–if you know anything at all wrong, or possibly wrong–to explain it. If you make a theory, for example, and advertise it, or put it out, then you must also put down all the facts that disagree with it, as well as those that agree with it. There is also a more subtle problem. When you have put a lot of ideas together to make an elaborate theory, you want to make sure, when explaining what it fits, that those things it fits are not just the things that gave you the idea for the theory; but that the finished theory makes something else come out right, in addition.

In summary, the idea is to give all of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another.

The easiest way to explain this idea is to contrast it, for example, with advertising. Last night I heard that Wesson oil doesn't soak through food. Well, that's true. It's not dishonest; but the thing I'm talking about is not just a matter of not being dishonest; it's a matter of scientific integrity, which is another level. The fact that should be added to that advertising statement is that no oils soak through food, if operated at a certain temperature. If operated at another temperature, they all will–including Wesson oil. So it's the implication which has been conveyed, not the fact, which is true, and the difference is what we have to deal with.

We've learned from experience that the truth will come out. Other experimenters will repeat your experiment and find out whether you were wrong or right. Nature's phenomena will agree or they'll disagree with your theory. And, although you may gain some temporary fame and excitement, you will not gain a good reputation as a scientist if you haven't tried to be very careful in this kind of work. And it's this type of integrity, this kind of care not to fool yourself, that is missing to a large extent in much of the research in "alternative science".

I would like to add something that's not essential to the science, but something I kind of believe, which is that you should not fool the layman when you're talking as a scientist. I'm talking about a specific, extra type of integrity that is not lying, but bending over backwards to show how you're maybe wrong, that you ought to have when acting as a scientist. And this is our responsibility as scientists, certainly to other scientists, and I think to laymen.

For example, I was a little surprised when I was talking to a friend who was going to go on the radio. He does work on cosmology and astronomy, and he wondered how he would explain what the applications of his work were. "Well," I said, "there aren't any." He said, "Yes, but then we won't get support for more research of this kind." I think that's kind of dishonest. If you're representing yourself as a scientist, then you should explain to the layman what you're doing–and if they don't support you under those circumstances, then that's their decision.

One example of the principle is this: If you've made up your mind to test a theory, or you want to explain some idea, you should always decide to publish it whichever way it comes out. If we only publish results of a certain kind, we can make the argument look good. We must publish BOTH kinds of results.

So I have just one wish for you–the good luck to be somewhere where you are free to maintain the kind of integrity I have described, and where you do not feel forced by a need to maintain your position in the organization, or financial support, or so on, to lose your integrity. May you have that freedom.

1.3 Large numbers

These notes deal with space and time. The first thing we notice about the universe around us is how big it is. In order to quantify things in cosmology very large numbers are required and the endless writing of zeroes quickly becomes tedious. Thus people invented what is called *the scientific notation* which is a way of avoiding writing many zeroes. For example the quantity 'one million' can be written as $1,000,000$ which is a one followed by six zeroes, this is abbreviated as 10^6 (the little number above the zero is called the *exponent* and denotes the number of zeroes after the one). In this way we have

$$\begin{aligned} \text{one million} &= 1,000,000 = 10^6 \\ \text{one billion} &= 1,000,000,000 = 10^9 \\ \text{one trillion} &= 1,000,000,000,000 = 10^{12}, \quad \text{etc.} \end{aligned} \tag{1.1}$$

So much for large numbers. There is a similar short-hand for small numbers, the only difference is that the exponent has a minus sign in front:

$$\begin{aligned} \text{one tenth} &= 0.1 = 10^{-1} \\ \text{one thousandth} &= 0.001 = 10^{-3} \\ \text{one millionth} &= 0.000001 = 10^{-6}, \quad \text{etc.} \end{aligned} \tag{1.2}$$

In order to get several times the above quantities one multiplies by ordinary numbers, so, for example, 8×10^6 =eight millions, 4×10^{-12} =four trillionths, etc.

This notation is a vast improvement also on the one devised by the Romans, and which was used up until the Renaissance. For example, our galaxy, the Milky Way, has a diameter of about 10^5 light years (a light year is the *distance* light travels in one year), in Roman numerals

$$10^5 = \qquad MMMMMMMMMMMMMMMMMMMM$$

$$\begin{array}{c}MMMMMMMMMMMMMMMMMMMM\\MMMMMMMMMMMMMMMMMMMM\\MMMMMMMMMMMMMMMMMMMM\\MMMMMMMMMMMMMMMMMMMM\end{array}$$

The Andromeda galaxy is about 2×10^6 (two million) light years from our galaxy, in Roman numerals writing this distance requires 40 lines.

Appendix: Examples of large numbers

Very small and very large numbers are not the sole property of cosmology, there are many cases where such numbers appear. What is hard to do is visualize the meaning of something like a million or a billion. Below I provide several examples of large and small numbers.

In the table for temperatures the values are given in degrees Kelvin; a degree Kelvin equals a degree Celsius, but zero degrees Kelvin corresponds to -273.16 degrees Celsius. In order to change to degrees Fahrenheit you need to do the following operation:

$$\text{Deg. Fahrenheit} = 1.8 \times \text{Deg. Kelvin} - 459.$$

Absolute zero, the temperature at which all systems reach their lowest energy level, corresponds to zero degrees Kelvin, and -459 degrees Fahrenheit.

Times (in seconds)

8.6×10^4	Earth rotation time
1.6×10^9	Time between Milky Way supernovae
3×10^{13}	Time for evolution of a species
7.3×10^{15}	Orbit time for sun around galaxy center
6×10^{16}	Time for galaxy to cross a cluster
1.1×10^{17}	Primeval slime to man time
1.5×10^{17}	Age of Earth and Sun
1.5×10^{17}	Uranium-238 half-life
3×10^{17}	Sun lifetime
3.8×10^{17}	Rough age of the Milky Way
4×10^{17}	Rough age of 47 Tucanae
4.1×10^{17}	Age of the universe

Distances (in meters)

1.8	Man
8 847	Height of Mount Everest
10 000	Neutron star radius
10 000	Typical comet radius
12 000	Typical airliner cruising altitude
3.2×10^6	Length of the Great Wall of China
6.3×10^6	Radius of the Earth
7.1×10^7	Radius of Jupiter
3.8×10^8	Distance to the Moon
7.0×10^8	Radius of the Sun
1.5×10^{11}	Earth/Sun mean distance
5×10^{11}	Radius of the supergiant star Betelgeuse
5.9×10^{12}	Pluto/Sun mean distance
9.46×10^{15}	1 light-year
4×10^{16}	Nearest non-solar star to Earth
4.5×10^{16}	Rough Crab Nebula radius
1.5×10^{18}	Typical globular cluster radius
5.2×10^{18}	Distance to the supergiant Betelgeuse
6.6×10^{19}	Distance to the Crab Nebula
1.2×10^{20}	Milky Way characteristic thickness
2.4×10^{20}	Distance from Sun to galactic center
3.9×10^{20}	Milky Way disk radius
3×10^{22}	Radius of the core of the Virgo cluster
7×10^{23}	Distance to the center of the Virgo cluster
1.3×10^{27}	Distance to the quasar PC 1247+3406

Velocities (in meters per second)

1.0×10^{-9}	Sea floor spreading rate
1.6×10^{-9}	Average slip rate of the San Andreas fault
2×10^{-8}	Grass growth rate
3×10^{-6}	Typical glacial advance rate
1.3	Human walking speed
25	Car speed
100	Speed of an electric nervous pulse
330	Sound speed in air
600	Fighter jet speed
2 380	Escape velocity from Moon's surface
11 000	Escape velocity from the Earth's surface
29 000	Earth's motion around the Sun
2.2×10^5	Velocity of the Sun around the Milky Way
3.1×10^5	Escape velocity from the Milky Way
6.2×10^5	Escape velocity from the Sun's surface
5×10^6	Young (months old) supernova ejecta
2×10^8	Escape velocity from neutron star surface
3×10^8	Light in a vacuum

Masses (in kilograms)

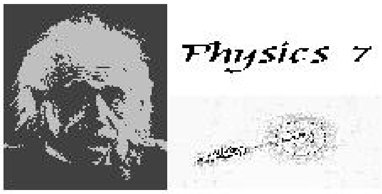
70	Lower limit to the allowed mass for a Sumo wrestler
1 000	Car
10 000	Tyrannosaurus Rex
1×10^{13}	Typical comet mass
3×10^{14}	Typical mountain mass
1.1×10^{16}	Superterranean biomass of Earth (ocean organisms are included)
5.3×10^{18}	Total mass of Earth's atmosphere
3×10^{19}	Typical asteroid mass
1.4×10^{21}	Total mass of Earth's oceans
7.3×10^{22}	Mass of the Moon
5.98×10^{24}	Mass of the Earth
1.9×10^{27}	Mass of Jupiter
1.99×10^{30}	Mass of the Sun
2.8×10^{30}	Maximum mass for a white dwarf star
6.0×10^{30}	Maximum mass for a neutron star
1.3×10^{44}	Rough mass of the stars in the Coma galaxy cluster
1.4×10^{49}	Rough total mass in spiral galaxies
2×10^{52}	Rough total mass of a critical density universe

Temperatures (in deg. Kelvin)

7×10^{-7}	Laser cooling of cesium atoms
2.17	Liquid ^{4}He superfluid transition temperature
2.726	Cosmic microwave background temperature today
273	Water freezing temperature
311	Human surface temperature
373	Water boiling temperature
506	Paper burning temperature
740	Typical surface temperature of Venus
1811	Melting temperature of iron
5770	Solar effective temperature
1.4×10^7	Center of the Sun
5×10^7	Typical gas temperature in a cluster of galaxies
3×10^{10}	Center of a supernova.

Monies (in 1994 US dollars)

9×10^7	Development and construction cost of the Keck telescope
1.5×10^8	Rough cost of a European Ariane rocket launch
2.1×10^8	Total spending in the 1994 U.S. senate election campaigns
9×10^8	Total cost of the *Magellan* probe
1.1×10^9	Worldwide Visa and MasterCard fraud in 1993
1.8×10^9	Amount of food stamp fraud in the USA in 1993
3.8×10^9	Microsoft revenue in 1993
1×10^{10}	Rough monetary losses associated with BCCI
1.3×10^{10}	Lockheed revenue in 1993
1.5×10^{10}	Rough United Nations yearly budget
2.8×10^{10}	Planned cost for the space station
2.6×10^{11}	United States 1994 military spending
2.6×10^{11}	United States 1994 predicted deficit
8×10^{11}	United States 1994 entitlement spending
1×10^{12}	Rough total United States health care spending in 1994
1.3×10^{12}	United States 1994 tax receipts
1.5×10^{12}	United States 1994 federal government spending
4.4×10^{12}	United States 1994 national debt
6.4×10^{12}	United States 1994 gross domestic product
1.4×10^{13}	United States 1994 unfunded liabilities for entitlement programs



Chapter 2

Greek cosmology

The first "cosmologies" were based on creation myths in which one or more deities made the universe out of sheer will, or out of their bodily fluids, or of the carcass of some god they defeated, etc. A few examples of such "theories" of the universe are provided in this chapter. These are hardly scientific theories in the sense that they have almost no support form observation and in that they predict very few things outside of the fact that there *is* a world (if everything is due to the whims of the Gods then there is very little one can predict). It is an interesting comment on the workings of the human mind that quite different cultures produced similar creation myths.

The first scientific cosmology was created by the Greeks more than 2000 years ago, and this chapter also describes these ideas and their origin. The Greeks used some of the knowledge accumulated by earlier civilizations, thus this chapter begins with a brief description of the achievements of the Egyptians and Babylonians. We then consider the highlights of Greek cosmology culminating with Ptolemy's system of the world.

2.1 Egypt and Babylon

2.1.1 Babylon

The Babylonians lived in Mesopotamia, a fertile plain between the Tigris and Euphrates rivers (see Fig. 2.1). They developed an abstract form of writing based on cuneiform (wedge-shaped) symbols. Their symbols were written on wet clay tablets which were baked in the sun; many thousands of these tablets have survived to this day; an example is shown in Fig. 2.1.

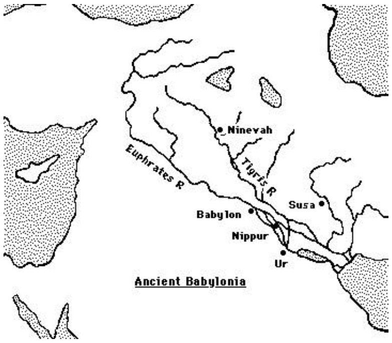


Figure 2.1: Left: Region dominated by the Babylonian civilization. Right: example of a cuneiform tablet containing Pythagorean triples.

The Babylonian apparently believed the Earth to be a big circular plane surrounded by a river beyond which lies an impassable mountain barrier, with the whole thing resting on a cosmic sea. No human may cross the river surrounding the Earth. The mountains support the vault of heaven, which is made of a very strong metal. There is a tunnel in the northern mountains that opens to the outer space and which also connects two doors, one in the East and one in the West. The sun comes out through the eastern door, travels below the metallic heavens and then exits through the western door; he spends the nights in the tunnel.

The creation myth is more lively than the Egyptian version. It imagines that the cosmic ocean *Apsu* mixed with chaos *Tiamat* and eventually generated life. For a while life was good for the gods but there came a time when Tiamat felt her domain was too small and made war against the other gods. All but *Marduk* were afraid of her, so Marduk, after getting all the powers from the frightened gods, fought Tiamat. When Tiamat opened her mouth to swallow him he thrust a bag full with hurricane winds into her so that she swelled and, taking advantage of her indisposition, Marduk pierced her with his lance and killed her. Then he split Tiamat's carcass making the lower half the earth and the upper the heavens. Finally Marduk mixed his bloodown blood with the earth to make men for the service of the gods.

Babylonians and Chaldeans observed the motion of the stars and planets from the earliest antiquity (since the middle of the 23rd century B.C.). They cataloged the motion of the stars and planets as well as the occurrence of eclipses and attempted to fit their behavior to some numerical theories. Many of these observations were used for astrological prophesying and, in fact, they were the originators of astrology. They believed that the motions and changes in the stars and planets determine (or so they believed) what occurs on this planet.

The Babylonians excelled in computational mathematics, they were able to solve algebraic equations of the first degree, understood the concept of function and realized the truth of Pythagoras' theorem (without furnishing an abstract proof). One of the clay tablets dated from between 1900 and 1600 B.C. contains answers to a problem containing Pythagorean triples, i.e. numbers a, b, c with $a^2 + b^2 = c^2$. It is said to be the oldest number theory document in existence. The Babylonians had an advanced number system with base 60 rather than the base 10 of common today. The Babylonians divided the day into 24 hours, each hour into 60 minutes, each minute into 60 seconds. This form of counting has survived for 40 centuries.

2.1.2 Egypt

The anciebloodnt Egyptians conceived the sky as a roof placed over the world supported by columns placed at the four cardinal points. The Earth was a flat rectangle, longer from north to south, whose surface bulges slightly and having (of course) the Nile as its center. On the south there was a river in the sky supported by mountains and on this river the sun god made his daily trip (this river was wide enough to allow the sun to vary its path as it is seen to do). The stars were suspended from the heavens by strong cables, but no apparent explanation was given for their movements.

There is no unique Egyptian creation myth, yet one of the most colorful versions states that at the beginning of the world,*Nuit*, the goddess of the night, was in a tight embrace with her husband *Sibû*, the earth god. Then one day, without an obvious reason, the god *Shû* grabed her and elevated her to the sky (to *become* the sky) despite the protests and painful squirmings of Sibû. But Shû has no sympathy for him and freezes Sibû even as he is thrashing about. And so he remains to this day, his twisted pose generating the irregularities we see on the Earth's surface (see Fig. 2.2). Nuit is supported by her arms and legs which become the columns holding the sky. The newly created world was divided into four regions or houses, each dominated by a god. Since the day of creation Sibû has been clothed in verdure and generations of animals prospered on his back, but his pain persists.

An extended version of this myth imagines that in the beginning the god *Tumu* suddenly cried "Come to me!" across the cosmic ocean, whence a giant lotus flower appeared which had the god *Ra* inside, then Ra separates Nuit and Sibû, and the story proceeds as above. It is noteworthy that creation did not come through muscular effort, but through Tumu's voiced

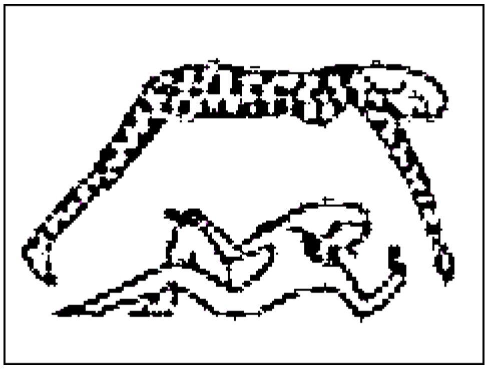


Figure 2.2: Nuit the sky above Sibû the Earth after being separated by Shû in a version of the Egyptian creation myth.

command. This later evolved into the belief that the creator made the world with a single word, then with a single sound (yet the creation through pure thought was not considered).

After creation the gods, especially *Thot* (Fig. 2.3), teach the arts and sciences to the Egyptians. In particular Thot taught the Egyptians how to observe the heavens and the manner in which the planets and the sun move, as well as the names of the (36) constellations (though he apparently neglected to tell them about eclipses which are never referred to).

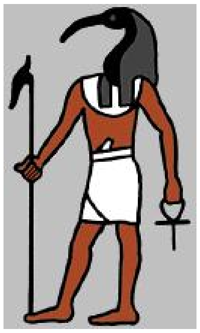


Figure 2.3: The Egyptian god Thot.

The study of the heavens was not made for altruistic purposes but with very practical aims: a good calendar was necessary in order to prepare for the regular flooding of the Nile as well as for religious purposes. The Egyptian calendar had a year of precisely 365 days and was used for many centuries; curiously they never corrected for the fact that the year is 365

1/4 days in length (this is why every four years we have a leap year and add a day to February) and so their time reckoning was off one day every four years. After 730 years this deficiency adds up to 6 months so that the calendar announced the arrival of summer at the beginning of winter. After 1460 years the Egyptian calendar came back on track and big celebrations ensued.

Egyptians knew and used the water clock whose origin is lost in the mists of time. the oldest clock in existence dates from the reign of the pharo Thutmose III (about 1450 B.C.) and is now in th Berlin Museum.

Most of Egyptian mathematics was aimed at practical calculations such as measuring the Earth (important as the periodic Nile floods erased property boundary marks) and business mathematics. Their number system was clumsy (addition was not too bad but multiplication is very cumbersome). To overcome this deficiency the Egyptians devised cunning ways to multiply numbers, the method, however, was very tedious: to obtain $41 \times 59 = 2419$, nine operations had to be performed (all additions and subtractions); yet they were able to calculate areas and estimate the number π. Examples of calculations have survived in several papyri (Fig. 2.4).

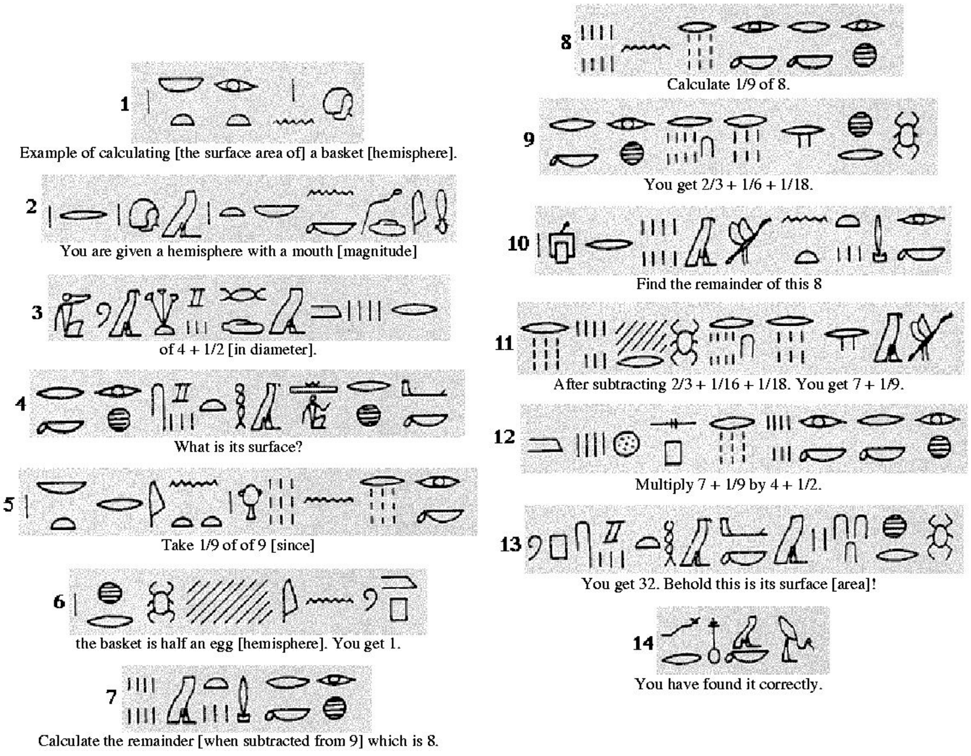
Unlike the Greeks who thought abstractly about mathematical ideas, the Egyptians were only concerned with practical arithmetic. In fact the Egyptians probably did not think of numbers as abstract quantities but always thought of a specific collection of objects when a number was mentioned.

2.2 Other nations

None of the early civilizations lacked a cosmology or creation myths. In this section a brief summary of some of these myths is presented.

2.2.1 India

The traditional Indian cosmology states that the universe undergoes cyclic periods of birth, development and decay, lasting 4.32×10^9 years, each of these periods is called a *Kalpa* or “day of Brahma”. During each Kalpa the universe develops by natural means and processes, and by natural means and processes it decays; the destruction of the universe is as certain as the death of a mouse (and equally important). Each Kalpa is divided into 1000 “great ages”, and each great age into 4 ages; during each age humankind deteriorates gradually (the present age will terminate in 426,902 years). These is no final purpose towards which the universe moves, there is no progress, only endless repetition. We do not know how the universe began, perhaps



1 Example of calculating [the surface area of] a basket [hemisphere].

2 You are given a hemisphere with a mouth [magnitude]

3 of 4 + 1/2 [in diameter].

4 What is its surface?

5 Take 1/9 of of 9 [since]

6 the basket is half an egg [hemisphere]. You get 1.

7 Calculate the remainder [when subtracted from 9] which is 8.

8 Calculate 1/9 of 8.

9 You get 2/3 + 1/6 + 1/18.

10 Find the remainder of this 8

11 After subtracting 2/3 + 1/16 + 1/18. You get 7 + 1/9.

12 Multiply 7 + 1/9 by 4 + 1/2.

13 You get 32. Behold this is its surface [area]!

14 You have found it correctly.

Figure 2.4: An example of Egyptian papyri, the Moscow papyrus and its translation; the text contains the estimate $\pi \simeq 256/81 = 3.1605$.

Brahma laid it as an egg and hatched it; perhaps it is but an error or a joke of the Maker.

This description of the universe is remarkable for the enormous numbers it uses. The currently accepted age of the universe is about 10^{18} seconds and this corresponds to about 7 Kalpas+335 great ages. A unique feature of Indian cosmology is that no other ancient cosmology manipulates such time periods.

In the Surya Siddanta it is stated that the stars revolved around the cosmic mountain Meru at whose summit dwell the gods. The Earth is a *sphere* divided into four continents. the planets move by the action of a cosmic wind and, in fact, the Vedic conception of nature attributes *all* motion to such a wind. It was noted that the planets do not move in perfect circles and this was attributed to "weather forms" whose hands were tied to the planets by "cords of wind"

2.2.2 China

The Chinese have a very long history of astronomical observations reaching back to the 13th century B.C. They noted solar eclipses as well as supernova events (exploding stars). The most impressive of these events was the observation on 1054 A.D. of such a supernova event which lasted for 2 years, after that the star dimmed and disappeared from view. The astronomical observations were sufficiently precise for later astronomers to determine that the location of that exploding star is now occupied by the crab nebula (Fig. 2.5); it was then shown that this nebula is expanding and, extrapolating backwards, that this expansion started in 1054 A.D.

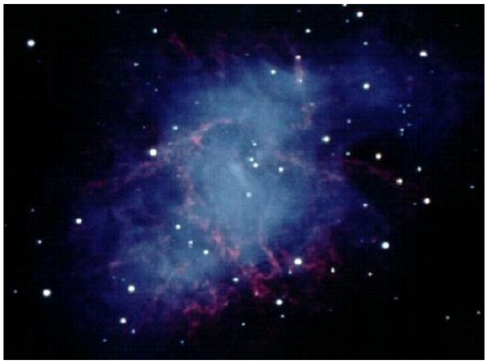


Figure 2.5: The Crab nebula, the remnant of a supernova.

The first Chinese cosmography imagines a round sky over a square Earth with the sun and heavens revolving around the Earth. Later this was replaced by a round Earth around which all heavenly bodies rotate. These theories propagated throughout Eastern Asia.

2.3 Early Greeks

The Greeks were apparently the first people to look upon the heavens as a set of phenomena amenable to human comprehension and separated from the sometimes fickle whims of the gods. They were able to extract an great amount of information using nothing but basic reasoning and very elementary observations. This makes their results all the more amazing.

In the earliest times their view of the world and its origin was firmly based on creation myths consolidated by Homer in the Iliad and Odyssey, as the culture evolved this view of the universe evolved and distanced itself from the purely religious outlook.

2.3.1 Mythology

A simplified version of the Greek creation myth follows.

In the beginning there was only chaos. Then out of the void appeared Night and Erebus, the unknowable place where death dwells. All else was empty, silent, endless, darkness. Then somehow Love (Eros) was born bringing a start of order. From Love came Light and Day. Once there was Light and Day, Gaea, the earth appeared.

Then Erebus slept with Night, who gave birth to Aether, the heavenly light, and to Day, the earthly light. Then Night alone produced Doom, Fate, Death, Sleep, Dreams, Nemesis, and others that come to man out of darkness.

Meanwhile Gaea alone gave birth to Uranus, the heavens. Uranus became Gaea's mate covering her on all sides. Together they produced the three Cyclops, the three Hecatoncheires, and twelve Titans.

However, Uranus was a bad father and husband. He hated the Hecatoncheires and imprisoned them by pushing them into the hidden places of the earth, Gaea's womb. This angered Gaea and she plotted against Uranus. She made a flint sickle and tried to get her children to attack Uranus. All were too afraid except, the youngest Titan, Cronus.

Gaea and Cronus set up an ambush of Uranus as he lay with Gaea at night. Cronus grabbed his father and castrated him, with the stone sickle, throwing the severed genitals into the ocean. The fate of Uranus is not clear. He either died, withdrew from the earth, or exiled himself to Italy. As he departed he promised that Cronus and the Titans would be punished. From his spilt blood came the Giants, the Ash Tree Nymphs, and the Erinyes. From the sea foam where his genitals fell came Aphrodite.

Cronus became the next ruler. He imprisoned the Cyclops and the Hecatoncheires in Tartarus. He married his sister Rhea, under his rule he and the other Titans had many offspring. He ruled for many ages. However, Gaea and Uranus both had prophesied that he would be overthrown by a son. To avoid this Cronus swallowed each of his children as they were born. Rhea was angry at the treatment of the children and plotted against Cronus. When it came time to give birth to her sixth child, Rhea hid herself, and after the birth she secretly left the child to be raised by nymphs. To conceal her act she wrapped a stone in swaddling clothes and passed it off as the baby to Cronus, who swallowed it.

This child was Zeus. He grew into a handsome youth on Crete. He consulted Metis on how to defeat Cronus. She prepared a drink for Cronus designed to make him vomit up the other children. Rhea convinced Cronus to accept his son and Zeus was allowed to return to Mount Olympus as Cronus's cup-bearer. This gave Zeus the opportunity to slip Cronus the specially prepared drink. This worked as planned and the other five children were vomited up. Being gods they were unharmed. They were thankful to Zeus and made him their leader.

Cronus was yet to be defeated. He and the Titans, except Prometheus, Epimetheus, and Oceanus, fought to retain their power. Atlas became their leader in battle and it looked for some time as though they would win and put the young gods down. However, Zeus was cunning. He went down to Tartarus and freed the Cyclops and the Hecatoncheires. Prometheus joined Zeus as well who returned to battle with his new allies. The Cyclops provided Zeus with lighting bolts for weapons. The Hecatoncheires he set in ambush armed with boulders. With the time right, Zeus retreated drawing the Titans into the Hecatoncheires's ambush. The Hecatoncheires rained down hundreds of boulders with such a fury the Titans thought the mountains were falling on them. They broke and ran giving Zeus victory.

Zeus exiled the Titans who had fought against him into Tartarus. Except for Atlas, who was singled out for the special punishment of holding the world on his shoulders.

However, even after this victory Zeus was not safe. Gaea angry that her children had been imprisoned gave birth to a last offspring, Typhoeus. Typhoeus was so fearsome that most of the gods fled. However, Zeus faced the monster and flinging his lighting bolts was able to kill it. Typhoeus was buried under Mount Etna in Sicily.

Much later a final challenge to Zeus rule was made by the Giants. They went so far as to attempt to invade Mount Olympus, piling mountain upon mountain in an effort to reach the top. But, the gods had grown strong and with the help of Heracles the Giants were subdued or killed.

One of the most significant features of the Greek mythology is the presence of the Fates: these were three goddesses who spend the time weaving a rug where all the affairs of men and gods appear. There is nothing that can be done to alter this rug, even the gods are powerless to do so, and it is this that is interesting. For the first time the idea appears of a force which rules *everything*, even the gods.

2.3.2 Early cosmology

In their many travels the early Greeks came into contact with older civilizations and learned their mathematics and cosmologies. Early sailors relied heavily on the celestial bodies for guidance and the observation that the heavens presented very clear regularities gave birth to the concept that these regularities resulted, not from the whims of the gods, but from physical laws. Similar conclusions must have been drawn from the regular change of the seasons. This realization was not sudden, but required a lapse of many centuries, yet its importance cannot be underestimated for it is the birth of modern science.

The earliest of the Greek cosmologies were intimately related to mythology: earth was surrounded by air above, water around and Hades below; ether surrounded the earth-water-Hades set (Fig. 2.6),

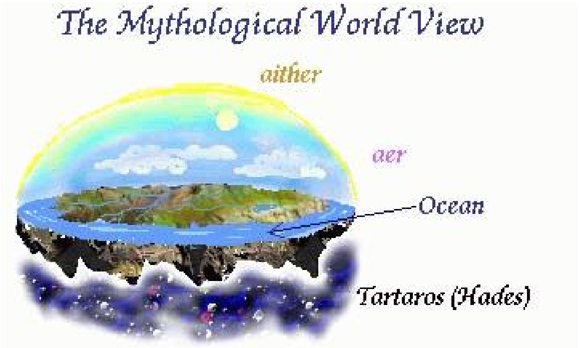
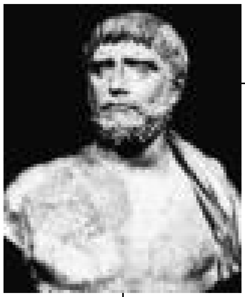


Figure 2.6: The universe according to Greek mythology.

This system was soon replaced by more sophisticated views on the nature of the cosmos. Two interesting examples were first the claim of Anaxagoras of Clazomenae that the Moon shines only through the light it reflects from the sun, and that that lunar eclipses are a result of the earth blocking the sunlight in its path to the moon; he also believed the Sun to be a ball of molten iron larger than the Peloponesus.

Another remarkable feat was the prediction of a solar eclipse by Thales in 585 B.C. (for which he used the data obtained by Babylonian astronomers). During this period other ideas were suggested, such as the possibility of an infinite, eternal universe (Democritus) and a spherical immovable Earth (Parmenides).



Thales of Miletus (624 B.C. - 546 B.C.). Born and died in Miletus, Turkey. Thales of Miletus was the first known Greek philosopher, scientist and mathematician. None of his writing survives so it is difficult to determine his views and to be certain about his mathematical discoveries. He is credited with five theorems of elementary geometry: *(i)* A circle is bisected by any diameter. *(ii)* The base angles of an isosceles triangle are equal. *(iii)* The angles between two intersecting straight lines are equal. *(iv)* Two triangles are congruent if they have two angles and one side equal. *(v)* An angle in a semi-circle is a right angle. Thales is believed to have been the teacher of Anaximander and he is the first natural philosopher in the Milesian School. [1].

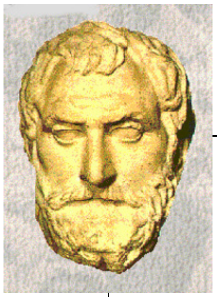
Despite these strikingly “modern” views about the sun and moon, the accepted cosmologies of the time were not so advanced. For example, Thales believed that the Earth floats on water (and earthquakes were the result of waves in this cosmic ocean), and all things come to be from this cosmic ocean. In particular the stars float in the upper waters which feed these celestial fires with their “exhalations”. The motion of the stars were assumed to be governed by (then unknown) laws which are responsible for the observed regularities.

A good example of the manner in which the Greeks drew logical conclusions from existing data is provided by the argument of Anaxagoras who pointed out that meteors, which are seen to fall from the heavens, are made of the same materials as found on Earth, and then hypothesized that the heavenly bodies were originally part of the Earth and were thrown out by the rapid rotation of the Earth; as the rapid rotation of these bodies decreases they are pulled back and fall as meteors. This conclusion is, of course, wrong, but the hypothesis proposed does demonstrate imagination as well as close adherence to the observed facts.

The early Greek cosmological theories *did* explain all the data available at the time (though they made no predictions). And, even with these deficiencies, this period is notable for the efforts made to understand the workings of Nature using a rational basis. This idea was later adopted by Plato and is the basis of all modern science.

There are many other early cosmologies, for example, Anaximander believed the Earth to be surrounded by a series of spheres made of mist and surrounded by a big fire; the Sun, Moon and stars are glimpses of this fire through the mist. In a different version of his cosmology he imagined the Earth to be a cylinder floating in space. In a more poetical vein, Empedocles believed the cosmos to be egg-shaped and governed by alternating reigns of love and hate.

Parmenides of Elea (515 B.C. - 445 B.C.). Born in Elea, a Greek city in southern Italy (today called Velia); almost certainly studied in Athens and there is ample evidence that he was a student of Anaximander and deeply influenced by the teachings of the Pythagoreans, whose religious and philosophical brotherhood he joined at their school in Crotona. All we have left of his writings are about 160 lines of a poem called Nature, written for his illustrious student Zeno and preserved in the writings of later philosophers such as Sextus Empiricus. His style influenced by Pythagorean mysticism.



Anaxagoras of Clazomenae (499 B.C. - 428 B.C.). Greek, born in Ionia, lived in Athens He was imprisoned for claiming that the sun was not a god and that the moon reflected the sun's light. While in prison he tried to solve the problem of squaring the circle, that is constructing with ruler and compasses a square with area equal to that of a given circle (this is the first record of this problem being studied). He was saved from prison by Pericles but had to leave Athens.

The early Greeks also considered the composition of things. It was during these times that it was first proposed (by Anaximines of Miletus, c. 525 B.C.) that everything was supposed to be made out of four "elements":

ings were supposed to
ade out of four
ents": earth, water,
d fire

earth, water, air and fire. This idea prevailed for many centuries. It was believed that earth was some sort of condensation of air, while fire was some sort of emission form air. When earth condenses out of air, fire is created in the process. Thus we have the first table of the elements (see Fig. 2.7)

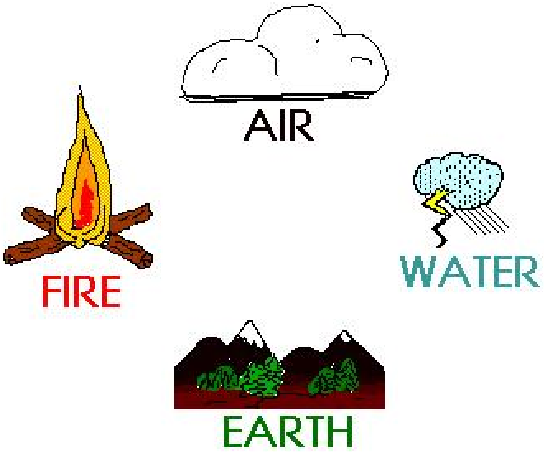
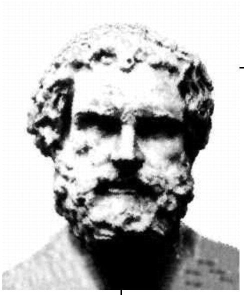


Figure 2.7: The earliest table of the elements.

This, however was not universally accepted. The most notable detractor was Democritus who postulated the existence of indestructible atoms (from the Greek *a-tome:* that which cannot be cut) of an infinite variety of shapes and sizes. He imagined an infinite universe containing an infinite number of such atoms, in between the atoms there is an absolute void.



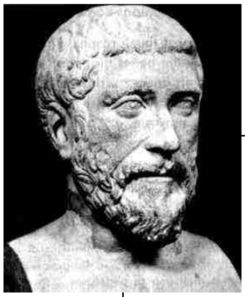
Democritus of Abdera (460 B.C. - 370 B.C.). Democritus is best known for his atomic theory but he was also an excellent geometer. Very little is known of his life but we know that Leucippus was his teacher. He's believed to have traveled widely, perhaps spent a considerable time in Egypt, and he certainly visited Persia. Democritus wrote many mathematical works but none survive. He claimed that the universe was a purely mechanical system obeying fixed laws. He explained the origin of the universe through atoms moving randomly and colliding to form larger bodies and worlds. He also believed that space is infinite and eternal, and that the number of atoms are infinite. Democritus's philosophy contains an early form of the conservation of energy.

2.3.3 The Pythagoreans

About five centuries B.C. the school founded by the Greek philosopher, mathematician and astronomer Pythagoras flourished in Samos, Greece. The Pythagoreans believed (but failed to prove) that the universe could be understood in terms of whole numbers. This belief stemmed from observations in music, mathematics and astronomy. For example, they noticed that vibrating strings produce harmonious tones when the ratios of the their lengths are whole numbers. From this first attempt to express the universe in terms of numbers the idea that the world could be understood through mathematics was born, a central concept in the development of mathematics and science.

The Pythagoreans originated the idea that the world could be understood through mathematics was born

The importance of pure numbers is central to the Pythagorean view of the world. A point was associated with 1, a line with 2 a surface with 3 and a solid with 4. Their sum, 10, was sacred and omnipotent [2].



Pythagoras of Samos (580–500 B.C.). Born Samos, Greece, died in Italy. Pythagoras was a Greek philosopher responsible for important developments in mathematics, astronomy, and the theory of music. He founded a philosophical and religious school in Croton that had many followers. Of his actual work nothing is directly known. His school practiced secrecy and communalism making it hard to distinguish between the work of Pythagoras and that of his followers.

Pythagoras also developed a rather sophisticated cosmology. He and his followers believed the earth to be perfectly spherical and that heavenly bodies, likewise perfect spheres, moved as the Earth around a central fire invisible to human eyes (this was *not* the sun for it *also* circled this central fire) as shown in Fig. 2.8. There were 10 objects circling the central fire which included a counter-earth assumed to be there to account from some eclipses but also because they believed the number 10 to be particularly

[2] Some relate this to the origin of the decimal system, but it seems to me more reasonable to associate the decimal system to our having ten fingers.

sacred. This is the first coherent system in which celestial bodies move in circles, an idea that was to survive for two thousand years.

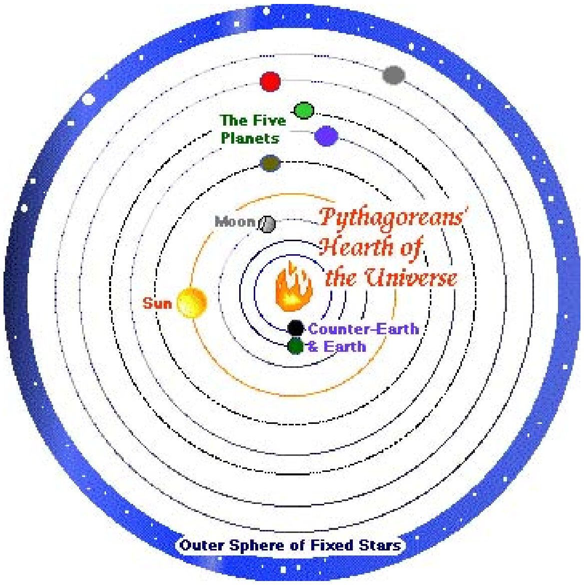

The Five Planets
Moon
Pythagoreans' Hearth of the Universe
Sun
Counter-Earth & Earth
Outer Sphere of Fixed Stars

Figure 2.8: The universe according to the Pythagoreans.

It was also stated that heavenly bodies give forth musical sounds "the harmony of the spheres" as they move in the cosmos, a music which we cannot discern, being used to it from childhood (a sort of background noise); though we would certainly notice if anything went wrong! The Pythagoreans did not believe that music, numbers and cosmos were just related, they believed that music *was* number and that the cosmos *was music*

Pythagoras is best known for the mathematical result (Pythagoras' theorem) that states that the sum of the squares of the sides of a right triangle equals the square of the diagonal; see Fig. 2.9. This result, although known to the Babylonians 1000 years earlier, was first proved by Pythagoras (allegedly: no manuscript remains). Pythagoras' theorem will be particularly important when we study relativity for, as it turns out, it is *not* valid in the vicinity of very massive bodies! Similar statements hold for Euclid's postulate that parallel lines never meet, see Sect. **??**.

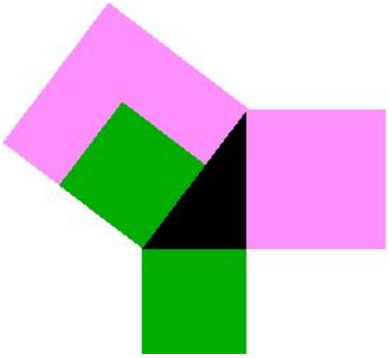
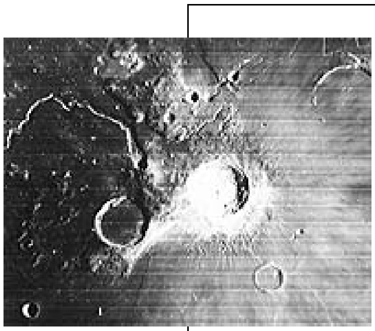


Figure 2.9: Pythagoras' theorem (the areas of the squares attached to the smaller sides of the triangle equal the area of the largest square).

2.4 Early heliocentric systems

By the IV century B.C. observations had shown that there are two types of stars: fixed stars whose relative position remained constant, and "wandering stars", or planets, whose position relative to the fixed stars changed regularly. Fixed stars moved as if fixed to a sphere that turned with the earth at its center, the planets moved about these fixed stars driven by an unknown agency. In fact, Plato regarded the investigation of the rules that determined the motion of the planets as a very pressing research problem.

A remarkable answer was provided by the heliocentric (!!) system of Aristarchus of Samos. Using a clever geometric argument Aristarchus estimated the size of the Sun and concluded it must be enormously larger than the Earth; he then argued that it was inconceivable that such a behemoth would slavishly circle a puny object like the Earth. Once he concluded this, he concluded that the Earth must rotate on its axis in order to explain the (apparent) motion of the stars. Thus Aristarchus conceived the main ingredients of the Copernican system 17 centuries before the birth of Copernicus! Unfortunately these views were soundly rejected by Aristotle: if the Earth is rotating, how is it that an object thrown upwards falls on the same place? How come this rotation does not generate a very strong wind? Due to arguments such as this the heliocentric theory was almost universally rejected until Copernicus' answered these criticisms.



Aristarchus of Samos (310 B.C. - 230 B.C.). Born and died in Greece. Aristarchus was a mathematician and astronomer who is celebrated as the exponent of a Sun-centered universe and for his pioneering attempt to determine the sizes and distances of the Sun and Moon. Aristarchus was a student of Strato of Lampsacus, head of Aristotle's Lyceum, coming between Euclid and Archimedes. Little evidence exists concerning the origin of his belief in a heliocentric system; the theory was not accepted by the Greeks and is known only because of a summary statement in Archimedes' *The Sand-Reckoner* and a reference by Plutarch. The only surviving work of Aristarchus, *On the Sizes and Distances of the Sun and Moon*, provides the details of his remarkable geometric argument, based on observation, whereby he determined that the Sun was about 20 times as distant from the Earth as the Moon, and 20 times the Moon's size. Both these estimates were an order of magnitude too small, but the fault was in Aristarchus' lack of accurate instruments rather than in his correct method of reasoning. Aristarchus also found an improved value for the length of the length of the solar year.

Astronomy also progressed, with the most striking result, due to Eratosthenes, was accurate measurement of the Earth's circumference [3] (the fact that the Earth is round was common knowledge) He noted that the distance from Alexandria to Aswan is 5,000 stadia and that when the sun casts no shadow in Alexandria it casts a shadow corresponding to an angle of 7.2^o (see Fig. 2.10). From this he determined the circumference of the Earth less than 2% accuracy!

The fact that the Earth is round was common knowledge

It is important to remember that the realization that the Earth was round was not lost to the following centuries, so that neither Columbus nor any of his (cultivated) contemporaries had any fear of falling off the edge of the world when traveling West trying to reach the Indies. The controversy surrounding Columbus' trip was due to a disagreement on the *size* of the Earth. Columbus had, in fact, seriously underestimated the radius of the Earth and so believed that the tiny ships he would command had a fair chance of getting to their destination. He was, of course, unaware of the interloping piece of land we now call America, had this continent not existed, Columbus and his crew would have perished miserably in the middle of the ocean.

[3] Aristotle had previously estimated a value of 400,000 stadia (1 stadium=157.5m) which is about 1.6 times its actual size.

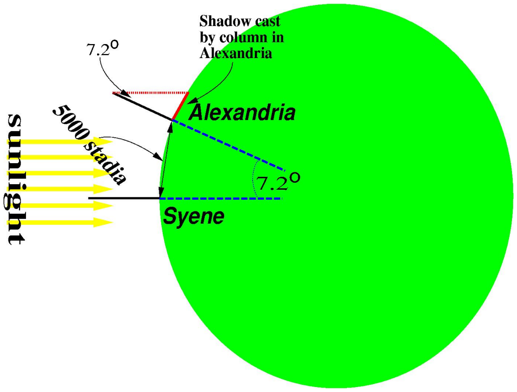


Figure 2.10: Description of Eratosthenes' procedure for measuring the Earth. He reasoned that the change in angle of the shadow was caused by moving about the Earth. By measuring the angle of the shadow at Seyene, and then in a city that was directly north of Seyene (Alexandria), he determined that the two cities were 7 degrees apart. That is to say, out of the 360 degrees needed to travel all the way around the world, the two cities were 360/7 of that distance. Since he knew that the two cities were about 500 miles apart, he concluded that the the Earth must be (360/7) × 500 miles in circumference, or roughly $25,000$ miles.

Eratosthenes of Cyrene (276 B.C. - 197 B.C.). Greek, lived in Alexandria He was born in Cyrene which is now in Libya. He worked on geometry and prime numbers. He is best remembered for his prime number sieve which, in modified form, is still an important tool in number theory research. Eratosthenes measured the tilt of the earth's axis with great accuracy and compiled a star catalogue containing 675 stars; he suggested that a leap day be added every fourth year and tried to construct an accurately-dated history. He became blind in his old age and is said to have committed suicide by starvation.

2.5 Aristotle and Ptolemy

There are few instances of philosophers that have had such a deep influence as Aristotle, or of cosmologists whose theories have endured as long as Ptolemy's. Aristotle's influence is enormous ranging form the sciences to logic. Many of his ideas have endured the test of the centuries. His cosmology, based on a geocentric system, is not one of them. In the words of W. Durant

> His curious mind is interested, to begin with, in the process and techniques of reasoning; and so acutely does he analyze these that his *Organon*, or Instrument–the name given after his death to his logical treatises–became the textbook of logic for two thousand years. He longs to think clearly, though he seldom, in extant works, succeeds; he spends half his time defining his terms, and then he feels that he has solved the problem.

It must me noted, however, that he forcefully argued for the sphericity of the Earth *based on data:* he noted that only a spherical Earth can account for the shadow seen on the Moon during a lunar eclipse

Ptolemy enlarged Aristotle's ideas creating a very involved model of the solar system which endured until the Copernican revolution of the middle 16th century. When comparing the Ptolemaic system with the Copernican heliocentric system Occam's razor (Sect. **??**) instantly tells us to consider the latter first: it provide a much simpler explanation (and, as it turns out, a much better one) that the former.

2.5.1 Aristotelian Cosmology

Aristotle's cosmological work *On The Heavens* is the most influential treatise of its kind in the history of humanity. It was accepted for more that 18 centuries from its inception (around 350 B.C.) until the works of Copernicus in the early 1500s. In this work Aristotle discussed the general nature of the cosmos and certain properties of individual bodies.

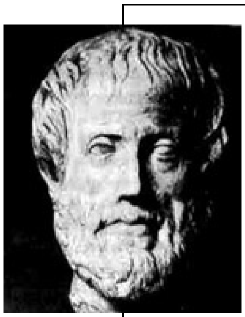
Aristotle believed that all bodies are made up of four elements: earth, water, air and fire (see Fig. 2.7). These elements naturally move up or down, fire being the lightest and earth the heaviest. A composite object will have the features of the element which dominates; most things are of this sort. But since the elements in, for example, a worm, are not where they belong (the fiery part is too low being bound by the earth part, which is

Aristotle believed that all bodies are made up of four elements: earth, water, air and fire

The elements naturally move up or down, fire being the lightest and earth the heaviest a bit too high), then the worm is imperfect. All things on earth are thus imperfect.

The idea that all bodies, *by their very nature*, have a natural way of moving is central to Aristotelian cosmology. Movement is *not*, he states, the result of the influence of one body on another

According to Aristotle all bodies, *by their very nature*, have a natural way of moving



Aristotle (384 B.C. - 322 B.C.). Born Stagirus, Greece, died Chalcis, Greece. In 367 Aristotle became a student at Plato's Academy in Athens. Soon he became a teacher at the Academy. After Plato's death in 347 B.C., Aristotle joined the court of Hermias of Atarneus. In 343 B.C. he became tutor to the young Alexander the Great at the court of Philip II of Macedonia. In 335 B.C. Aristotle founded his own school the Lyceum in Athens. The Academy had become narrow in its interests after Plato's death but the Lyceum under Aristotle pursued a broader range of subjects. Prominence was given to the detailed study of nature. After the death of Alexander the Great in 323 B.C., anti-Macedonian feeling in Athens made Aristotle retire to Chalcis where he died the following year. Aristotle was not primarily a mathematician but made important contributions by systematizing deductive logic. He wrote on physical subjects; some parts of his *Analytica Posteriora* show an unusual grasp of mathematics. He also had a strong interest in anatomy and the structure of living things in general which helped him to develop a remarkable talent for observation.

Some bodies naturally move in straight lines, others naturally stay put. But there is yet another natural movement: the circular motion. Since to each motion there must correspond a substance, there ought to be some things that naturally move in circles. Aristotle then states that such things are the heavenly bodies which are made of a more exalted and perfect substance than all earthly objects.

Since the stars and planets are made of this exalted substance and then move in circles, it is also natural, according to Aristotle, for these objects to be spheres also. The cosmos is then made of a central earth (which he accepted as spherical) surrounded by the moon, sun and stars all moving in circles around it. This conglomerate he called "the world". Note the strange idea that all celestial bodies are perfect, yet they must circle the imperfect Earth. The initial motion of these spheres was caused by the action of a "prime mover" which (who?) acts on the outermost sphere of the fixed stars; the motion then trickles down to the other spheres through a dragging force.

Aristotle also addresses the question whether this world is unique or not; he argues that it *is* unique. The argument goes as follows: earth (the substance) moves naturally to the center, if the world is not unique there ought to be at least two centers, but then, how can earth know to which of the two centers to go? But since "earthy" objects have no trouble deciding how to move he concludes that there can only be one center (the Earth) circled endlessly by all heavenly bodies. The clearest counterexample was found by Galileo when he saw Jupiter and its miniature satellite system (see Fig. 2.11), which looks like a copy of our "world". Aristotle was wrong not in the logic, but in the initial assumptions: things do *not* have a natural motion.

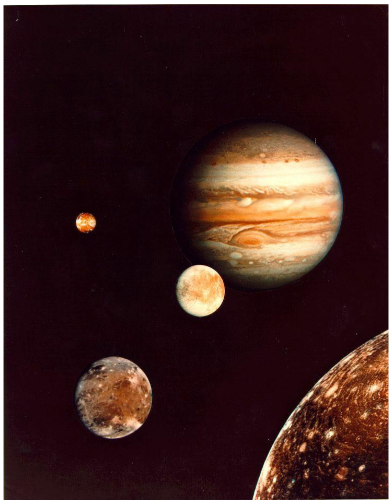


Figure 2.11: Montage of Jupiter and the Galilean satellites, Io, Europa, Ganymede, and Callisto.

It is interesting to note that Aristotle asserts that the world did not come into being at one point, but that it has existed, unchanged, for all eternity (it had to be that way since it was "perfect"); the universe is in a kind of "steady state scenario". Still, since he believed that the sphere was the most perfect of the geometrical shapes, the universe did have a center (the Earth) and its "material" part had an edge, which was "gradual" starting in the lunar and ending in the fixed star sphere. Beyond the sphere of the stars the universe continued into the spiritual realm where material things

Aristotle asserts that the world did not come into being at one point, but that it has existed, unchanged, for all eternity cannot be (Fig. 2.12). This is in direct conflict with the Biblical description of creation, and an enormous amount of effort was spent by the medieval philosophers in trying to reconcile these views.

Figure 2.12: A pictorial view of the Aristotelian model of the cosmos.

On the specific description of the heavens, Aristotle created a complex system containing 55 spheres(!) which, despite it complexity, had the virtue of explaining *and predicting* most of the observed motions of the stars and planets. Thus, despite all the bad publicity it has received, this model had all the characteristics of a scientific theory (see Sect. **??**): starting from the hypothesis that heavenly bodies move in spheres around the Earth, Aristotle painstakingly modified this idea, matching it to the observations, until all data could be accurately explained. He then used this theory to make predictions (such as where will Mars be a year from now) which were confirmed by subsequent observations.

2.5.2 The motion according to Aristotle

One of the fundamental propositions of Aristotelian philosophy is that there is no effect without a cause

One of the fundamental propositions of Aristotelian philosophy is that there is no effect without a cause. Applied to moving bodies, this proposition dictates that there is no motion without a force. Speed, then is proportional to force and inversely proportional to resistance

$$force = (resistance) \times (speed)$$

(though none of these quantities were unambiguously defined). This notion is not at all unreasonable if one takes as one's defining case of motion, say, an ox pulling a cart: the cart only moves if the ox pulls, and when the ox stops pulling the cart stops.

Aristotle's law of motion. (from *Physics*, book VII, chapter 5). Then, A the movement have moved B a distance Γ in a time Δ, then in the same time the same force A will move $\frac{1}{2}B$ twice the distance Γ, and in $\frac{1}{2}\Delta$ it will move $\frac{1}{2}B$ the whole distance for Γ: thus the rules of proportion will be observed.
The translation into modern concepts is $A \to F$ =force, $B \to m =$ mass, $\Gamma \to d$ =distance, and $\Delta \to t$ =time.
The statements then mean

- The distance is determined by the force F, the mass m and the time t
- Given a force F which moves a mass m a distance d in a time t, it will also move half the mass by twice the distance in the same time.
- Given a force F which moves a mass m a distance d in a time t, it will also move half the mass the same distance in half the time.

These three rules imply that the product of the mass and the average speed depends only on the force. For example, a body of constant mass under the action of a constant force will have a constant speed. This is wrong: the speed increases with time.

Qualitatively this implies that a body will traverse a thinner medium in a shorter time than a thicker medium (of the same length): things will go faster through air than through water. A natural (though erroneous) conclusion is that there could be no vacuum in Nature, for if the resistance became vanishingly small, a tiny force would produce a very large "motion"; in the limit where there is no resistance any force on any body would produce an infinite speed. This conclusion put him in direct contradiction with the ideas of the atomists such as Democritus (see Sect. 2.3.2). Aristotle (of course) concluded the atomists were wrong, stating that matter is in fact continuous and infinitely divisible.

Aristotle argued that there could be no vacuum in Nature

For falling bodies, the force is the weight pulling down a body and the resistance is that of the medium (air, water, etc.). Aristotle noted that a falling object gains speed, which he then attributed to a gain in weight. If weight determines the speed of fall, then when two different weights are dropped from a high place the heavier will fall faster and the lighter slower, in proportion to the two weights. A ten pound weight would reach the Earth by the time a one-pound weight had fallen one-tenth as far.

Aristotle asserted that when two different weights are dropped from a high place the heavier will fall faster and the lighter slower

2.5.3 Ptolemy

The Aristotelian system was modified by Hipparchus whose ideas were popularized and perfected by Ptolemy. In his treatise the *Almagest* ("The Great System") Ptolemy provided a mathematical theory of the motions of the Sun, Moon, and planets. Ptolemy vision (based on previous work by Hipparchus) was to envision the Earth surrounded by circles, on these circles he imagined other (smaller) circles moving, and the planets, Sun, etc. moving on these smaller circles. This model remained unchallenged for 14 centuries.

The system of circles upon circles was called a system of *epicycles* (see Fig. 2.14). It was extremely complicated (requiring several correction factors) but it did account for all the observations of the time, including the peculiar behavior of the planets as illustrated in Fig. 2.15. The Almagest was not superseded until a century after Copernicus presented his heliocentric theory in Copernicus' *De Revolutionibus* of 1543.

Ptolemy (100 - 170). Born in Ptolemais Hermii, Egypt, died Alexandria, Egypt. One of the most influential Greek astronomers and geographers of his time, Ptolemy propounded the geocentric theory that prevailed for 1400 years. Ptolemy made astronomical observations from Alexandria Egypt during the years A.D. 127-41. He probably spent most of his life in Alexandria. He used his observations to construct a geometric model of the universe which accurately predicted the positions of all significant planets and stars. This model employed combinations of circles known as epicycles, within the framework of the basic Earth-centered system supplied by Aristotle. His model is presented in his treatise *Almagest*. In a book entitled *Analemma* he discussed the projection of points on the celestial sphere. In *Planisphaerium* he is concerned with stereographic projection. He also devised a calendar that was followed for many centuries. There where problems with it, however, and this required corrections of about 1 month every 6 years. This generated a lot of problems in particular in agriculture and religion!

This model was devised in order to explain the motion of certain planets. Imagine that the stars are a fixed background in which the planets move, then you can imagine tracing a curve which joins the positions of a given planet everyday at midnight (a "join the dots" game); see, for example Fig 2.13. Most of the planets move in one direction, but Mars does not, its motion over several months is seen sometimes to backtrack (the same behavior would have been observed for other celestial objects had Ptolemy had the necessary precision instruments).

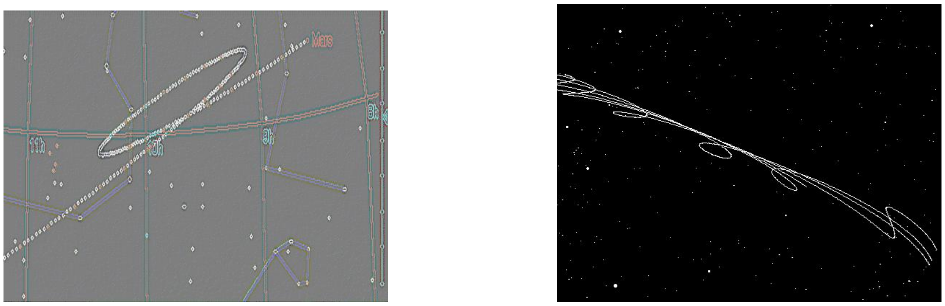


Figure 2.13: This computer simulations shows the retrograde motion of Mars (left) and the asteroid Vesta (right). Vesta's trajectory is followed over several years; it moves from right to left (west to east), and each loop occurs once per year. The shape of the retrograde loop depends on where Vesta is with respect to Earth.

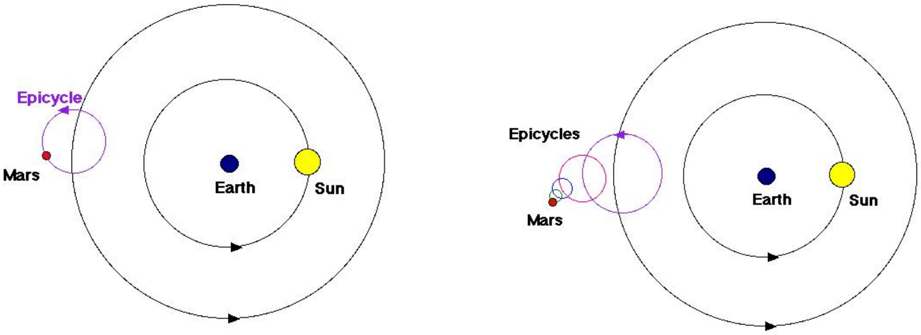


Figure 2.14: The simplest form of an epicycle (left) and the actual form required to explain the details of the motion of the planets (right).

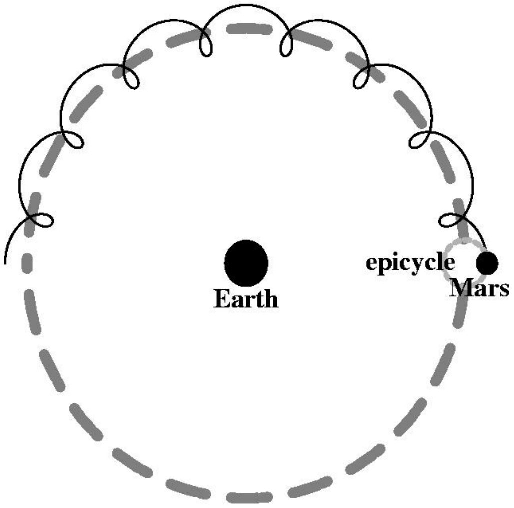


Figure 2.15: Example of how a system of epicycles can account for the backtracking in the motion of a planet. The solid line corresponds to the motion of Mars as it goes around the epicycle, while the epicycle itself goes around the Earth. As seen from Earth, Mars would move back and forth with respect to the background stars.

Chapter 3

From the Middle Ages to Heliocentrism

3.1 Preamble

The Roman empire produced no scientific progress in the area of cosmology, and the Church tainted it during most of the Middle Ages. Europe forgot most of the discoveries of the Greeks until they were reintroduced by Arab astronomers in the XII-th century through the Crusades and other less distressing contacts. The Renaissance brought a breath of fresh air to this situation, and allowed for the heretofore untouchable dogmas to be re-examined, yet, even in this progressive climate, the influence of the Church was still enormous and this hampered progress.

In the XVI-th century the Copernican view of the solar system saw the light. In this same era the quality of astronomical observations improved significantly and Kepler used these data to determine his famous three laws describing the motion of the planets. These discoveries laid the foundation for the enormous progress to be achieved by Galileo and then Newton.

3.2 The Middle Ages.

The development of new scientific theories came almost to a stop during the centuries covering the Roman Empire and the Middle Ages. During this long period there was a gradual emergence of irrational theories that threatened to engulf the whole of science: astrology challenged astronomy, magic insinuated itself into medicine and alchemy infiltrated natural science. The beginning of the Christian era, when Oriental mysticism became the rage in Greece and Rome, witnessed the appearance of exotic sects such as the Gnostics and the Hermetics who propagated distorted and over-simplified cosmologies ostensibly given to them by God [1].

During the Middle Ages European mental efforts were directed towards non-scientific pursuits. This attitude was perpetuated by the absence of libraries and the scarcity of books (both a consequence of the economic depression suffered by Europe at that time), and by the constraints imposed by the Church which forbade various ares of investigations as they were felt to be against the teachings of the Bible.

These problems did not permeate the whole world, however, and, in fact, Arab science flourished during this time devising the now-common Arab numerals, increasingly accurate time-keeping devices and astronomical instruments, and providing corrections to Ptolemy's observations. Later, through the close contacts generated by the Crusades, Arab knowledge was carried to Europe.

The scientific climate in Europe improved by the XIII century with the creation of the first universities. It was during this last part of the Middle Ages that the 3 dimensional nature of space was determined and the concept of force was made precise. The experimental basis of scientific inquiry was recognized as well as the need for internal logical consistency. With these developments the field was ready for the scientific developments of the Renaissance.

Through all these medieval tribulations Ptolemy's magnum opus, the *Alamgest*, together with Aristotle's *On The Heavens* survived as *the* cosmological treatises. Their influence became widespread after translations into Latin became readily available (at least at the universities). There was much discussion on the reconciliation of Aristotle's view of the world and the descriptions found in the Bible. Issues such as whether the universe is infinite and whether God can create an infinite object were the subject of heated discussions.

Sometimes the conclusions reached by the philosophers were not satisfactory to the theologians of the era and, in fact, in 1277 the bishop of Paris collected a list of 219 propositions connected with Aristotle's doctrine which no-one could teach, discuss or consider in any light under penalty of excommunication. For example,

- Aristotle argued against the possibility of there being other worlds, that is, copies of his set of spheres which are supposed to describe our

[1] From, *A History of Science* by H. Smith Williams.

world; these arguments can be interpreted as stating that God does not have the power to create such other worlds, an idea unacceptable to the Church.

- Aristotle assumed matter to be eternal and this contradicted the creation of matter, and in fact, of the whole universe by the will of God.
- Aristotelian advocates believed in the eternal pre-ordained motion of heavenly bodies which *nothing* could alter, this again implied limits on the powers of God.

In my opinion there is an interesting issue connected with the conflict between the Bible and Aristotle. It was Aristotle's belief that there are rules which objects are, by their very nature, forced to obey without the need for divine intervention. It is this idea that is prevalent in science today: there are natural laws that determine the behavior of inanimate objects without the intervention of higher authority. It is always possible to argue who or what determines these natural laws, whether there is some underlying will behind all of this. But that lies beyond the reach of science (at least in its present form), not because the question is of no interest, but because it cannot be probed using the reliable framework provided by the scientific method (Sect. **??**).

It was Aristotle's belief that there are natural laws that determine the behavior of inanimate objects without the intervention of higher authority

The problems with the theory of the universe perfected by Ptolemy were not apparent due to deficiencies in the instruments of the time. First was the problem of keeping time accurately: there were no precise clocks (a problem solved only when Galileo discovered the pendulum clock); a state of the art time-keeping mechanism of that time, the water-clock, is illustrated in Fig. 3.1; such mechanisms were not significant better than the water clocks used in Egypt starting form 1600 B.C. Secondly there was a notational problem: large numbers were very cumbersome to write since only Roman numerals were known (this notation has no notion of zero and of positional value; see Sect. **??** for a comparison between modern and Roman numerals).

These problems were recognized and (eventually) solved. The Arabic number system was slowly accepted in the Western world after its first introduction around 1100 A.D. during the Crusades. The discoveries of the other Greek scientists (not belonging to the Ptolemaic school) were also introduced in the West during this period in the same way. The first mechanical clocks waere developed in Europe in the XIII-th century. They worked using pulleys and weights but were still very inaccurate: the best ones were able only to give the nearest hour!

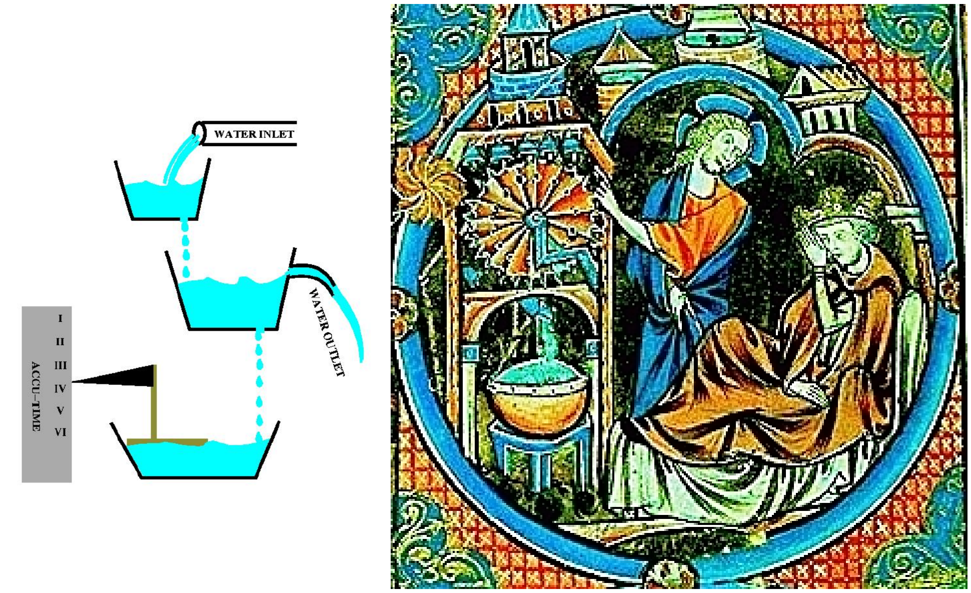


Figure 3.1: Illustrations of a water clock (left) and its use (right).

Despite the bad connotation the Middle Ages have, not all aspects of life during that time were horrible. In fact the basic ideas behind the universe in this time were very comforting to Jews, Christians and Muslims. These ideas provided a stable framework where most people had a (reasonably) clear view of their place in society, their duties and expectations.

The universe had the Earth at its center with all heavenly bodies circling it. Beyond the last sphere (that of the fixed stars) lay paradise, hell was in the bowels of the Earth (a sort of "under-Earth"), and purgatory was in the regions between Earth and the Moon (Fig. 3.2). One of the main architects of this vision was Thomas Aquinas whose view was adopted by Dante in his Divine Comedy.

The Middle Ages provided the gestation period during which the necessary conditions for the Renaissance were created. This is witnessed by the writings of various visionaries, with Roger Bacon as the best example. Bacon believed that Nature can be described using mathematics and required that all accepted theories be based on experimental evidence, not merely as conclusions drawn from ancient treatises (which themselves have not been tested). Many of these ideas were, of course, of Greek ancestry.

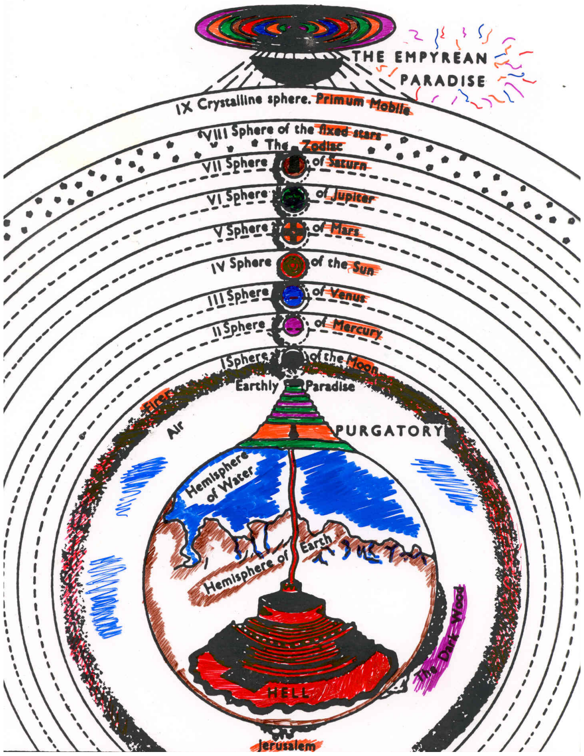


Figure 3.2: Illustration of a typical Medieval cosmological model.

Roger Bacon (1220–1292). He is remembered for his work in mathematics, and as a early advocate of the scientific method. He was a student at the university in Paris and later at Oxford in England. He became a Franciscan friar during the 1250s. His works include writings in mathematics, alchemy, and optics. He is known to have authored *Compendium of the Study of Philosophy* (1272) and *Compendium of the Study of Theology* (1292). During his life time he experimented with ideas about the development of gunpowder, flying machines, motorized vehicles, and telescopes.

Also worth of mentioning is William of Ockham, who parted from Plato's claim that ideas are *the* true and eternal reality (we only see imperfect shadows cast by these ideas, and this taints our perception of Nature). Ockham argued in his famous "razor" statement that this is an unnecessary complication in the description of Nature: *pluralitas non est ponenda sine neccesitate*, entities must not be needlessly multiplied, which was discussed extensively in Sect. **??**.

Entities must not be needlessly multiplied

William of Ockham (1285-1349). Born in Ockham – near Ripley, Surrey – England, died in Munich, Bavaria – now Germany. Ockham's early Franciscan education concentrated on logic. He studied theology at Oxford and between 1317 and 1319 he lectured on the Sentences , the standard theology text used in universities up to the 1600's. His opinions aroused strong opposition and he left Oxford without his Master's Degree. He continued studying mathematical logic and made important contributions to it. He considered a three valued logic where propositions can take one of three truth values. This became important for mathematics in the 20th Century but it is remarkable that it was first studied by Ockham 600 years earlier. Ockham went to France and was denounced by the Pope. He was excommunicated and in 1328 he fled seeking the protection of Louis IV in Bavaria (Louis had also been excommunicated). He continued to attack papal power always employing logical reasoning in his arguments until his death.

Yet the great majority of intellectuals accepted Ptolemy's model of the world. But, was this acceptance based on a belief that this was an accurate description of nature, or just on the fact that there no superior models to replace Ptolemy's? Some astronomers were of the second opinion, for example, the Arab astronomer Averroes declared (in his commentary on Aristotle's works) *"we find nothing in the mathematical sciences that would lead us to believe that eccentrics and epicycles exist"* and *"actually in our time astronomy is nonexistent; what we have is something that fits calculation but does not agree with what is"*. Similarly, Bacon believed that epicycles were a convenient mathematical description of the universe, but had no physical reality. Another notable exception to the general acceptance of Ptolemy's model was, perhaps not surprisingly, Leonardo da Vinci who at the dawn of the Renaissance concluded that the Earth moves (which implies that the Sun does not).

Leonardo da Vinci (1452-1519). Born in Vinci – near Empolia, Italy – died in Cloux – Amboise, France. Leonardo da Vinci had many talents in addition to his painting. He worked in mechanics but geometry was his main love. Received the usual elementary education of reading, writing and arithmetic at his father's house. From 1467 to 1477 he was an apprentice learning painting, sculpture and acquiring technical and mechanical skills; accepted into the painters' guild in Florence in 1472. From that time he worked for himself in Florence as a painter. During this time he sketched pumps, military weapons and other machines.

Was in the service of the Duke of Milan (1482–1499) as a painter and engineer. Completed six paintings and advised on architecture, fortifications and military matters. Also considered a hydraulic and mechanical engineer. During this time he became interested in geometry to the point of being neglectful of his paintings.

Leonardo studied Euclid and Pacioli's Suma and began his own geometry research, sometimes giving mechanical solutions, for example gave several such methods of squaring the circle. Wrote a book on the elementary theory of mechanics which appeared in Milan around 1498.

Leonardo certainly realized the possibility of constructing a telescope (as verified by sections of Codex Atlanticus and Codex Arundul). He understood the fact that the Moon shone with reflected light from the Sun. He believed the Moon to be similar to the Earth with seas and areas of solid ground.

In 1499 the French armies entered Milan and the Duke was defeated. Leonardo then left Milan, traveled to Mantua, Venice and finally Florence. Although he was under constant pressure to paint, but kept his mathematical studies; for a time was employed by Cesare Borgia as a senior military architect and general engineer.

By 1503 he was back in Florence advising on the project to divert the River Arno behind Pisa to help with the siege then suffered by the city. He then produced plans for a canal to allow Florence access to the sea (neither was carried out).

In 1506 Leonardo returned for a second period in Milan. again his scientific work took precedence over his painting and he was involved in hydrodynamics, anatomy, mechanics, mathematics and optics.

In 1513 the French were removed from Milan and Leonardo moved again, this time to Rome. Appears to have led there a lonely life more devoted to mathematical studies and technical experiments in his studio than to painting. After three years of unhappiness Leonardo accepted an invitation from King Francis I to enter his service in France. The French King gave Leonardo the title of first painter, architect, and mechanic of the King but seems to have left him to do as he pleased. This means that Leonardo did no painting except to finish off some works he had with him, St. John the Baptist, Mona Lisa and the Virgin and Child with St Anne. Leonardo spent most of his time arranging and editing his scientific studies. He died in 1519.

Finally I'd like to mention a peculiar alternative to the Aristotle+Ptolemy view of the world: the "Dairy cosmology", due to an Italian miller called Domenico Scandella (1532-1599/1600?), called Menoccio. Scandella believed that God and the angels were spontaneously generated by nature from the original chaos *"just as worms are produced from a cheese"*. The chaos was made of the four elements air, water, earth and fire, and out of them a mass formed *"just as cheese forms from milk"*. Within this mass of cheese, worms appeared, and *"the most holy majesty declared that these should be God and the angels."* Menoccio was tried by the Inquisition, found guilty and executed in 1599 or 1600.

3.3 The Copernican Revolution

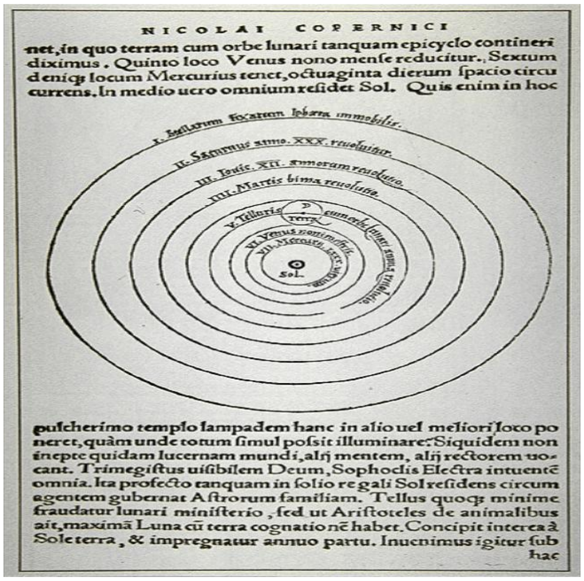
The 16th century finally saw what came to be a watershed in the development of Cosmology. In 1543 Nicolas Copernicus published his treatise *De Revolutionibus Orbium Coelestium* (The Revolution of Celestial Spheres) where a new view of the world is presented: the heliocentric model.

It is hard to underestimate the importance of this work: it challenged the age long views of the way the universe worked and the preponderance of the Earth and, by extension, of human beings. The realization that we, our planet, and indeed our solar system (and even our galaxy) are quite common in the heavens and reproduced by myriads of planetary systems provided a sobering (though unsettling) view of the universe. All the reassurances of the cosmology of the Middle Ages were gone, and a new view of the world, less secure and comfortable, came into being. Despite these "problems" and the many critics the model attracted, the system was soon accepted by the best minds of the time such as Galileo

Copenicus' model, a rediscovery of the one proposed by Aristarchus centuries before (see Sect. **??**), explained the observed motions of the planets (eg. the peculiar motions of Mars; see Fig. **??**) more simply than Ptolemy's by assuming a central sun around which all planets rotated, with the slower planets having orbits farther from the sun. Superimposed on this motion, the planets rotate around their axes. Note that Copernicus was not completely divorced from the old Aristotelian views: the planets are assumed to move in circles around the sun (Fig. 3.3).

Copernicus' model consisted of a central sun around which all planets rotated, with the slower planets having orbits farther from the sun

Nicholas Copernicus (Feb 19 1473 - May 24 1543). Born Torun, Poland, died Frauenburg, Poland. Copernicus studied first at the University of Krakow which was famous for mathematics, philosophy, and astronomy. Copernicus then studied liberal arts at Bologna from 1496 to 1501, medicine at Padua, and law at the University of Ferrara. He graduated in 1503 with a doctorate in canon law. He then took up duties at the cathedral in Frauenberg. during this period Copernicus performed his ecclesiastical duties, practiced medicine, wrote a treatise on monetary reform, and became interested in astronomy. In May 1514 Copernicus circulated in manuscript *Commentariolus*, the first outline of his heliocentric model; a complete description of which was provided in *De Revolutionibus Orbium Coelestium* in 1543. Copernicus suffered a stroke in 1542 and was bedridden by the time his magnum opus was published, legend has it that he saw the first copies (with an unauthorized preface by Osiander which tried to placate the Church's criticisms) the day he died.



NICOLAI COPERNICI

net, in quo terram cum orbe lunari tanquam epicyclo contineri diximus. Quinto loco Venus nono menſe reducitur., Sextum deniqꝫ locum Mercurius tenet, octuaginta dierum ſpacio circu currens. In medio uero omnium reſidet Sol. Quis enim in hoc

I. Stellarum fixarum ſphæra immobilis.
II. Saturnus anno. XXX. reuoluitur.
III. Iouis. XII. annorum reuolutio.
IIII. Martis bima reuolutio.
V. Telluris cum orbe lunari annua reuolutio.
☽
terra
VI. Venus nonimestris.
VII. Mercurii. LXXX. dierum.
Sol.

pulcherimo templo lampadem hanc in alio uel meliori loco poneret, quàm unde totum ſimul poſsit illuminare? Siquidem non inepte quidam lucernam mundi, alij mentem, alij rectorem uocant. Trimegiſtus uiſibilem Deum, Sophoclis Electra intuentē omnia. Ita profecto tanquam in ſolio regali Sol reſidens circum agentem gubernat Aſtrorum familiam. Tellus quoqꝫ minime fraudatur lunari miniſterio, ſed ut Ariſtoteles de animalibus ait, maximā Luna cū terra cognationē habet. Concipit interea à Sole terra, & impregnatur annuo partu. Inuenimus igitur ſub hac

Figure 3.3: The page in Copernicus' book *De Revolutionibus Orbium Coelestium* outlining the heliocentric model.

It must be noted that Copernicus not only put forth the heliocentric idea, but also calculated various effects that his model predicted (thus following the steps outlined in Sect. **??**). The presentation of the results was made to follow Ptolemy's *Almagest* step by step, chapter by chapter. Copernicus' results were quite as good as Ptolemy's and his model was simpler; but its predictions were not superior (since the planets do not actually move in circles but follow another – though closely related – curve, the ellipse); in order to achieve the same accuracy as Ptolemy, Copernicus also used epicycles, but now in the motion of the planets around the Sun. The traditional criticisms to the heliocentric model he answered thusly,

- To the objection that a moving Earth would experience an enormous centrifugal force which would tear it to pieces, Copernicus answered that the same would be true of, say, Mars in the Ptolemaic system, and worse for Saturn since the velocity is much larger.
- To the question of how can one explain that things fall downwards without using the Aristotelian idea that all things move towards the center, Copernicus stated that that gravity is just the tendency of things to the place from which they have been separated; hence a rock on Earth falls towards the Earth, but one near the Moon would fall there. Thus he flatly contradicted one of the basic claims of Aristotle regarding motion.
- To the objection that any object thrown upward would be "left behind" if the Earth moves, and would never fall in the same place, Copernicus argued that this will not occur as all objects in the Earth's vicinity participate in its motion and are being carried by it.

Copernicus was aware that these ideas would inevitably create conflicts with the Church, and they did. Though he informally discussed his ideas he waited until he was about to die to publish his magnum opus, of which he only printed a few hundred copies. Nonetheless this work was far from ignored and in fact was the first (and perhaps the strongest) blow to the Medieval cosmology. His caution did not save him from pointed criticisms, for example, Luther pointed out (from his *Tabletalk*)

> *There was mention of a certain new astrologer who wanted to prove that the Earth moves and not the sky, the Sun, and the Moon. This would be as if somebody were riding on a cart or in a ship and imagined that he was standing still while the Earth and the trees were moving* [2]*. So it goes now. Whoever wants to*

[2] This was a prescient remark, see Sect. **??** and Chap. 6.

> *be clever must agree with nothing that others esteem. He must do something of his own. This is what that fellow does who wishes to turn the whole astronomy upside down. Even in these things that are thrown into disorder I believe the Holy Scriptures, for Joshua commanded the Sun to stand still and not the Earth.* [3]

The Pope Paul III was not very critical, but his bishops and cardinals agreed with Luther and the model was condemned by the Church.

The heliocentric model was eventually universally accepted by the scientific community, but it spread quite slowly. There were several reasons for this, on the one hand there certainly was a reticence to oppose the authority of the Church and of Aristotle, but there was also the fact that the heliocentric model apparently contradicted the evidence of the senses. Nonetheless the model became better known and was even improved. For example, Copernicus' version had the fixed stars attached to an immovable sphere surrounding the Sun, but its generalizations did and assumed them to be dispersed throughout the universe (Fig. 3.4); Giordano Bruno even proposed that the universe is infinite containing many worlds like ours where intelligent beings live.

In fact it was Bruno's advocacy of the Copernican system that produced one of the strongest reactions by the Church: Bruno advocated not only the heliocentric model, but denied that objects posses a natural motion, denied the existence of a center of the universe, denying even the Sun of a privileged place in the cosmos. Bruno was executed by the Inquisition in 1600.

[3] This statement was produced during an informal after-dinner conversation and was published after Luther's death; it should therefore be taken with caution.

Giordano Bruno (1548–1600). Born in Nola, near Naples. He became a Dominican monk and learned Aristotelian philosophy and he was attracted to "unorthodox" streams of thought (eg. Plato). Left Naples (1576) and later Rome (1577) to escape the Inquisition. Lived in France until 1583 and in London until 1585. In 1584 he published *Cena de le Ceneri* (The Ash Wednesday Supper) and *De l'Infinito, Universo e Mondi* (On the Infinite Universe and Worlds). In the first he defended the heliocentric theory (though he was clearly confused on several points); in the second he argued that the universe was infinite, containing an infinite number of worlds inhabited by intelligent beings. Wherever he went, Bruno's passionate utterings led to opposition; he lived off the munificence of patrons, whom he finally outraged. In 1591 he moved to Venice where he was arrested by the Inquisition and tried; he recanted but was sent to Rome for another trial, he did not recant a second time. He was kept imprisoned and repeatedly interrogated until 1600 when he was declared a heretic and burned at the stake. It is often maintained that Bruno was executed because of his Copernicanism and his belief in the infinity of inhabited worlds. In fact, we do not know the exact grounds on which he was declared a heretic because his file is missing from the records. Scientists such as Galileo and Johannes Kepler were not sympathetic to Bruno in their writings.

The slow progress of the heliocentric model was also apparent among part of the scientific community of the time; in particular Tycho Brahe, the best astronomer of the late 16th century, was opposed to it. He proposed instead a "compromise": the earth moves around the sun, but the rest of the planets move around the Earth (Fig. 3.5). Brahe's argument against the Copernican system was roughly the following: if the Earth moves in circles around the Sun, nearby stars will appear in different positions at different times of the year. Since the stars are fixed they must be very far away but then they should be enormous and this is "unreasonable" (of course they only need to be enormously bright!)

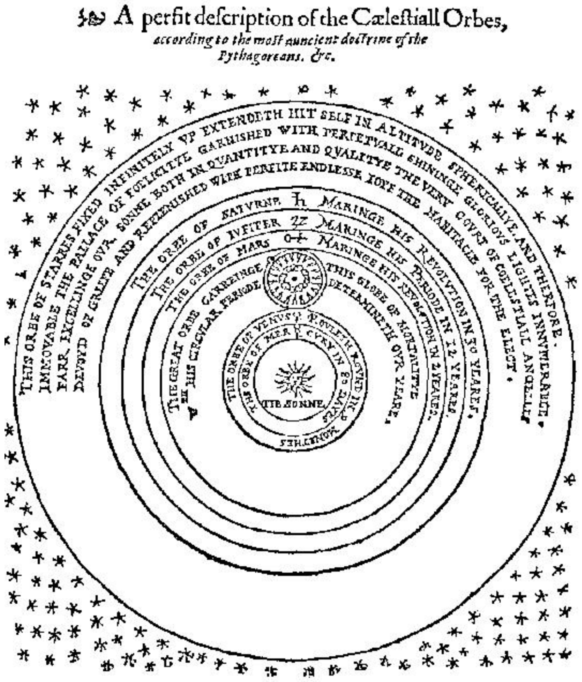

A perfit deſcription of the Cæleſtiall Orbes,
according to the moſt auncient doctrine of the Pythagoreans. &c.
THIS ORBE OF STARRES FIXED INFINITELY VP EXTENDETH HIT SELF IN ALTITVDE SPHERICALLYE, AND THERFORE IMMOVABLE THE PALLACE OF FOELICITYE GARNISHED WITH PERPETVALL SHININGE GLORIOVS LIGHTES INNVMERABLE, FARR EXCELLINGE OVR SONNE BOTH IN QVANTITYE AND QVALITYE THE VERY COVRT OF COELESTIALL ANGELLES, DEVOYD OF GREEFE AND REPLENISHED WITH PERFITE ENDLESSE IOYE THE HABITACLE FOR THE ELECT.
THE ORBE OF SATVRNE MAKINGE HIS REVOLVTION IN 30 YEARES.
THE ORBE OF IVPITER MAKINGE HIS PERIODE IN 12 YEARES.
THE ORBE OF MARS MAKINGE HIS REVOLVTION IN 2 YEARES.
THE GREAT ORBE CARRYINGE THIS GLOBE OF MORTALITYE DETERMINETH OVR YEARE, IN HIS CIRCULAR PERIODE
THE ORBE OF VENVS ROVLETH ROVND IN 9 MONETHES
THE ORBE OF MERCVRY IN 80 DAYES
THE SONNE

Figure 3.4: The heliocentric model of Thomas Digges (1546-1595) who enlarged the Copernican system by asserting that the stars are not fixed in a celestial orb, but dispersed throughout the universe.

Tycho Brahe (14 Dec 1546 - 24 Oct 1601). Born in Denmark he was fascinated by astronomy and, being a wealthy man (and being helped by the Danish monarchy), was able to devote a lot of time to the meticulous recording of the observed trajectories of the planets. He rented the island of Hven from the king of Denmark and set up a state of the art observatory there (without telescopes, they did not appear for 100 years). He later had to leave this island having has a disagreement with the king over religious matters. He then went to Prague as Imperial Mathematician and it was there that he interacted with Kepler. He did not adopt the Copernican system

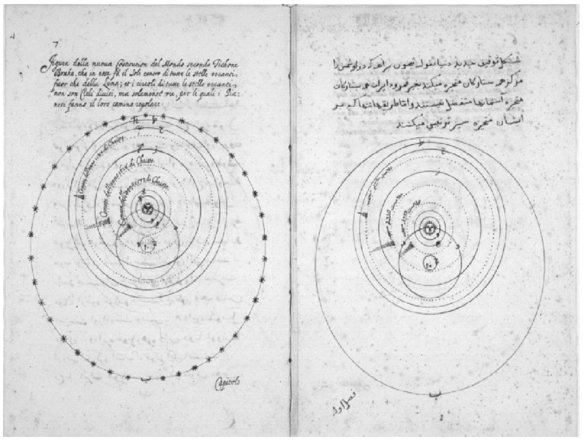


Figura della nuova costitution del Mondo secondo Tichone Brahe, che in essa fà il Sole centro di tutte le stelle erranti, fuor che della Luna; et i circoli di tutte le stelle erranti non son Cieli divisi, ma solamente vie, per le quali i Pianeti fanno il loro camino regolare.
Cometa dell'anno 1618 di Christo
Cometa dell'anno 1577 di Christo
Capricorno

Figure 3.5: Brahe's model of the universe: a central Earth around which the sun moves surrounded by the other planets [From the *Compendio di un trattato del Padre Christoforo Borro Giesuita della nuova costitution del mondo secondo Tichone Brahe e gli altri astologi moderni* (Compendium of a treatise of Father Christoforo Borri, S.J. on the new model of the universe according to Tycho Brahe and the other modern astronomers) by Pietro della Valle, Risalah- i Padri Khristafarus Burris Isavi dar tufiq-i jadid dunya.

3.3.1 Aristotle in the 16th century

In 1572 Tycho observed a star which suddenly appeared in the heavens (we now recognize this as an exploding star: a supernova). He noted that this "new star" did not change in position with respect to the other stars and should therefore be in the outer sphere of Aristotle's universe. But this was supposed to be an eternal, unchanging sphere! He published these observations in *The Nova Stella* in 1574. The same type of problems arose due to his observations of a comet which appeared in 1577, for he could determine that this object was farther than Venus again contradicting the Aristotelian idea that the universe beyond the Moon was perfect, eternal and unchanging. This is a case where better observations when pitted against the best theory of the time produced discrepancies which, in time, proved to be fatal to the current model and would eventually give rise to a better, more precise theory of the universe (see Sect. **??**).

By this time also most of the Medieval approach to physics had been shed, though not completely. For example, the motion of a projectile was thought to be composed of an initial violent part (when thrown) and a subsequent natural part (which returns it to the ground). Still it was during this time that the importance of velocity and force in determining the motion of objects was realized.

The birth of new theories is not easy, however. In this case it was not until the late 17th century that a complete new view of the universe was polished and could be used as a tool for investigating Nature. By this time the Aristotelian doctrine was, finally, set aside. The first step in this long road was taken by Copernicus, the next by Johaness Kepler in his investigations of the motion of the planets and then by Galielo through his investigations on the nature of motion and his description of the solar system.

3.4 Kepler

Johaness Kepler readily accepted the Copernican model, but his first attempts to understand the motion of the planets were still tied to the Aristotelian idea that planets "must" move on spheres. Thus his first model of the solar system was based on the following reasoning: there are, he argued, six planets (Uranus, Neptune and Pluto would not be discovered for almost 300 years) which move on the surfaces of spheres. There are also five perfect geometric figures, the Platonic solids: cube, tetrahedron, octahedron, icosahedron and dodecahedron. Then, he argued that the relative sizes of the spheres on which planets move can be obtained as follows (see Fig. 3.6)

- Take the Earth's sphere and put a dodecahedron around it.
- Put a sphere around this dodecahedron, Mars will move on it.
- Put a tetrahedron around Mars' sphere and surround it by a sphere, Jupiter will move on it.
- Put a cube around Jupiter's sphere and surround it by a sphere, Saturn will move on it.
- Put an icosahedron inside the Earth's sphere, then Venus will move on a sphere contained in it.
- Put a octahedron inside Venus sphere, then Mercury will move on a sphere just contained in it.

Therefore the ordering is octahedron, icosahedron, dodecahedron, tetrahedron, cube (8-faces, 20-faces, 12-faces, 4-faces, 6-faces). He spent 20 years trying to make this model work...and failed: the data would just not agree

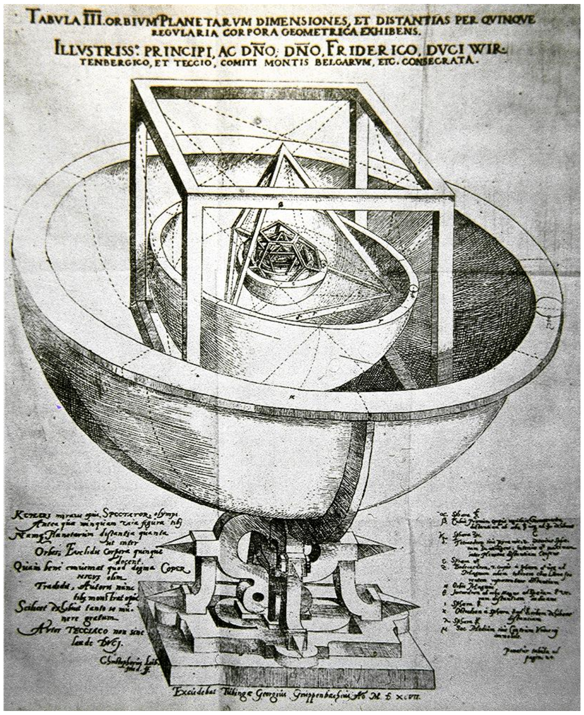

TABVLA III. ORBIVM PLANETARVM DIMENSIONES, ET DISTANTIAS PER QVINQVE REGVLARIA CORPORA GEOMETRICA EXHIBENS.
ILLVSTRISS: PRINCIPI, AC DNO. DNO. FRIDERICO, DVCI WIRTENBERGICO, ET TECCIO, COMITI MONTIS BELGARVM, ETC. CONSECRATA.

Figure 3.6: Illustration of Kepler's geometrical model of the solar system

with the model. Hard as this was, he dropped this line of investigation. This work, however, was of some use: he was recognized as "someone" and, in 1600, was hired by Tycho Brahe (then in Prague) as an assistant (at miserly wages). Tycho was very reluctant to share his data with Kepler (who was also made fun for being provincial); Tycho died in 1601 and the king appointed Kepler as successor (at a much smaller salary which was irregularly paid).

For many years thereafter Kepler studied Tycho's data using the heliocentric model as a hypothesis. In 1609 he determined that Mars does not move in a circle but in an ellipse with the sun in one of the foci and that in so moving it sweeps equal areas in equal times. This later blossomed into his first and second laws of planetary motion. Ten years after he discovered his third law: the cube of the average distance of a planet to the sun is proportional to the square of its period. All this was very important: Tycho's data, thanks to Kepler's persistence and genius, finally disproved the epicyclic theory and, on top of this, the idea that planets must move in circles.

This is a good example of the evolution of a scientific theory (see Sect. **??**). The data required Kepler to modify the original hypothesis (planets move in circles with the sun at the center) to a new hypothesis (planets move in ellipses with a sun at one focus). He showed that this was the case for Mars, and then checked whether it was also true for the other planets (it was).

Johannes Kepler. (Dec 27 1571 - Nov 15 1630). Born Weil der Stadt, Germany. Died Regensburg, Germany. Educated in Tübingen where he became acquainted with the Copernican system, which he embraced and sought to perfect; in 1596 he published *Mysterium Cosmographicum* in which he defended the Copernican theory and described his ideas on the structure of the planetary system. He was a devout Lutheran but inclined towards Pythagorean mysticism. He was intoxicated by numbers and searched for simple mathematical harmonies in the physical world; in particular he believed that the planets emit music as they travel and he even gave the various tunes. In 1609 he published *Astronomia Nova* ("New Astronomy") which contained his first two laws. In 1619 he published *Harmonices Mundi* (Harmonies of the World) in which he stated his third law.

The three laws obtained by Kepler are

1. Planets move in ellipses with the sun at one focus; see Fig. 3.7.
2. Planets sweep equal areas in equal times in their motion around the sun; see Fig. 3.8.
3. The average distance to the sun cubed is proportional to the period squared; see Table 3.1 for the data which led Kepler to this conclusion.

Kepler's 1st law: Planets move in ellipses with the sun at one focus

Kepler's 2nd law: Planets sweep equal areas in equal times in their motion around the sun

Kepler's 3rd law: The average distance to the sun cubed is proportional to the period squared

The first two laws describe the motion of single planets, the third law relates the properties of the orbits of different planets.

Kepler did not know why planets behaved in this way. It was only about 50 years later that Newton explained these laws in terms of his universal law of gravitation. In modern language these results imply the following (discovered by Newton): the planets move the way they do because they experience a force from the sun, this force is directed along the line from the planet to the sun, it is attractive and decreases as the square of the distance.

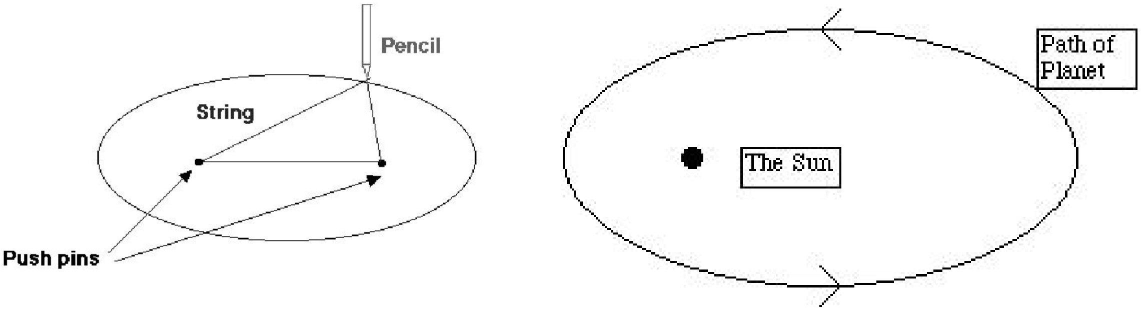


Figure 3.7: How to draw and ellipse (left) and the elliptical orbit of planets (right)

Planet	Period (years)	Avg. dist. (AU)	Period ^{2}	Dist^3
Mercury	0.24	0.39	0.06	0.06
Venus	0.62	0.72	0.39	0.37
Earth	1.00	1.00	1.00	1.00
Mars	1.88	1.52	3.53	3.51
Jupiter	11.9	5.20	142	141
Saturn	29.5	9.54	870	868

Table 3.1: Period and average distancs for the innermost five planets, a plot of the last two columns gives a straight line as claimed by Kepler's third law.

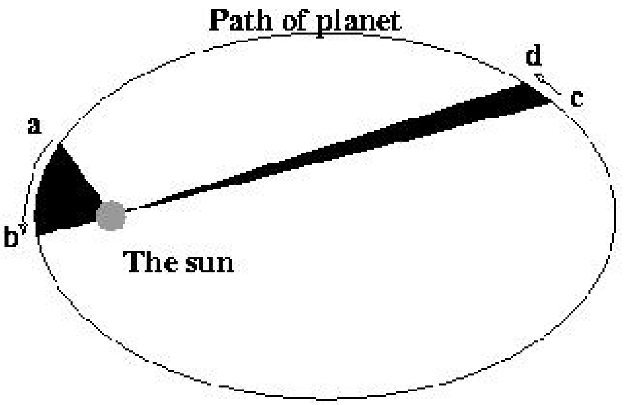
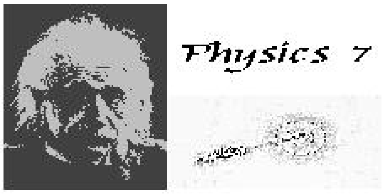


Figure 3.8: The planet in a given time moves from **a** to **b** sometime later it reaches **c** and it takes the *same* time to go from **c** to **d**. Kepler's second law states that the shaded areas are equal



Chapter 4

Galileo and Newton

4.1 Introduction

The discoveries of Kepler, and the paradigm of the solar system of Copernicus provided a very solid framework for the works of Newton and Galileo. The resulting theories changed the way we do science to this day and some of their ideas have withstood the passing of time with little change.

4.2 Galileo Galilei

Only rarely humankind is fortunate to witness the birth and flourishing of a mind as keen and fertile as Galileo's. To him we owe our current notions about motion and the concepts of velocity and acceleration. He was the first to use the telescope as an astronomical tool. Galileo was also creative in devising practical machines: he invented the first accurate clock, an efficient water pump, a precision compass and a thermometer. These achievements distinguish him as the preeminent scietist of his time.

Galileo's research in the exact sciences banished the last vestiges of Aristotelian "science" and replaced it with a framework within which the whole of physics would be constructed. These changes were not achieved without pain: Galileo was judged and condemned by the Inquisition and died while under house arrest after being forced to recant his Copernican beliefs.

Underlying all the discoveries made by Galileo there was a modern philosophy of science. He strongly believed, along the Pythagorean tradition, that the universe should be described by mathematics. He also adopted the view, following Ockham's razor (Sect. **??**), that given various explanations of a phenomenon, the most succinct and economic one was more likely to be the correct one. Still any model must be tested again and again against experiment: no matter how beautifuland economical a theory is, should it fail to describe the data, it is useless except, perhaps, as a lesson.

4.2.1 Galilean relativity

Imagine a person inside a ship which is sailing on a perfectly smooth lake at constant speed. This passeneger is in the ship's windowless hull and, despite it being a fine day, is engaged in doing mechanical experiments (such as studying the behavior of pendula and the trajectories of falling bodies). A simple question one can ask of this researcher is whether she can determine that the ship is moving (with respect to the lake shore) *without going on deck or looking out a porthole*.

Since the ship is moving at constant speed and direction she will not *feel* the motion of the ship. This is the same situation as when flying on a plane: one cannot tell, without looking out one of the windows, that the plane is moving once it reaches cruising altitutde (at which point the plane is flying at constant speed and direction). Still one might wonder whether the experiments being done in the ship's hull will give some indication of the its motion. Based on his experiments Galileo concluded that this is in fact impossible: all mechanical experiments done inside a ship moving at constant speed in a constant direction would give precisely the same results as similar experiments done on shore.

The conclusion is that one observer in a house by the shore and another in the ship will not be able to determine that the ship is moving by comparing the results of experiments done inside the house and ship. In order to determine motion these observers must look at each other. It is important important to note that this is true *only* if the ship is sailing at constant speed and direction, should it speed up, slow down or turn the researcher inside *can* tell that the ship is moving. For example, if the ship turns you can see all things hanging from the roof (such as a lamp) tilting with respect to the floor

Generalizing these observations Galileo postulated his **relativity hypothesis**:

> any two observers moving at constant speed and direction with respect to one another will obtain the same results for all mechanical experiments

(it is understood that the apparatuses they use for these experiments move with them).

In pursuing these ideas Galileo used the scientific method (Sec. **??**): he derived consequences of this hypothesis and determined whether they agree with the predictions.

Velocity is *not* absolute

This idea has a very important consequence: *velocity is* not *absolute.* This means that velocity can only be measured in reference to some object(s), and that the result of this measurment changes if we decide to measure the velocity with respect to a diferent refernce point(s). Imagine an observer traveling inside a windowless spaceship moving away from the sun at constant velocity. Galileo asserted that there are no mechanical experiments that can be made inside the rocket that will tell the occupants that the rocket is moving . In fact, the question "are we moving" has no meaning unless we specify a reference point ("are we moving with respect to that star" *is meaningful*). This fact, formulated in the 1600's remains very true today and is one of the cornerstones of Einstein's theories of relativity.

> *Turbulence. (from Relativity and Its Roots,* by B. Hoffmann*).* Although this question will seem silly, consider it anyway: Why do the flight attendants on an airplane not serve meals when the air is turbulent but wait until the turbulence has passed?
> The reason is obvious. If you tried to drink a cup of coffee during a turbulent flight, you would probably spill it all over the place.
> The question may seem utterly inane. But even so, let us not be satisfied with only a partial answer. The question has a second part: Why is it all right for the flight attendants to serve meals when the turbulence has passed?
> Again the reason is obvious. When the plane is in smooth flight, we can eat and drink in it as easily as we could if it were at rest on the ground.
> Yes indeed! And *that* is a most remarkable fact of experience. Think of it.

A concept associated with these ideas is the one of a "frame of reference". We intuitively know that the position of a small body relative to a reference point is determined by three numbers. Indeed consider three long rods at 90^o from one another, the position of an object is uniquely determined by the distance along each of the corresponding three directions one must travel in order to get from the point where the rods join to the object (Fig. 4.1)

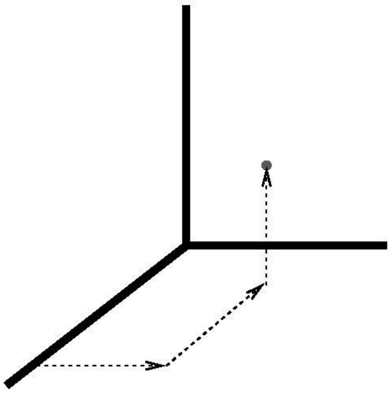


Figure 4.1: A frame of reference.

Thus anyone can determine positions and, if he/she carries clocks, motion of particles accurately by using these rods and good clocks. This set of rods and clock is called a *reference frame*. In short: a reference frame determines the where and when of anything with respect to a reference point.

A reference frame determines the where and when of anything with respect to a reference point.

A prediction of Galileo's principle of relativity is that free objects will move in straight lines at constant speed. A free object does not suffer form interactions from other bodies or agencies, so if it is at one time at rest in some reference frame, it will remain at rest forever in this frame. Now, imagine observing the body form another reference frame moving at constant speed and direction with respect to the first. In this second frame the free body is seen to move at constant speed and (opposite) direction. Still nothing has been done to the body itself, we are merely looking at it from another reference frame. So, in one frame the body is stationary, in another frame it moves at constant speed and direction. On the other hand if the body is influenced by something or other it will change its motion by speeding up, slowing down or turning. In this case either speed or direction are not constant as observed in ¡EM¿any¡/EM¿ reference frame. From these arguments Galileo concluded that free bodies are uniquely characterized by moving at constant speed (which might be zero) and direction.

An interesting sideline about Galilean relativity is the following. Up to that time the perennial question was, what kept a body moving? Galileo realized that this was the *wrong question*, since uniform motion in a straight line is not an absolute concept. The right question is, what keeps a body from moving uniformly in a straight line? The answer to that is "forces" (which are defined by these statements). This illustrates a big problem in physics, we have at our disposal all the answers (Nature is before us), but only when the right questions are asked the regularity of the answers before us becomes apparent. Einstein was able to ask a different set of questions and this lead to perhaps the most beautiful insights into the workings of Nature that have been obtained.

Galilean relativity predicts that free motion is in a straight line at constant speed. This important conclusion cannot be accepted without experimental evidence. Though everyday experience seems to contradict this conclusion (for example, if we kick a ball, it will eventually stop), Galileo realized that this is due to the fact that in such motions the objects are *not* left alone: they are affected by friction. He then performed a series of experiments in which he determined that frictionless motion would indeed be in a straight line at constant speed. Consider a ball rolling in a smooth bowl (Fig. 4.2).

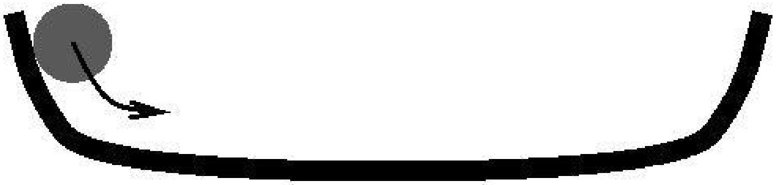


Figure 4.2: Illustration of Galileo's experiments with friction

The ball rolls from it's release point to the opposite end and back to a certain place slightly below the initial point. As the surfaces of the bowl and ball are made smoother and smoother the ball returns to a point closer and closer to the initial one. In the limit of zero friction, he concluded, the ball would endlessly go back and forth in this bowl.

Following this reasoning and "abstracting away" frictional effects he concluded that

Free horizontal motion is constant in speed and direction.

Free horizontal motion is constant in speed and direction

This directly contradicts the Aristotelian philosophy which claimed that

- all objects on Earth, being imperfect, will naturally slow down,
- that in a vacuum infinite speeds would ensue,
- and that perfect celestial bodies must move in circles.

In fact objects on Earth slow down due to friction, an object at rest would stay at rest even if in vacuum, and celestial bodies, as anything else, move in a straight line at constant speed or remain at rest unless acted by forces.

4.2.2 Mechanics

Most of Galileo's investigations in physics had to do with the motion of bodies; these investigations lead him to the modern description of motion in terms of position and time. He realized that two important quantities that describe the motion of all bodies are velocity (which determines how position changes with time) and acceleration (which determines how velocity changes with time)

Velocity tells how position changes with time

Acceleration tells how velocity changes with time

Two important definitions:

Velocity: the rate of change of position, (how position changes with time).	*Acceleration*: the rate of change of velocity, (how velocity changes with time).

The motion of falling bodies

Galileo realized, even during his earliest studies (published in his book *On motion*) that the speed of a falling body is *independent* of its weight [1]. He argued as follows: suppose, as Aristotle did, that the manner in which a body falls does depend on it weight (or on some other quality, such as its "fiery" or "earthy" character), then, for example, a two pound rock should fall faster than a one pound rock. But if we take a two pound rock, split it in half and join the halves by a light string then one the one hand this contraption should fall as fast as a two pound rock, but on the other hand it should fall as fast as a one-pound rock (see Fig. 4.3). Since any object should have a definite speed as it falls, this argument shows that the Aristotle's assumption that the speed of falling bodies is determined by their weight is inconsistent; it is simply wrong. Two bodies released from a given height will reach the ground (in general) at different times not because they have different "earthliness" and "fiery" characteristics, but merely because they are affected by air friction differently. If the experiment is tried in vacuum *any* two objects when released from a given height, will reach the ground simultaneously (this was verified by the Apollo astronauts on the Moon using a feather and a wrench).

[1] Galileo allegedly demonstrated his conclusions by dropping weights from the leaning tower of Pisa though this has been doubted by historians.

This result is peculiar to gravity, other forces do not beahve like this at all. For example, if you kick two objects (thus applying a force to them) the heavier one will move more slowly than the lighter one. In contrast, objects being affected by gravity (and starting with the same speed) will have the same speed at all times. This unique property of gravity was one of the motivations for Einstein's general theory of relativity (Chap. **??**).

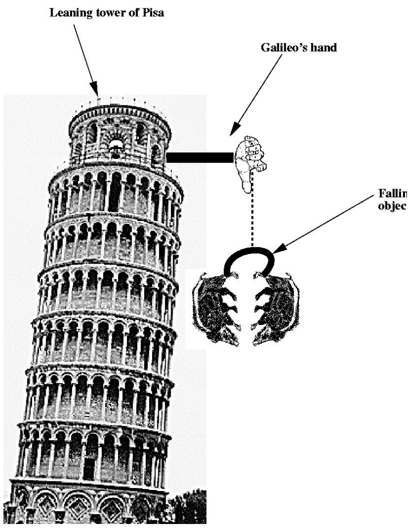


Figure 4.3: Illustration of Galileo's experiments with falling bodies.

Also in his investigations of falling bodies Galileo determined that the acceleration of these bodies is constant. He demonstrated that an object released from a height starts with zero velocity and increases its speed with time (before him it was thought that bodies when released acquire instantaneously a velocity which remained constant but was larger the heavier the object was). Experimenting with inclined planes, and measuring a ball's positions after equal time intervals Galileo discovered the mathematical expression of the law of falling bodies: the distance traveled increases as the *square* of the time.

The motion of projectiles

Galileo also considered the motion of projectiles. He showed that their motion can be decomposed in a motion along a vertical and horizontal directions. Thus if a ball is thrown horizontally (and air friction is ignored) it will move in the horizontal direction with constant speed; in the vertical direction it will experience the pull of gravity and will undergo free fall. The use of this can be illustrated by the following situation. Suppose a ball is let fall from a height h and is found to take t seconds to reach the ground. Now suppose that the ball is instead thrown horizontally with speed v, what distance will it cover? The answer is vt because the ball, even though it is moving horizontally, in the vertical direction is still freely falling: the two motions are completely independent! (see Fig. 4.4). This, of course, was of great use in warfare.

Motions along perpendicular directions are completely independent

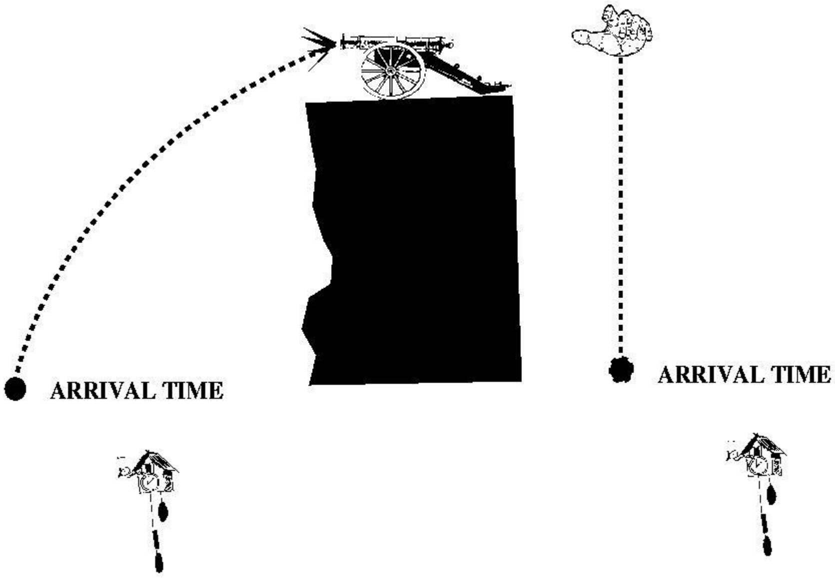


Figure 4.4: Horizontal and vertical motion are independent: the cannon shoots the ball *horizontally* at the same time the hand drops its ball; they both hit the ground at the same time.

As another experiment consider the “shoot the monkey” demonstration (Fig. 4.5). The setup is the following: a hunter wants to shoot a monkey who is hanging from a branch. As soon as he shoots the monkey lets go of the branch (thinking that the hunter aimed at the branch, he believes that the bullet will miss him). But the bullet, to the monkey’s surprise (and distress), does hit him! [2]

The reason is the following: if there were no forces the bullet would go in a straight line (as indicated by the dotted line in the figure) and the monkey would not fall. So the bullet would hit the monkey. Now, since

[2] No real animals were hurt in this demonstration.

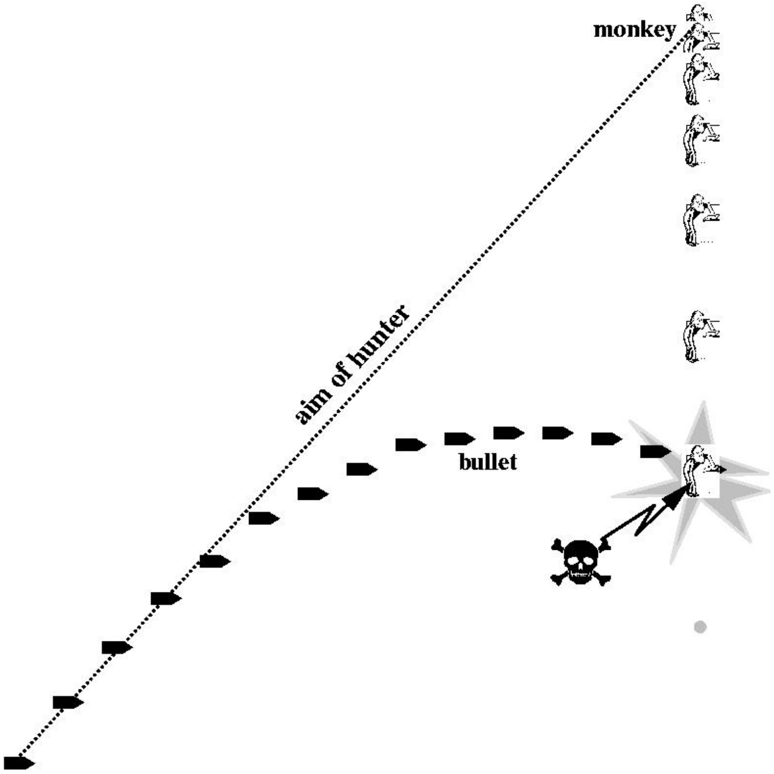


Figure 4.5: Shoot the “monkey”: an illustration of motion in two dimensions.

we have a force acting on the system (gravity) the monkey will not stay at rest but will accelerate downward. But precisely the same force acts on the bullet in precisely the same way, hence the bullet will not go in a straight line but will follow the curve indicated in the figure. The deviation from their force-free motions (rest for the monkey, straight line for the bullet) are produced by a force which generates the same acceleration in both objects, hence these deviations are precisely matched in such a way that the bullet hits the monkey.

Now, given a force of *constant strength*, it will affect bodies in varying degrees; the more massive the object the smaller the effect: a blow from a hammer will send a small ball flying, the same blow will hardly affect a planet. On the other hand gravity produces the same *acceleration* on the monkey and the bullet; that is why the monkey is hit. Since the mass of the monkey is very different from that of the bullet we conclude that gravity’s *force* is very different for each of them. The fact that the accelerations are independent of the mass but the force is not is actually a very profound fact: the whole of general relativity is based on it (Chap. **??**).

4.2.3 Astronomy

Throughout his life Galileo would provide some of the most compelling arguments in favor of the heliocentric model; though this brought him endless trouble in his lifetime, he was vindicated by all subsequent investigators. The beginnings of Galileo's astronomical studies were quite dramatic: in 1604 a "new star" (a supernova—an exploding star) was observed,. Galileo demonstrated that this object must lie beyond the Moon, contradicting the Aristotelian doctrine which claimed that the region beyond the Moon was perfect and unchanging. Yet here was a star that was not there before and would soon disappear!

A few years later he learned about the discovery of the telescope. He quickly realized its potential as a tool in astronomical research, and constructed several of them (Fig. 4.6), which he used to investigate the heavens.

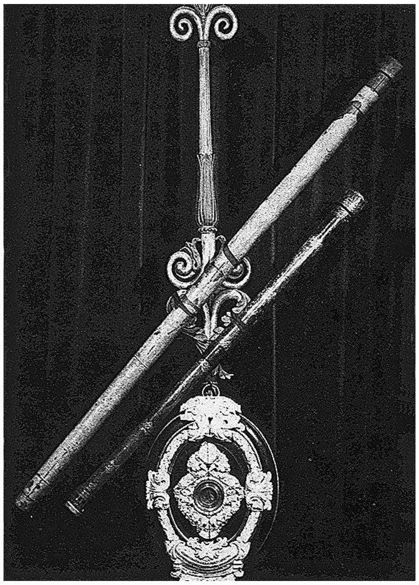


Figure 4.6: One of Galileo's telescopes

The first object which he studied with his telescope was the Moon of which he made many drawings (Fig. 4.7) some of which are quite accurate. He found that the surface of the Moon was heavily scarred, and identified some of the dark features he observed as shadows. The Moon was not exactly spherical and hardly perfect.

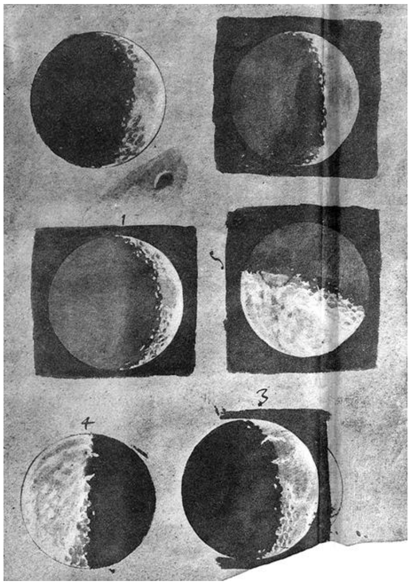


Figure 4.7: Galileo's drawings of the Moon.

Galileo was the first person to discover that Venus, like the Moon, shows periodic phases (Fig. 4.8). The simplest explanation is that this planet goes around the sun in accordance with the Copernican system. Galileo's astronomical observations were later verified by the Jesuit mathematicians of the Collegio Romano (although they did not necessarily agree with Galileo's interpretation!).

But the most dramatic of Galileo's astronomical discoveries was that of Jupiter's satellites (1610) [3]. He found that Jupiter was surrounded by a swarm of bodies that circled *it* and not Earth! These satellites, together with Jupiter, formed a mini-version of the Copernican model of the solar system with Jupiter taking the place of the Sun and it's satellites the places of the planets. All this was in blatant contradiction of the Aristotelian model; any remaining doubts which he might have had in his belief of the heliocentric model vanished.

Jupiter and its satellites formed a mini-version of the Copernican model of the solar system with Jupiter taking the place of the Sun and it's satellites the places of the planets

In 1613, in a book on sunspots, Galileo openly declared the Earth to circle the Sun. But by then the Church was getting worried about these

[3] This landed him a permanent position as "Chief Mathematician of the University of Pisa and Philosopher and Mathematician to the Grand Duke of Tuscany"

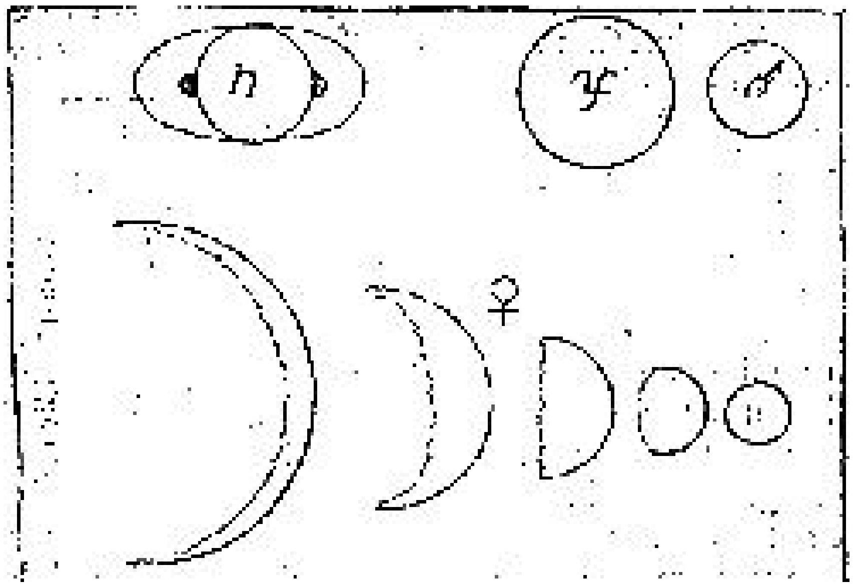


Figure 4.8: Galileo's drawings of the phases of Venus.

ideas: in 1616 Pope Pius V declared the Earth to be at rest and labeled the heliocentric model heretical, Copernicus' magnum opus was black-listed (where it remained until 1822!), and Galileo was called to Rome and told not to defend Copernicus' ideas.

In 1632 Galileo published his book on the Copernican and Ptolemaic systems *Dialogue Concerning the Two Chief Systems of the World* (in Italian so everyone could understand it). This was originally condoned by the Church, but the Pope Urban VIII had a change of heart and forbade the distribution of the book. Galileo was summoned to appear before the Roman Inquisition where, in a penitential garb and on one knee, he was made to swear on the Bible that he

> "...abjured, cursed, and detested the error and heresy that the Sun is fixed and the Earth moves"

and that he would no longer support this idea in any manner. He was put under house arrest and was made to recite the seven penitential psalms weekly for three years. This, of course, did not change the fact that the planets do move around the sun, but it embittered Galielo's last years.

4.2.4 Galileo and the Inquisition

Being one of the most renowned scientist of his time Galileo's opinions were scrutinized not only be his peers, but by also by Church officials and the public in general. This made Galileo the lightning-rod of many complaints against the Copernican doctrine (and also some against Galileo himself). He did not come out unscathed out of these encounters.

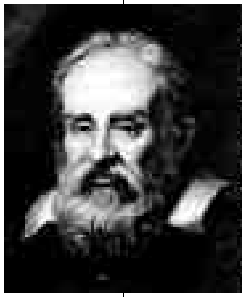
In 1611 Galileo came to the attention of the Inquisition for the first time for his Copernican views. Four years later a Dominican friar, Niccolo Lorini, who had earlier criticized Galileo's view in private conversations, files a written complaint with the Inquisition against Galileo's Copernican views. Galileo subsequently writes a long letter defending his views to Monsignor Piero Dini, a well connected official in the Vatican, he then writes his *Letter to the Grand Duchess Christina* arguing for freedom of inquiry and travels to Rome to defend his ideas

In 1616 a committee of consultants declares to the Inquisition that the propositions that the Sun is the center of the universe and that the Earth has an annual motion are absurd in philosophy, at least erroneous in theology, and formally a heresy. On orders of the Pope Paul V, Cardinal Bellarmine calls Galileo to his residence and administers a warning not to hold or defend the Copernican theory; Galileo is also forbidden to discuss the theory orally or in writing. Yet he is reassured by Pope Paul V and by Cardinal Bellarmine that he has not been on trial nor being condemned by the Inquisition.

In 1624 Galileo meets repeatedly with his (at that time) friend and patron Pope Urban VIII, he is allowed to write about the Copernican theory as long as he treated it as a mathematical hypothesis.

In 1625 a complaint against Galileo's publication *The Assayer* is lodged at the Inquisition by a person unknown. The complaint charges that the atomistic theory embraced in this book cannot be reconciled with the official church doctrine regarding the Eucharist, in which bread and wine are "transubstantiated" into Christ's flesh and blood. After an investigation by the Inquisition, Galileo is cleared.

In 1630 he completed his book *Dialogue Concerning the Two Chief World Systems* in which the Ptolemaic and Copernican models are discussed and compared and was cleared (conditionally) to publish it by the Vatican. The book was printed in 1632 but Pope Urban VIII, convinced by the arguments of various Church officials, stopped its distribution; the case is referred to the Inquisition and Galileo was summoned to Rome despite his infirmities.



Galileo Galilei (Feb. 15, 1564–1642). Born near Pisa, Italy, died near Florence, Italy. In 1581 he matriculates as a student of the Arts at the University of Pisa (his father's wish is that he study medicine) and he is first introduced to Euclid's Elements while studying in Florence under the court mathematician Ostilio Ricci. In 1585 he returns to Florence without a degree. He gives private lessons in mathematics until 1589; he begins his studies in physics. In 1588 he obtained a lectureship of mathematics at the Univ. of Pisa where he taught until 1592; he publishes *On motion*. In 1592 Galileo obtains the chair of mathematics at the University of Padua in the Venetian Republic where he remains until 1610.

In 1599 he enters a relationship with Marina Gamba with whom he had three children, two daughters and one son. The daughters were placed in a convent as Galileo could not provide adequate dowries; he eventually managed to have his son legitimated. In 1613 Marina Gamba married Giovanni Bartoluzzi, it appears that Galileo kept cordial relations with Gamba and Bartoluzzi.

In 1609, he observes (using telescopes of his construction) the Moon, and discovers 4 satellites around Jupiter. In this year he was also appointed (for life) "Chief Mathematician of the University of Pisa and Philosopher and Mathematician to the Grand Duke of Tuscany". In 1611 he is admitted to the Lycean Academy and came to the attention of the Inquisition for the first time . In 1615 he is denounced to the the Inquisition, he defends himself in the *Letter to the Grand Duchess Christina*. In 1616 the Copernican doctrine is declared heretical, Galileo is warned against supporting this theory either orally, but he is allowed to write about it as a mathematical hypothesis. In 1621 Galileo is elected Consul of the Accademia Fiorentino. In 1625 a complaint to the Inquisition against Galileo's publication *The Assayer* is lodged by a person unknown; the complaint charges that the atomistic theory embraced in this book is heretical; Galileo is cleared.

In 1630 completes his book *Dialogue Concerning the Two Chief World Systems* contrasting the Ptolemaic and Copernican models. The book was printed in 1632 but the Pope Urban VIII stopped its distribution; the case is referred to the Inquisition and Galileo was summoned to Rome despite his physical infirmities. A year later Galileo is formally interrogated by the Inquisition. He recants of his support of the Copernican model and is ordered held under house arrest where he would remain until his death; also in 1633 he begins writing his *Discourse on Two New Sciences*. His health deteriorates steadily, in 1634 he suffers a painful hernia, by 1638 he is totally blind. Galileo dies in Arcetri on 8 January 1642.

Galileo also invented several objects of great practical interest such as an hydrostatic balance (1608), a horse-driven water pump (1593), a geometric and military compass (1597), various telescopes (1609) and a thermometer (1606). In 1641 he conceives of the application of the pendulum to clocks.

In 1633 Galileo was formally interrogated for 18 days and on April 30 Galileo confesses that he may have made the Copernican case in the Dialogue too strong and offers to refute it in his next book. Unmoved, the Pope decides that Galileo should be imprisoned indefinitely. Soon after, with a formal threat of torture, Galileo is examined by the Inquisition and sentenced to prison and religious penances, the sentence is signed by 6 of the 10 inquisitors. In a formal ceremony at a the church of Santa Maria Sofia Minerva, Galileo abjures his errors. He is then put in house arrest in Sienna. After these tribulations he begins writing his *Discourse on Two New Sciences*.

Galileo remained under house arrest, despite many medical problems and a deteriorating state of health, until his death in 1642. The Church finally accepted that Galileo might be right in 1983.

4.3 Isaac Newton

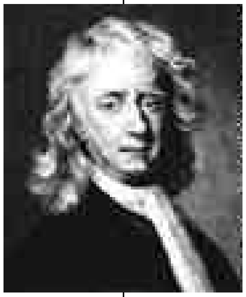
On Christmas day 1642, in the manor house of Woolsthorpe, a weak child was born and christened Isaac. He was to become the most influential scientist of the next 250 years. Isaac Newton discovered the laws that explained all phenomena known at the time, form the motion of the stars to the behavior of dust particles. It was his extremely successful model that lead people to believe that humanity was on the verge of understanding the whole of Nature.

Newton's life can be divided into three quite distinct periods. The first is his boyhood days from 1642 up 1665 when the Plague forced him to leave Cambridge. The second period from 1665 to 1687 was the highly productive period in which he became Lucasian professor at Cambridge. The third period (nearly as long as the other two combined) saw Newton as a highly paid government official in London with little further interest in science and mathematics.

I will talk about Newton quite a bit because his view of the world together with the mathematical formalism he developed lasted for 200 years: the first experimental results incompatible with it were obtained at the end of the XIX-th century and the whole structure was shown not to be fundamentally correct by 1925. One nonetheless should be aware of the fact that, while not perfectly correct, the results using the Newtonian are exceedingly accurate in all every-day applications. Newton's theory is not "wrong" it's just that it has a limited range of validity.

Isaac Newton (1643–1727). Born in the manor house of Woolsthorpe, near Grantham in Lincolnshire on Christmas Day 1642. Newton came from a family of farmers; his father died before he was born. His mother remarried, moved to a nearby village, and left him in the care of his grandmother. Upon the death of his stepfather in 1656, Newton's mother removed him from grammar school in Grantham where he had shown little promise in academic work. His school reports described him as 'idle' and 'inattentive'. Legend has it that one day the student just ahead of him in class kicked him in the stomach, Newton won the fight and he also decided to get ahead of this student in class ranking. He succeeded admirably. An uncle decided that he should be prepared for the university, and he entered his uncle's old College, Trinity College, Cambridge, in June 1661.

Instruction at Cambridge was dominated by the philosophy of Aristotle but some freedom of study was allowed in the third year of study. Newton's aim at Cambridge was a law degree, yet he also studied the philosophy and analytical geometry of Descartes, Boyle's works, and the mechanics of the heliocentric astronomy of Galileo.



His scientific genius flourished suddenly when the "Black Death" plague closed the University in the summer of 1665 and he had to return to Lincolnshire. There, in a period of less than two years, while Newton was still under 25 years old, he began revolutionary advances in optics, physics, and astronomy. In mathematics he laid the foundation for differential and integral calculus several years before its independent discovery by Leibniz. (this work, *De Methodis Serierum et Fluxionum*, was written in 1671 but appeared only 60 years later).

Impressed with Newton's abilities, Barrow resigned the Lucasian chair in 1669 recommending that Newton (still only 27 years old) be appointed in his place. Newton's first work as Lucasian Professor was on optics. Newton was elected a fellow of the Royal Society in 1672 after donating a reflecting telescope. In that year he published his first scientific paper on light and color in the Philosophical Transactions of the Royal Society.

Newton's relations with the influential scientist Robert Hooke deteriorated and Newton turned away from the Royal Society and mainstream science; he delayed the publication of a full account of his optical researches until after Hooke's death in 1703: Newton's *Opticks* appeared in 1704.

Newton's greatest achievement was his work in physics and celestial mechanics, which culminated in the theory of universal gravitation. His results are summarized in his treatise of physics *Philosophiae Naturalis Principia Mathematica* which appeared in 1687.

After suffering a nervous breakdown in 1693, Newton retired from research to take up a government position in London becoming Warden of the Royal Mint (1696) and Master (1699). In 1703 he was elected president of the Royal Society and was re-elected each year until his death. He was knighted in 1708 by Queen Anne, the first scientist to be so honored for his work. Newton died in 1727; his tomb in Westminster Abbey is inscribed with these words: " Mortals! Rejoice at so great an ornament to the human race!"

4.3.1 Mechanics.

During the years of the Plague Newton constructed what was to become an remarkably successful model of Nature. In it he proposed three laws that describe the motion of all material bodies (at least for all phenomena within reach at the time). These were not mere descriptions but actual calculational tools, and the enormous accuracy in the predictions achieved by this theory resulted in its universal acceptance that lasted more than two centuries...until Einstein came along.

After returning to Cambridge, Newton lost interest in mechanics until 1684. In this year Halley, tired of Hooke's boasting, asked Newton whether he could prove Hooke's conjecture that planets moved in ellipses because the sun attracted them with a force decreasing as the square of the distance. Newton told him that he had indeed solved this problem five years earlier, but had now mislaid the proof. At Halley's urging Newton reproduced the proofs and expanded them into a paper on the laws of motion and problems of orbital mechanics. Halley then persuaded Newton to write a full treatment of his new physics and its application to astronomy. Over a year later (in 1687) Newton published the *Philosophiae Naturalis Principia Mathematica* or the *Principia* as it is commonly known. It is one of the greatest scientific books ever written.

Newton laid in his Principia three laws which describe the motion of bodies. These laws have an immense range of applicability, failing only at very small distances (of 10^{-8}cm or less), for very strong gravitational fields (about 10^8 stronger than the Sun's), or for very large speeds (near 10^8 m/s).

The first of Newton's laws addresses the motion of free bodies. The second law states quantitatively how a motion differs form free motion. The third law states the effect experienced by a body when exerting a force on another object.

- *1st law.* Every body continues its state of rest or uniform motion in a straight line unless it is compelled to change this state by forces acting on it.

Free bodies move in straight lines or remain at rest

- *2nd law.* The effect of a force F on the motion of a body of mass m is given by the relation

  $$F = ma$$

  where a is the acceleration: a body in the presence of a force F attains an acceleration equal to F/m.

Force=mass×acceleration

- *3rd law.* Every body exerting a force on another, experiences a force exerted by the second body equal in magnitude and in opposite direction.

Action=reaction

These three laws constitute Newton's basic hypothesis He asserted that they are valid in all circumstances and to all bodies, in particular for heavenly bodies as well as for earth objects; this marks the final passing of Aristotelian physics. All experimental evidence of the time (and for the next two centuries) was to support these hypothesis, Newton's theory became *the* theory of Nature.

I will now discuss some of the features of these laws.

1st Law and Newtonian space and time.

One of the most important consequences of the First Law is that it *defines* what we mean by an inertial frame of reference.

> An inertial reference frame is a reference frame where isolated bodies are seen to move in straight lines at constant velocity.

An inertial reference frame is a reference frame where isolated bodies are seen to move in straight lines at constant velocity

An observer at rest with respect to an inertial frame of reference is called an *inertial observer*. The laws of physics devised by Newton take a particularly simple form when expressed in terms of quantities measured by an inertial observer (such as positions, velocities, etc.). For example, an inertial observer will find that a body on which no forces act moves in a straight line at constant speed or is at rest.

All motion occurs in space and is measured by time. In Newton's model both space and time are unaffected by the presence or absence of objects. That is *space and time are absolute*, an arena where the play of Nature unfolds. In Newton's words,

Newton assumed that space and time are absolute

> *Absolute space in its own nature, without relation to anything external, remains always similar and immovable.*
>
> *...absolute and mathematical time, of itself, and from its own nature, flows equally without relation to anything external, and by another name is called duration.*

Space and time were taken to be featureless objects which served as a universal and preferred reference frame (see Fig. 4.9 for an illustration). A consequence of this is that a given distance will be agreed upon by any two observers at rest with respect to each other or in uniform relative motion, for, after all, they are just measuring the separation between two immovable points in eternal space. In the same way a time interval will be agreed upon by *any* two observers for they are just marking two notches on eternal time.

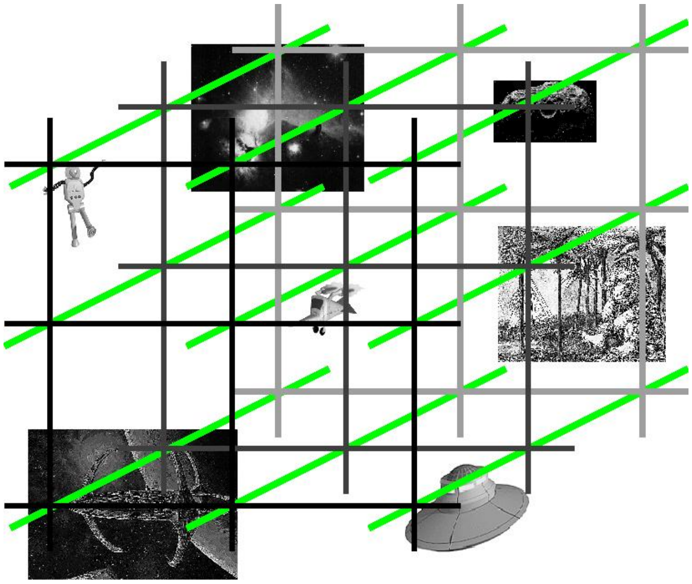


Figure 4.9: Illustration of Newton's concept of space. The grids represent space which are unaffected by the presence and properties of the objects in it.

Newton's assumptions about space and time are the foundation of his theory of Nature and were accepted due to the enormous successes of the predictions. Eventually, however, experimental results appeared which disagreed with the predictions derived from Newton's theory. These problems were traced to the fact that these basic assumptions are not accurate descriptions of space and time (though they do represent a very good approximation): space and time are not absolute (Chaps. **??**, **??**) [4]. The realization that Newton's theory required revisions came to a head at the beginning of the XXth century. In the two decades from 1905 to 1925 a completely new framework was constructed and has now replaced Newton's ideas. These theories comprise the special and general theories of relativity and quantum mechanics.

[4] $F = ma$ is also not universally valid but deviations from this expression occur only at very small distances and can be understood in the framework of Quantum Mechanics.

Do we know that the current theories of space and time are <u>the</u> truth? The answer is no: we do know that the current theories explain all the data (including the one explained by Newton and more), but we cannot determine whether they represent the ultimate theories of Nature. In fact, we expect them not to be the last word as there are many unexplained questions; for example, why should the proton be precisely 1836.153 times heavier than the electron? Why should space have 3 and not 25 dimensions? etc.

But in the 17th century there was no inkling of these problems and very few scientist questioned Newton's hypothesis. In particular Newton constructed his mechanics to comply with Galilean relativity: an observer in uniform motion with respect to another cannot, without looking outside his laboratory, determine whether he is at rest or not. And even if he looks outside, he cannot decide whether he is in motion or the other observer is. In fact for two inertial observers moving relative to each other the question, "which of us is moving?" is un-answerable and meaningless. The only thing to be said is that they have a certain relative velocity.

2nd Law

The second law is of great practical use. One can use experiments to determine the manner in which the force depends on the position and velocity of the bodies and then use calculus (which was also invented by Newton) to determine the motion of the bodies by obtaining the position as a function of time using the known form of F and the equation $F = ma$. Note that in this equation m measures how strongly a body responds to a given force (the larger m is the less it will be accelerated); m measures the inertia of the body.

Suppose we choose a test body of mass, say, 1gm. By measuring its motion one can obtain its acceleration and, using $F = ma$, determine the force. Once F is known the motion of *any* body is predicted: by measuring the falling an apple you can predict the motion of the Moon.

3rd Law

The third law is, at first sight, almost unbelievable: if I kick a ball, the ball kicks me back? But in fact it *is* so: suppose I push a friend while we are both standing on ice (to minimize friction), then he/she will move in the direction of the push, but I will move backward! What happens when I kick a ball is that the push backward is countered by the friction between my other leg and the ground, and because of this no motion backward ensues.

It is interesting to do the kick-the-ball experiment on ice, you should try it.

4.3.2 Optics

Newton's first work as Lucasian Professor was on optics. Every scientist since Aristotle had believed light to be a simple entity, but Newton, through his experience when building telescopes, believed otherwise: it is often found that the observed images have colored rings around them (in fact, he devised the reflecting telescope, Fig. 4.10, to minimize this effect). His crucial experiment showing that white light is composite consisted in taking beam of white light and passing it through a prism; the result is a wide beam displaying a spectrum of colors. If this wide beam is made to pass through a second prism, the output is again a narrow beam of white light. If, however, only one color is allowed to pass (using a screen), the beam after the second prism has this one color again. Newton concluded that white light is really a mixture of many different types of colored rays, and that these colored rays are not composed of more basic entities (see Fig. 4.11).

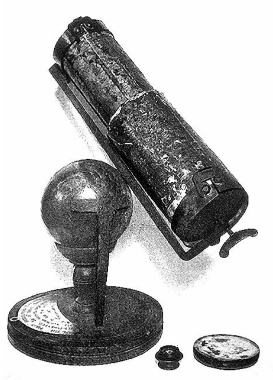


Figure 4.10: Newton's first reflective telescope.

Concerning the nature of light. Newton believed that it consists of a stream of small particles (or corpuscules) rather than waves. Perhaps because of Newton's already high reputation this "corpuscular" theory was accepted until the wave theory of light was revived in the 19th C.

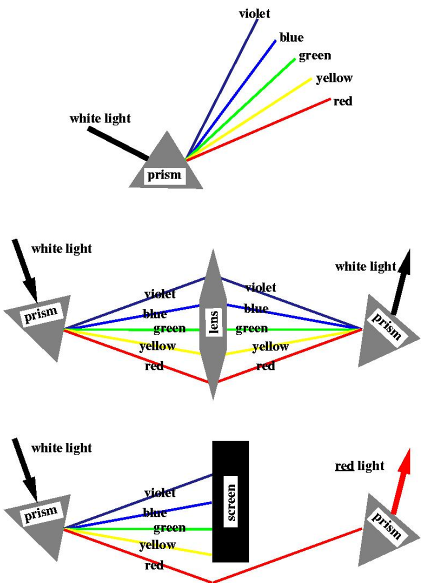


Figure 4.11: Diagram of Newton's experiments on the composition of white light.

4.3.3 Gravitation.

One of Newton's greatest achievements was on the field of celestial mechanics where he produced the first synthesis in the theories describing Nature: he realized that the same force that makes things fall, gravity, is responsible for the motion of the Moon around the Earth and the planets around the Sun.

He reasoned (more or less) as follows. Suppose I let an apple fall form a very high tower, it will take, say, t seconds to reach the ground. Now suppose I throw it very hard, then again it will take t seconds to reach the ground *provided* I assume the Earth is flat. But the Earth *isn't* flat and has curved from beneath the apple! Hence the apple will take longer to hit the ground. By throwing the apple with increasing force one reaches a point where the apple never hits the ground as the distance it falls equals the distance the earth has curved under it: the apple is in orbit! (see Fig. 4.12)

With this thought experiment Newton convincingly argued that an apple can behave in the same way as the Moon, and, because of this it is the very same force, gravity, which makes the apple fall and the Moon orbit the Earth. This is consistent with the hypothesis that gravitation is universal. In a way it represents the unification a several physical effects which appear

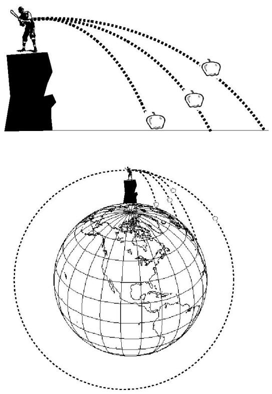


Figure 4.12: Newton's explanation of the equivalence between the force making apples fall and the one responsible for the Moon orbiting the Earth.

unrelated at first sight: the falling of apples and the orbiting of planets.

Having realized this he then used the results of Kepler and showed that if the planets and the sun are assumed to be point-like, the gravitational force drops as the inverse distance squared: the gravitational force between two bodies of masses m and M separated by a distance r is attractive and directed along the line joining the bodies, its value is

$$F_{\text{grav}} = \frac{mMG}{r^2}$$

where G is a universal constant, in words,

> *all matter attracts all other matter with a force proportional to the product of their masses and inversely proportional to the square of the distance between them.*

All matter attracts all other matter with a force proportional to the product of their masses and inversely proportional to the square of the distance between them

Having discovered this Newton was able to explain a wide range of previously unrelated phenomena: the eccentric orbits of comets, the tides and their variations, the precession of the Earth's axis, and motion of the Moon as perturbed by the gravity of the Sun. It also predicts the position of the planets for thousands of years so that the occurrence of eclipses can be foretold with exquisite accuracy, Moon landings can be planned without uncertainties, etc.

Consider now the application of the second law to the case of the gravitational force.

$$\frac{mMG}{r^2} = F_{\text{grav}} = ma$$

so that the factors of m cancel (!) This implies that the motion of a body generated by the gravitational force is *independent* of the mass of the body (!!), (just as Galileo had observed). This unique feature results from $F_{\rm grav}$ being precisely proportional to m. So m is seen made to play two roles:

- On the one hand m in $F = ma$ is a measure of how strongly is a body accelerated by a given force: it is a measure of the body's inertia. In this role m is called the *inertial mass.*
- On the other hand m in $F_{\rm grav}$ is a measure of how strongly is a body affected by the force of gravity and also how strong a gravitational force is generated by m; in this role it is called the *gravitational mass.*

These two quantities refer to different properties of a body and need not be equal. Extremely precise measurements, however, indicate that they *are* equal (at least to one part in ten parts per trillion). Newton just stated that this was the way of the world and kept going. Einstein, in contrast, noted this as a very important fact of nature, which he used to give birth to his General Theory of Relativity (Chap. **??**).

Concerning the nature of gravitation. there is another interesting feature of $F_{\rm grav}$: it is time independent. this implies that if a body moves, this change is perceived instantaneously by all the bodies throughout the universe. Leibnitz (among others) criticized Newton's hypothesis along these lines, and was disregarded. But this only due to the enormous success of Newtoninan gravity in making predictions of the motions of the bodies in the solar system. In fact we will see that this is not correct, and that the effect spreads out from the body at a finite speed (Chap. **??**).

To give an idea of the trust and excellent successes of Newtonian gravity consider the story of the discovery of Neptune. In 1843 a young astronomer at Cambridge, J.C. Adams discovered an anomaly in the orbit of Uranus and by the end of 1845 had concluded that this was due, not to a failure of Newton's law of gravity, but to the presence of a new planet. Adams submitted his results to G. Airy, his boss, who was unconvinced and dropped the matter. Meanwhile U. Leverrier in France had done a similar set of calculations independently, he published in 1846. This spurred Airy into action, but the Cambridge Observatory lacked an up to date chart of the region of the sky were the new planet was supposed to have resided at the time. During that time Leverrier wrote to J.G. Galle at the Berlin Observatory who promptly located the new planet. After much discussion this planet was called Neptune.

Chapter 5

The Clouds Gather

For more than two centuries after its inception the Newtonian view of the world ruled supreme, to the point that scientists developed an almost blind faith in this theory. And for good reason: there were very few problems which could not be accounted for using this approach. Nonetheless, by the end of the 19th century new experimental evidence difficult to explain using the Newtonian theory began to accumulate, and the novel theories required to explain this data would soon replace Newtonian physics. In 1884 Lord Kelvin in his Baltimore lectures already mentions the presence of "Nineteenth Century Clouds" over the physics of the time, referring to certain problems that had resisted explanation using the Newtonian approach. Among the problems of the time (not all were mentioned by Kelvin) were

- Light had been recognized as a wave, but the properties (and the very existence!) of the medium that conveys light appeared inconsistent.

- The equations describing electricity and magnetism were inconsistent with Newton's description of space and time (Sect **??**).

- The orbit of Mercury, which could be predicted very accurately using Newton's equations, presented a small but disturbing unexplained discrepancy between the observations and the calculations.

- Materials at very low temperatures do not behave according to the predictions of Newtonian physics.

- Newtonian physics predicts that an oven at a stable constant temperature has infinite energy.

The first quarter of the 20-th century witnessed the creation of the revolutionary theories which explained these phenomena. They also completely changed the way we understand Nature. The first two problems require the introduction of the Special Theory of Relativity. The third item requires the introduction of the General Theory of Relativity. The last two items can be understood only through the introduction of a completely new mechanics: quantum mechanics.

As a result of these developments the formalism developed by Newton lost its fundamental character. It is of course still a perfectly good theory *but* with a very well defined range of applicability. As mentioned previously, this does not imply that Newton was "wrong", it merely implies that his theories, although accurately describing Nature in an impressive range of phenomena, do not describe *all* of it. The new theories that superseded Newton's have the virtue of explaining everything Newtonian mechanics did (with even greater accuracy) while extending our understanding to an even wider range of phenomena. In this chapter I will describe the growth of the theory of electricity and magnetism which was to be fundamental to the development of Special Relativity.

The replacement of Newtonian mechanics was driven by the *data* that required the replacement of Newtonian physics by these more fundamental ones; the theories of relativity and quantum mechanics together explain all the phenomena probed to date, but they might be replaced in the future by others providing a yet deeper understanding of nature. These new theories will have to explain everything relativity and quantum mechanics do *and* provide experimentally verifiable predictions which are subsequently confirmed.

5.1 Electricity and magnetism

5.1.1 Electricity

It was known to the ancient Greeks as long ago as 600 B.C. that amber, rubbed with wool, acquired the property of attracting light objects. In describing this property today, we say that the amber is electrified, (from the Greek, *elektron*: amber), possesses an electric charge, or is electrically *charged*. It is possible to put an electric charge on any solid material by rubbing it with any other material (rubbing brings many points of the surfaces into good contact, so that, at the atomic level, electrons are ripped from one material and transferred to the other). Thus, an automobile becomes charged when it moves through the air, a comb is electrified in passing through dry hair, etc.

By the end of the 18th century it was known that electricity comes in two flavors: positive and negative; and that equal charges repel while unequal charges attract. The manner in which this attraction and repulsion occurs was discovered by Coulomb in 1785. He found that the force between them is very similar in form to the gravitational force: it is proportional to the charges of each body, directed along the line joining them, and decreases like the distance squared. There is, however, an important difference: this electric force can be attractive or repulsive; the gravitational force is *always* attractive.

electricity comes in two flavors: positive and negative

The electric force is proportional to the charges of each body, directed along the line joining them, and decreases like the distance squared

Charles Augustin de Coulomb (June 14, 1736-Aug 23 1806)). Born in Angouleme, France; died in Paris, France. Coulomb spent 9 years as a military engineer in the West Indies but his health suffered so, when the French Revolution began, he retired to the country to do scientific research. He worked on applied mechanics, but he is best known for his work on electricity and magnetism. He established experimentally the inverse square law for the force between two charges which became the basis of Poisson's mathematical theory of magnetism. Coulomb also wrote on structural analysis, the fracture of beams, the fracture of columns, the thrust of arches and the thrust of the soil.

5.1.2 Magnetism

The earliest observations on magnets can also be traced back to the early Greeks (eg. Thales of Miletus; see Sect. **??**). The Chinese literature also has extensive references to naturally occurring magnets (then called loadstones). The fact that magnets align in a unique way, together with the fact that the Earth itself is a magnet, lead to the discovery of the compass. This was of paramount importance to the development of civilization. The earliest known compass appeared in China by the first century A.D.; it arrived in Europe by the twelfth century A.D.

William Gilbert (1544-1603). Born in Colchester, England, into a middle class family of some wealth. Entered St. John's College, Cambridge in 1558, and obtained his B.A. (1561), M.A. (1564) and M.D. (1569). Became a senior fellow of the college, holding several offices and set up a medical practice in London becoming a member of the Royal College of Physicians. He never married.
He published *De Magnete* (On the Magnet) in 1600 which became the standard work throughout Europe on electrical and magnetic phenomena. It is a comprehensive review of what was known about the nature of magnetism, and Gilbert added much knowledge through his own experiments. He built a philosophy where magnetism was the soul of the Earth; he believed that a perfectly spherical lodestone, when aligned with the Earth's poles, would spin on its axis, just as the Earth spins on its axis in 24 hours.

According to thirteenth-century philosophy, the compass needle points towards the North star which, unlike all other stars, in the night sky, appears to be fixed. Thus, philosophers reasoned that the lodestone obtained its "virtue" from this star. Better observations, however, showed that the needle does not point exactly to the North Star and eventually it was shown that it is the Earth that affects the compass. Apart from the roundness of the Earth, magnetism was the first property to be attributed to the body of the Earth as a whole:

> Magnus magnes ipse est globus terrestris [the whole Earth is a magnet]. *William Gilbert*

By the early 17th century the properties of magnets were well known and many folk tales (such as the anti-magnetic properties of garlic) had been debunked. Magnetism was believed to be an effect different from electricity, their intimate relationship had not been discovered.

Careful experimentation with magnets came to a head in the late 19th century. By then reliable batteries had been developed and the electric current was recognized as a stream of charged particles. In 1870 Ørsted noted that a compass needle placed near a wire was deflected when a current was turned on, that such a deflection also occurs when the wire is moved, and he concluded that moving charges generate magnetic effects. These results were furthered by Ampère and who rendered them into a precise mathematical formulation.

Moving charges generate magnetic effects

Hans Christian Ørsted (Aug. 14, 1777 – March 9, 1851). In 1806 Ørsted became a professor at the University of Copenhagen, where his first physical researches dealt with electric currents and acoustics. During an evening lecture in April 1820, Ørsted discovered that a magnetic needle aligns itself perpendicularly to a current-carrying wire, definite experimental evidence of the relationship between electricity and magnetism (this phenomenon had been first discovered by the Italian jurist Gian Domenico Romagnosi in 1802, but his announcement was ignored).
Ørsted's discovery, in 1820, of piperine, one of the pungent components of pepper, was an important contribution to chemistry, as was his preparation of metallic aluminum in 1825. In 1824 he founded a society devoted to the spread of scientific knowledge among the general public. Since 1908 this society has awarded an Ørsted Medal for outstanding contributions by Danish physical scientists. In 1932 the name oersted was adopted for the physical unit of magnetic field strength.

André Marie Ampère (Jan. 20 1775-June 10 1836). Born in Lyon, France, died in Marseilles, France. André Ampère was a Professor at the École Polytechnique from 1814 to 1828 and then at Université de France from 1826 until his death. He worked on electromagnetism and analysis. He also made contributions to line geometry extending ideas of Binet. Ampère attempted to give a combined theory of electricity and magnetism in the early 1820's. He formulated a circuit force law and treated magnetism by postulating small closed circuits inside the magnetized substance. This approach became fundamental for the 19th Century. Ampère's most important publication is Memoir on the Mathematical Theory of Electrodynamic Phenomena, Uniquely Deduced from Experience (1827).

During the same period Faraday made various experiments with moving magnets (as opposed to moving wires). He found that a magnet moving in a coil of wire generates a current: moving magnets generate currents. This

Moving magnets generate currents

result provides the principle behind electric generators, be it small household ones, or the giant ones found in Hoover Dam. The fact that charges in motion create magnets and that moving magnets generate currents demontrates the intimate connection between electric and magnetic phenomena.

Michael Faraday (Sept. 22, 1791 – August 25, 1867). Michael Faraday became one of the greatest scientists of the 19th century. He began his career as a chemist; wrote an important manual of practical chemistry, and discovered a number of new organic compounds, among them benzene. He was the first to liquefy a "permanent" gas (*i.e.*, one that was believed to be incapable of liquefaction).

His major contributions were in the field of electricity and magnetism. He was the first to produce an electric current from a magnetic field, invented the first electric motor and dynamo. He provided the experimental, and a good deal of the theoretical, foundation upon which Maxwell erected classical electromagnetic field theory.

Faraday created the concept of a field. He imagined that any magnet or charged object generates an influence that permeates space, such emanation is called a field. If another magnet or charged object draws near, it is the interaction between this field and the new charged object or magnet which the latter feels as a force. He also showed that charge is never destroyed not created.

James Clerk Maxwell (June 13 1831-Nov 5 1879). Born in Edinburgh, Scotland, died in Cambridge, Cambridgeshire, England. Maxwell attended Edinburgh Academy where he had the nickname 'Dafty'. While still at school he had two papers published by the Royal Society of Edinburgh. Maxwell then went to Peterhouse Cambridge but moved to Trinity where it was easier to obtain a fellowship. Maxwell graduated with a degree in mathematics from Trinity College in 1854. He held chairs at Marischal College in Aberdeen (1856) and married the daughter of the Principal. However in 1860 Marischal College and King's College combined and Maxwell, as the junior of the department had to seek another post. After failing to gain an appointment to a vacant chair at Edinburgh he was appointed to King's College in London (1860). He made periodic trips to Cambridge and, rather reluctantly, accepted an offer from Cambridge to be the first Cavendish Professor of Physics in 1871. He designed the Cavendish laboratory and helped set it up.

Maxwell's first major contribution to science was to show that Saturn's rings must consist of many solid particles (confirmed by the Voyager spacecraft), this result won him the Adams Prize at Cambridge. Maxwell next considered the theory of gases and showed that temperature and heat are related to the motion of gas molecules.

Maxwell's most important achievement was his extension and mathematical formulation of Faraday's theories of electricity and magnetism. His paper on Faraday's theory was read to the Cambridge Philosophical Society in two parts, 1855 and 1856. Maxwell showed that a few relatively simple mathematical equations could express the behavior of electric and magnetic fields and their interrelation. The four equations (now known as Maxwell's equations), first appeared in fully developed form in his book *Electricity and Magnetism* (1873). They are one of the great achievements of 19th-century mathematics. Maxwell showed that an electromagnetic disturbance travels at a speed of light (1862) and concluded that light is an electromagnetic phenomenon.

Charge is conserved

Faraday also showed that charge is conserved. That is, the amount of positive charge minus that of negative charge is always the same.

The results of all these investigations can be summarized in a series of four equations. These were studied extensively by Maxwell who noted that they are inconsistent with charge conservation, but Maxwell himself realized that a slight modification in one equation would get rid of this problem. The modification proposed by Maxwell is simple, but the results are so momentous that the modified set of four equations are known as

Maxwell's equations. Why are Maxwell's equations so important? There are four reasons:

- They describe all electromagnetic phenomena with perfect accuracy for distances larger than about 10^{-8}cm.
- They are inconsistent with Newtonian mechanics, and so present the first solid evidence for the modification of Newton's theory.
- There are solutions of the equations which describe waves traveling at speed $c = 299,792$km/s (which is also the speed of light).

The last point leads to the inescapable conclusion is that light is precisely the object that was described by the wave-like solution of Maxwell's equations (without his modification there are no wave-like solutions); in Maxwell's own words

> We can scarcely avoid the conclusion that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.

It is in this way that the next unification in physics occurred: light, electricity and magnetism are different aspects of the same set of phenomena and are described by a single theory. Because of this we now speak of *electromagnetism* and not of electric and magnetic phenomena separately.

Light, electricity and magnetism are different aspects of the same set of phenomena and are described by a single theory

5.2 Waves vs. particles

I mentioned above the word "wave" in several occasions. Since waves will appear repeatedly in the following I will take a short detour to explain what waves are and what are their properties. The *American Heritage Dictionary* defines wave as

> A disturbance or oscillation propagated form point to point in a medium or in space

Thus when a stone is dropped on a calm pond we see a series of circular waves emanating form the spot where the stone hit the water, spreading out at a certain speed. If a bigger stone is used the water the waves become more pronounced, the distance form crest to trough becomes larger. If instead of dropping a stone we attach it to a rod and move it up and down we find that the faster we move it the closer together the crests and troughs of the waves; so that if we look at one point on the pond's surface we will see the water swelling and ebbing faster.

These characteristics of the waves have definite names; see Fig. 5.1,

- The *frequency* is the number of wave-crests that go through a point on the pond every second.
- The *wavelength* is the distance between two crests.
- The *amplitude* is the distance between crest and trough.

The *frequency* is the number of wave-crests that go through a point on the pond every second

The *wavelength* is the distance between two crests

The *amplitude* is the distance between crest and trough

These properties, together with the speed at which the wave spreads characterize the waves.

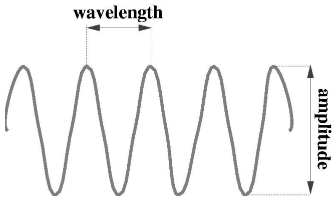


Figure 5.1: Definition of the wavelength and amplitude of a wave.

Imagine a cork floating on the pond. As the wave goes by the place where the cork is floating it will boob up and down. Suppose that you measure the time it takes for it to go down from its highest point, down to its lowest and then back to its highest point again, then the frequency is the *inverse* of this time. So if the cork takes 0.5 seconds to go up and down and back up, the frequency would be $\frac{1}{0.5\,sec}$ or 2 inverse-seconds. This is just a way of counting the number of oscillations per second: if each oscillation takes half a second, there will be two oscillations per second, and so the frequency is two inverse-seconds; a frequency of 7 inverse seconds indicates that there are seven oscillations each second, etc. There are many kinds of waves: water waves on a pond, sound waves in air (or water or any other medium), electromagnetic waves, etc.

Imagine now a calm pond with a few leaves floating on the surface. At one time a child drops a stone which makes a series of expanding circular waves. As they spread the waves eventually come to the floating leaves which bob up and down. The notable thing about this detail is that the leaves do not change position, even though the wave spreads, it does not carry the leaves with it. The same thing can be said of the water itself, the waves spread though it but do not carry the water along with them. In fact, if you look closely at the particles suspended in water (ponds usually have many of those) as the waves pass, they make circular motions about their initial positions but are not carried along. These waves use water as their *propagation medium*, in the same way as sound waves use air (or water or other materials) to propagate in. Without a medium these waves simply do not propagate: there is no sound in the vacuum. A reasonable question in connection with these observations is whether *all* waves need a medium to propagate in, the answer is (perhaps surprisingly) **no!**, and the way this was discovered is the subject of many of the following sections

A particle is characterized by its mass and other measurable properties (for example, its charge). I will assume that this is intuitively clear. Ordinary everyday experience shows that waves behave very differently from particles [1]. For example, if you are taking cover behind a wall form a person shooting peas at you, you will not be hit; yet when she screams that you are a chicken, you hear her perfectly well. Sound waves (and all waves in general) have the ability to go around obstacles (up to a certain extent: if the wall is very tall and wide the insults will not reach you); particles have no such ability.

The above properties of sound waves are well known. But, if light is a wave, should it not behave in the same way? And if it does, how come we do not see a person standing behind a wall (whom we can clearly hear)? I will now consider this (apparent) paradox.

5.3 Light

It is now known that under all common circumstances light behaves as a wave propagating at a speed close to $300,000$km/s. This, however, is a recent realization; in fact, whether light traveled at finite or infinite speed was the subject of much debate was left unanswered for a long time. Galileo tried to measure the speed of light by experiment: he put two persons on hills (separated by a bit less than a mile), and then told one open a lantern, the other was to raise his/her hand when he/she sees the light and the first notes any lapse between his/her opening the lantern and seeing the raised hand. No time delay was observed (which is not unnatural, the lapse is about 10^{-5}s!). So the question remained unanswered [2].

[1] This is not true when phenomena at very short distances are examined, at distances below 10^{-8}cm (atomic size) the difference between waves and particles becomes blurred.

[2] One can, however, use this result to get a limit on the speed of light. If the human response time is, say, half a second, then this experiment shows that light travels faster

In 1670 the Danish mathematician Olaus Rømer observed that the eclipses of Jupiter's moons were 11 minutes ahead of schedule when the Earth was closer to Jupiter, and they lagged behind (also by 11 minutes) when the Earth was farthest from Jupiter. Assuming that there are no problems with the predictions of Newtonian physics concerning the motion of Jupiter's moons, he concluded that the discrepancy was due to the different times light takes to get to Earth at the two extremes of its orbit (Jupiter moves very little during one year, it takes 12 years for it to circle the sun), see Fig. 5.2. Rømer then calculated that the speed of light would be $210,000$km/s. The modern value of the speed of light is $299,792$km/s.

This is, of course, not the only possible explanation, Rømer could have argued, for example, that Newton's equations could not account for Jupiter's motion. Still the hypothesis that light travels at a finite speed furnished the simplest explanation and, following Ockham's razor (Sect. **??**) it is the one which ought to be examined first. Soon after Rømer's argument was made public the fact that light travels at finite speed was demonstrated in various experiments and was universally accepted.

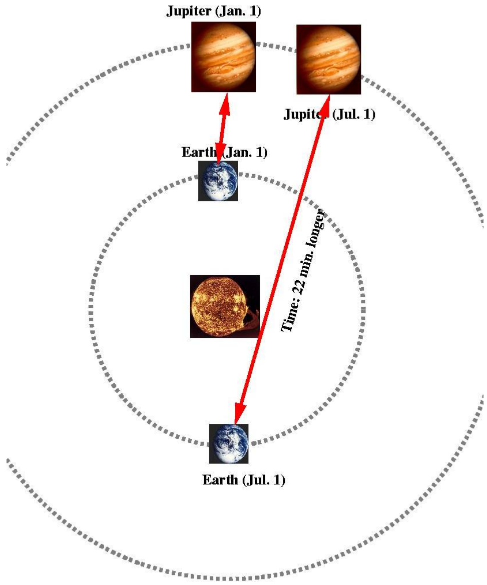


Figure 5.2: Diagram of the reasoning used by Rømer to determine the speed of light.

So light propagates at a finite speed. What is it made of? Newton

than 2miles per second.

believed that light was made of corpuscles, but even the weight of Newton's opinions could not withstand the experimental evidence showing that light behaves as a wave. This sounds preposterous: a wave, such as sound, will "go around corners" but light does nothing of the kind...or does it? In fact, it does! If you look very closely at a very sharp edged screen you will see that some light actually goes behind the screen: light does behave as a wave (see Fig. 5.3). This is not common knowledge because it is a small effect, light dies out almost as it turns the corner, if the corner is not very sharp, light is scattered in many ways and the effects disappears; in other words, for light, almost any obstruction is a very tall wall.

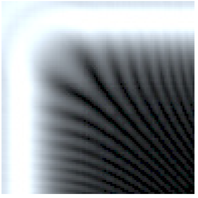


Figure 5.3: Picture of the shadow cast by the corner of a screen. Noote that the shadow region is not completely dark.

The wave theory of light leads to some surprising consequences. For example, it predicts that the shadow cast by a dark circular screen should have a bright spot in its center, and this would be absurd were it not for the fact that the bright spot is indeed there! (see Fig: 5.4)

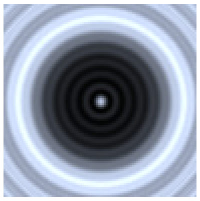


Figure 5.4: Shadow cast by a small opaque disk. Note the bright spot in the center of the shadow.

By the beginning of the 19th century the hypothesis that light is a wave traveling at large (by our standards) but finite speed [3] was proven and was universally accepted. Being a wave we can ask what is its wavelength, amplitude, frequency, etc; it turns out that visible light has very small wavelength,

[3] The speed depends on the medium in which light travels; the value given above corresponds to the speed in space.

about 10^{-5}cm. Another natural question is then, do electromagnetic waves with larger and smaller wavelengths exist?

The answer is yes. Visible light is but a member of a large family of waves; they are all electromagnetic waves, and they are all described by the Maxwell's equations. For historical reasons waves of different wavelengths have different names (see Fig. 5.5). Thus we have (the symbol $\sim$ means "about")

Wavelengths of electromagnetic waves

Name	*Wavelength*
Radio	$\sim$ 10cm or larger
Microwave	$\sim$ 1cm
Infrared	$\sim 10^{-3}$cm
Visible	$\sim 10^{-5}$cm
Ultraviolet	$\sim 10^{-6}$cm
X-rays	$\sim 10^{-8}$cm
Gamma-rays	$\sim 10^{-9}$cm or smaller

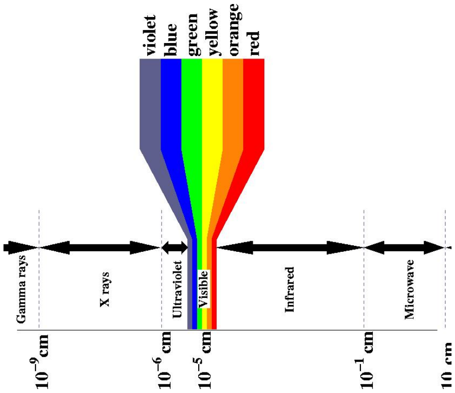


Figure 5.5: The electromagnetic spectrum.

All of these are common names. Every one of these waves travels at the same speed in vacuum [4] equal to the speed of light (called "visible" above) in vacuum; the only difference between them is the wavelength, the distance between two consecutive crests in the corresponding wave trains.

So light is a wave, similar then to sound waves, or water waves. But all these waves are produced by the undulations of some medium: water for

[4] In a medium there is some interaction between the atoms and the waves and the speed can be different.

water-waves, air (for example) for sound, etc. Thus it was *postulated* that the medium in which light undulates is called *ether.*

5.4 Problems

The end of the 19th century witnessed the growth of evidence against the classical physics based on Newton's theory. I will discuss two such problems, the first concerns the ether, which appeared to have inconsistent properties; the second refers to an apparent contradiction between Galilean relativity and the theory of electromagnetism. The resolution of these conflicts cannot be achieved within Newtonian physics: it requires the theory of relativity.

5.4.1 Ether

Having postulated the existence of the ether as the medium in which light travels it becomes interesting to determine the properties of this material. First and foremost, since the light from distant stars does reach us, we must assume that the ether permeates the whole universe up to its farthest reaches. We must then imagine that the Earth plunges through this ether as it circles the Sun. The ether must then be very tenuous, for otherwise the friction would have stopped the Earth long ago. Let us now derive some other predictions derived from the ether hypothesis.

As the Earth moves through this ether a kind of "ether wind" must be present on Earth's surface. To see why this should happen consider the following analogy. Imagine a windless day in which you take a ride in your red convertible which, unfortunately, has no windshield. As you speed up you will feel the air blowing, the faster you go, the stronger this wind is. In the same way, replacing *air* → *ether* and *red convertible* → *earth*, a very sensitive apparatus on the surface of the earth should detect and ether wind.

So, can the ether wind be detected? Apparently yes! The idea for the first experiments is based on the following argument. Imagine yourself back in your convertible (with no windshield) taking your nagging grandmother to the store; she sits in the back seat...it's safer. She talks all the time, but, fortunately, her words get blown back by the wind. In contrast she hears *everything* you say, for your words get blown back by the wind, right into her ears (good grief!). In the same way, as we stand on Earth, the ether wind should blow back the light coming from the stars. At different times of the year, the ether wind blows in different directions since the earth is moving in different directions, hence the observed positions of the stars should change (see Fig. 5.6)... and they do!

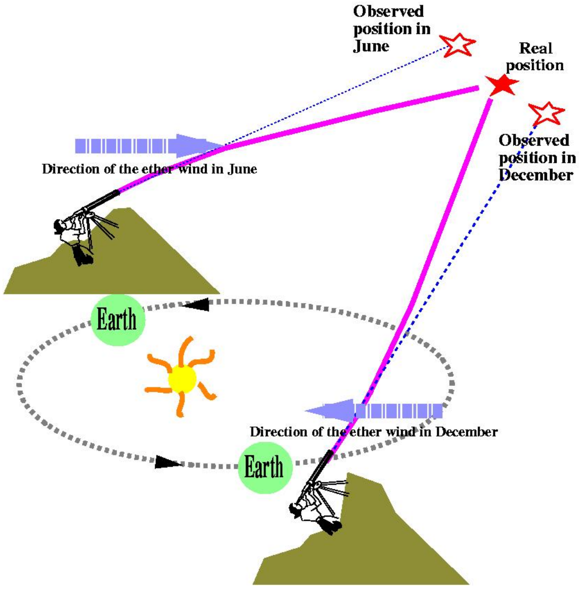

Observed position in June
Real position
Observed position in December
Direction of the ether wind in June
Earth
Direction of the ether wind in December
Earth

Figure 5.6: The shift in the observed position of the stars caused by the ether wind.

But, wouldn't the earth drag with it some of the ether in its vicinity? Well, since this peculiar behavior of the images of the stars were observed, the earth must not drag the ether with it: ether goes through the earth "much as the wind goes through a grove of trees" (as described by T. Young.)

This consequence of the ether wind is not the only prediction of the ether hypothesis; in order to derive other consequences we need to go back briefly to Newtonian mechanics. Suppose you are in a train moving at a speed of 1m/s with respect to a train station. Suppose now you kick a ball in the direction of the train's velocity and which, as a result of your action moves at 2m/s as measured in the train. Then an observer in the station will see the ball move at $1 + 2 = 3$m/s (see Fig. 5.7).

Thus the two parallel velocities (the train's and the ball's with respect to the train) add up. In contrast if the ball were thrown up both observers would measure the same (vertical) velocity. Consider now the same situation but with light replacing the ball. If the train moves at speed v then light traveling forward will move at speed $v + c$. If the light-beam is directed upward both observers would measure the same vertical speed c. These conclusions are inescapable from the Newtonian standpoint and, because they are wrong, constitute some of the most important nails in the coffin

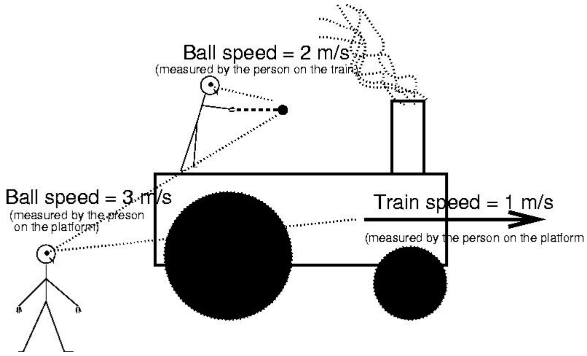

Ball speed = 2 m/s
(measured by the person on the train)
Ball speed = 3 m/s
(measured by the preson on the platform)
Train speed = 1 m/s
(measured by the person on the platform

Figure 5.7: Addition of velocities according to Newtonian mechanics

of Newtonian mechanics. Let me examine first the following consequence derived from it.

Suppose you consider light going in air and that the same beam is made to enter a piece of glass. In air light will have a speed $c_{\rm air}$, while in the glass it will have speed $c_{\rm glass}$; these two quantities being measured at rest with respect to the ether. The experiment I want to discuss measures the ratio of speeds in glass and air. Now, if there is an ether, and the earth is moving at a speed v with respect to it, then one can select the orientation for the apparatus such that the beam happens to lie along the velocity v [5]. In this case the speed of light in air and in glass will be altered, they become $c_{\rm air} + v$ and $c_{\rm glass} + v$ respectively; the experiment should give the result $(c_{\rm air} + v) / (c_{\rm glass} + v)$. If the beam is rotated 180^o then the direction of the ether wind is reversed and the experiment ought to produce the value $(c_{\rm air} - v) / (c_{\rm glass} - v)$. The amazing thing is that, as first shown by Arago, that this experiment gives the *same* value no matter how it is oriented with respect to the motion of the Earth through the ether. In order to explain this Fresnel suggested that transparent substances trapped some of the ether and dragged it along, and the amount and manner of trapping was "just-so" that the above experiment does not exhibit any effect. Of course the shift in the position of the stars would then imply that the air does not trap ether at all.

Curiouser and curiouser: the speed of light in glass depends on the color of light, nonetheless the above experiment gives no effect for any color. Therefore the ether trapped in glass should undulate with light precisely so

[5] In practice the experiment is set on a rotating table and is repeated for a variety of orientations.

as to compensate for this difference in speeds (note that the ether trapped with the glass travels with it).

So the ether is a medium which goes through all objects, but some of it is trapped by transparent substances and whose elasticity depends on the color of light going through it. In order to test this Fizeau performed a very important experiment. He sent light through tubes with water flowing in different directions. The water was supposed to drag at least some ether, which would then alter the speed of light. The results were positive and in accordance with Fresnel's hypothesis. So we have a big contradiction: the observation of starlight requires the Earth *and* the Earth's air not to drag any ether. But the Fizeau experiment requires transparent media to drag a significant (and measurable) amount of ether.

The most famous of the experiments made to detect the motion through the ether was the Michelson-Morley (or M&M)experiment. This is a very clear experiment. The idea is to send to take a light beam, to split it in two and send the daughter beams in perpendicular directions, these are then reflected back and recombined. The distances traveled by the daughter beams will be different and so there will be a mismatch between the two light wave trains resulting in a pattern of light and dark fringes after they are recombined (see Fig. 5.8)

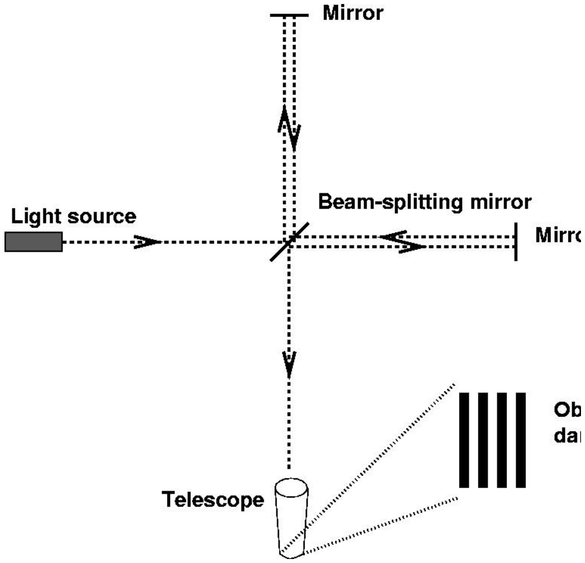


Figure 5.8: A diagram of the Michelson–Morely interferometer

Now suppose we rotate the table where the experiment is placed. The speeds of the two beams with respect to the ether will change, and so will the times taken for the beams to recombine. Because of this the mismatch between troughs and crests in the two wave trains also changes and a shift in the pattern of dark and bright lines should be seen...except that it wasn't! No detection of the motion through the ether could be measured.

It was then claimed that the only thing proved was that the ether in the basement where the experiment was done was dragged along with the air. But the experiment was repeated a large number of times, in particular it was done on a hilltop: no effects were ever obtained.

This last result was the death blow to the ether theory: M&M's experiment showed that the ether must be dragged along by the air, while stellar observations denied precisely that!

5.4.2 Galilean Relativity

Maxwell's equations do not specify in which frame the speed of light equals c

Galileo formulated his principle of relativity by stating that one cannot use any mechanical experiment to determine absolute constant uniform velocity. Now Maxwell's equations contain a velocity c but *they do not specify with respect to what is this velocity to be measured!*. We must conclude that either absolute velocities can be determined using experiments involving light, or else light must move at speed c in *all* reference frames.

But this is impossible to accept within Newtonian mechanics, for within this theory velocities simply add. If we then have a source of light moving at speed v, the light form it ought to travel at speed $c+v$ in direct contradition to Maxwell's equations which predice that light travels with speed c, no matter how fast the speed of the source.

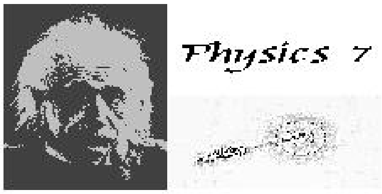
5.5 Prelude to relativity

The properties of the ether were inconsistent

Newtonian physics was inconsistent with the results of electromagnetism

So this was the situation before 1905: the ether was postulated, but its properties were inconsistent. Newton was believed to be right, but the corresponding mechanics was inconsistent with the results of electromagnetism. Was Newton's theory correct and all the experiments in electricity and magnetism wrong? If Newton was wrong, how can all the successes of his theory be understood? How can one understand light as a wave if the thing in which it travels cannot be described consistently?

All these problems were solved with the advent of the Special Theory of Relativity to which I now turn.



Chapter 6

The Special Theory of Relativity

6.1 Introduction

The puzzling properties of light and the ether remained through the turn of the century and up to 1904: the speed of light (as described by the equations of electromagnetism) did not depend on the motion of the observer and, stranger still, the medium in which light propagates could not be described consistently.

A final effort was made in order to understand in a "fundamental" way the negative result of the Michelson-Morley experiment. It was postulated (independently) by Fitz-Gerald and by Lorentz that matter moving through the ether is compressed, the degree of compression being just so that there is a negative result in the M&M experiment. The claim was that the ether wind does slow down and speed up light, but it also contracts all objects and these two effects conspire to give no effect in all experiments.

A calculation shows that an object of length ℓ moving with velocity v with respect to the ether should be contracted to length ℓ' given by

$$\ell' = \ell\,\sqrt{1 - \frac{v^2}{c^2}}$$

(where c is the speed of light) in order to get the null result required.

So in order to understand the gamut of experimental results the ether had to be a very tenuous medium that could not be felt or tasted, nonetheless the strongest materials would be squashed by it by an amount which makes it impossible to see the ether's effects. The amount a material would be squashed, though admittedly very small, would always be there and is independent of the composition of the object going through the ether (see Fig. 6.1). This is a situation like the one I used in the " little green men on the moon" example (see Sect. **??**): the ether has was awarded the property that no experiment could determine its presence; the ether hypothesis is not falsifiable.

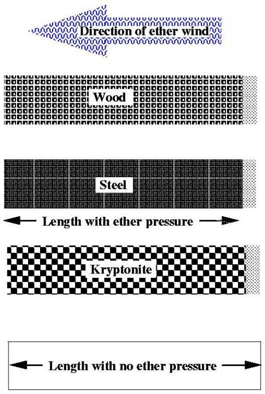


Figure 6.1: The idea behind the Lorentz–Fitz-Gerald contraction.

6.2 Enter Einstein

In 1905 Einstein published three papers. The first (dealing with the so-called "photoelectric effect") gave a very strong impulse to quantum theory, and got him the Nobel prize in 1921. The second dealt with the movement of small particles in a fluid (Brownian motion).

The third paper (Fig. 6.3) of 1905 was called *On the electrodynamics of moving bodies,* it changed the face of physics and the way we understand nature.

This paper starts with a very simple (and well known) example: if a magnet is moved inside a coil a current is generated, if the magnet is kept fixed and the coil is moved again the same current is produced (Fig. 6.4). This, together with the difficulties in detecting the motion with respect to the ether, led Einstein to postulate that

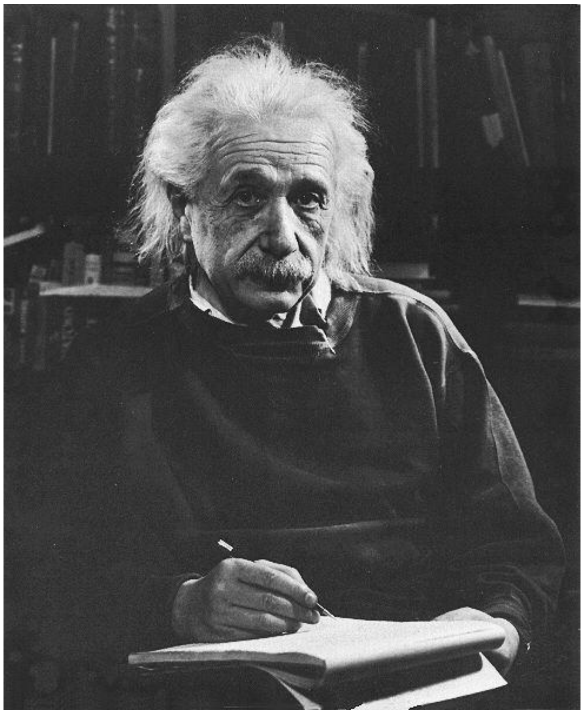


Figure 6.2: Albert Einstein (in his later years)

> *the same laws of electrodynamics and optics will be valid for all frames of reference for which the laws of mechanics hold good*

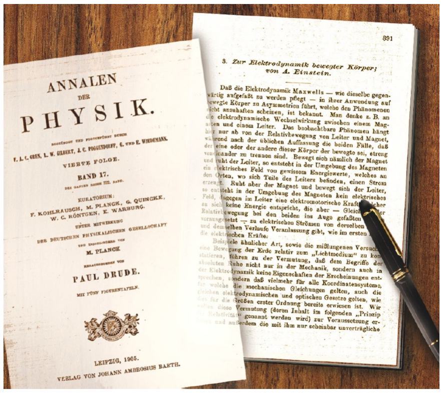
which is known as the *Principle of Relativity*.

In order to understand the implications of the Principle of Relativity we need (again) the concept of an inertial observer (see Sec. **??**). This is a person which, when observing an object on which no forces act, finds that it moves with constant speed in a straight line, or else is at rest. In terms of inertial observers we can restate the Principle of Relativity:

> *all the laws of physics are the same for all inertial observers.*

All the laws of physics are the same for all inertial observers

Galileo made a very similar statement but he referred only to the laws of mechanics, Einstein's achievement was not only to provide a generalization, but to derive a host of strange, surprising, unexpected and wonderful consequences from it.



ANNALEN
DER
PHYSIK.

BEGRÜNDET UND FORTGEFÜHRT DURCH
F. A. C. GREN, L. W. GILBERT, J. C. POGGENDORFF, G. UND E. WIEDEMANN.

VIERTE FOLGE.

BAND 17.
DER GANZEN REIHE 322. BAND.

KURATORIUM:
F. KOHLRAUSCH, M. PLANCK, G. QUINCKE,
W. C. RÖNTGEN, E. WARBURG.

UNTER MITWIRKUNG
DER DEUTSCHEN PHYSIKALISCHEN GESELLSCHAFT
UND INSBESONDERE VON
M. PLANCK

HERAUSGEGEBEN VON
PAUL DRUDE.

MIT FÜNF FIGURENTAFELN.

LEIPZIG, 1905.
VERLAG VON JOHANN AMBROSIUS BARTH.

891

3. *Zur Elektrodynamik bewegter Körper;*
von A. Einstein.

Daß die Elektrodynamik Maxwells — wie dieselbe gegenwärtig aufgefaßt zu werden pflegt — in ihrer Anwendung auf bewegte Körper zu Asymmetrien führt, welche den Phänomenen nicht anzuhaften scheinen, ist bekannt. Man denke z. B. an die elektrodynamische Wechselwirkung zwischen einem Magneten und einem Leiter. Das beobachtbare Phänomen hängt hier nur ab von der Relativbewegung von Leiter und Magnet, während nach der üblichen Auffassung die beiden Fälle, daß der eine oder der andere dieser Körper der bewegte sei, streng voneinander zu trennen sind. Bewegt sich nämlich der Magnet und ruht der Leiter, so entsteht in der Umgebung des Magneten ein elektrisches Feld von gewissem Energiewerte, welches an den Orten, wo sich Teile des Leiters befinden, einen Strom erzeugt. Ruht aber der Magnet und bewegt sich der Leiter, so entsteht in der Umgebung des Magneten kein elektrisches Feld, dagegen im Leiter eine elektromotorische Kraft, welcher an sich keine Energie entspricht, die aber — Gleichheit der Relativbewegung bei den beiden ins Auge gefaßten Fällen vorausgesetzt — zu elektrischen Strömen von derselben Größe und demselben Verlaufe Veranlassung gibt, wie im ersten Falle die elektrischen Kräfte.

Beispiele ähnlicher Art, sowie die mißlungenen Versuche, eine Bewegung der Erde relativ zum „Lichtmedium“ zu konstatieren, führen zu der Vermutung, daß dem Begriffe der absoluten Ruhe nicht nur in der Mechanik, sondern auch in der Elektrodynamik keine Eigenschaften der Erscheinungen entsprechen, sondern daß vielmehr für alle Koordinatensysteme, für welche die mechanischen Gleichungen gelten, auch die gleichen elektrodynamischen und optischen Gesetze gelten, wie dies für die Größen erster Ordnung bereits erwiesen ist. Wir wollen diese Vermutung (deren Inhalt im folgenden „Prinzip der Relativität“ genannt werden wird) zur Voraussetzung erheben und außerdem die mit ihm nur scheinbar unverträgliche

Figure 6.3: The 1905 paper on Special Relativity

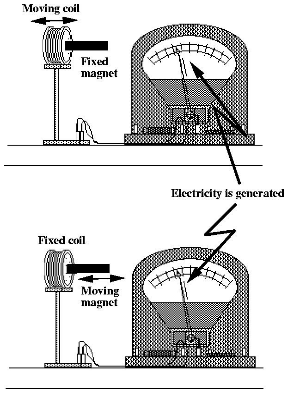


Figure 6.4: Illustration of one of the experimental facts that lead Einstein to the Principle of Relativity.

6.2.1 The first prediction: the speed of light and the demise of Newton's mechanics

Now that we have stated the Principle of Relativity we can examine its implications, and almost immediately we find reason to worry.

Maxwell's equations (the equations of electromagnetism, see page **??**) contain a quantity we called c, the speed of light, which is given without reference to any inertial observer. So, if we accept the Principle of Relativity *and* trust Maxwell's equations, we must conclude that c is the same for all inertial observers. So if Jack measures the speed of a beam of light while sitting at the top of the hill, and Jill also measures the speed of the same beam of light while running up the hill, they should get exactly the same

The speed of light c is the same for all inertial observers answer, no matter how fast Jill runs. It is often said that Einstein "proved that everything is relative" but, in fact, his first conclusion was that the speed of light is *absolute*.

This property of light is very different from, say, the properties of peas as described by the mechanics of Newton: if a person rides on a scooter and shoots peas, these move faster than the peas shot by a person standing by (see Sect. **??**). In contrast if the person on the scooter turns on a laser and the person standing by does the same when they coincide on the street, these two laser beams will reach Pluto at the same time (Fig. 6.5); this happens even if the scooter moves at 99% of the speed of light.

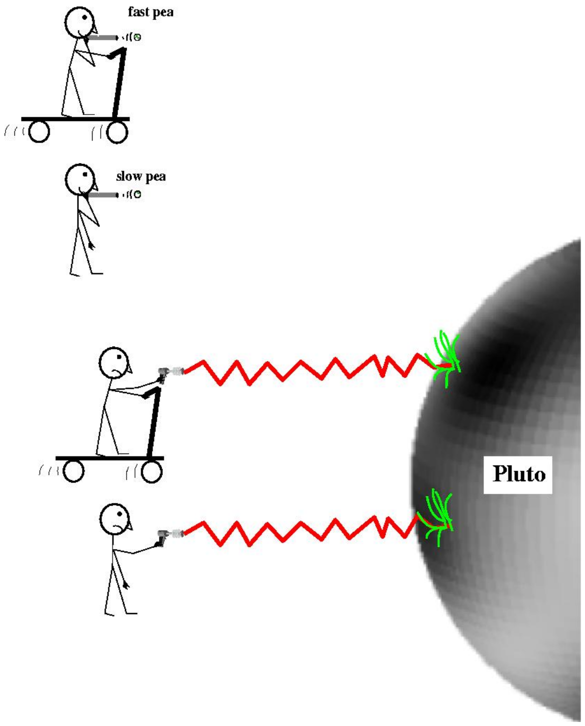


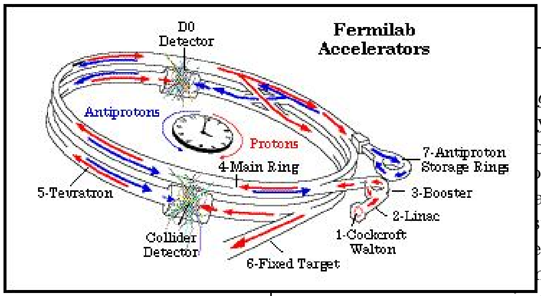
Figure 6.5: The pea shot from the scooter moves faster, yet both laser beams get to Pluto (it is really a photograph of Pluto) at the same time.

Newton would be horrified by this behavior of light beams: according to his mechanics velocities add, so that the laser beam from the scooter should reach Pluto sooner.

Thus, once Einstein adopted his Principle of Relativity, he was faced with a choice: either dismiss Newtonian mechanics or dismiss Maxwell's equations. It was impossible for them both to be right. Newton's mechanics had survived for about 250 years, it was universally accepted in the physics community, and its predictions agreed with all experiments (done up to 1905). Maxwell's equations, in contrast, were rather new, were not tested as thoroughly as Newton's, and were not universally accepted. Nonetheless Einstein took the daring path of siding with Maxwell and so challenged the whole edifice of the Newtonian theory. He was right.

Having chosen sides, Einstein assumed that Newton's mechanics were not a good description of Nature under *all* circumstances: it must then be only a good approximation. Einstein's work was then cut out for him: he needed to find a generalization of Newton's mechanics which is consistent with the Principle of Relativity, *and* which agrees with experiment as well as (or better than) Newton's theory. He was successful.

Significant discrepancies between Newton's and Einstein's mechanics become noticeable only at speeds close to c which explains why no problems were detected with Newton's theory before 1905: all experiments were done at speed very small compared to c. In this century a wealth of experimental evidence has been gathered which supports Einstein's mechanics in favor of Newton's. The best examples appear in experiments done since the 1950's using subatomic particles which are relatively easily accelerated to speeds approaching c. The behavior of such experiments completely vindicates Einstein's approach while being inexplicable from the Newtonian viewpoint.



gh energy accelerators. Most of the studies in subatomic ysics are done in enormous machines commonly called olliders" where electrically charged particles such as elec-ons and protons are accelerated to speeds very close to at of light and then forced to crash into each other. The sulting debris provides important clues as to the funda-ental structure of matter. A popular design for a collider nsists of one or more concentric rings in which the colliding particles are piped and accelerated using electric and magnetic fields. Given the enormous speeds of the particles the design must be extremely accurate, even a very small error can send all the particles crashing into the walls of the ring. All calculations are done using Einstein's mechanics, and the behavior of the particles perfectly matches the predictions of the theory; a design of a collider using Newtonian mechanics would lead to a useless machine.

Concerning the addition of velocities, Newton's formula is, strictly speaking, not correct even for slow moving obejcts. The corrections are, however, very samll when the speeds are small compared to that of light. For example for the case of the passenger in a train in Fig. **??** if the speed of the ball is u and that of the train is v the speed measured from the platform is not $u+v$ as Newton would claim, but

$$(u+v) \times c^2/(c^2+uv)$$

that is, there is a small correction factor $c^2/(c^2+uv)$ which, for ordinary velocities is very small indeed, for example for the example $u = 1$m/s, $v = 2$m/s, this factor is 0.9999999998 (Newton would have predicted 1 instead). On the other hand, if both u and v are half the speed of light, the speed seen from the platform would be 80% of the speed of light (and not c as Newton would have expected). For the extreme case where either u or v (or both) are equal to c, the speed seen from the platform would again be c.

In conclusion: the Principle of Relativity together with Maxwell's equations imply that there is a universal speed whose value is the same to all inertial observers. This fact required several fundamental changes in the manner we understand the world.

6.2.2 The second prediction: Simultaneity is relative

One concept which is radically modified by the Principle of Relativity is that of simultaneity. Every-day experience indicates that the statement "two events happened at the same time" (*i.e.* they are simultaneous) is universal, and would be verified by any one looking into the matter. Thus I can say, "I got home at the same time you got to work" and nobody (usually) wonders about the consistency of such statement.

The surprising result is that two FBI agents looking into the matter but moving with respect to each other (and having very accurate clocks) would get conflicting answers. In order to illustrate this result we will consider two murder mysteries, one set in Victorian England which is analyzed using Newton's ideas, the other is set in outer space and is studied following Einstein's guidance.

The first murder mystery (*ca.* 1890)

Sherlock Holmes is called to investigate a murder: a man was found shot in a train car, with two bullets in his head. After much investigation Sherlock finds a hobo who was at a station as the train wheezed by. This man saw two men come in from opposite sides of a wagon and simultaneously fire their revolvers at a chap sitting right in the middle of the train car. Being a Newton acolyte, Holmes is a firm believer that simultaneity is a universal concept, and concludes that both men fired at the same time as an absolute fact. Inspector Lestrade (from Scotland Yard) manages to find both men, who are found guilty of the crime and die in the gallows.

The second murder mystery (*ca.* 2330)

A murdered man is found in the cargo bay of the starship Enterprise with two head wounds caused by laser beams. The tragedy was observed from three places: a space station, the cargo bay itself, and a Klingon ship (a "bird of prey"). At the time of the crime the Enterprise was moving at a speed $c/2$ with respect to the space station; the bird of prey was moving in the same direction as the Enterprise at a speed $3c/4$ with respect to the space station (and was ahead of the Enterprise). To simplify the language we will say that both ships as seen from the space station were moving to the right (see Fig. 6.6).

Everyone agrees that the dead man was hit on the head by two laser beams simultaneously. These beams were fired by a klingon at the back of the cargo bay, and by a human at the front. They shot while they stood at the same distance from the victim. Both life-forms are arrested and put to trial.

Captain Kirk, then at the space station, acts as the human's lawyer. Kirk points out that the klingon must have fired first. Indeed, at the time of the murder the klingon was placed in such a way that the Enterprise carried the victim away from his laser bolt; in contrast, the ship carried the victim towards the human's laser bolt (Fig. 6.7). Since both bolts hit at the same time, *and they travel at the same speed c for* <u>*all*</u> *observers*, the klingon must have fired first. "The klingon's guilt is the greater one!" Kirk shouted dramatically, and sat down.

The captain of the bird-of-prey, who is (of course) acting as the klingon's lawyer, disagrees. His ship was moving to the right of the space station, but much faster than the Enterprise, hence, with respect to this ship, the Enterprise was moving to the *left*. "I can then use my esteemed colleague's arguments and categorically state that it was the *human* that fired first (see Fig. 6.8), it is *her* guilt that is the greatest."

Dr. McCoy happened to be in the cargo bay at the time of the shooting and testifies that he saw both the human and the klingon fire at the same time: since the beams hit the victim at the same time, and they were at the same distance, they must have fired at the same time (Fig. 6.9).

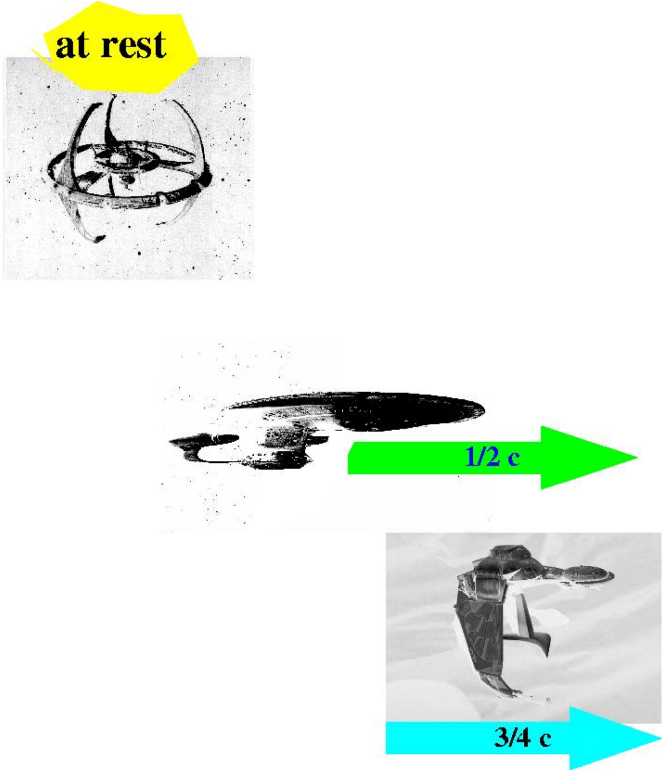


Figure 6.6: The setup for the second murder mystery. The velocities are measured with respect to the space station (labeled “at rest”).

Now, the law (in this story) states that the guilty party is the one who fired first, but deciding who did fire first is impossible! This is so because events occurring at different places will not be simultaneous to all observers. The fact that c is the same for all observers implies that if two events separated by some distance (such as the firing of the lasers) are simultaneous to one observer (such as McCoy) they will *not be simultaneous* to observers moving relative to the first (such as Kirk and the Klingon captain). Even the ordering in time of these events is relative

Events occurring at different places will not be simultaneous to all observers

> *Simultaneity is relative for events separated by a non-zero distance.* [1]

Let me use a short-hand and let **K** be the statement “the klingon shoots”, while **H** denotes “the human shoots”. Then

[1] This was explained by Spock to Kirk...at great length.

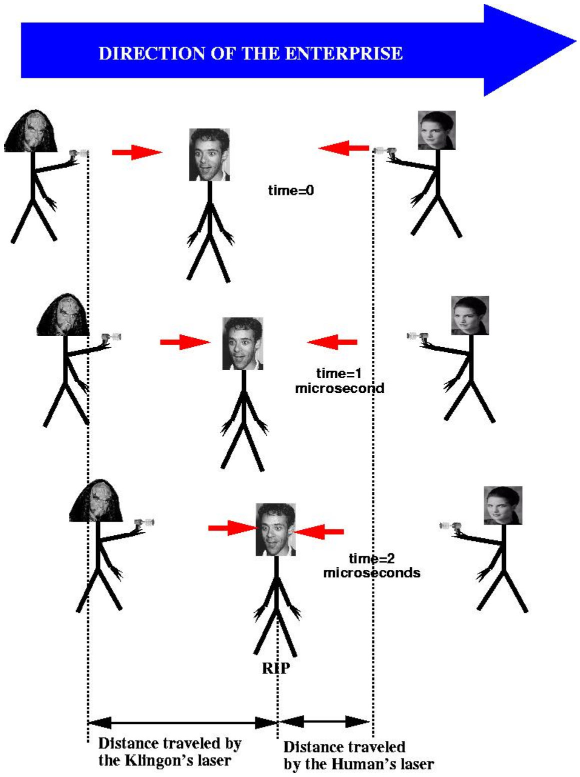


Figure 6.7: Illustration of Kirk's argument (the murder as seen from the space station)

Summary of the arguments				
K	happens before	**H**	as seen from the space station	(Kirk's argument)
H	happens before	**K**	as seen from space station	(Klingon capt.'s argument)
K	simultaneous with	**H**	as seen from Enterprise	(McCoy's argument)

So the Principle of Relativity forces us to conclude that in this situation the ordering of events in time is *relative*. But, this better not be true for *all* events: if the Principle of Relativity would predict that all time orderings are relative we could then imagine an observer who sees you, the reader, being born before your parents!

So there are events such as birth and death of a person which should occur in succession with the *same* ordering for any observer. And there are other events, such those in the shooting mystery, whose ordering in time is observer dependent. What is their difference?

The one clue is the following: in the story the assassins came in from

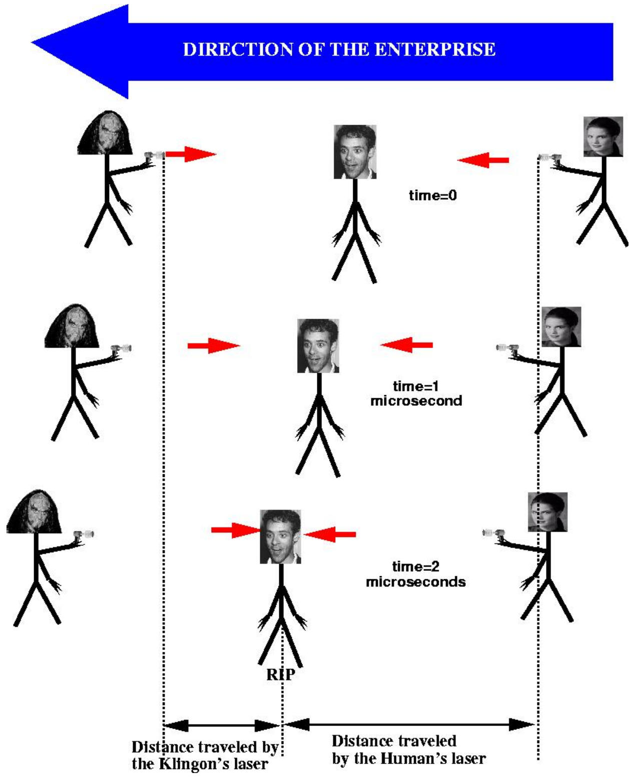


Figure 6.8: Illustration of Klingon captain's argument (the murder as seen from the bird of prey)

opposite sides of a cargo bay and shot the victim. Since lasers travel at the speed of light, the human will receive the image of the klingon shooting only after she herself has fired (in order to see anything we must receive light from some source); the same is true for the klingon. So, *when they fired they could not have been aware of each other's action.*

This is not the same as for birth and death: a cat is born and then the dog eats the cat. It is then possible for you to tell your dog, that is, to send him a signal, that the cat was born. This signal reaches the dog before he performs his grim action (Fig. 6.10)

So two events A (cat is born) and B (dog eats cat) are ordered in the same way in time for all observers if we can send a signal at the time one event occurs (A) which will reach an observer who will witness the second event (B). In this case everyone will agree that A occurs before B, no matter what the relative speed of the observer. An extreme case consists of those

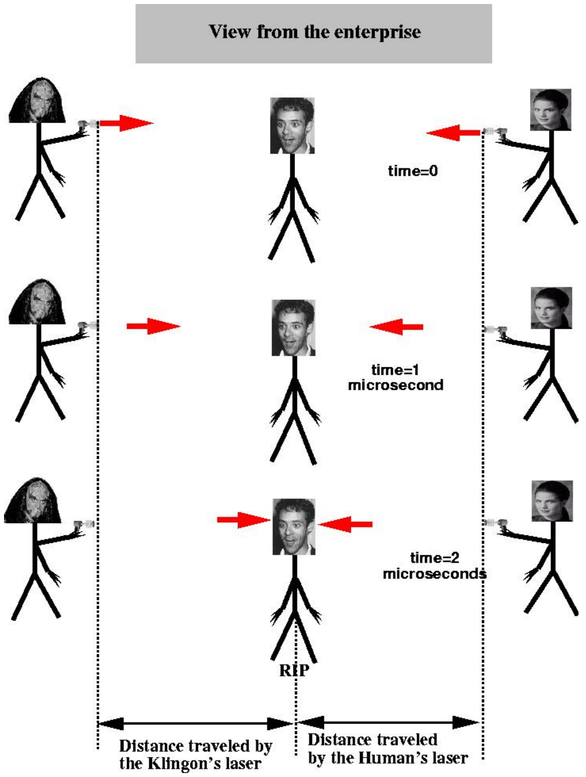


Figure 6.9: Illustration of McCoy's argument (the murder as seen from the Enterprise).

events occurring at the same time at the same place will be seen to occur at the same time by all observers (everyone agrees that the laser beams hit the victim at the same time).

In contrast if no signals sent at the time A occurs can reach an observer before B happens, then the ordering in time of A and B depends on the relative velocity of the observer.

So there is no hope of going back in time with the winning Loto number and becoming a millionaire. If you think about it, the number of paradoxes which would arise if all time orderings were relative would be enormous: if you could go back in time, there would be two of you: one a pauper and the other a millionaire...but which one *is* you? Fortunately the Special Theory of Relativity simply disallows such situations.

Why did all this happen? Because the speed of light is always c. Both laser bolts will be seen to travel at the same speed by all observers, and

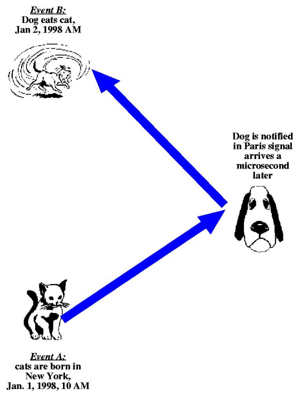


Figure 6.10: Illustration of events whose time ordering is *the same* for all observers.

because c is not infinite, the time it takes to reach a target depends on how the target is moving.

I will emphasize again the conclusions. Since the speed of light is the same for every observer in an inertial frame of reference, two things that are simultaneous to one observer will not be so according to other observers. The inescapable conclusion is that simultaneity is not an absolute concept: the statement "two events at different places occurred at the same time" is true only in a certain inertial reference frame and will be found to be incorrect in other frames.

Despite this there *are* events that everyone will agree are simultaneous: any two events happening at the same time and *at the same spot* will be seen to coincide by any observer. It is when the events are separated by a distance that simultaneity is relative. If events occurring at the same time and place for one observer were seen to occur at different times by another observer one can imagine going to a reference frame where the bullet that killed Lincoln went by his seat one hour before the president sat down. In this frame he was never assassinated!

One thing that Principle of Relativity does not permit is for some events which occur sequentially and such that the first affects the second to be inverted in order. For example it is impossible to go to a frame of reference in which the end of an exam occurs before it begins. It is only events that are mutually independent whose ordering in time can be inverted: two babies could be seen to be born one before the other or vice-versa, but only if they are not born at the same time at the same spot, so Jacob could not be the first born to Isaac (as opposed to Essau) in some frame of reference...the

Bible's story is, in this sense, frame independent.

6.2.3 The third prediction: The demise of Universal Time

Another peculiar and surprising consequence of the Principle of Relativity is that time intervals are no longer universal but depend on the frame of reference. Consider, for example, a clock consisting of a light source and detector. The source emits a light pulse, the pulse goes up and is reflected at a height h by a mirror. It is then detected and this determines one unit of time. See Fig 6.11.

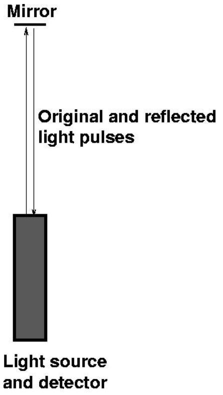

Mirror
Original and reflected light pulses
Light source and detector

Figure 6.11: A clock at rest with respect to the observer

The time it takes the light pulse to come and go is $t_0 = 2h/c$. This is precisely the time it would be measured by any observer carrying any other clock as long as this observer is not moving with respect to the above timepiece.

Now let's consider what an observer moving with respect to this simple clock sees. This is shown in Fig. 6.12

It is clear that the distance traveled by the beam is larger than the up-down trip observed by the first person. But since the speed of the light beam is the same for both observers, the time measured by the second observer will be *larger*. If we have two such clocks one is at rest with respect to us and the other is moving, we find that the moving clock slows down, moreover, the faster it moves the slower it ticks. This is called *time dilation*: a moving clock ticks slower.

Time dilation: a moving clock ticks slower

This argument was based on the simple clock of Fig. 6.11, will it be true for *all* clocks? To examine this question let's assume we have another clock (a Rollex, for example) which gives ticks *same* way no matter how

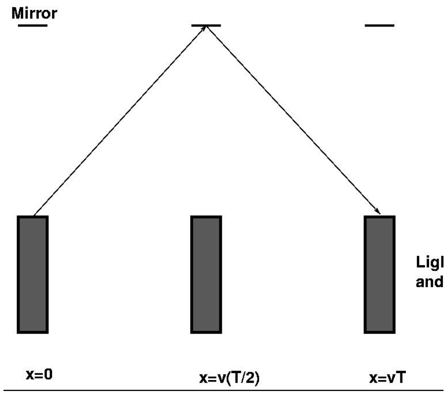
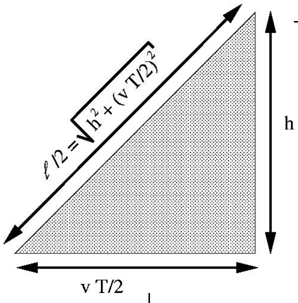


Figure 6.12: A clock moving with speed v to the right with respect to the observer

it moves. You go on a long trip to a near-by star taking the Rollex with you *and* also a clock like the one in Fig. 6.11. Your spaceship, you will notice, has no windows (they had to cut the budget *somewhere!*), but you go anyway. You experience the effects of lift-off but after a while you appear to be at a standstill: you are then moving at a constant speed with respect to Earth. But remember we assumed that the Rollex still ticks the same way as the clocks on Earth, and we have proved that your light-clock does not. So you will see a mismatch between the Rollex and the light-clock: this is an experiment which is done completely inside the spaceship and which determines whether you are moving. If there were such a Rollex the Principle of Relativity would be violated.

If we accept the Principle of Relativity we must conclude that time dilation will occur for *any* clocks, be it a Rollex, a biological clock or a Cartier. Note that this follows from the Principle of Relativity and the validity of Maxwells' equations, no additional assumptions are required.

If an observer at rest with respect to a clock, finds that she is pregnant and eventually delivers, the whole process taking precisely nine months, another observer moving with respect to her (and the simple clock) will find this claim to be wrong, he will state that she had a longer pregnancy (or a very long delivery) but that in any case the whole thing took longer than nine months.



Time dilation and Pythagoras' theorem. The distance the light has to travel in Fig. 6.12 can be determined by using Pythagoras' theorem.
In this reference frame light travels along the long sides of the triangles, each has a length which I call $\ell/2$; let's call T the time it takes to complete the trip, by Pythagoras' theorem $\ell/2 = \sqrt{h^2 + (vT/2)^2}$. On the other hand $\ell = cT$ since light moves at speed c for any observer and it takes a time T (according to the moving observer!) for it to get back to the detector. Solving for T we get

$$T = \frac{(2h/c)}{\sqrt{1-(v/c)^2}} = \frac{T_0}{\sqrt{1-(v/c)^2}}.$$

Thus the observer in motion with respect to the clock will measure a time T greater than T_0, the precise expression being given by the above formula.

So how come we do not see this in ordinary life? The reason is that the effect is very small in everyday occurrences. To be precise it an observer at rest with respect to the clock in Fig. 6.11 measures a time T_0 then the observer which sees the clock move at speed v (and sees the situation depicted in Fig. 6.12) will measure a time T, where $T = T_0/\sqrt{1-v^2/c^2}$ (see the box above). So the effect reduces to the appearance of the factor $1/\sqrt{1-v^2/c^2}$ which in usual circumstances is very close to one (so that T is almost equal to T_0). For example an ordinary man moving at, say 90miles/hr (trying to get his wife to the hospital before she delivers), $v/c = 0.0000001 = 10^{-7}$ (approximately) so that the above factor is essentially one (up to a few hundredths of a trillionth). This is typical of the magnitude of the new effects predicted by Einstein's theory for everyday situations: they are in general very small since the velocities of things are usually very small compared to c.

There are some instances, however, in which the effects are observable. There are subatomic particles which are unstable and decay (the process by which they decay is irrelevant) in a very small time interval when measured in the laboratory. It has also been found that high intensity radiation coming from space and hitting the upper atmosphere generates these same particles (again the process is immaterial). To the initial surprise of the experimenters, these particles survive the trip down to surface of the earth, which takes longer, *as measured on the Earth*, than the particle's lifetime!

The surprise evaporated when it was noted that the particles are moving very fast with respect to the Earth, almost at the speed of light, so that a time interval which is very short when measured at rest with respect to the particle will be much longer when measured in the laboratory.

So the rate of all clocks depends on their state of motion. In this sense

The rate of all clocks depends on their state of motion.

Time is relative.

And while the effect is small in many cases, it is spectacular in others. This is a surprising consequence of the Principle of Relativity and requires a complete divorce from Newton's concept of time (which he assumed to flow evenly under all circumstances, see Sect. **??**): time intervals depend on the motion of the observer, there is no "universal" time.

Time dilation is a *prediction* of the theory which must not be accepted as dogma but should be verified experimentally. All experiments do agree with this prediction. The fact that the theory of relativity makes predictions which can be tested experimentally, is what makes this an honest theory: it is falsifiable. It has been accpeted not because of its beauty, but because these predictions have been verified.

6.2.4 Length contraction

So time is relative, what about distance? In order to think about this note that when we say that the distance between two objects is ℓ we imagine measuring the position of these objects simultaneously...but simultaneity is relative, so we can expect distance to be a relative concept also.

To see this consider the above subatomic particles. As mentioned they are moving very fast but we can still imagine Superman (an unbiased observer if there is one) riding along with them. So we have two pictures: from the observer on earth Superman's clocks (accompanying the particle) are very slow, and so he/she can understand why it takes so long for the particle to decay. But for Superman the particle is at rest and so it must decay in its usual short time...the fact remains, however, that the particle does reach the earth. How can this be? Only if the distance which the particle traveled as measured in the frame of reference in which it is at rest is very short. This is the only way the observation that the particle reaches the earth's surface can be explained: for the observer on the earth this is because of time dilation, for the observer riding along with the particle, this is because of length contraction, see Fig. 6.13.

But we do not require peculiar subatomic particles in order to demonstrate length contraction (though the Principle of Relativity requires that if

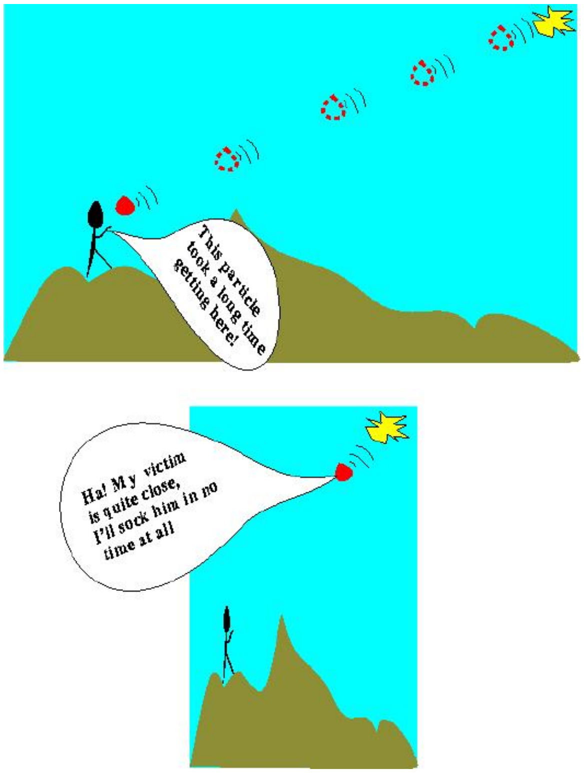

This particle took a long time getting here!
Ha! My victim is quite close, I'll sock him in no time at all

Figure 6.13: An observer measures a long life-time for the particles due to time dilation. The particles measures a short distance between itself and the observer due to length contraction.

it occurs for the example above it should occur in all systems, otherwise we could determine by comparison which system has an absolute motion). So consider the previous experiment with the moving clock (Fig 6.12).

- The observer watching the clock move with velocity v notes that in a time T the clock moves a distance $\ell = vT$.

- The observer riding with the clock notes that the same distance is covered in a time T_0; therefore the length measured by him/her is $\ell_0 = vT_0$ (He also sees the other observer receding with speed v.)

- Therefore we have $\ell = vT = vT_0/\sqrt{1-(v/c)^2} = \ell_0/\sqrt{1-(v/c)^2}$. Thus, the observer moving with the clock will measure a shorter length compared to the one measured by the other observer.

It is important to note that these expressions are *not* to be interpreted as “illusions”, the an observer in motion with respect to a ruler will, when measuring its length, find a result smaller than the result of an observer at rest with respect to the ruler. An observer in motion with respect to a clock will measure a time larger than the ones measured by an observer at rest with the clock.

The question, “what is ‘really’ the length of a ruler?” has no answer for this length depends on the relative velocity of the ruler to the measuring device [2]. The same as with velocity, specifying lengths requires the framework provided by a frame of reference,

Length is relative.

Length is relative

Note that this peculiar effect occurs only for lengths measured along the direction of motion and will *not* occur for lengths perpendicular to it. To see this imagine two identical trees, we sit at base of one and we observe the other move at constant speed with respect to us, its direction of motion is perpendicular to the trunk. In this setup as the roots of both trees coincide also will their tops, and so in *both* frames of reference we can *simultaneously* determine whether they have the same height; and they do.

This implies that a moving object will be seen thinner (due to length contraction) but not shorter. Thin fellows will look positively gaunt at speeds close to that of light.

These conclusions require we also abandon Newton’s description of space: distances are observer-dependent, no longer notches in absolute space.

6.2.5 Paradoxes.

The above conclusions can be very confusing so it might be worthwhile to discuss the a bit.

Take for example length contraction: the Principle of Relativity implies that if we measure the length some rod while at rest with respect to it, and then we measure it when it is moving along its length, the second measurement yields a smaller value. The crucial point to keep in mind is the condition that the first measurement is made at rest with respect to the rod.

[2] One can, of course, say that *the* length of a ruler is the one measured while at rest with respect to it...but this is only a convention. Once the result of any length measurement is known (for any relative speed between ruler and measuring device), special relativity determines unambiguously what any other observer would measure.

Similarly suppose we have two clocks labeled 1 and 2. which are in perfect agreement when they are at rest with respect to each other. Suppose now these clocks are endowed with a relative velocity. Then when we look at clock 2 in the frame of reference in which clock 1 is at rest, clock 2 will be measured to tick slower compared to clock 1. Similarly, in the frame of reference in which clock 2 is stationary, clock 1 will run slower compared to clock 2.

These results can be traced back to the fact that simultaneous events are not preserved when we go from one reference frame to another.

There are many "paradoxes" which appear to imply that the Principle of Relativity is wrong. The do not, of course, but it is interesting to see how the Principle of Relativity defends itself.

1. Consider a man running with a ladder of length ℓ (measured at rest) and a barn also of length ℓ (again, when measured at rest). The barn has two doors and there are two persons standing at each of them; the door nearer to the ladder is open the farthest is closed. Now the man with the ladder runs fast towards the barn while the door persons have agreed to close the first door and open the second door as soon as the rear of the ladder goes through the first door.
   This is a paradox for the following reason. The ladder guy is in a frame of reference in which the ladder is at rest but the barn is moving toward him, hence he will find the length of the barn shortened (shorter than his ladder), and will conclude that the front of the ladder will hit the second door before the first door is closed.
   The barn people in contrast find the ladder shortened and will conclude that it will fit comfortably. There will even be a short lapse between the closing of the first door and the opening of the second, there will be no crash and the ladder guy will sail through.
   So who is right?
   The answer can be found by remembering that an even simultaneous for the barn people (the closing and opening of the doors) will not be simultaneous for the ladder guy. So, while for the door person the opening of the rear door and closing of the front occur at the same time, the ladder guy will see the person at the second door open it *before* the person at his rear closes that door and so he will sail through but only because, he would argue, the door guards were not synchronized.

2. There is an astronaut whose length is 6 ft and he sees a big slab of metal with which he/she is going to crash. This piece of metal has

a square hole of length 6 ft. (measured at rest with respect to the slab). From the point of view of the astronaut the hole is shrunk and so he will be hit...and die! From the point of view of an observer on the shuttle the plate is falling toward earth and the astronaut moving at right angles toward it, hence this observer would measure a short astronaut (5 ft) [3] and conclude that he/she will not be harmed (see Fig. 6.14). What does really happen?

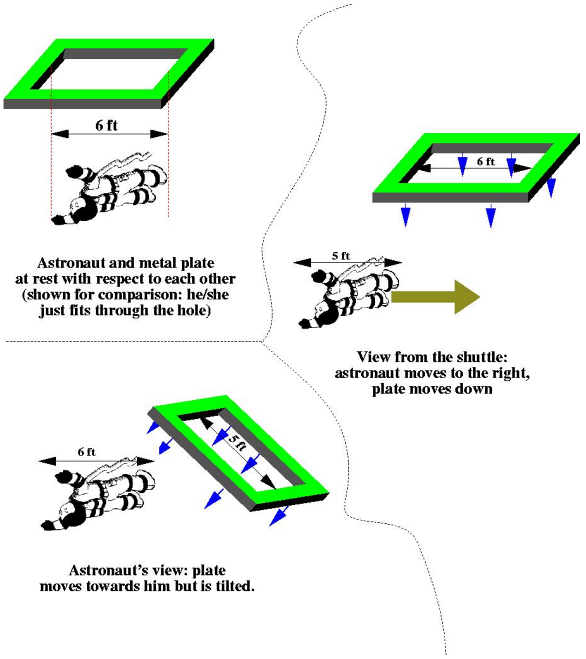

6 ft
Astronaut and metal plate
at rest with respect to each other
(shown for comparison: he/she
just fits through the hole)
6 ft
5 ft
View from the shuttle:
astronaut moves to the right,
plate moves down
6 ft
5 ft
Astronaut's view: plate
moves towards him but is tilted.

Figure 6.14: An astronaut's close encounter with a metal plate

The problem is solved in the same way as above. For the astronaut to be hit a simultaneous coincidence of his head and legs with the two extremes of the slab's hole should occur. In fact he is not hit. What is more peculiar is what he sees: he will see the slab tilt in such a way that he goes through the hole with no problem!

This story illustrates the peculiar look which big objects acquire at very large speeds. For example, a kettle moving close to the speed of light with respect to, say, the Mad Hatter will be observed to twist in

[3] This corresponds to an astronaut moving at about half the speed of light toward the plate.

a very curious way indeed, see Fig. 6.15.

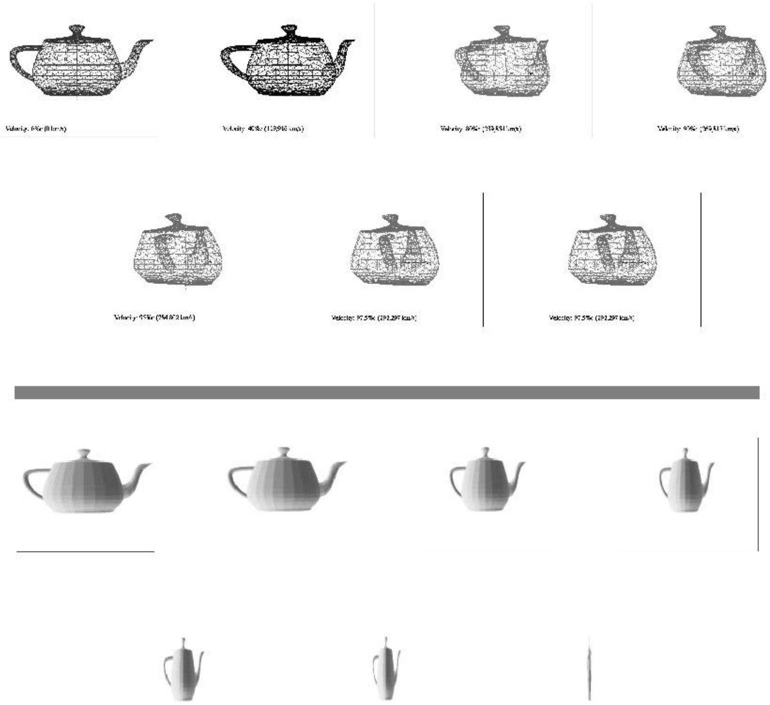

Velocity: 0%c (0 km/s)
Velocity: 40%c (119,916 km/s)
Velocity: 80%c (239,834 km/s)
Velocity: 90%c (269,813 km/s)
Velocity: 95%c (284,802 km/s)
Velocity: 97.5%c (292,297 km/s)
Velocity: 97.5%c (292,297 km/s)

Figure 6.15: A relativistic kettle. The top view shows how the three dimensional view is distorted due to relativistic effects. The bottom view shows the corresponding behavior of a flat kettle which exhibits only length contraction.

Just as for the case of length contraction and time dilation, the effect on the kettle is *not* an optical illusion, but any unbiased observer (such as a photographic camera) would detect the above images precisely as shown. If the relative velocity between the observer and the kettle is known, one can use the formulas of special relativity to determine the shape of the object when at rest with respect to it...and we would obtain the first of the images: a nice kettle

3. Consider two identical twins. One goes to space on a round trip to Alpha-Centauri (the star nearest to the Sun) traveling at speeds very

close to c. The round trip takes 10 years as clocked on Earth [4]. As seen by the twin remaining on Earth all clocks on the ship slow down, including the biological clocks. Therefore he expects his traveling twin to age less than 10 years (about 4.5 years for these speeds; the difference is large since the speed is close to c).

On the other hand the twin in the spacecraft sees his brother (a together with the rest of the solar system) traveling backwards also at speeds close to c and he argues that Einstein requires the twin on Earth to age less than 10 years. Thus each one states that the other will be younger when they meet again!

The solution lies in the fact that the traveling twin is not always in an inertial frame of reference: he must decelerate as he reaches Alpha-Centauri and then accelerate back. Because of this the expressions for time dilation as measured by the traveling twin will not coincide with the ones given above (which are true only for observers in different inertial frames). It is the traveling twin that will be younger.

6.2.6 Space and Time

All events we witness are labeled by a series of numbers, three to tell us where it happened, and one to determine *when* it happened. All in all four numbers are needed. These numbers are determined by some measuring devices such as measuring rods and clocks.

According to Newton (see Sect. **??**) the properties of measuring rods and clocks can be made completely independent of the system which they measure (if it does not look like that, you can buy a higher quality device which will satisfy this criterion). But Einstein showed this is *not* the case: even Cartier watches slow down when compared to Seiko watches when they move with respect to each other. Even high density steel beams will be measured to be shorter than wimpy papers when their relative velocity is non-zero.

The measurements obtained by two observers in motion relative to each other are not identical, but they *are* related. For example, the times measured by two clocks are related by the time-dilation formula given earlier. Suppose observer A measures the location and time at which an event occurs: spider-man ran the 100 yard dash in 3 seconds flat. Now observer B, moving with respect to A, wants a description of this feat in his own coordinates. In order to find how many yards spider man ran *as measured by B*

[4] This corresponds to a speed of 90% that of light

this observer needs to know his velocity with respect to A, the distance spidy ran as measured by A (100 yds) and how long did he take as measured by A's clock (3 sec); it is *not* enough to know the distance and relative velocity, the *time* it took is also needed.

The fact that in order to compare results from different observers both position and time are required is completely foreign to Newtonian mechanics. Yet this is the way the universe is organized. Far from being independent, space (that is, position) and time are interlinked. In fact, the mathematical description of the Special Theory of Relativity is most naturally expressed by combining space and time into one object: *space-time*. A point in space-time determines the position *and* time of occurrence of an event.

Within Special Relativity space-time is unaltered by whatever is in it. There are rules that state how the measurements of two observers are related, but these rules are unaltered by the objects (and beings) that populate space-time, they are the same whether we look at a pea, an elephant or a star millions of times more massive than the Sun. Space and time are still the arena where Nature unfolds.

We will see when we describe the General Theory of Relativity (Chap. **??**) that space-time is far from being this imperturbable object where things just happen, it is in fact a dynamical system which affects and is affected by the matter in it. The development of our ideas of space and time from being independent of each other and imperturbable, to being meshed into space-time system, to being a dynamical object is one of the most profound developments derived from the general and special theories of relativity.

6.2.7 The top speed.

In all the above discussion all the effects would go away is the speed of light were infinite. If there is a top speed, which by definition has and absolute value (the same for all observers), then all the above effects return. It is because the equations found by Maxwell involve an absolute speed, and because they agree with experiment, and because nothing has been found to travel faster than light in vacuum, and because all the consequences of the Principle of Relativity are verified again and again with the top speed equal to c, that we believe this top speed to be precisely c.

Imagine, as Einstein did when a teenager, what would happen is you could move at the speed of light. As you go by a village (for example) you'd move at the same speed as all the light coming form that village. So, if you look around, you would see the same things all the time, nothing would ever change since you are riding along with a single image: the one carried by the light from the village at the time you passed it. In your frame of reference time would stand still! (we will see, however, that it is impossible for something having mass –such as you– to move at the speed of light. You can reach speeds very close to c but never reach the speed of light itself).

Imagine now what would happen if, for example, a rat manages to travel at a speed greater than c. Let's imagine that as the rat travels by you, you send a short laser light pulse after it. According to you the rat will gain on the light pulse steadily. Since the distance between the rat's tail and the front edge of the pulse increases the rat would think that the pulse is moving in the opposite direction. So you and the rat would disagree even on the direction along which the light is traveling. This, of course, contradicts the Principle of Relativity and/or Maxwell's equations and it shows that the Principle of Relativity together with Maxwells' equations imply that nothing can move faster than light.

This is a good feature: if a faster-than-light-rat could be found, the vermin farme of reference would have time flowing backwards. To see this imagine the rat going by the same village mentioned above. Since the rat moves faster than light it will steadily gain on the light beams than come from the village. As it looks around the rat will see the church clock strike 12, and, as it gains on some earlier images, the rat would wee the clock strike 11, etc. So events whose time orderings wer aboslute would no longer occur in the correct order in this frame.

6.2.8 Mass and energy.

How could it be that we cannot accelerate something to go faster than light? Surely we could kick a ball again and again and again until it travels faster right? No! and the reason is quite interesting.

As something is moving with respect to another object we say that the moving thing has a certain amount of energy by virtue of its motion. Energy is the ability to do some work, and, indeed, a moving thing can be lassoed and made to do some work, like pulling a car (of course in so doing it looses energy and slows down).

Now, when we have the above object moving, it will have a certain amount of energy. Einstein argued, the only way we can insure that it cannot be accelerated indefinitely, is if there is a universal equivalence between mass and energy. The more energy an object has, the heavier it will be. When we speed it up a little bit it becomes a bit heavier, and so it also becomes a bit harder to speed it up further. In fact, the closer we are to the speed of light, the larger the force is needed to accelerate the object; an infinite force is needed to speed up a material object to the speed of light: it never happens!

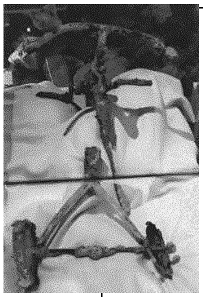
But there is more to the equivalence of mass and energy, for it also implies that an object of mass m has energy, just by virtue of its existence; the specific relationship is

$$E = mc^2.$$

This formula plays a basic role in nuclear reactions (and in atom bombs, for that matter): in these processes an atomic nucleus of initial mass M is transformed (either because the environment is tailored to insure this or because is is unstable and disintegrates spontaneously) into another object of smaller mass m. The difference in mass is released as energy in the amount $(M - m)c^2$.

To give an idea of how powerful this is, suppose we initially have a sheet of paper weighing 6gr, and that at the end we have something weighing half this amount. The energy released is then so big as to turn on a light bulb of 100W for about 86,000 years, or run a hair-drier for about 4000 years.

The energy released through the transformation of mass is also capable of destroying a whole planet (or at 'least' all life on it). Einstein was not aware of this application until much later in his life.



Shin's tricycle. Shin-ichi was a three year old boy who loved his tricycle. When the bomb was dropped, he was playing with his best friend, Kimiko. They died. They were buried in the garden of Shin-ichi's house together. In July 1985, 40 years later, their parents decided to move them to a proper grave.
From the story of "*Shin's Tricycle*" (Translation by Kazuko Hokumen-Jones and Jacky Copson):
Early in the morning, I began to dig open the grave with Kimi's mother, who had come to help. After digging for a while a rusty pipe began to show. "Oh! It's the tricycle!" Before I realized it I had started to sob. To tell you the truth, I'd forgotten all about the tricycle.
"Look! There's something white," someone cried. I felt like ice. Carefully we uncovered the bones using chop-sticks and brushes. There were a number of tiny bones.
"Shin-ichi, Shin-ichi." "Kimiko." Everyone's eyes were glued to the little white hands of the two children. They were still holding hands....

The principle $E = mc^2$ was used during the Second World War to develop what is now known as atomic weapons (Fig. 6.16). Shortly thereafter it was used to develop the hydrogen bomb. Atomic bombs were used during the Second World War in two Japanese cities, Hiroshima and Nagasaki. Hundreds of thousands of people died. The creation of nuclear weapons was one of the watersheds of the 20-th century, and it marks one of the most dramatic instances in which physics has affected the social structure of the planet. Yet the very same formulas also suggest the possibility of obtaining vast amounts of energy which can be used for constructive purposes. It is a burden of post-second World War physicists to deal with this issue, and to strive for decent and environmentally safe applications of nuclear power.

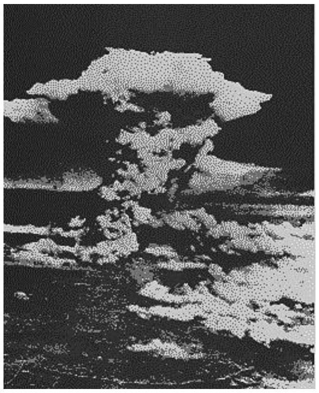


Figure 6.16: An atomic explosion.

Chapter 7

The General Theory of Relativity

The General Theory of Relativity is, as the name indicates, a generalization of the Special Theory of Relativity. It is certainly one of the most remarkable achievements of science to date, it was developed by Einstein with little or no experimental motivation but driven instead by philosophical questions: Why are inertial frames of reference so special? Why is it we do not feel gravity's pull when we are freely falling? Why should absolute velocities be forbidden but absolute accelerations by accepted?

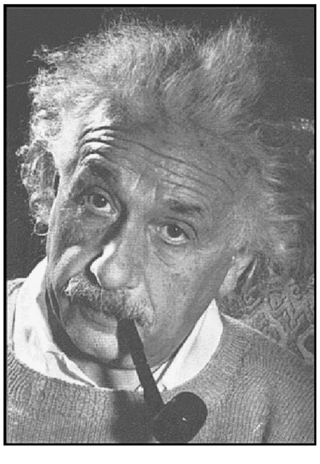


Figure 7.1: Einstein

7.1 The happiest thought of my life.

In 1907, only two years after the publication of his Special Theory of Relativity, Einstein wrote a paper attempting to modify Newton's theory of gravitation to fit special relativity. Was this modification necessary? Most emphatically yes! The reason lies at the heart of the Special Theory of Relativity: Newton's expression for the gravitational force between two objects depends on the masses and on the distance separating the bodies, but makes no mention of time at all. In this view of the world if one mass is moved, the other perceives the change (as a decrease or increase of the gravitational force) *instantaneously*. If exactly true this would be a physical effect which travels faster than light (in fact, at infinite speed), and would be inconsistent with the Special Theory of Relativity (see Sect. **??**). The only way out of this problem is by concluding that Newton's gravitational equations are not strictly correct. As in previous occasions this does not imply that they are "wrong", it only means that they are not accurate under certain circumstances: situations where large velocities (and, as we will see, large masses) are involved cannot be described accurately by these equations.

In 1920 Einstein commented that a thought came into his mind when writing the above-mentioned paper he called it "the happiest thought of my life":

> *The gravitational field has only a relative existence...* Because for an observer freely falling from the roof of a house – *at least in his immediate surroundings* – there exists no gravitational field.

Let's imagine the unfortunate Wile E. Coyote falling from an immense height [1]. As he starts falling he lets go of the bomb he was about to drop on the Road Runner way below. The bomb does not gain on Wile nor does ? it lag behind. If he were to push the bomb away he would see it move with constant speed in a fixed direction. This realization is important because this is exactly what an astronaut would experience in outer space, far away from all bodies (we have good evidence for this: the Apollo 10–13 spacecrafts did travel far from Earth into regions where the gravitational forces are quite weak).

Mr. Coyote is fated to repeat the experience with many other things: rocks, magnets, harpoons, anvils, etc. In all cases the same results are obtained: with respect to him all objects, irrespective of composition, mass,

[1] I ignore air resistance

etc. behave as if in free space. So, if he should fall inside a closed box, he would not be able to tell whether he was plunging to his death (or, at least, severe discomfort), or whether he was in outer space on his way to Pluto at constant speed.

This is reminiscent of Galileo's argument: the observer lets go of some objects which remain in a state of uniform motion (with respect to him!). This behavior is independent of their chemical or physical nature (as above, air resistance is ignored). The observer (Wile), as long as he confines his/her observations to his/her immediate vicinity (that is, as long as he/she does not look down) has the right to interpret his state as 'at rest'. Just as Galileo argued that experiments in a closed box cannot determine the state of uniform motion of the box, Einstein argued that experiments in a freely falling small[2] closed box cannot be used to determine whether the box is in the grip of a gravitational force or not.

Why would this be true? The answer can be traced back to the way in which gravity affects bodies. Remember (see Sect. **??**) that the quantity we called m (the mass) played two different roles in Newton's equations. One is to determine, given a force, what the acceleration of the body would be: $F = ma$ (the inertial mass). The other is to determine the intensity with which the said body experiences a gravitational force: $F = mMG/r^2$ (the gravitational mass). As mentioned before these two quantities need not be equal: the first "job" of m is to tell a body how much to accelerate given *any* force, a kick, an electric force (should the body be charged), etc. The second "job" tells the body how much of the gravitational force should it experience and also determines how strong a gravitational force it generates. But, in fact, both numbers are equal (to a precision of ten parts per billion).

What does this imply? Well, from Newton's equations we get

$$\frac{mMG}{r^2} = ma \quad \text{so that} \quad \frac{MG}{r^2} = a;$$

this equation determines how a body moves, which trajectory it follows, how long does it take to move from one position to another, etc. *and is independent of* m! Two bodies of different masses, composition, origin and guise will follow the same trajectory: beans, bats and boulders will move in the same way.

So the equality of the two m's was upgraded by Einstein to a postulate: the *Principle of Equivalence*; this one statement (that the m in ma and the m in mMG/r^2 are identical) implies an incredible amount of new and

[2] The reasons behind the requirement that the box be small will become clear soon.

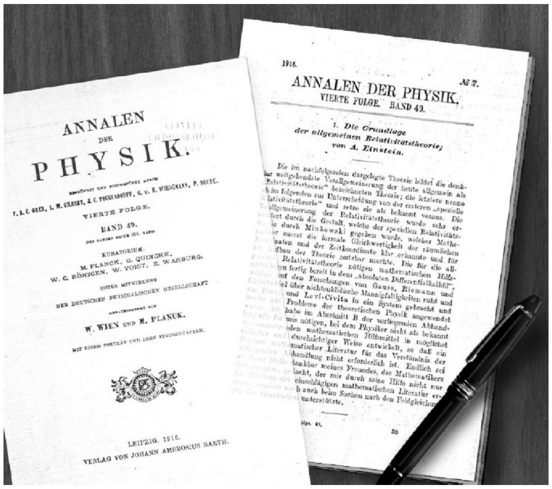
The m in $F = ma$ is called the *inertial mass* and the m in mMG/r^2 the *gravitational mass*

The inertial and gravitational masses are identical

surprising effects. The m in $F = ma$ is called the *inertial mass* and the m in mMG/r^2 the *gravitational mass*. Then the Principle of Equivalence states that the inertial and gravitational masses are identical.

The whole of the General Theory of Relativity rests on this postulate, and will fail if one can find a material for which the inertial and gravitational masses have different values. One might think that this represents a defect of the theory, its Achilles heel. In one sense this is true since a single experiment has the potential of demolishing the whole of the theory (people have tried...hard, but all experiments have validated the principle of equivalence). On the other hand one can argue that a theory which is based on a minimum of postulates is a better theory (since there are less assumptions involved in its construction); from this point of view the General Theory of Relativity is a gem [3].

The completed formulation of the General Theory of Relativity was published in 1916 (Fig. 7.2).



ANNALEN
DER
PHYSIK.

BEGRÜNDET UND FORTGEFÜHRT DURCH
F. A. C. GREN, L. W. GILBERT, J. C. POGGENDORFF, G. u. E. WIEDEMANN, P. DRUDE.

VIERTE FOLGE.

BAND 49.
DER GANZEN REIHE 354. BAND.

KURATORIUM:
M. PLANCK, G. QUINCKE,
W. C. RÖNTGEN, W. VOIGT, E. WARBURG.

UNTER MITWIRKUNG
DER DEUTSCHEN PHYSIKALISCHEN GESELLSCHAFT
HERAUSGEGEBEN VON
W. WIEN UND M. PLANCK.

MIT EINEM PORTRÄT UND ZEHN FIGURENTAFELN.

LEIPZIG, 1916.
VERLAG VON JOHANN AMBROSIUS BARTH.

1916. ANNALEN DER PHYSIK. № 7.
VIERTE FOLGE. BAND 49.

1. *Die Grundlage der allgemeinen Relativitätstheorie; von A. Einstein.*

Die im nachfolgenden dargelegte Theorie bildet die denkbar weitgehendste Verallgemeinerung der heute allgemein als „Relativitätstheorie" bezeichneten Theorie; die letztere nenne ich im folgenden zur Unterscheidung von der ersteren „spezielle Relativitätstheorie" und setze sie als bekannt voraus. Die Verallgemeinerung der Relativitätstheorie wurde sehr erleichtert durch die Gestalt, welche der speziellen Relativitätstheorie durch Minkowski gegeben wurde, welcher Mathematiker zuerst die formale Gleichwertigkeit der räumlichen Koordinaten und der Zeitkoordinate klar erkannte und für den Aufbau der Theorie nutzbar machte. Die für die allgemeine Relativitätstheorie nötigen mathematischen Hilfsmittel lagen fertig bereit in dem „absoluten Differentialkalkül", welcher auf den Forschungen von Gauss, Riemann und Christoffel über nichteuklidische Mannigfaltigkeiten ruht und von Ricci und Levi-Civita in ein System gebracht und bereits auf Probleme der theoretischen Physik angewendet wurde. Ich habe im Abschnitt B der vorliegenden Abhandlung alle für uns nötigen, bei dem Physiker nicht als bekannt vorauszusetzenden mathematischen Hilfsmittel in möglichst einfacher und durchsichtiger Weise entwickelt, so daß ein Studium mathematischer Literatur für das Verständnis der vorliegenden Abhandlung nicht erforderlich ist. Endlich sei an dieser Stelle dankbar meines Freundes, des Mathematikers Grossmann, gedacht, der mir durch seine Hilfe nicht nur das Studium der einschlägigen mathematischen Literatur ersparte, sondern mich auch beim Suchen nach den Feldgleichun der Gravitation unterstützte.

olge. 49. 50

Figure 7.2: Einstein's General Theory of Relativity paper.

[3] The Special Theory of Relativity is equally nice, it is based on the one statement that all inertial frames of reference are equivalent.

7.1.1 Newton vs. Einstein

I have stated that Newton's mechanics and his theory of gravitation are but approximations to reality and whose limitations are now known [4]. So it might be questionable to use $F = ma$ and $F_{\rm grav} = mMG/r^2$ as basis to any argument as was done above. Einstein was careful to use these expressions only in situations where they are extremely accurate (small speeds compared to c and small gravitational forces). In these cases the inertial and gravitational masses are identical, as shown by experiment. Then he postulated that the same would be true under *all* circumstances. This statement, while consistent with Newton's equations, cannot, in a strict logical sense, be derived from them.

7.2 Gravitation vs acceleration

Consider the following experiment: a person is put in a room-size box high above the moon (chosen because there is no air and hence no air friction) with a bunch of measuring devices. This box is then taken high above the lunar surface and then let go: the box is then freely falling. The question is now, can the observer determine whether he/she is falling or whether he/she is in empty space unaffected by external forces (of course the answer is supposed to come before the box hits the surface). The answer to that is a definite NO! The observer can do experiments by looking at how objects move when initially at rest and when given a kick, he/she will find that they appear to move as is there were no gravitational forces at all! Similarly any experiment in physics, biology, etc. done solely inside the box will be unable to determine whether the box is freely falling or in empty space.

Why is that? Because of the equality of the gravitational and inertial masses. All objects are falling together and are assumed to be rather close to each other (the box is not immense) hence the paths they will follow will be essentially the same for each of them. So if the observer lets go of an apple, the apple and the observer follow essentially the same trajectory, and this implies that the observer will not see the apple move with respect to him. In fact, *if we accept the priniciple of equivalence*, nothing can be done to determine the fact that the observer is falling towards the Moon, for this can be done only if we could find some object which behaved differently from all the rest, and this can happen only if its gravitational and inertial masses

[4] For all we know our present theories of mechanics and gravitation may also be invalid under certain conditions.

are different. The principle of equivalence then implies that the observer will believe that he/she is an inertial frame of reference...until disabused of the notion by the crash with the surface.

The principle of equivalence is of interest neither because its simplicity, nor because it leads to philosophically satisfying conclsions. It's importance is based on the enormous experimental evidence which confirms it; as with the Special Theory, the General Theory of Relativity is falsifiable.

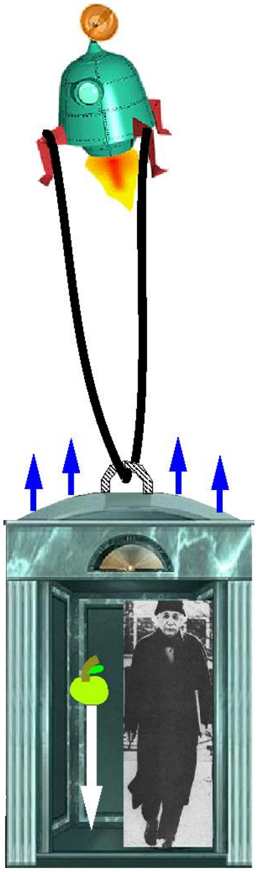


Figure 7.3: An observer cannot distinguish between acceleration produced by a rocket and the acceleration produced by gravity.

The lesson is that for any gravitational force we can always choose a frame of reference in which an observer will not experience any gravitational effects in his/her immediate vicinity (the reason for this last qualification will become clear below). Such a frame of reference is, as stated above,

For any gravitational force we can always choose a frame of reference in which an observer will not experience any gravitational effects in his/her immediate vicinity *freely falling.*

Conversely one can take the box an attach it to a machine that accelerates it (Fig. 7.3). If an observer drops an apple in such an accelerated box he/she will see the apple drop to the floor, the observer will also feel hi/her-self pressed against the bottom of the box, etc. The observer *cannot* distinguish between this situation and the one he/she would experience in the presence of gravitational forces! As long as we do experiments in a small region, the effects produced by a gravitational force are indistinguishable from those present in an accelerated reference frame.

In a small region the effects produced by a gravitational force are indistinguishable from those present in an accelerated reference frame

Does this mean that the gravitational forces are a chimera, an illusion? Of course no. Consider for example Fig. 7.4, two apples fall to the Moon inside a box which is also falling. If they are separated by a sufficiently large distance an observer falling with the apples and box will find that the distance between the apples shortens as time goes on: this cannot be an inertial frame he argues (or else it is, but there is some force acting on the apples).

This same set-up can be used to distinguish between a box under the influence of a gravitational force and one being pulled by a machine; again we need a very big box (planet-sized). An observer places an two apples at the top of the box and releases them, he/she carefully measures its initial separation. The apples fall to the bottom of the box and the observer measures their separation there. If it is the same as above, and is the same irrespective of their initial separation, the observer is being pulled by a machine (box and all). If the separation is different, he/she can conclude that he/she is experiencing the effects of a gravitational force.

7.3 Light

A very surprising corollary of the above is that light paths are bent by gravitational forces! I will argue this is true in a slightly round-about way.

Consider an elevator being pulled by a crane so that it moves with constant acceleration (that is its velocity increases uniformly with time). Suppose that a laser beam propagating perpendicular to the elevator's direction of motion enters the elevator through a hole on the left wall and strikes the right wall. The idea is to compare what the crane operator and the elevator passenger see.

The crane operator, who is in an inertial frame of reference, will see the sequence of events given in Fig. 7.5. Note, that according to him/her, light travels in a straight line (as it must be since he/she is in an inertial frame!).

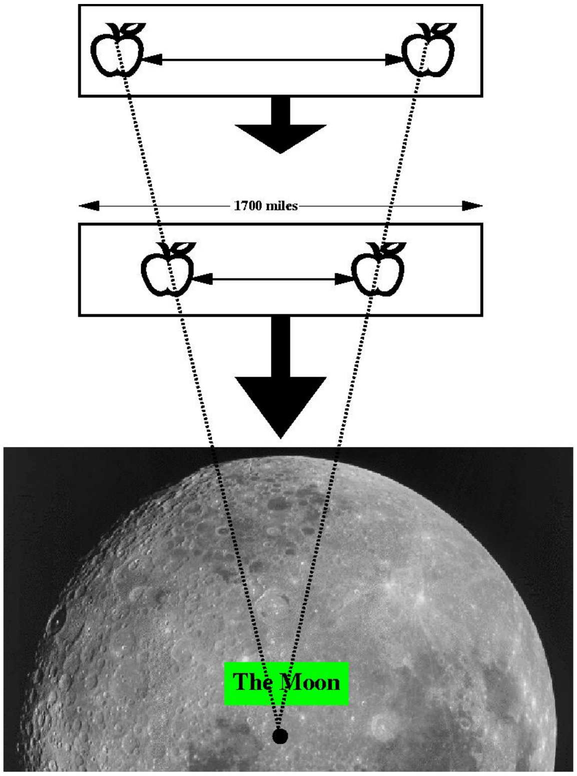


Figure 7.4: Experiment that differentiates between a gravitational effect and the effects of uniform acceleration: for an observer in the box the apples will draw closer.

The elevator passenger will see something very different as shown in Fig. 7.5: the light-path is curved! Thus for this simple thought experiment light paths will be curved for observers inside the elevator.

Now we apply the equivalence principle which implies that we cannot distinguish between an elevator accelerated by a machine and an elevator experiencing a constant gravitational force. It follows that the *same* effect should be observed if we place the elevator in the presence of a gravitational force: *light paths are curved by gravity*

Light light paths are curved by gravity

That gravity affects the paths of planets, satellites, etc. is not something strange. But we tend to think of light as being different somehow. The above argument shows that light is not so different from other things and is indeed affected by gravity in a very mundane manner (the same elevator experiment could be done by looking at a ball instead of a beam of light and the same

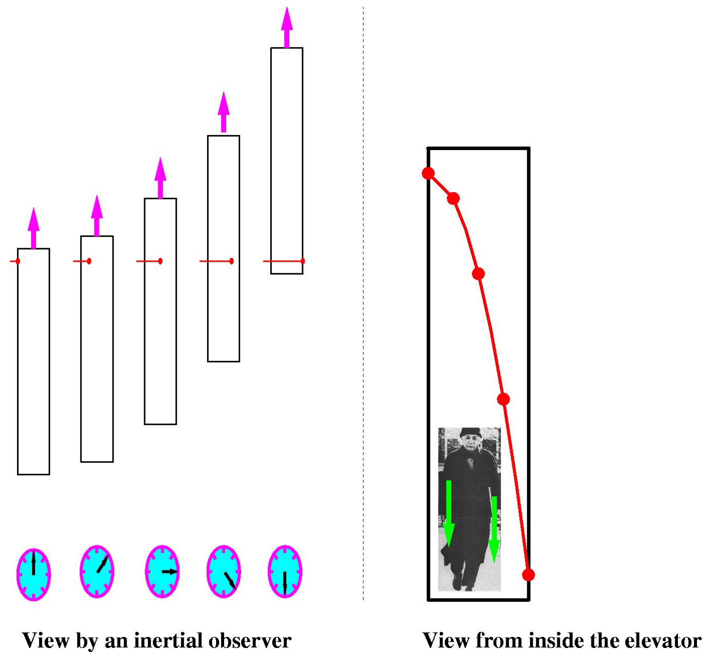


Figure 7.5: **Left:** sequence of events seen by an crane operator lifting an elevator at constant acceleration (the speed increases uniformly with time). The short horizontal line indicates a laser pulse which, at the initial time, enters through an opening on the left-hand side of the elevator. At the final time the light beam hits the back wall of the elevator. **Right:** same sequence of events seen by a passenger in an elevator being hoisted by a crane. The line joining the dots indicates the path of a laser which, at the initial time, enters through an opening on the left-hand side of the elevator. At the final time the light beam hits the back wall of the elevator.

sort of picture would result).

Ex A natural question is then, why do we not see light fall when we ride an elevator? The answer is that the effect in ordinary life is very small. Suppose that the height of the elevator in Fig. 7.5 is 8 ft. and its width is 5 ft; if the upward acceleration is 25% that of gravity on Earth then the distance light falls is less than a millionth of the radius of a hydrogen atom (the smallest of the atoms). For the dramatic effect shown in the figure the acceleration must be enormous, more than 10^{16} times the acceleration of gravity on Earth (this implies that the passenger, who weights 70 kg on Earth, will weigh more than 1,000 trillion tons in the elevator).

This does not mean, however, that this effect is completely unobservable (it is small for the case of the elevator because elevators are designed for very small accelerations, but one can imagine other situations). Consider from example a beam of light coming from a distant star towards Earth (Fig. 7.6) which along the way comes close to a very massive dark object. The arguments above require the light beam to bend; and the same thing will happen for any other beam originating in the distant star. Suppose that the star and the opaque object are both prefect spheres, then an astronomer on Earth will see, not the original star, but a *ring* of stars (often called an "Einstein ring). If either the star or the massive dark object are not perfect spheres then an astronomer would see several images instead of a ring (Fig. 7.7). This effect has been christened *gravitational lensing* since gravity acts here as a lens making light beams converge.

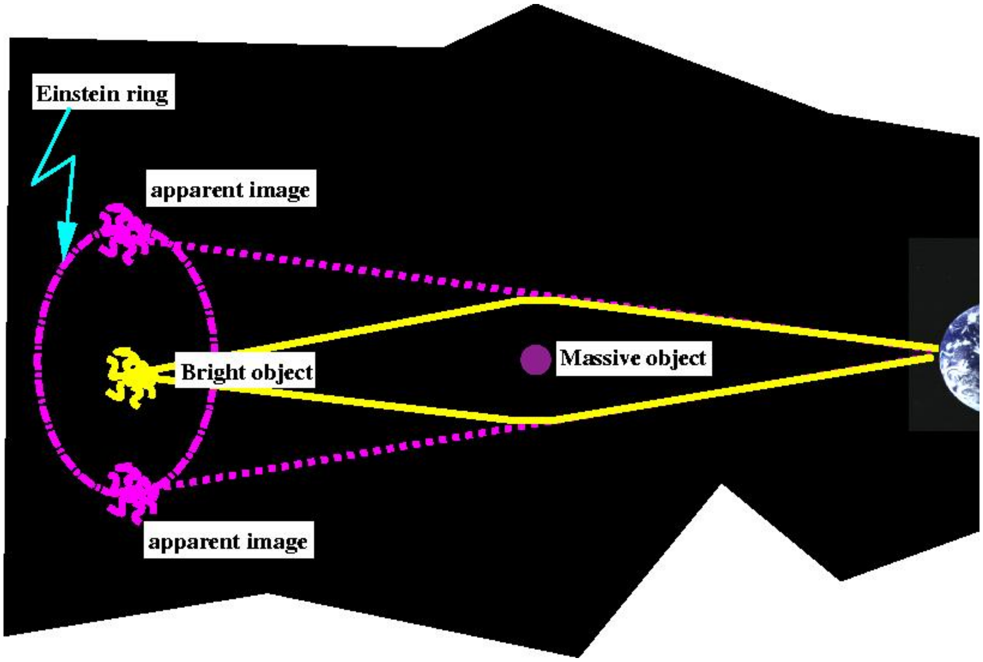


Figure 7.6: Diagram illustrating the bending of light from a star by a massive compact object. If both the bright objects and the massive object are prefect sphere, there will be an apparent image for every point on the "Einstein ring".

How do we know that the multiple images which are sometimes seen (Fig. 7.7) are a result of the bending of light? The argument is by contradiction:

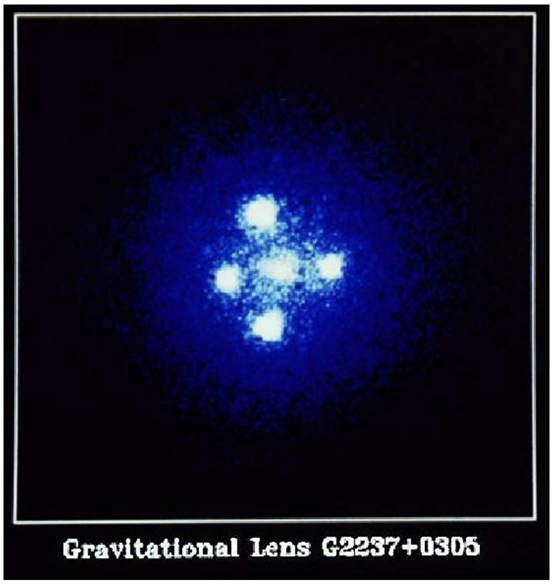


Figure 7.7: The Einstein Cross: four images of a quasar GR2237+0305 (a very distant, very bright object) appear around the central glow. The splitting of the central image is due to the gravitational lensing effect produced by a nearby galaxy. The central image is visible because the galaxy does not lie on a straight line from the quasar to Earth. The Einstein Cross is only visible from the southern hemisphere.

suppose they are not, that is suppose, that the images we see correspond to different stars. Using standard astronomical tools one can estimate the distance between these stars; it is found that they are separated by thousands of light years, yet it is observed that if one of the stars change, all the others exhibit the same change instantaneously! Being so far apart precludes the possibility of communication between them; the simplest explanation is the one provided by the bending of light. It is, of course, possible to ascribe these correlations as results of coincidences, but, since these correlations are observed in many images, one would have to invoke a "coincidence" for hundreds of observations in different parts of the universe.

The bending of light was one of the most dramatic predictions of the General Theory of Relativity, it was one of the first predictions that were verified as we will discuss below in Sect. 7.12.

7.4 Clocks in a gravitational force.

When comparing a clock under the influence of gravitational forces with one very far from such influences it is found that the first clock is slow compared to the second. To see this consider the same clock we used in the Special Theory of Relativity. For this experiment, however, imagine that the clock is being accelerated upward, being pulled by a crane. The clock gives off a short light pulse which moves towards the mirror at the top of the box, at the same time the mirror recedes from the pulse with even increasing speed (since the box accelerates). Still the pulse eventually gets to the mirror where it is reflected, now it travels downward where the floor of the box is moving up also with ever increasing velocity (see Fig. 7.8).

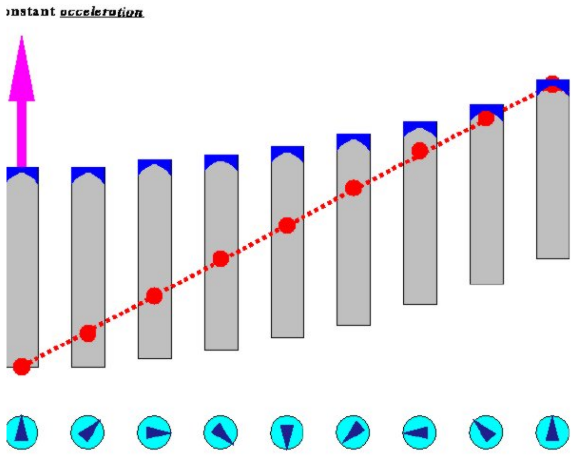


Figure 7.8: An accelerated clock. The circle denotes a pulse of light which at the initial is sent from a source; after a time it reaches the top of the the box and is reflected. The time it takes to do the trip is *longer* than for a clock at rest.

On the trip up the distance covered by light is larger than the height of the box at rest, on the trip down the distance is smaller. A calculation shows that the whole distance covered in the trip by the pulse is larger than twice the height of the box, which is the distance covered by a light pulse when the clock is at rest.

Since light always travels at the same speed, it follows that the time it takes for the pulse to go the round trip is longer when accelerating than when at rest: *clocks slow down whenever gravitational forces are present.*

This has an amazing consequence: imagine a laser on the surface of a very massive and compact planet (so that the gravitational field is very strong). An experimenter on the planet times the interval between two crests of the laser light waves and gets, say, a millionth of a second. His clock , however, is slow with respect to the clock of an observer far away in deep space, this observer will find that the time between two crests is larger. This implies that the frequency of the laser is larger on the planet than in deep space: *light leaving a region where gravity is strong reddens*. This is called the *gravitational red-shift* (see Fig. 7.9).

Light leaving a region where gravity is strong reddens

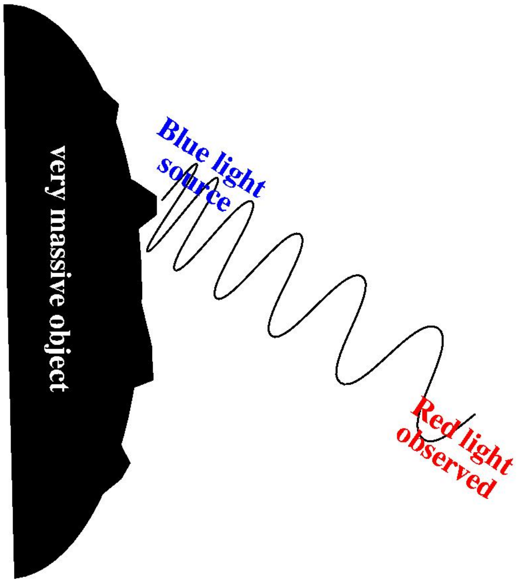


Figure 7.9: The gravitational redshidft. Since clocks slow down in a strong gravitational field then light, whose oscillations can be used as clocks, will be shifter towards the red as it leaves a region where gravity is strong.

As for time dilation, the slowing down of clocks in the presence of gravitational forces affects *all* clocks, including biological ones. A twin trveling to a region where gravity is very strong will come back a younger than the twin left in a rocket in empty space. This is an effect on top of the one produced by time dilation due to the motion of the clocks. The gravitational forces required for a sizable effect, however, are enormous. So the twin will return younger...provided she survives.

7.5 Black holes

So gravity pulls on light just as on rocks. We also know that we can put rocks in orbit, can we put light in orbit? Yes! but we need a very heavy object whose radius is very small, for example, we need something as heavy as the sun but squashed to a radius of less than about 3km. Given such an object, light moving towards it in the right direction will, if it comes close enough land in an orbit around it. If you place yourself in the path of light as it orbits the object, you'd be able to see your back.

But we can go farther and imagine an object so massive and compact that if we turn on a laser beam on its surface gravity's pull will bend it back towards the surface. Think what this means: since no light can leave this object it will appear perfectly black, this is a *black hole.* An object which comes sufficiently close to a black hole will also disappear into it (since nothing moves faster than light if an object traps light it will also trap everything else).

The effect of a black holes, like all gravitational effects, decreases with distance. This means that there will be a "boundary" surrounding the black hole such that anything crossing it will be unable to leave the region near the black hole; this boundary is called the *black-hole horizon* see Fig. 7.10 Anything crossing the horizon is permanently trapped. Black holes are prefect roach motels: once you check in (by crossing the horizon), you never check out.

The distance from the black hole to the horizon is determined by the mass of the black hole: the larger the mass the mode distant is the horizon from the center. For a black hole with the same mass as out sun the horizon is about 3 km from the center; for black holes with a billion solar masses (yes there *are* such things) this is increased to 3×10^9 km, about the distance from the sun to Uranus. For very massive black holes the horizon is so far away from the center that an observer crossing it might not realize what has just happened, only later, when all efforts to leave the area prove futile, the dreadful realization of what happened will set in.

Imagine a brave (dumb?) astronaut who decides to through the horizon

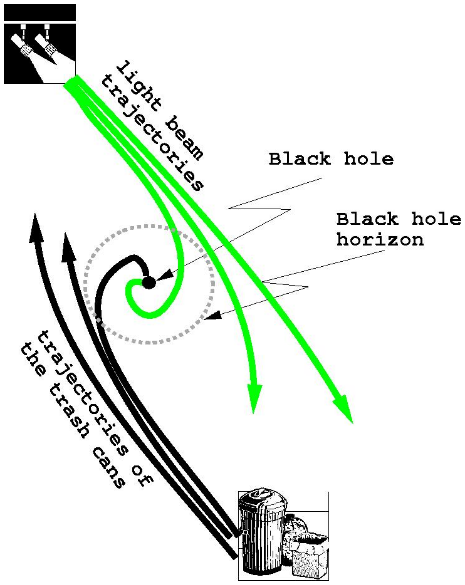


Figure 7.10: Illustration of the horizon surrounding the black hole. The black holes is represented by the small heavy dot, the light rays or particle trajectories which cross the dotted line cannot cross it again.

and into the nearest black hole and let us follow his observations. The first effects that becomes noticeable as he approaches the event horizon is that his clock ticks slower and slower with respect to the clocks on his spaceship very far from the black hole (see Sect. 7.4) to the point that it will take infinite spaceship time for him to cross the horizon. In contrast it will take a finite amount of astronaut time to cross the horizon, an extreme case of the relativity of time.

As the astronaut approaches the horizon the light he emits will be more and more shifted towards the red (see Sect. 7.4) eventually reaching the infrared, then microwaves, then radio, etc. In order to see him the spaceship will eventually have to detect first infrared light, then radio waves, then microwaves, etc.

After crossing the horizon the astronaut stays inside. Even though the crossing of the horizon might not be a traumatic experience the same cannot be said for his ultimate fate. Suppose he decides to fall feet first, then, when sufficiently close to the black hole, the gravitational pull on his feet will be much larger than that on his head and he will be literally ripped to pieces.

So far black holes appear an unfalsifiable conclusion of the General Theory of Relativity: their properties are such that no radiation comes out of them so they cannot be detected from a distance, and if you should decide to go, you cannot come back to tell your pals whether it really was a black hole or whether you died in a freak accident. Doesn't this contradict the basic requirement that a scientific theory be falsifiable (Sect. **??**)?

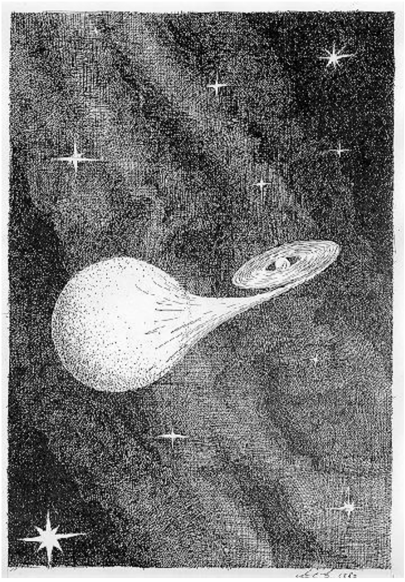


Figure 7.11: Artist's version of a black hole accreting matter from a companion star. The Star is on the left of the picture and is significantly deformed by the gravitational pull of the black hole; the object on the right represents the matter which surrounds the black hole and which is being sucked into it. The black hole is too small to be seen on the scale of this picture

Well, no, General Theory of Relativity even in this one of its most extreme predictions *is* falsifiable. The saving circumstance is provided by the matter surrounding the black hole. All such stuff is continuously being dragged into the hole (see Fig. 7.11) and devoured, but in the process it gets extremely hot and radiates light, ultraviolet radiation and X rays. Moreover, this cosmic Maelstrom is so chaotic that the radiation changes very rapidly, sometimes very intense, sometimes much weaker, and these changes come very rapidly (see Fig. 7.12). From this changes one can estimate the size of the object generating the radiation.

On the other hand astronomers can see the gravitational effects on nearby stars of whatever is making the radiation. And from these effects they

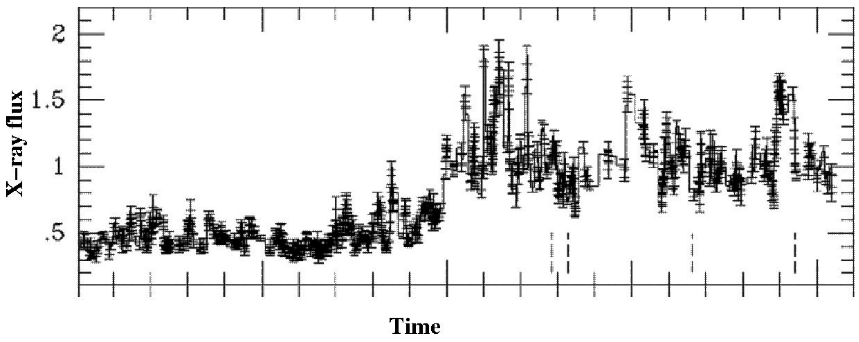


Figure 7.12: X-ray emission from a black hole candidate (Cygnus X1)

can estimate the mass of the beast. Knowing then the size, the manner in which matter radiates when it comes near, and the mass one can compare this to the predictions of General Theory of Relativity and decide whether this is a black hole or not. The best candidate for a black hole found in this way is called Cygnus X1 (the first observed X ray source in the constellation Cygnus, the swan).

All the ways we have of detecting black holes depend on the manner in which they affect the matter surrounding them. The most striking example is provided by some observation of very distant X-ray sources which are known to be relatively compact (galaxy size) and very far away. Then the very fact that we can see them implies that they are extremely bright objects, so bright that we know of only one source that can fuel them: the radiation given off by matter while being swallowed by a black hole [5]. So the picture we have of these objects, generically called active galactic nuclei, is that of a supermassive (a billion solar masses or so) black hole assimilating many stars per second, and in disappearing these stars give off the energy that announces their demise.

All this from the (apparently) innocent principle of equivalence.

7.6 Gravitation and energy

Consider a beam of sunlight falling on your skin; after a while your skin warms and, eventually, will burn: light carries energy (which is absorbed by your skin thus increasing its temperature). Recall also that a body with

[5] This is much more efficient than nuclear power which would be incapable of driving such bright sources.

mass m, by its very existence, carries and energy mc^2 (Sec. **??**). There is no way, however, in which we can associate a mass with light; for example, we can always change the speed of a mass (even if only a little bit), but this cannot be done with light.

The force of gravity affects both light and all material bodies; since both carry energy, but only the bodies carry mass, it follows that gravity *will affect anything carrying energy*. This conclusion lies at the root of the construction of Einstein's equations which describe gravity.

Gravity will affect anything carrying energy

Note that this conclusion has some rather strange consequences. Consider for example a satellite in orbit around the Earth, when the Sun shines on it it will increase its energy (it warms up), and gravity's pull with it. When the satellite is in darkness it will radiate heat, lose energy and the force of gravity on it will decrease [6].

Again let me emphasize that this argument is *not* intended to imply that light carries mass, but that gravity will affect anything that carries energy.

7.7 Space and time.

When considering the Special Theory of Relativity we concluded that the state of motion of an observer with respect to, say, a laboratory, determines the rate at which his/her clocks tick with respect to the lab's clocks (see Sect.**??**). Thus, in this sense, time and space mingle: the position of the observer (with respect to the lab's measuring devices) determines, as time evolves, his/her state of motion, and this in turn determines the rate at which his/her clocks tick with respect to the lab's.

Now consider what happens to objects moving under the influence of a gravitational force: if initially the objects set out at the same spot with the same speed they will follow the same path (as required by the principle of equivalence). So what!? To see what conclusions can be obtain let me draw a parallel, using another murder mystery.

Suppose there is a closed room and a line of people waiting to go in. The first person goes in and precisely two minutes afterward, is expelled through a back door, dead; it is determined that he died of a blow to the head. The police concedes that the room is worth investigating, but procrastinates, alleging that the person was probably careless and his death was accidental. Soon after, however, a second victim enters the room with precisely the same results, she also dies of an identical blow to the head; the police claims an astounding coincidence: two accidental deaths. This goes on for many

[6] Needless to say this is a very small effect, of the order of one part in a trillion.

hours, each time the victim dies of the same thing irrespective of his/her age, occupation, habits, color, political persuasion or taste in Pepsi vs. Coke; animals suffer the same fate, being insects of whales. If a rock is sent flying in, it comes out with a dent of the same characteristics as the ones suffered by the people and animals.

The police finally shrewdly concludes that there is something in the room that is killing people, they go in and... But the result is not important, what is important for this course is the following. We have a room containing something which inflicts a certain kind of blow to everything going through the room, I can then say that this inflicting of blows is a property of the room.

Consider now a region of empty space relatively near some stars. Assume that the only force felt in this region is the gravitational pull of these stars, hence all objects, people, animals, etc. going into this region will accelerate in precisely the same way. Then I can state that the region in space has a property which generates this acceleration[7].

Remember however that the region considered was in empty space (it only contains the objects we send into it), yet some property of this region determines the motion of anything that goes through it; moreover this property is a result of the gravitational pull of nearby heavy objects. The conclusion is then that *gravity alters the properties of space*, we also saw that the rates of clocks are altered under the influence of a gravitational force, it follows that *gravity alters the properties of space and time.* Space and time is in fact very far from the unchanging arena envisaged by Newton, they are dynamical objects whose properties are affected by matter and energy. These changes or deformations of space and time in turn determine the subsequent motion of the bodies in space time: matter tells space-time how to curve and space-time tells matter how to move (Fig. 7.13).

Gravity alters the properties of space and time

Matter tells space-time how to curve and space-time tells matter how to move

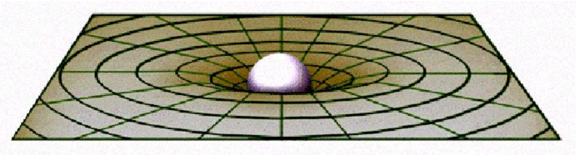


Figure 7.13: An illustration of the bending of space produced by a massive object

[7] I assume that the objects coming into this region are not too heavy, so that their gravitational forces can be ignored and that the start from the same spot with identical velocities.

7.8 Properties of space and time.

Up to here I've talked little of the implications of the Special Theory of Relativity on the General Theory of Relativity, I have only argued that in special relativity time and space are interconnected. In a separate discussion I argued that gravity alters space. In this section I will use what we know about length contraction together with the equivalence principle to determine how space is altered by gravity and to show that it is this deformation of space that is responsible for the gravitational force.

G+ Consider two identical disks one of which is made to rotate uniformly as in Fig. 7.14. In the non-rotating disk we select a small segment of its circumference of length ℓ_o. For the rotating disk this same segment will be measured to have length ℓ which is *smaller*, due to length contraction (Sec. **??**), than ℓ_o. Since the little segment we focused on is no different from any other (small) segment of the circumference, we can conclude that the circumference of the rotating disk is smaller than the circumference of the non-rotating disk.

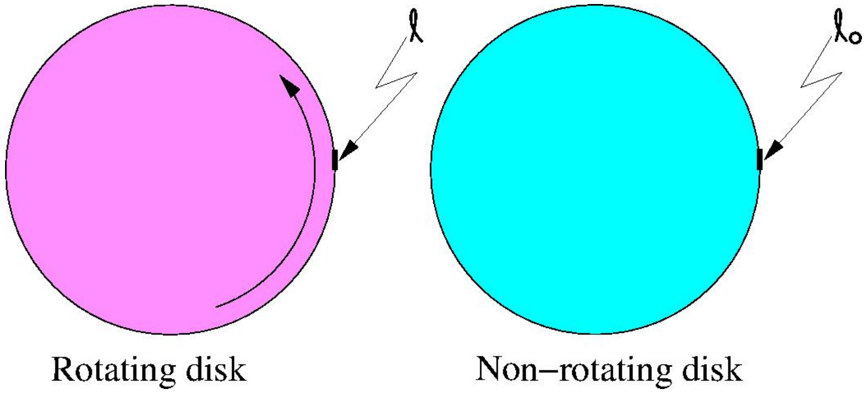


Figure 7.14: A rotating vs a non-rotating disk. The bit labeled ℓ in the rotating disk is shorter, due to length contraction to the corresponding bit ℓ_o in the non-rotating disk.

Consider now a radius of the disks. This is a length that is always perpendicular to the velocity of the disk and it is unaffected by the rotation. thus both disks will continue to have the same radius (see Sect. **??**).

So now we have one non-rotating disk whose circumference is related to the radius by the usual formula, *circumference* $= 2\pi\times$ *radius*, and a rotating disk whose circumference is *smaller* than this number!

How can this be? Isn't it true that the perimeter always equals 2π

radius? The answer to the last question is yes...provided you draw the circle on a flat sheet of paper. Suppose however that you are constrained to draw circles on a sphere, and that you are forced to measure distances only on the sphere. Then you find that the perimeter *measured along the sphere* is smaller than $2\pi\times$*radius* (with the radius also measured along the sphere, see Fig. 7.15).

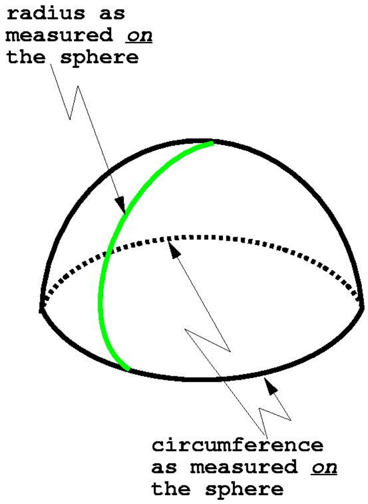

radius as measured *on* the sphere
circumference as measured *on* the sphere

Figure 7.15: The distance from the equator to the pole on a sphere is larger than the radius. For being constrained to move on the surface of the sphere this distance is what *they* would call the radius of their universe, thus for them the circumference is smaller than $2\pi\times$radius and they can conclude that they live in curved space.

We conclude that the uniformly rotating disk behaves as a (piece of a) sphere due to length contraction. So much for the effects of special relativity.

Now let us go back to the principle of equivalence. One of its consequences is that, by doing experiments in a small region one cannot distinguish between a gravitational force and an accelerated system. So if we attach a small laboratory of length ℓ_0 (at rest) to the small section of the perimeter, experiments done there will not be able to tell whether the lab. is in a rotating disk or experiences a gravitational force (remember that a rotating object is changing its velocity – in direction – and it is therefore accelerating!).

Putting together the above two arguments we get

Gravitation curves space and time.

Gravitation curves space and time

Conversely curved space and time generate effects which are equivalent to gravitational effects. In order to visualize this imagine a world where all things can only move on the surface of a sphere. Consider two beings labeled **A** and **B** as in Fig. 7.16, which are fated live on the surface of this sphere. On a bright morning they both start from the equator moving in a direction perpendicular to it (that is, they don't meander about but follow a line perpendicular to the equator).

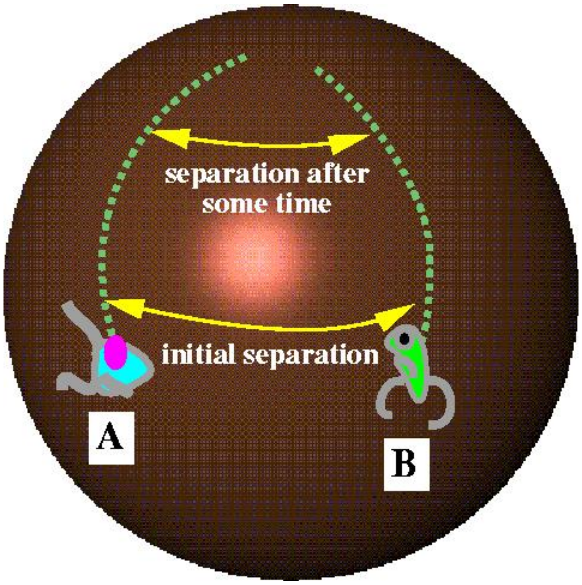


Figure 7.16: Two beings moving on a sphere are bound to come closer just as they would under the effects of gravity

As time goes on the two beings will come closer and closer. This effect is similar to the experiment done with two apples falling towards the moon (Fig. 7.4): an observer falling with them will find their distance decreases as time progresses; sentient apples would find that they come closer as time goes on.

So we have two descriptions of the same effect: on the one hand gravitational forces make the apples approach each other; on the other hand the fact that a sphere is curved makes the two beings approach each other; mathematically both effects are, in fact, identical. In view of this the conclusion that gravity curves space might not be so peculiar after all; moreover, in this picture the equivalence principle is very natural: bodies move the way they do due to the way in which space is curved and so the motion is independent of their characteristics [8], in particular the mass of the body does not affect its motion.

Bodies move the way they do due to the way in which space is curved and so the motion is independent of their characteristics

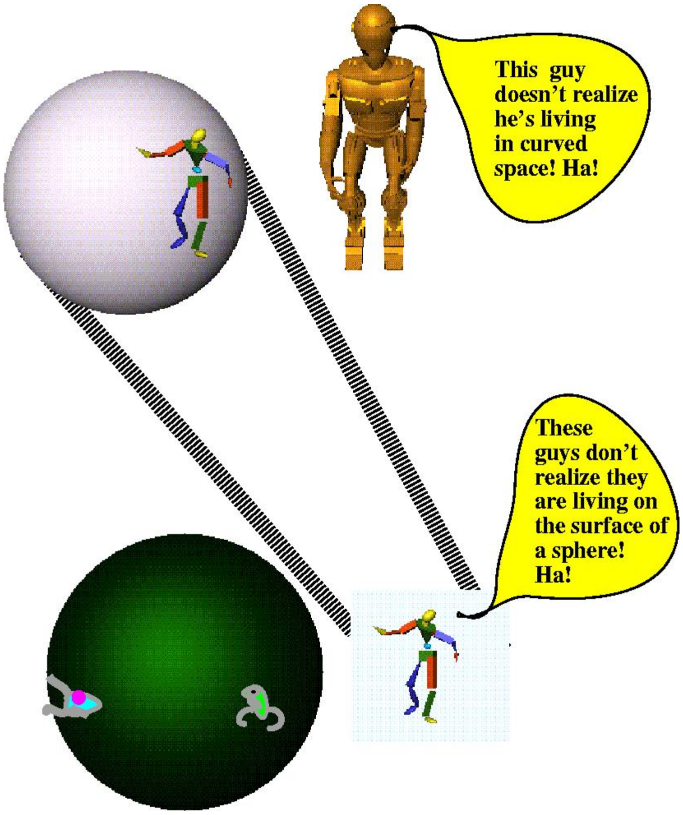

This guy doesn't realize he's living in curved space! Ha!
These guys don't realize they are living on the surface of a sphere! Ha!

Figure 7.17: Just as bugs fated to live on the surface of a sphere might find it peculiar to learn their world is curved, so we might find it hard to realize that our space is also curved.

Now the big step is to accept that the same thing that happened to the above beings is happening to us all the time. So how come we don't see that the space around us is really curved? The answer is gotten by going back to the beings **A** and **B**: they cannot "look out" away from the sphere where they live, they have no perception of the perpendicular dimension to this sphere, and so they cannot "see it from outside" and realize it is curved. The same thing happens to us, we are inside space, in order to see it curved we would have to imagine our space in a larger space of more dimensions

[8] I am assuming here that the moving things are not massive enough to noticeably curve space on their own.

and *then* we could see that space is curved; Fig. 7.17 gives a cartoon version of this.

7.9 Curvature

When considering the beings living on a sphere it is easy for us to differentiate between the sphere and some plane surface: we actually see the sphere being curved. But when it comes to us, and our curved space, we cannot see it since this would entail our standing outside space and looking down on it. Can we then determine whether space is curved by doing measurements inside it?

To see that this can be done let's go back to the beings on the sphere. Suppose they make a triangle by the following procedure: they go form the equator to the north pole along a great circle (or meridian) of the sphere, at the north pole they turn 90^o to the right and go down another great circle until they get to the equator, then they make another 90^o turn to the right until they get to the starting point (see Fig. 7.18). They find that all three lines make 90^o angles with each other, so that the sum of the angles of this triangle is 270^o, knowing that angles in all flat triangles always add up to 180^o they conclude that the world they live on is not a flat one. Pythagoras' theorem only holds on flat surfaces

We can do the same thing: by measuring very carefully angles and distances we can determine whether a certain region of space is curved or not. In general the curvature is very slight and so the distances we need to cover to observe it are quite impractical (several light years), still there are some special cases where the curvature of space is observed: if space were flat light would travel in straight lines, but we observe that light does no such thing in regions where the gravitational forces are large; I will discuss this further when we get to the tests of the General Theory of Relativity in the following sections.

The curvature of space is real and is generated by the mass of the bodies in it. Correspondingly the curvature of space determines the trajectories of all bodies moving in it. The Einstein equations are the mathematical embodiment of this idea. Their solutions predict, given the initial positions and velocities of all bodies, their future relative positions and velocities. In the limit where the energies are not too large and when the velocities are significantly below c the predictions of Einstein's equations are indistinguishable from those obtained using Newton's theory. At large speeds and/or energies significant deviations occur, and Einstein's theory, not Newton's, describes

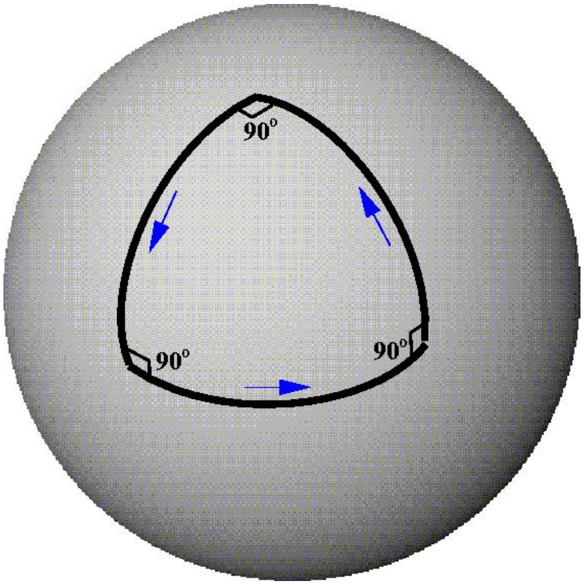


Figure 7.18: A path followed by a determined being living on the surface of a sphere; each turn is at right angles to the previous direction, the sum of the angles in this triangle is then 270^o indicating that the surface in which the bug lives is not flat.

the observations.

7.10 Waves

A classical way of picturing the manner in which heavy bodies curve space is to imagine a rubber sheet. When a small metal ball is made to roll on it it will go in a straight line at constant speed (neglecting friction). Now imagine that a heavy metal ball is placed in the middle of the sheet; because of its weight the sheet will be depressed in the middle (Fig 7.13). When a small ball is set rolling it will no longer follow a straight line, its path will be curved and, in fact, it will tend to circle the depression made by the heavy ball. The small ball can even be made to orbit the heavy one (it will eventually spiral in and hit the heavy ball, but that is due to friction, if the sheet is well oiled it takes a long time for it to happen). This toy then realizes what was said above: a heavy mass distorts space (just as the heavy ball distorts the rubber sheet). Any body moving through space experiences this distortion and reacts accordingly.

Now imagine what happens if we drop a ball in the middle of the sheet. It will send out ripples which spread out and gradually decrease in strength. Could something similar happen in real life? The answer is yes! When there is a rapid change in a system of heavy bodies a large amount of gravitational waves are produced. These waves are ripples in space which spread out form their source at the speed of light carrying energy away with them.

A computer simulation of a gravitational wave is given in Fig. 7.19. The big troughs denote regions where the wave is very intense, the black dot at the center denotes a black hole, the ring around the hole represents the black hole's horizon.

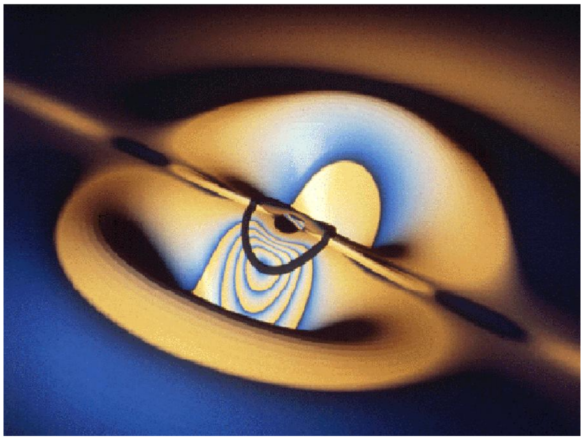


Figure 7.19: A computer simulation of a gravitational wave generated by a collision of two black holes, which have now merged and are represented by the heavy black dot in the middle.

Can we see gravitational waves? Not yet directly, but we have very strong indirect evidence of their effects. Several systems which according to the General Theory of Relativity ought to lose energy by giving off gravitational waves have been observed. The observations show that these systems lose energy, and the rate at which this happens coincides precisely with the predictions from the theory.

Observing gravitational waves directly requires very precise experiments. The reason is that, as one gets farther and farther away from the source these waves decrease in strength very rapidly. Still, if a relatively strong gravitational wave were to go by, say, a metal rod, its shape would be deformed by being stretched and lengthened periodically for a certain time. By accurately measuring the length of rods we can hope to detect these changes.

The technical problems, however, are enormous: the expected variation is of a fraction of the size of an atom! Nonetheless experiments are under way.

Gravitational waves are generated appreciably only in the most violent of cosmic events. During the last stages in the life of a star heavier than 3 solar masses, most of the stellar material collapses violently and inexorably to form a black hole (n the rubber sheet picture this corresponds to dropping a very small and very heavy object on the sheet). The corresponding deformation of space travels forth from this site site as a gravitational wave. High intensity gravitational waves are also produced during the collision of two black holes or any sufficiently massive compact objects.

7.11 Summary.

The conclusions to be drawn from all these arguments are,

- *All* frames of reference are equivalent, provided we are willing to include possible gravitational effects (in non-inertial or accelerated frames forces will appear which are indistinguishable from gravitational forces).
- Space-time is a dynamic object: matter curves it, and the way in which it is curved determines the motion of matter in it. Since all bodies are affected in the same way by the curvature of space and time the effects of gravity are independent of the nature of the body. Changes in the distribution of matter change space-time deforming it, and, in some instances, making it oscillate.

7.12 Tests of general relativity.

After Einstein first published the General Theory of Relativity there was a very strong drive to test its consequences; Einstein himself used his equations to explained a tiny discrepancy in the motion of Mercury. Yet he most dramatic effect was the shifting of the positions of the stars (see below). Since 1916 there have been many measurements which agree with the General Theory of Relativity to the available accuracy. Here I will concentrate on the “classical” tests of the thoery.

7.12.1 Precession of the perihelion of Mercury

A long-standing problem in the study of the Solar System was that the orbit of Mercury did not behave as required by Newton’s equations.

To understand what the problem is let me describe the way Mercury's orbit looks. As it orbits the Sun, this planet follows an ellipse...but only approximately: it is found that the point of closest approach of Mercury to the sun does not always occur at the same place but that it slowly moves around the sun (see Fig. 7.20). This rotation of the orbit is called a *precession*.

The precession of the orbit is not peculiar to Mercury, *all* the planetary orbits precess. In fact, Newton's theory predicts these effects, as being produced by the pull of the planets on one another. The question is whether Newton's predictions agree with the *amount* an orbit precesses; it is not enough to understand qualitatively what is the origin of an effect, such arguments must be backed by hard numbers to give them credence. The precession of the orbits of all planets *except* for Mercury's can, in fact, be understood using Newton;s equations. But Mercury seemed to be an exception.

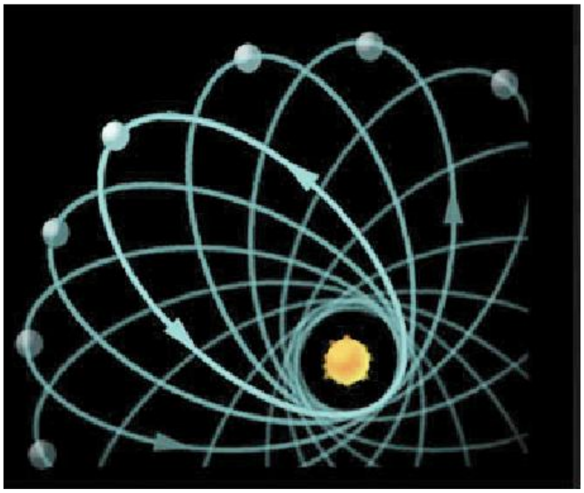


Figure 7.20: Artist's version of the precession of Mercury's orbit. Most of the effect is due to the pull from the other planets but there is a measurable effect due to the corrections to Newton's theory predicted by the General Theory of Relativity.

As seen from Earth the precession of Mercury's orbit is measured to be 5600 seconds of arc per century (one second of arc= 1/3600 degrees). Newton's equations, taking into account all the effects from the other planets (as well as a very slight deformation of the sun due to its rotation) and the fact that the Earth is not an inertial frame of reference, predicts a precession of 5557 seconds of arc per century. There is a discrepancy of 43 seconds of arc per century.

This discrepancy cannot be accounted for using Newton's formalism. Many *ad-hoc* fixes were devised (such as assuming there was a certain amount of dust between the Sun and Mercury) but none were consistent with other observations (for example, no evidence of dust was found when the region between Mercury and the Sun was carefully scrutinized). In contrast, Einstein was able to *predict*, without any adjustments whatsoever, that the orbit of Mercury should precess by an extra 43 seconds of arc per century should the General Theory of Relativity be correct.

7.12.2 Gravitational red-shift.

We saw in Sec. 7.4 that light leaving a region where the gravitational force is large will be shifted towards the red (its wavelength increases; see Figs. 7.21,7.9); similarly, light falling into a region where the gravitational pull is larger will be shifted towards the blue. This prediction was tested in Harvard by looking at light as it fell from a tower (an experiment requiring enormous precision since the changes in the gravitational force from the top to the bottom of a tower are minute) and the results agree with the predictions from the General Theory of Relativity.

The gravitational red-shift was also tested by looking at the light from a type of stars which are very very well-studied. The observations showed that the light received on Earth was slightly redder than expected and that the reddening is also in agreement with the predictions from the General Theory of Relativity.

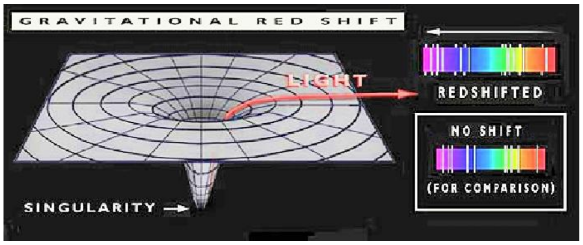


Figure 7.21: Illustration of the gravitational red-shift predicted by the General Theory of Relativity. A heavy object is denoted by a deformation of space represented by the funnel. As light leaves the vicinity of this object it is shifted towards the red: for a sufficiently compact and massive object a blue laser on the surface will be seen as red in outer space.

7.12.3 Light bending

If we imagine observing a beam of light in an accelerated elevator we will see that the light path is curved. By the equivalence principle the same must be true for light whenever gravitational forces are present. This was tested by carefully recording the position of stars near the rim of the sun during an eclipse (see Fig. 7.23) and then observing the same stars half a year later when there is no eclipse.

During the eclipse the observed starlight reaches us only after passing through a region where gravitational effects from the sun are very strong (that is why only stars near the rim are used), but the observation half a year later are done at a time where the gravitational effects of the sun on starlight is negligible.

It is found that the position of the stars are displaced when photographs of both situations are compared (see Fig. 7.23). The deviations are the same as the ones predicted by General Relativity. Eddington first observed this effect in 1919 during a solar eclipse. The early 20th century telegram (see Fig. **??**) announcing this observation for the frist time marks the change in our views about the structure of space time.

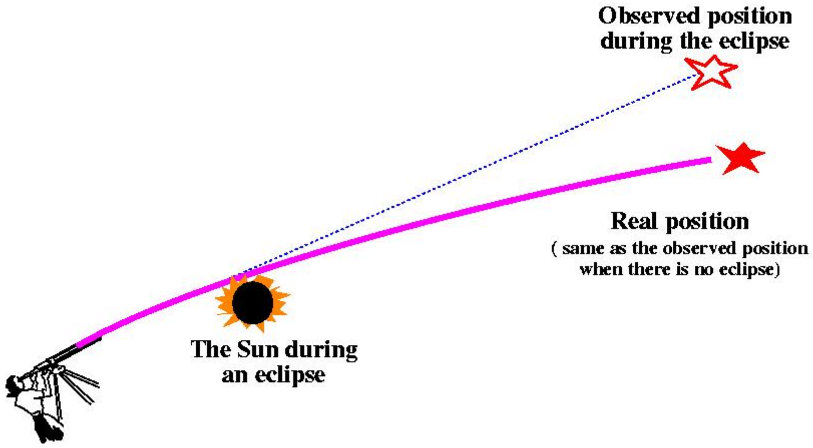
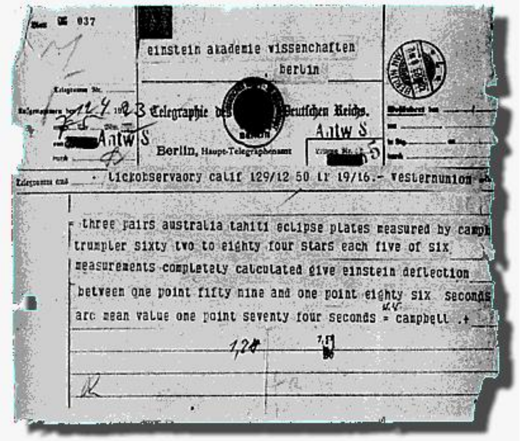


Figure 7.22: Illustration of the effects of the gravitational bending of light: during an eclipse the observed positions of the stars will be shifted away from the Sun.



037

einstein akademie vissenschaften
berbin

12.4.23 Telegraphie des Deutschen Reichs.

[redacted] Antw S

Berlin, Haupt-Telegraphenamt

Antw S

[redacted] 5

lickobservaory calif 129/12 50 lr 19/16.- vesternunion

three pairs australia tahiti eclipse plates measured by camph
trumpler sixty two to eighty four stars each five of six
measurements completely calculated give einstein deflection
between one point fifty nine and one point eighty six seconds
arc mean value one point seventy four seconds = campbell .+

1,24

Figure 7.23: Eddington's telegram to Einstein announcing the observation of the bengin of light by a gravitational force as predicted by the General Theory of Relativity.

7.12.4 The double pulsar

There are certain kind of stars which are called *pulsars* (see Sect. **??**). These are very compact objects (they have a diameter of about 10km but are several times heavier than the sun) which emit radio pulses at very regular intervals.

In the early 80's, Taylor and Hules (recent Nobel prize winners for this work) discovered a system where one pulsar circles another compact object. Because the pulsar pulses occur at very regular intervals, they can be used as a clock. Moreover there are several physical effects which can be used to determine the shape of the orbits of the pulsar and the compact object. It was found that these objects are slowly spiraling into each other, indicating that the system is losing energy in some way.

This system can also be studied using the General Theory of Relativity which predicts that the system should radiate gravitational waves carrying energy with them and producing the observed changes. These predictions are in perfect agreement with the observations. This is the first test of General Theory of Relativity using objects outside our solar system.

Chapter 8

The universe: size, origins, contents

8.1 Introduction

The general and special theories of relativity discussed in the previous chapters are the tools currently used in the investigation and description of the universe. Most of the objects in the universe are somewhat mundane: stars, planets, rocks and gas clouds. Yet in many respects the universe is far from being a placid and peaceful place. There are stars which explode with the energy of a billion suns, black holes with millions of times the mass of our sun which devour whole planetary systems, generating in one day as much energy as our galaxy puts out in two years. There are enormous dust cluds where shock waves trigger the birth of new stars. There are intense bursts of gamma rays whose origin is still uncertain.

These phenomena are not infrequent, but appear to be so due to the immense distances which separate stars and galaxies; for one of the most impressive properties of the universe is its size. The universe is so large that just measuring it is very difficult, and finding out the distance to various objects we observe can be a very complicated proposition.

In order to extract information about the universe a toolbox of methods has been devised through the years. I will first discuss the most important of these methods, and with these I will describe how measure the universe and discuss its evolution. We need to determine sizes and distances because, as we will see, they provide basic information about the history of the universe.

Most of the data we get from the universe comes in the form of light (by which I mean all sorts of electromagnetic radiation: from radio waves to gamma rays). It is quite remarkable that using only the light we can determine many properties of the objects we observe, such as, for example, their chemical composition and their velocity (with respect to us). In the first two sections below we consider the manner in which we can extract information from the light we receive.

But detecting light is not the only way to obtain information from the universe, we also detect high-energy protons and neutrons (forming the majority of cosmic rays). The information carried by these particles concerns either our local neighborhood, or else is less directly connected with the sources: isolated neutrons are not stable (they live about 10 minutes), so those arriving on Earth come from a relatively close neighborhood (this despite time dilation - Sect. **??**). Protons, on the other hand are very stable (the limit on their lifetime is more than 10^{32} years!), but they are charged; this means that they are affected by the magnetic fields of the planets and the galaxy, and so we cannot tell where they came from. Nonetheless the more energetic of these particles provide some information about the most violent processes in the universe.

In the future we will use yet other sources of information. Both gravitational wave detectors and neutrino telescopes will be operational within the next few years. Neutrinos are subatomic particles which are copiously produced in many nuclear reactions, hence most stars (including our Sun) are sources of neutrinos. These particles interact very very weakly, and because of this they are very hard to detect. On the other hand, the very fact that they interact so weakly means that they can travel through very hostile regions undisturbed. Neutrinos generated in the vicinity of a black hole horizon can leave their native land unaffected and carry back to Earth information about the environment in which they were born.

8.2 Light revisited

In this section I will describe two properties of the light we receive and the manner in which it can be used to extract information about its sources.

8.2.1 The inverse-square law

A source of light will look dimmer the farther it is. Similarly the farther away a star is the fainter it will look; using geometry we can determine just how a star dims with distance

Imagine constructing two spheres around a given star, one ten times farther from the star than the other (if the radius of the inner sphere is R, the radius of the outer sphere is $10R$). Now let us subdivide each sphere into little squares, 1 square foot in area, and assume than on the inner sphere I could fit one million such squares. Since the area of a sphere increases as the square of the radius, the second sphere will accommodate 100 times the number of squares on the first sphere, that is, 100 million squares (all 1 square foot in area). Now, since all the light from the star goes through both spheres, the amount of light going through one little square in the inner sphere must be spread out among 100 similar squares on the outer sphere. This implies that the brightness of the star drops by a factor of 100, when we go from the distance R to the distance $10R$ (see Fig. 8.1).

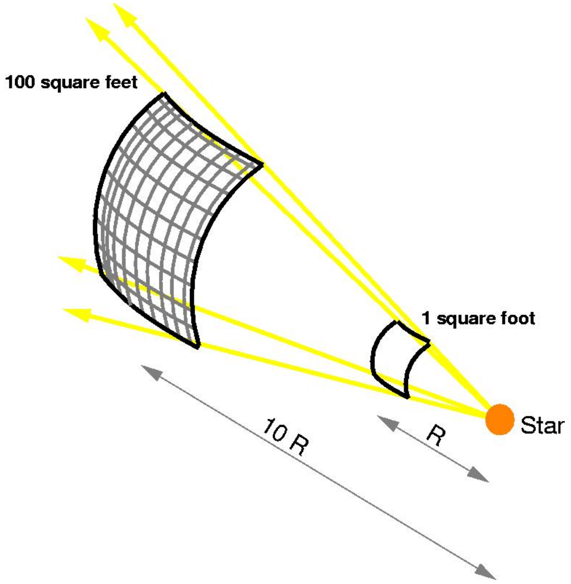


Figure 8.1: Illustration of the inverse-square law: all the light trough the 1 square-foot first area goes through the second one, which is 100 times larger, hence the light intensity *per square foot* is 100 times smaller in the second area. The intensity drops as $1/R^2$.

If we go to a distance of $20R$ the brightness would drop by a factor of 400, which is the square of 20, for $30R$ there would be a decrease by a factor of $900 = (30)^2$, etc. Thus we conclude that

The brightness drops as $1/\left(\mathit{distance}\right)^2$.

Light intensity drops as $1/\left(\text{distance}\right)^2$.

This fact will be used repeatedly below.

8.2.2 The Doppler effect

We have seen that light always travels at the same speed of about 300,000km/s; in particular light emitted by a sources in relative motion to an observer travels at this speed. Yet there is one effect on light which shows that its source is moving with respect to the observer: its color changes.

Imagine standing by the train tracks and listening to the train's horn. As the train approaches the pitch of the blast is higher and it becomes lower as the train recedes from you. This implies that the frequency of the sound waves changes depending on the velocity of the source with respect to you, as the train approaches the pitch is higher indicating a higher frequency and smaller wavelength, as the train recedes from you the pitch is lower corresponding to a smaller frequency and a correspondingly larger wavelength.

This fact, called the *Doppler effect*, is common to *all* waves, including light waves. Imagine a light bulb giving off pure yellow light; when it moves towards you the light that reaches you eye will be bluer, when the bulb moves away form you the light reaching your eye will be redder. If you have a source of light of a known (and pure) color, you can determine its velocity with respect to you by measuring the color you observe. Qualitatively, if one observes a redder color (longer wavelength than the one you know is being emitted) then the source is moving away from you, if bluer (shorter wavelength that the one you know is being emitted) the source is moving toward you (see Fig. 8.2).

If one observes a redder color (longer wavelength than the one you know is being emitted) then the source is moving away from you, if bluer (shorter wavelength that the one you know is being emitted) the source is moving toward you

The important point here is that knowing the frequency at the source and measuring the observed frequency one can deduce the velocity of the source [1] If the source is moving sufficiently fast towards you the yellow light will be received as, for example, X-rays; in this case, however, the source must move at 99.99995% of the speed of light. For most sources the shift in frequency is small.

8.2.3 Emission and absorption lines

When heated every element gives off light. When this light is decomposed using a prism it is found to be made up of a series of "lines", that is, the output from the prism is not a smooth spectrum of colors, but only a few of them show up. This set of colors is unique to each element and provides a unique fingerprint: if you know the color lines which make up a beam of light (and you find this out using a prism), you can determine which elements were heated up in order to produce this light.

[1] More precisely this is the velocity along the line of sight,

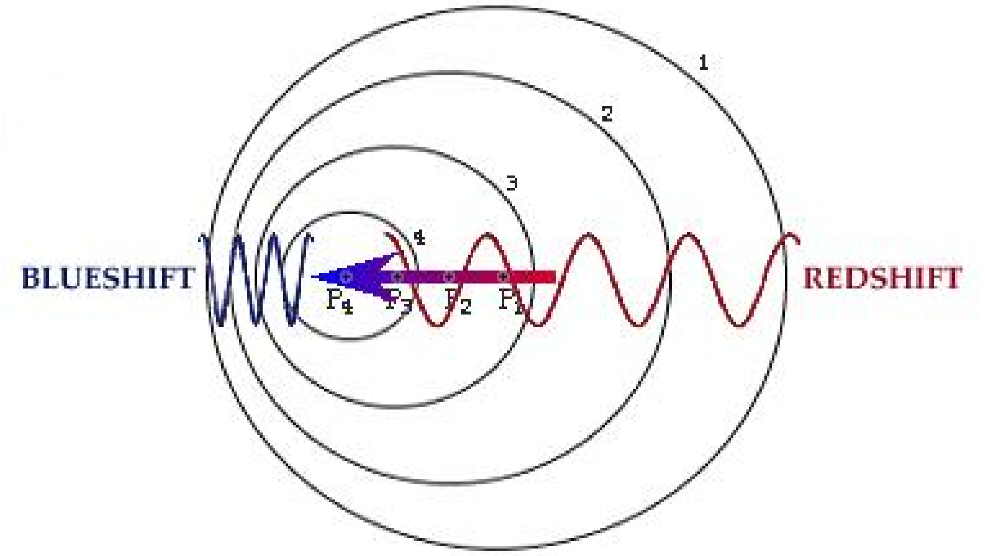


Figure 8.2: Diagram illustrating the Doppler effect. The source is moving to the left hence a receiver on the right will see a red-shifted light while a receiver on the left will see a blue-shifted one. .

Similarly, when you shine white light through a cold gas of a given element, the gas blocks some colors; when the "filtered" light is decomposed using a prism the spectrum is not full but shows a series of black lines (corresponding to the colors blocked by the gas); see Fig. 8.3. For a given element the colors blocked when cold are exactly the same as the ones emitted when hot.

The picture in Fig. 8.3 corresponds to a single element. For a realistic situation the decomposed light can be very complex indeed, containing emission and absorption lines of very many elements. An example is given in Fig. 8.4.

Each element presents a unique set of lines

After the discovery of emission and absorption lines scientist came to rely heavily on the fact that each element presents a unique set of lines: it is its inimitable signature. In fact, when observing the lines from the solar light, it was found that some, which are very noticeable, did not correspond to any known element. Using this observation it was then predicted that a new element existed whose absorption lines corresponded to the ones observed in sunlight. This element was later isolated on Earth, it is called Helium (from *helios*: sun).

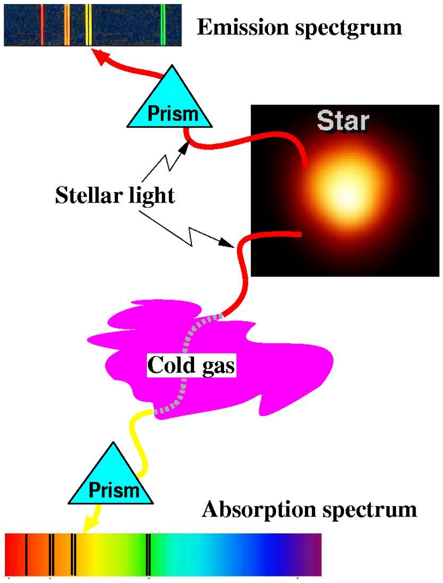


Figure 8.3: Diagram illustrating emission and absorption lines: when light given off by hot gas is decomposed using a prism it is shown to be made up of colored lines (emission lines). When white light shines trough a cold gas the resulting light , when decomposed is shown to have dark lines (absorption lines). The emission and absorption lines for the same element match.

In following this line of argument one has to be very careful that the lines are not produced by *any* other element. This is complicated by the fact that some lines are observable only under extreme circumstances and one has to take them into consideration as well. For example, after the success of the discovery of Helium, *another* set of lines (not so prominent) was isolated and associated with yet another element, "coronium". It was later shown that the coronium lines were in fact iron lines, which are clearly observable only in the extreme conditions present in the sun (one can also see them in the laboratory, it's just hard to do so).

8.2.4 A happy marriage

When observing stellar light from various distant stars (decomposed using a prism) it was found that, just as for the sun, they presented lines. But, curiously, these lines corresponded to no known element! This may imply that each star carries a new set of elements, but the simplest hypothesis (which should be investigated first, see Sect. **??**) is that the mismatch between the

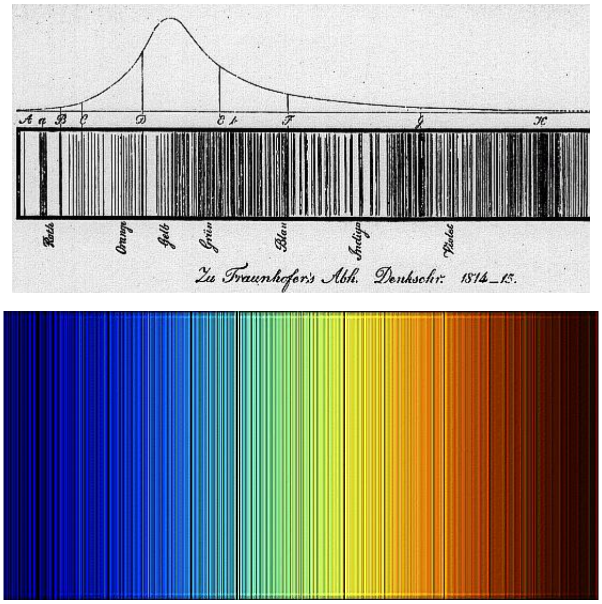


Figure 8.4: Solar light decomposed by a prism exhibiting the emission and absorption lines. At the top is one of the first of such measurements (1817); the curve above the lines denotes the intensity of the various colors, as expected it is largest in the yellow. The second figure is a modern photograph of the solar absorption lines.

laboratory and stellar lines is due to the Doppler effect which will shift the lines towards the red or blue according to the motion of the star (which is the source in this case) with respect to Earth. One can then use the shift in the observed stellar lines to determine the velocity at which the star is moving (with respect to us) and also the elements in it. In one fell swoop we determine the constitution and the speed of the stars using only the light we get from it.

Using spectral lines we can determine both the speed of the star and the elements in it

8.3 Cosmic distance ladder

Another important piece of information regarding objects in the universe is their distance to us. This is not an easy thing to measure since these objects are usually very far apart. I will measure distances in light years: one light year is the distance covered by light during one year, which is about 9.5 trillion kilometers, or about 6 trillion miles.

In order to understand why several steps are needed in measuring distances it is useful to consider a simple example. A student is in her room sitting at her desk and would like to find the distance to the window; she gets a ruler and laboriously measures this distance to be 3 feet. This I will call "the first rung in the student's distance ladder"

Her next task is to find the distance to a building which she can see through the window. This building is too far away for her to use her 12 inch ruler. What she does is to use sound: she notices that when she claps her hands outside her window there is an echo produced by the sound bouncing off the building in front of her. She has a good watch and so she can determine the time it takes for the sound to get from her window to the building and back. Now, if she can determine the speed of sound, she could use the formula $distance = speed \times time$ to get the distance. In order to measure the speed of sound she closes her window and times the echo from her desk to the window. Since we already knows the distance to the window (which she measured using her ruler) and she now knows the time it takes sound to go from her desk to the window and back she can determine the speed of sound. So, *using the first measurement* she determines the speed of sound and this allows her to measure things that are much farther away. In this way she has "constructed" the second rung in her distance ladder.

The same idea is used when measuring far away things in space: one finds a reliable method to determine the distances to near-by stars (the equivalent of using the ruler). Then one devises another method which requires a sort of calibration (the equivalent of determining the speed of sound); once this calibration is achieved the second method can be used to find distances to objects that are outside the range of the first method. Similarly a third, fourth, etc. methods are constructed, each based on the previous ones.

Step 1: distances up to 100 $\ell.y.$

For near-by stars their distance is measured by parallax: the star is observed in, say, December and then in June, and the direction of the star with respect to the sun is measured in both cases. Knowing these angles and the diameter of the orbit of the Earth around the sun, one can determine the distance to the star (see Fig. 8.5).

As we look at farther and farther stars the angles measured come closer and closer to 90^o. For stars more than 100 $\ell.y.$ from Earth one cannot distinguish the angles from right angles and the method fails.

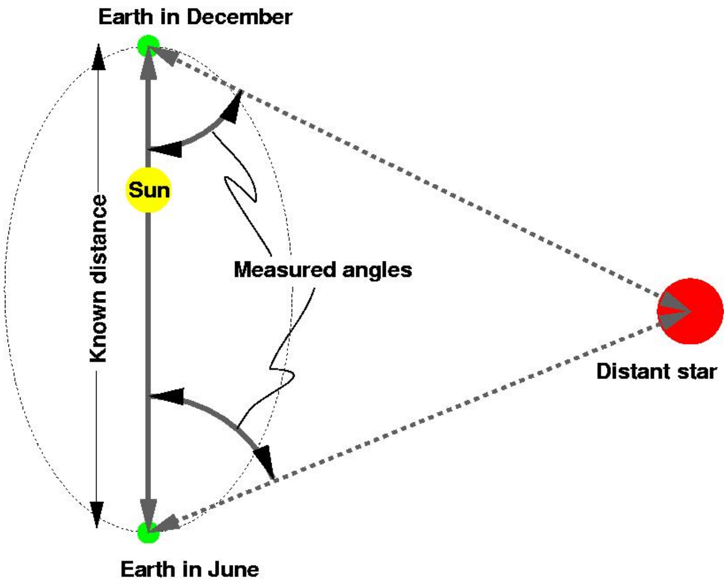


Figure 8.5: Knowing the size of Earth's orbit and measuring the angles of the light from the star at two points in the orbit, the distance to the star can be derived. The farther the star is, the smaller the angles.

Step 2: distances up to 300,000 $\ell.y.$

In the decade 1905-1915 Hertzprung and Russel observed a group of near-by stars whose distances they knew (using parallax). For each star they recorded its color and calculated its brightness as it would be measured at a distance of 1 $\ell.y.$ (using the $1/(\text{distance})^2$ law, see Sect. 8.2.1). Then they plotted this brightness versus the color; what they found is that most stars (90% of them) lie on a narrow band in this type of plot which they called the *main sequence* (see Fig. 8.6).

Suppose we now obtain the HR plot for stars which are far away, say on the other side of the galaxy, about 10^5 light years (10^{18} km). If we choose these stars such that they are not too far apart (there are good astronomical indicators for this) the distance from Earth to any such star will be more or less the same. It is found that, as for the near-by stars, 90% of these far stars will again fall on a main-sequence strip in the color vs. brightness plot.

On the other hand all these stars are dimmer than the near-by stars originally used by H&R; the decrease in brightness is due to the fact that brightness drops as the square of the distance (Sect 8.2.1). Comparing the two main sequences (for near and far stars) as in Fig. 8.7, we can extract

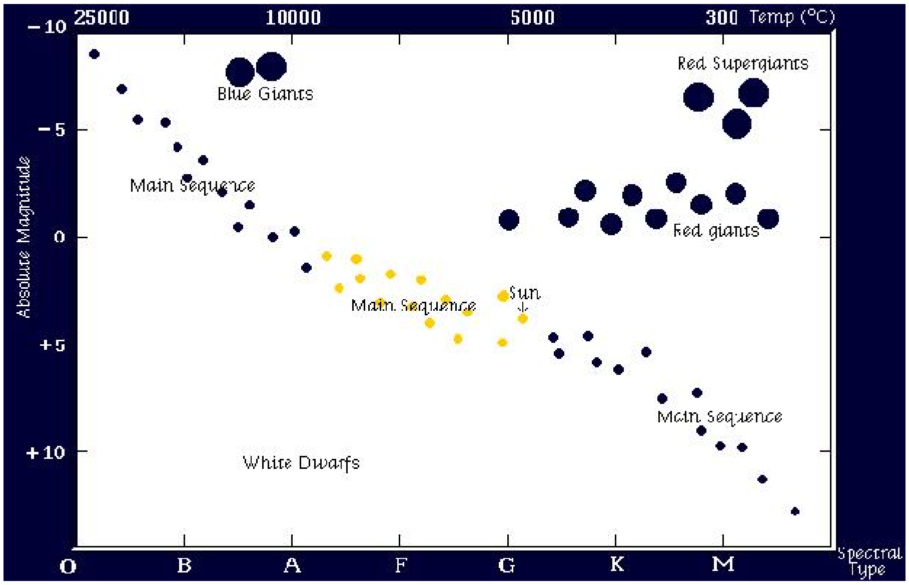


Figure 8.6: The Hertzprung-Russel diagram. The horizontal axis corresponds to the color of the star: blue to the left, red to the right. The vertical axis corresponds to the star's brightness (brighter stars are plotted higher). Though the diagram does not represent it, the groups labeled red supergiants, red giants, blue giants and white dwarfs, are but a small fraction of the whole stellar population, most stars are in the main sequence.

the distance to these far-away stars. This method can be used to determine distances up to 300,000 $\ell.y.$; for larger distances the main sequence stars are too dim to obtain a reliable estimate of their brightness.

Step 3: distances up to 13,000,000 $\ell.y.$

In 1912 Henrietta Swan Leavitt noted that 25 stars, called Cepheid stars [2] (their location in the HR diagram is given in Fig. 8.10), in the Magellanic cloud [3] (see Fig, 8.8) are variable, that is, they brighten and dim periodically. Many stars are variable, but the Cepheids are special because their period (the time they for them to brighten, dim and brighten again, see Fig. 8.9) is

[2] The name derives from the constellation in which they were first observed.

[3] This is a small galaxy (of only 10^8 stars) bound to the Milky Way.

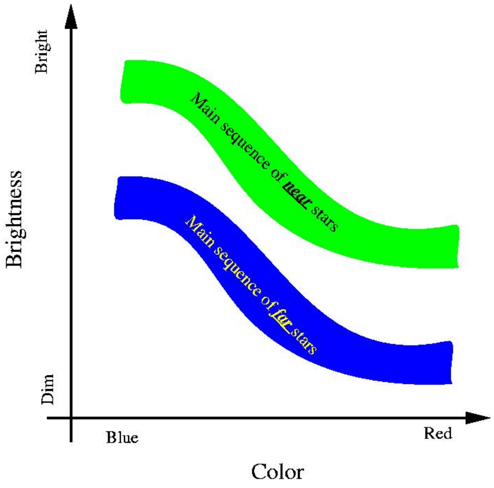


Figure 8.7: The Hertzprung-Russel diagram for the main sequence of near and far stars. The comparison is used to determine the distance to the far stars.

i) regular (that is, does not change with time), and

ii) a uniform function of their brightness (at a 1 light-year distance). That is, there is relation between the period and brightness such that once the period is known, the brightness can be inferred.

Leavitt was able to measure the period by just looking at the stars and timing the ups and downs in brightness,

But in order to obtain the brightness at the distance of one light year she needed to fist measure the maximum brightness on Earth and then, using the HR method, determine the distance from Earth to these stars (as it turns out, the Magellanic cloud is about 10^5 light years away from us).

What she obtained is that the brighter the Cepheid the longer its period, and that the relation between brightens and period was very simple: a straight line (Fig. 8.11). This means that the period and brightness are proportional to each other

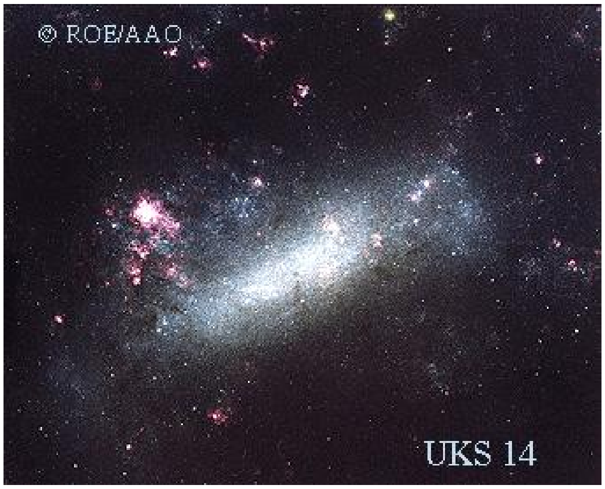


Figure 8.8: The Magellanic Cloud

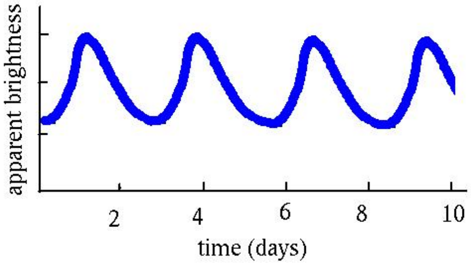


Figure 8.9: Illustration of the brightening and dimming of a variable star.

Measuring properties of the Cepheid variables. The color is no problem, you just observe the starlight through different color filters and observe the intensity; 'the' color of the star corresponds to the filter which lets pass the highest intensity light. The intensity at the distance of one light year is obtained by measuring the intensity on Earth and calculating the distance to the star, then one uses the fact that the intensity drops as the square of the distance. For example, suppose we observe a star which has intensity of 1 (in some units), and which we know is at a 10 $\ell.y.$ from Earth, then at a distance of 1 $\ell.y.$ (which is 10 times smaller) the intensity will be 100 times larger (the square of 10 is 100) and so the intensity at the distance of 1 $\ell.y.$ will be 100 (in the same units as before).

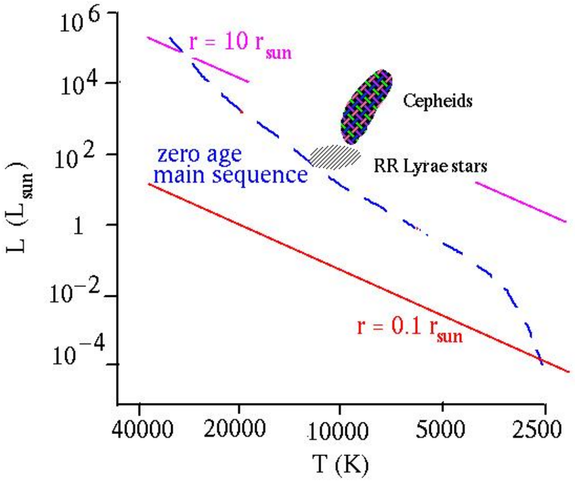


Figure 8.10: The location of the Cepheid variable stars in the Hertzprung-Russel plot.

These stars are quite distinct, reasonably abundant and very bright. One can identify them not only in our galaxy, but in many other galaxies as well.

If one requires the distance to a given galaxy one first locates the Cepheid variables in this galaxy. From these observations one determines the period of each of these stars. Leavitt's data states that a given period has a unique brightness associated to it. So form the period and Leavitt's plot we get the brightness at the distance of one light year. We can also measure the brightness on Earth. The brightness at the distance of one light year will be larger than the observed brightness due to the fact that this quantity drops like the square of the distance (Sect. 8.2.1). From these numbers one can extract the distance to the stars. This method works up to 13 million $\ell.y.$ when Earth-bound telescopes are used; for larger distances these stars become too dim to be observed.

Much more recently the Hubble telescope has used this same type of indicators to much farther distances (the Hubble is outside the Earth's atmosphere and can detect much fainter stars). Looking at a galaxy in the Virgo cluster (the galaxy is "called" M100), Wendy L. Freedman found (1994-5) that the Cepheid variables in this galaxy could be used to determine its distance; the result is 56 million $\ell.y.$.

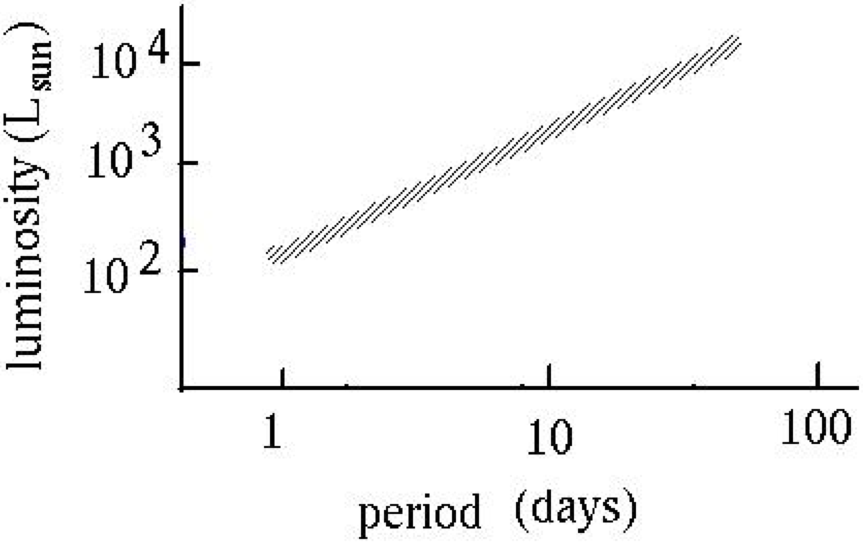


Figure 8.11: Relationship between the brightness and period of the Cepheid variable stars.

Young Cepheids. Recent observations of Cepheid variables in the galaxy M100 from the Hubble telescope have generated some puzzling questions. Using these observations (and the General Theory of Relativity) it is possible to determine the age of the universe: we measure both the distance and the velocity of these objects (with respect to us) and we can calculate the rate of expansion of the universe, from this we get the time it took to get to its present size. Curiously enough the age is in conflict with some other age determinations: some stars are older than the number obtained!

How can this be resolved? There are several possibilities. one of the most likely ones is that, since the Cepheid is observed by the Hubble telescope are very far away, the light we get was sent out when the stars were quite young. But it has not been shown that Leavitt's data is also valid for such teenage stars. It is quite possible that these stars have a different behavior and only settle into regular predictability only as they become middle-aged.

Step 4: distances up to 1,000,000,000 $\ell.y.$

For larger distances run of the mill stars are of no use: they are too dim. There are, however, some stars which at the end of their life blow themselves apart and, in doing so, become anomalously bright (out-shining a galaxy in many cases) for a brief period of time (less than a month); such an object is called a supernova (for more details see Sect. **??**). The unique characteristics and enormous brightness of a certain type of supernova can be used to determine distances beyond the reach of the previous methods.

There have been many measurements of the manner in which a supernova, whose distance to Earth is known (using one of the previous methods), increases its brightness and then dims into oblivion. There is one type (called type Ia) for which this brightening and dimming is very regular: when the maximum brightness at a distance of 1 $\ell.y.$ is calculated (using the known distance and the 1/distance^2 rule), it is found to be the same for all cases [4].

If the distance to a far away galaxy is required, one must first locate a type Ia supernova in it (which do occur regularly) and then measure its observed brightness. Comparing this result with the known maximum brightness (at a $1\ell.y.$ distance) achieved by all such supernovae one can determine the distance to the galaxy in question (again using the 1/distance^2 rule). Since supernovae are extremely bright this method is useful to very large distances, up to 10^9 $\ell.y.$.

Step 5: distances beyond 1,000,000,000 $\ell.y.$

For very far objects none of the above methods work. The reason is interesting: since we are looking at very distant objects their light has taken a very long time to reach us, so the light we get must have left the object a long time ago. Because of this the farther we look the earlier the images we get: looking far away is equivalent to looking back in time. When we look at the farthest obects we can see, what we get are images of their early stages of their development.

In addition, since the brightness drops as the square of the distance, these far objects must also be very bright. From this it follows that the most distant objects we see are necessarily very bright and very young.

In order to determine the distances with any degree of accuracy we need to know the brightness at a distance of 1 $\ell.y.$, but here we hit a stone wall: the only objects we see are much older than the ones we are interested in,

[4] In doing so astronomers must select type Ia supernovae that exhibit no abnormalities, else the measurements might be corrupted.

and we do not have a reliable theory of the way in which these things evolve, we have no way of calibrating our observations using any near-by objects.

It is here, in the observation of the universe at large, that the General Theory of Relativity must be used to measure distances. How this is done is described in the next section.

8.4 The relativistic universe

In everyday life there are many forces that strongly affect the world around us: friction, electric, magnetic, etc. But in the universe at large there is only one predominant force: gravity. It is gravity that determines the structure of the universe at large.

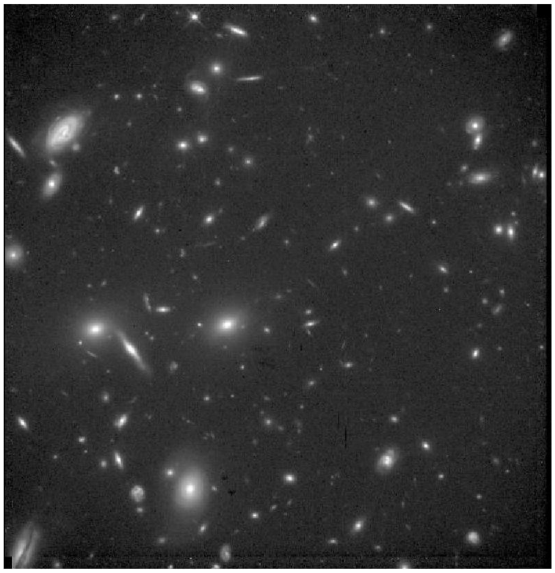


Figure 8.12: NASA Hubble Space Telescope image of the central portion of a remote cluster of galaxies (CL 0939+4713).

The (visible) universe is filled with galaxies (see Fig.8.12) each containing a billion suns (more or less) tightly bound by their mutual gravitational atraction. Because of this we can think of a galaxy as a solid object of a given mass (in the same way that when you look at the gravitational pull of the Earth on the Moon you don't have to worry about the fact that they are made of atoms; the stars are the "atoms" which make up galaxies).

Magnitudes. The typical galaxy like the Milky Way has most of its stars in a central bulge of 10^4 $\ell.y.$ diameter or less, where about 10^{11} sun-sized stars are concentrated. The pull of these stars on out Sun is 10^4 times stronger than that of the nearest sizable galaxy (Andromeda) which is at about 2×10^6 $\ell.y.$ away and also has about 10^{11} sun-sized stars.

In this simplified picture the visible matter in the universe (that which shines) is concentrated in a dusting of galaxies. In addition the universe can contain matter which does not shine, such as planet-sized objects, cold dust and, perhaps, other more exotic objects (see Sect. 8.5.1). The universe also contains electromagnetic radiation: for example, stars continuously give off light and heat (infrared radiation) which then disperses throughout the universe (this is why we can *see* them!). Finally the universe contains a significant amount of microwaves (see Sect. 8.4.2) and neutrinos, (see Sect. 8.5.1), both relics from a very early time.

The first person to look at the cosmos through the eyes of the General Theory of Relativity was Einstein himself. He took the above picture of a universe filled with matter and radiation he added two assumptions

- *Homogeneity:* on average the universe looks the same from every vantage point.
- *Isotropy:* on average the universe looks the same in every direction

These assumptions, though reasonable, still require justification; I will come back to them. With these preliminaries one can solve the equations of the General Theory of Relativity and find a description of the universe and the manner in which it evolves.

To Einstein's initial surprise there were no steady solutions: the universe according to the General Theory of Relativity *must* expand or contract. He compared this result with the best observational data of the time and found, to his dismay, that the observations strongly favored a steady universe. He then made what he called "the greatest scientific blunder of my life": he modified the equations of the General Theory of Relativity by adding a term that countered the expansion or contraction present in his initial solutions [5]. With this *ad hoc* modification he did find a steady universe and was (temporarily) satisfied.

[5] The modification amounts to the inclusion of a uniform cosmic pressure which balances the tendency to the universe to expand.

Not long afterwards Hubble published his now famous observations that demonstrated that our universe is, in fact, expanding; and the manner in which it expands agrees with the predictions of the solutions first obtained by Einstein. It was then that Einstein, to his satisfaction, dropped his modification of the equations. But this was not the end of this saga: the added term, like the genie from the bottle, refused to disappear, showing up in many models (recent observations suggest that it must be included in order to account for the observations). I will come back to this in Sect. 8.5.2.

What Hubble did was to measure the red-shift of a group of galaxies whose distances he knew (there were no blue-shifted galaxies, which means that these galaxies were receding from the Milky Way). Using the measured red-shift and the formulas for the Doppler effect, he found the speed at which they receded. Then he made a plot (called now a "Hubble plot") of velocity vs. distance and found that, *as predicted by the General Theory of Relativity* all points fall in a straight line (see Fig. 8.13); the slope of this line is called *Hubble's constant*. General Relativity then predicts that the distance d to an object is related to its velocity v (both measured with respect to the Earth) by

$$v = H_o \, d$$

General Relativity then predicts that the distance d to an object is related to its velocity v both measured with respect to the Earth by $v = H_o \, d$ which is called *Hubble's law* and H_o is Hubble's constant

which is called *Hubble's law* and H_o is Hubble's constant, its value is approximately

$$H_o = \frac{1}{1.5 \times 10^{10} \text{years}}.$$

It is the above relation between distance and velocity that is used to measure distances beyond 10^9 $\ell.y.$: the final step in the cosmic distance ladder. Needless to say astronomers have verified Hubble's law for distances below 10^9 $\ell.y.$ using supernovae (Sect. 8.3). In order to find the distance to the farthest objects in the universe one first obtains their redshift and, using Doppler's formulas, derives the velocity v of the object. The distance is then v/H_o.

8.4.1 The expanding universe

All of Hubble's (and subsequent) measurements indicate hat all galaxies are receding from the Milky way and its neighbors. One might think that we are being ostracized by the universe as a whole, that the Milky Way has become a cosmic pariah; but a little thought shows that this is not the case. According to General Relativity the universe is expanding, but this does

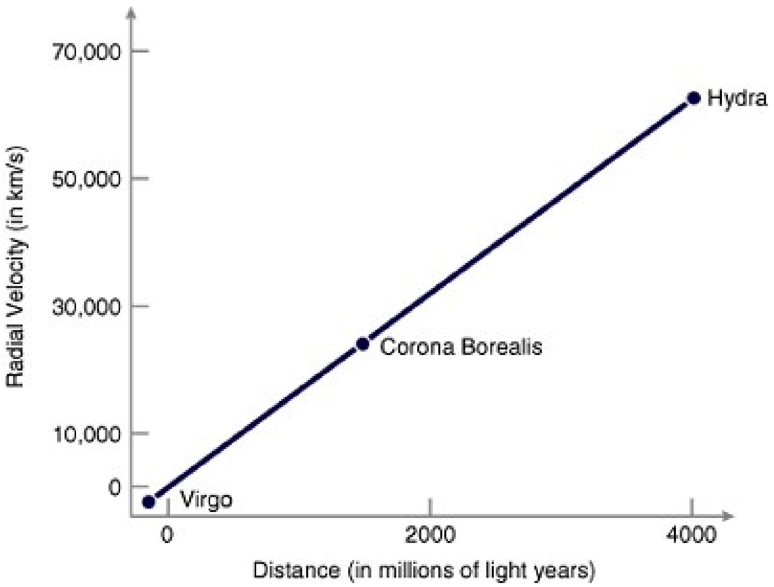


Figure 8.13: Illustration of Hubble's law.

not mean that the galaxies and such are flying out into space, it means that *space itself* is growing, and in so doing, it increases the separation between the galaxies. The classical example is to imagine a balloon with dots drawn on it; the balloon's latex represents space, the dots represent the galaxies. As the balloon is inflated (space grows) the distance between the dots (the galaxies) increases. An observer in any one dot would see the other dots receding from him/her (just as we see distant galaxies receding from us).

This universal expansion represents only the average motion of the galaxies, the motion of a given galaxy can present deviations from this average. For example, galaxies which are close together are bound by their mutual gravitational pull and this distorts the Hubble flow.

The General Theory of Relativity predicts that the universe is not static, and observations confirm this indicating that it is expanding. Thus the universe must have been smaller in the past, and, following this idea to its limit, must have been a point in its inception. Thus the universe began at a point, in the distant past and has been expanding ever since. The event marking this beginning is known (with a characteristic scientific flair for words) as the *Big Bang*.

Just after the Big Bang the universe contained an extremely hot and dense soup of matter and energy (which are equivalent in the sense of the

Special Theory of Relativity) under which conditions any kind of object would melt almost instantaeously into its components. Yet the universe expanded and cooled accordingly, and this cooling allowed for the formation of more and more complicated structures, ranging from atoms ($300,000$ years after the Big Bang) to Galaxies (10^9 years after the Big Bang) (see Fig. 8.16).

It must be remembered that the Big Bang represent the creation of the universe, *including* space and time. The Big Bang is *not* to be pictured as a big explosion somewhere out in space with galaxies being spewed out from the explosion region. Instead the picture provided by General Relativity is of the whole universe, including space, appearing at the Big Bang and expanding after that (like the balloon model described above). In this picture the Big Bang occurred *everywhere*.

The Big Bang occurred *everywhere*.

And now what?

The universe expanding, but what will become of it? There are three possible solutions to the equations of the General Theory of Relativity which represent homogeneous and isotropic universes: either it will continue its expansion forever, or it will eventually stop and re-contract or it will expand slowing down to a stop at infinite time. The contents of the universe (matter and radiation) determine which of these is realized in *our* universe. In all three cases the shape of space remains the same as the universe expands (or in the second case, as it expands and contracts).

The universe will continue its expansion forever, or it will eventually stop and re-contract or it will expand slowing down to a stop at infinite time.

That the shape of space is determined by the amounts of matter and energy in the universe is not surprising as it is matter and energy which determine the curvature of space (see Sect. **??**).

- Space in an eternally expanding or *open* universe is shaped like a 3-dimensional horse saddle. In this case the angles in a triangle add up to *less* than 180^o.
- Space in a *closed* universe which will eventually re-contract is shaped like a 3-dimensional sphere. In this case the angles in a triangle add up to *more* than 180^o.
- Space in a *flat* universe which expands slowing down to a stop at infinite time is shaped like a 3-dimensional plane. In this case the angles in a triangle add up to 180^o.

Space in an eternally expanding or *open* universe is shaped like a 3-dimensional horse saddle

Space in a *closed* universe which will eventually re-contract is shaped like a 3-dimensional sphere

Space in a *flat* universe which expands slowing down to a stop at infinite time is shaped like a 3-dimensional plane

These possibilities are illustrated in Fig. 8.14.

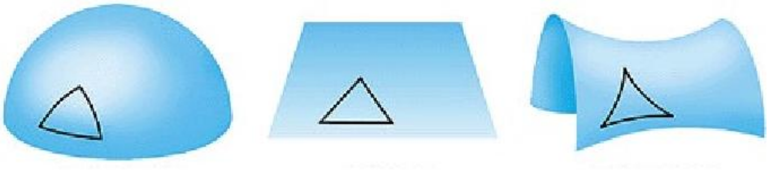


Figure 8.14: The three possible shapes of a homogeneous and isotropic universe: a closed universe (left), a flat universe (center) and open universe (right). See the text for an explanation.

These three possibilities give the *average* shape of space. Individual masses produce local bumps and troughs. This is similar to the way we talk about the Earth: we say it is a sphere, though we know it is full of bumps (for example, Himalayas) and troughs (the Dead Sea, for example).

Of these possibilities the one corresponding to our universe is determined by the amount of matter in the cosmos. If there is very little the initial thrust from the Big Bang will never be stopped, if however there is a large amount of matter, the mutual gravitational pull will be sufficient to break the expansion and eventually cause a re-contraction. Hence there is a critical amount of matter such that if our universe has more it will re-contract, if less it will expand forever (if it has precisely the critical amount it will expand forever slowing down to a stop at infinite time). These possibilities are illustrated in Fig. 8.15.

The obvious question is then: how much stuff is in the universe? And to that we can say: we don't know. If we count all the matter that shines (stars and such) we get a number very low compared to the critical value. But, is most of the matter shining? Could it not be that there is a lot of dust out there? The latest results suggest that the universe will expand forever, but at present its ultimate fate is unknown.

8.4.2 The Microwave Background Radiation

General Relativity not only provides a nice history of the universe, but it also points out viable measurements which can support its validity. The most important is the so-called *Microwave Background Radiation*.

When the universe began the density and temperature of the initial fireball was so high that all matter dissociated into its primary components. Note also that in this initial setting the force of gravity was enormous. As the expansion progressed the universe cooled and the initial fundamental

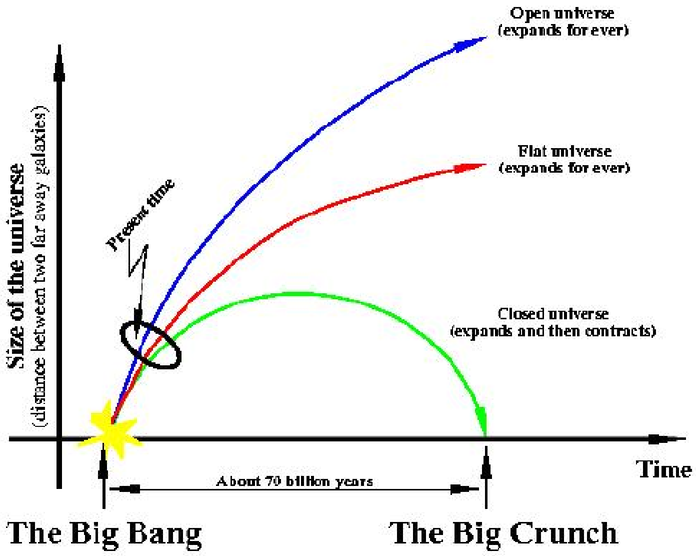


Figure 8.15: The universe might expand forever or will re-contract

constituents formed increasingly more complicated objects. This is so because when the temperature is very high everything is jiggling very fast and anything that can be dissociated will; as the temperature drops so does the jiggling and, eventually, composite structures can form and survive. Thus, if we had been able to film the contents of the universe as it cooled, and then run the film backwards we would first see atoms which are then broken apart into nuclei and electrons by the intense heat, then we would see the nuclei themselves decomposing into protons and neutrons, then the protons and neutrons decomposing into quarks [6]. The microwave background radiation is a messenger from this primordial soup.

To understand how why is this microwave radiation present and how it was generated I need to talk a bit about the way charged bodies interact with light. Remember now that light is described by the same equations that describe the physics of electric charges (Maxwell's equations), this suggests (and it is true) that light will interact with charged objects. In fact this is how your skin gets hot when exposed to the sun: your skin is composed of molecules which are made of atoms. Atoms in their turn are composed of a small heavy nucleus (with positive charge) surrounded by a cloud of

[6] There are many hypotheses about the way the universe looked at times before that of quark formation, but none has been accepted yet this is an area of active research.

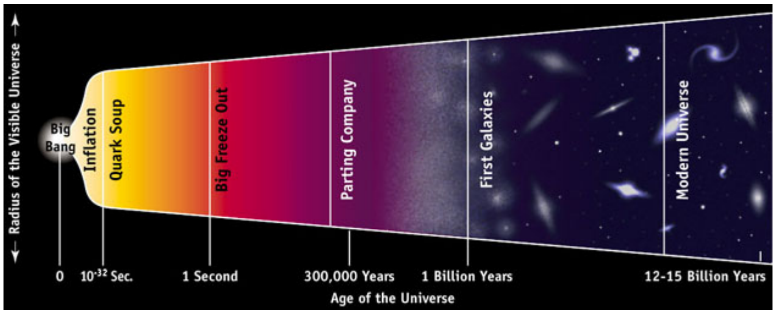


Figure 8.16: Abbreviated history of the universe according to the Big Bang model.

negatively charged light particles, the electrons. When light shines on your skin it is absorbed by the electrons which get agitated, and it is this agitation which you perceive as heat. This is not as efficient as it might be because the electrons are not free, they are inside atoms, so that *on average* the atoms are neutral. Much more light would be absorbed by a set of free electrons. This also works in the reverse: if you jiggle electrons sufficiently rapidly they will give off light, this is how a light-bulb works.

Suppose now that you have a box with perfectly reflecting walls and which are kept very hot. Into that box we introduce a bunch of electrons and nuclei and also light. Assume that the system is so hot that the electrons are not bound to the nuclei: as soon as they come close they are wrenched apart by the intense heat of the environment. So, on average, what you see is a bunch of charged particles and light running amok. In this case light is constantly being absorbed and emitted by the electrons and nuclei.

Now imagine that you cool the box by making it larger. Eventually things will get cold enough for the electrons to stay attached to the nuclei, the heat is not sufficiently high for them to be wrenched apart. At this point the rate at which light is absorbed and emitted drops rather suddenly for now the particles in the box are neutral (on average). From this point on light will just stream forth unimpeded (until it is reflected by a wall).

This is precisely what happens in the universe. After the big bang there came a point where electrons and nuclei were formed. They were immersed in intense electromagnetic radiation (light, X-rays, gamma rays, etc.). As time progressed and the expansion of the universe continued, the system became cooler (much as for the box when we increased its size). Eventually a point was reached where the universe was cool enough for atoms to form and from this moment on most of the radiation just streamed forth unimpeded.

This happened when the universe was a mere 300,000 years old.

So, can we see this relic of the ancient universe? The answer is yes! But before we look for it one thing must be kept in mind. The universe has been getting bigger and bigger and less and less dense. This implies that the average gravitational force is getting smaller with time. So the radiation, from the moment it no longer interacted with the newly formed atoms has been shifting from an environment where gravity's force is large to that where gravity is small and, using (again!) General Relativity, it must be red-shifted. In fact the *prediction* of General Relativity is that this radiation should be seen mostly as microwaves...and it *has* been seen. This prediction is not only of the existence of this relic radiation, but also how this radiation depends frequency . These predictions have been confirmed to great accuray (see Fig. 8.17). This ubiquitous sea of radiation that permeates the cosmos is called the *microwave background radiation*.

The microwave background radiation was created in approximately the same environment everywhere (remember that it came from an epoch in which everything was a very homogeneous hot mixture of nuclei and electrons) and because of this we expect it to look the same in every direction. This is precisely what happens, but, as it turns out, it is too much of a good thing: the microwave background radiation is the same everywhere to a precision of 0.1%, and understanding this presents problems, see Sect. 8.5.3.

But one can go even farther. Even though the microwave background radiation is very homogeneous, there *are* small deviations. These represent inhomogeneities in the universe at the time radiation and atoms stopped interacting strongly. These inhomogeneities provide a picture of the universe in its most tender infancy, see Fig. 8.18. As the universe expands and cools atoms will conglomerate into stars and stars into galaxies; the initial seeds for this process to start are these inhomogeneities. They correspond to regions where the matter was slightly mode dense than the average, and will, in the eons that follow, attract other matter to form the structures we see today.

It is very hard to explain the microwave background radiation by any theory other than the Big Bang. It represent one of its biggest successes.

8.4.3 Nucleosynthesis

The most abundant element in the universe is Hydrogen, the second most abundant element is Helium. A great success of the Big Bang theory is to be able to predict the relative amounts of these elements: after the universe

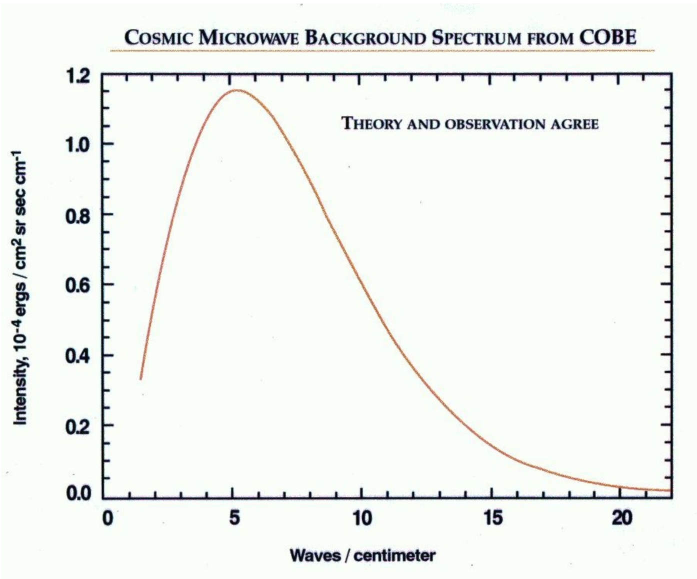


Figure 8.17: Radiation relics from the epoch shortly after the Big Bang. The horizontal axis corresponds to the frequency of the radiation, the vertical axis to the intensity. The measurements fall precisely on the curve.

cooled down sufficiently protons and neutrons were able, after a collision, to remain in the form of heavier atomic nuclei, in this manner Helium and Lithium were created, and also Deuterium (whose nucleus has one proton and one neutron). The universe was 1s old, its temperature was 10^{10o}K.

It was initially thought that *all* elements would be generated by the Big Bang, but this is not the case: even at the extreme temperatures available when Helium and Lithium nuclei were crated, this was not enough to smash two Helium nuclei to create something heavier, the creation of the remaining elements of the periodic table had to await the appearance of the first stars (see Sect. **??**). Deuterium and Lithium, while used up in stars through the nuclear reactions that make them shine (see Sect. **??**), are very rarely created by them . Whatever Deuterium and Lithium we see in nature was created about 15 billion years ago. Most of the Helium we observe (even though it is manufactured in stars) also came from that epoch.

The Big Bang theory predicts is the relative amounts of Helium and Lithium and Deuterium and Hydrogen. And the observations match the predictions; for example there are about 4 atoms of Hydrogen for each one of Helium. These same calcualations predict that there are 3 light neutrinos,

The Big Bang theory predicts is the relative amounts of Helium and Lithium and Deuterium and Hydrogen

The Big Bang theory predicts that there are 3 light neutrinos

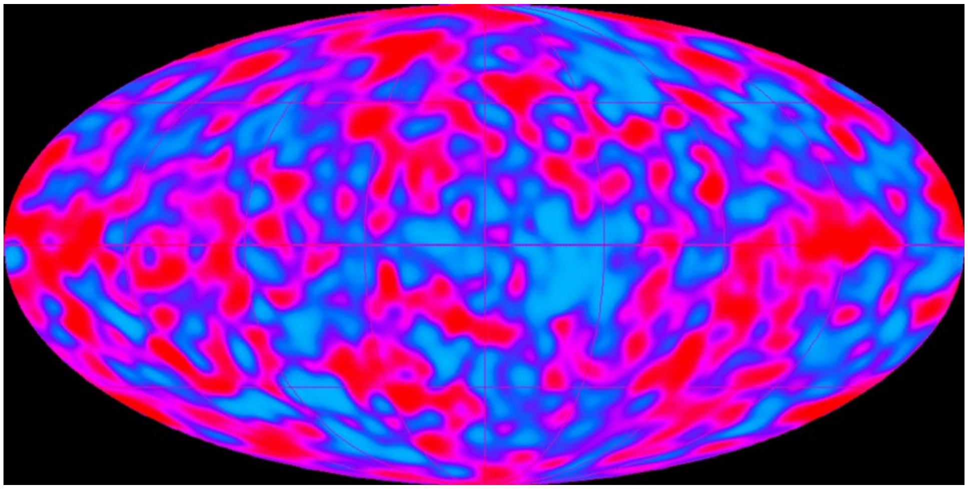


Figure 8.18: Inhomogeneities in the microwave background radiation. These give an idea of the way the universe looked shortly after the Big Bang.

again confirmed by observation.

Coupled with our understanding of stellar processes and evolution (Sects. **??** and **??**) we now understand the manner in which *all* elements in the periodic table were created. This is one of the most important predictions of modern cosmology.

We understand the manner in which *all* elements in the periodic table were created

8.5 At the cutting edge

Up to now all the results presented are well accepted and verified. There is little doubt that the General Theory of Relativity provides an excellent description of the universe at large, nor that the universe is currently expanding. Yet there are several puzzling results...

8.5.1 Dark matter

When considering the universe we observe only what we can see. Nonetheless there are strong indications that there is something more. Suppose you look at how stars in the outskirts of a galaxy move. Since gravity decreases with distance one would expect that the stars would slow down as the distance to the galaxy center increases, but this is *not* what is seen: the speed of these outlying stars appears to be constant (see Fig. 8.19). This is explained by assuming that the galaxy is in fact surrounded by a mass of matter which emits no or very little light, the so-called *dark matter*. In fact, calculations show that if this hypothesis is correct, this kind of matter is the main ingredient of galaxies, and perhaps the whole universe; an illustration of the "dark matter halo" surrounding a typical galaxy is given in Fig. 8.20

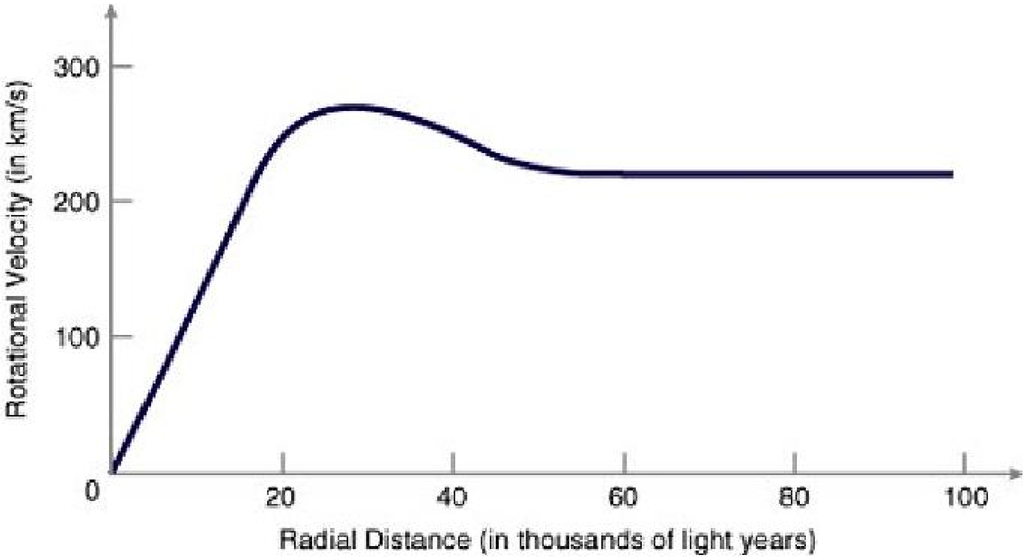


Figure 8.19: Rotation curve for stars in the Andromeda galaxy. The velocity becomes constant far away from the center suggesting the presence of dark matter.

What is this dark matter? No one knows! Is it perhaps a very large number of rocks, or planets? Is it something else? Or, maybe, is there a completely new effect which we interpret as dark matter while in reality there are new forces in action? The only recent answer is that there are strong indication that there are large numbers of planet-like objects in the vicinity of our galaxy. But these are not nearly enough to account for the whole effect. Many experiments are under way aiming at detecting the nature of dark matter (and it very existence).

Neutrinos

The early universe produced electromagnetic radiation which reaches us in the form of microwaves. This radiation was the result of the electromagnetic interactions among charged particles. There are, however, other types of interactions. We already met the gravitational interaction, and there are

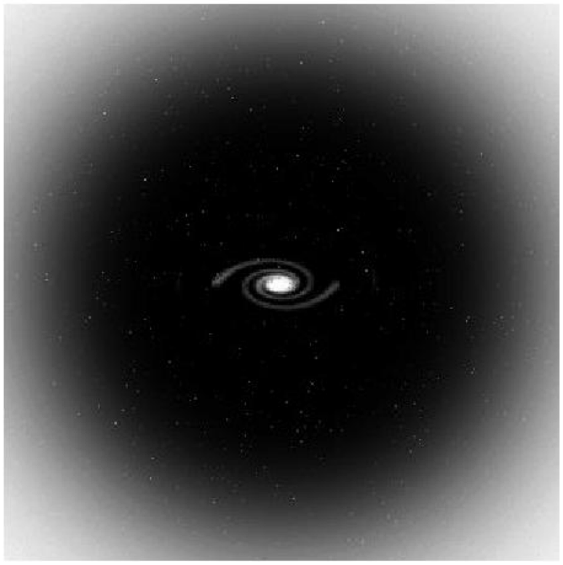


Figure 8.20: Illustration of the dark matter halo surrounding a typical galaxy.

two others called (again with a flair for words) the strong and the weak interactions.

Strong interactions are the ones responsible for nuclear forces between protons and neutrons (the constituents of atomic nuclei), and we will come back to them when we look at the evolution of a star (Sect. **??**). The remaining type, the weak forces, are experienced by *all* types of matter, but they are usually overwhelmed by the electromagnetic and strong forces because the weak interactions are, well, *weak!*

One is used to hear about electrons and protons and, perhaps to a lesser extent, neutrons. All these are constituents of atoms and atomic nuclei. But nature has a much richer population, and among its citizens one of the most intriguing are the neutrinos.

Neutrinos are very light particles [7] and experience *only* the weak interactions and it is because of this that they are rarely affected by other types of matter. Only in the densest of environments are neutrinos strongly disturbed. These occur in the center of neutron stars (Sec. **??**) or in the early universe. In this last case neutrinos were originally extremely energetic

[7] It had been assumed for a long time that they were massless, recent results however, indicate that neutrinos have a very small mass, of a billionth of a proton mass or less.

but, just as in the case of radiation, there came a time when the universe expanded to the point that the environment wasn't dense enough for the neutrinos to be affected by it. From that point on the neutrinos have been just cruising along, interacting only very rarely.

Initially these neutrinos lived in a very hot environment, which implies that each of them had a lot of energy *and* they were in a situation where very large gravitational forces were present. Nowadays they are in an environment where the gravitational forces are very weak. To understand what this implies consider the following analogy.

Imagine that you throw a ball up from the earth: initially the ball has a lot of kinetic energy, that is, energy due to its motion, but as it rises it slows down losing kinetic energy. Of course, this energy does not disappear, it is stored in potential energy (see Fig. 8.21). As the ball falls it will pick up speed so that when you catch it will be moving at the initial velocity (or close to it). In the same way the neutrinos in the present universe will have lost most of their kinetic energy.

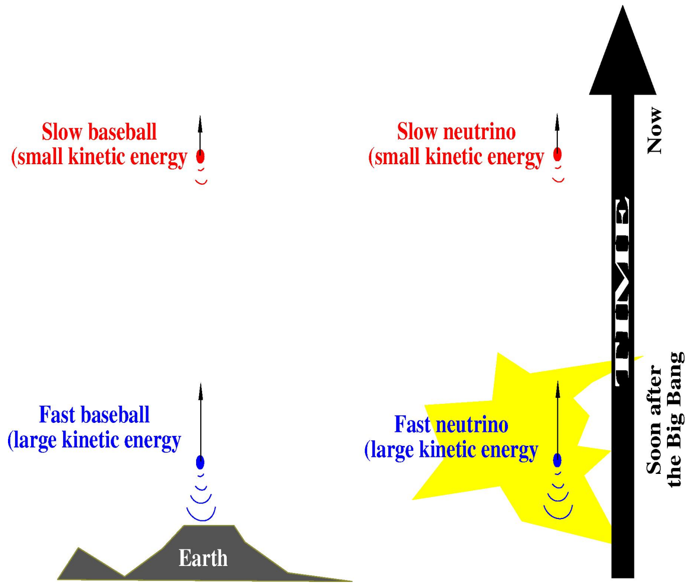


Figure 8.21: Neutrinos from the early universe have smaller kinetic energy now than in earlier epochs just as a baseball has lower kinetic energy the farther it is from Earth.

So another prediction of the Big Bang theory is that the universe is filled with neutrinos of very small kinetic energy. Unfortunately, out current technology is not sufficiently sophisticated to be able to detect them directly, but this might improve in the future.

8.5.2 The cosmological constant

When Einstein first studied the universe at large using the General Theory of Relativity he discovered that his equations predicted a universe which was either expanding or contracting, and this was contradicted with the best astronomical observations at the time. He then modified his equations to satisfy the observations. This modification corresponds to the assumption that the whole universe is permeated with a constant pressure (which in his case balanced the expansion yielding a steady universe). this universal pressure is called the *cosmological constant*

Though subsequently the data showed that the universe is in fact expanding and Einstein rejected the modification, on a philosophical basis the question still remains whether the *measured* cosmological constant is indeed zero (remember that on philosophical grounds Aristotle rejected heliocentrism: one must eventually back assumptions with observations). For many years the best value for the cosmological was assumed to be zero since no measurement gave positive indication to the contrary. Yet even a very small pressure can be important if it permeates the whole universe.

For many years the best value for the cosmological was assumed to be zero since no measurement gave positive indication to the contrary. Yet even a very small pressure can be important if it permeates the whole universe.

Recent measurements of the expansion rate of the universe (see Sect. 8.4.1) using type Ia supernovae (Sect. 8.4.1) favor an open universe with a small but non-zero cosmological constant. If these results are confirmed, Einstein's "blunder" will prove to be one more piece in the jigsaw of nature.:

8.5.3 Homogeneity and isotropy

One of the central simplifying assumptions of Einstein's cosmology is that, on average, the universe is the same in every direction (isotropy) and in every location (homogeneity). this does not mean, however, that the universe is a boring tapioca-like thing. The distribution of galaxies is far from smooth with most of them concentrated in relatively narrow sheets separated by large voids, see Fig. 8.22. The situation is reminiscent of a series of soap bubbles where the soapy water corresponds to the galaxies, the air inside the bubbles to the voids.

There are a few hipotheses which explain the origin of this type of structure. These must account not only for the voids, but also for the inhomogeneities in the comsic background radiation; and they must also predict a reasonable time-line for the development of galaxies. All these constraints are difficult to satisfy, making this an area of very active current research.

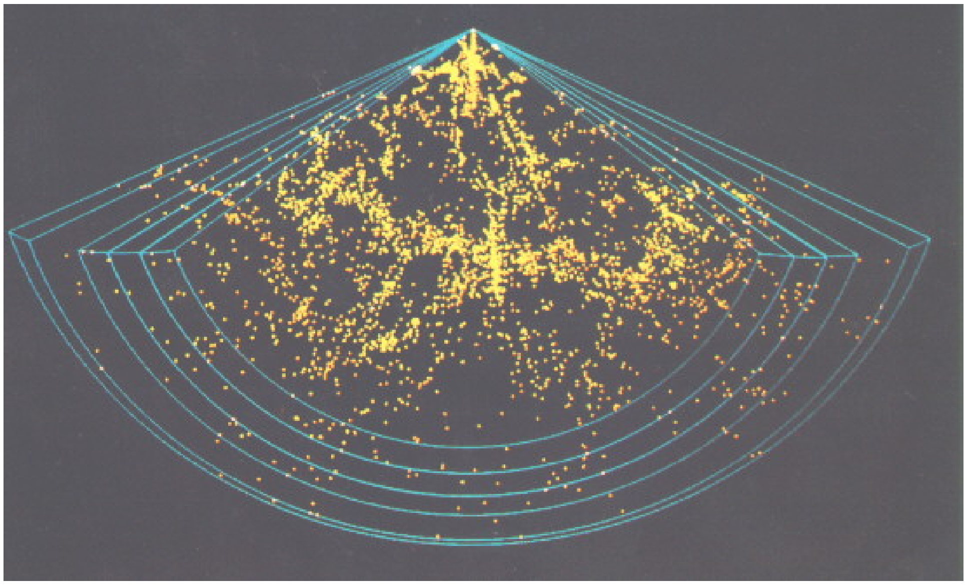


Figure 8.22: Large scale bubble-like structures in the universe. The image contains about 4000 galaxies each representing one luminous point.

Inflation

When we look at the microwave background radiation it looks the same in every direction, even from opposite sides of the sky, to a precision of 0.1%. Since they are so nicely correlated one would naturally assume that at some time all points in the observable universe were in close contact with each other, for otherwise it would be an unbelievable coincidence for all of them to look so much the same (at least through a microwave detector).

Now, a perfectly reasonable question is whether the Big Bang model has this property: will the Big Bang model predict not only the existence of the microwave background radiation, but also its exquisite uniformity? The answer is "yes" but only with additional assumptions.

This seems confusing: is the Big Bang theory to be modified and tuned every time a new piece of data comes along which does not agree with its predictions? Isn't this cheating? Doesn't this sound like Ptolemy adding epicycles every time things weren't quite accurate?

Fortunately this is not the case. The Big Bang theory determines the evolution of the universe *provided* the matter and energy content is known, *and* their behavior at very extreme conditions is well understood. The fact is, however, that we are not certain of all the matter and energy in the universe, nor do we know, for example, how they behave at temperatures above 10^{15} ^{o}K. Hence these "modifications" of the Big Bang theory correspond to different hypothesis of the behavior of matter at very high temperatures and densities, not of the general description provided by the General Theory of Relativity.

The simplest version of the Big Bang model which predicts a very uniforms microwave backgound goes by the way of *Inflation*. The idea is the following: the simplest way of getting uniform background radiation is if all the observable universe was in very close contact at an early time. Granted that, inflation provides a mechanism for increasing the size of this initially tiny region to the very large universe we see. Though mathematically involved what is assumed is that at a *very* time (about 10^{-35}s after the Big Bang) a new force comes into play which forces an exponential increase in the size of the universe (hence the name 'inflation'). After a fraction of a second this force is balanced by other interactions and the universe resumes a more dignified, if ponderous, expansion (see Fig. 8.23).

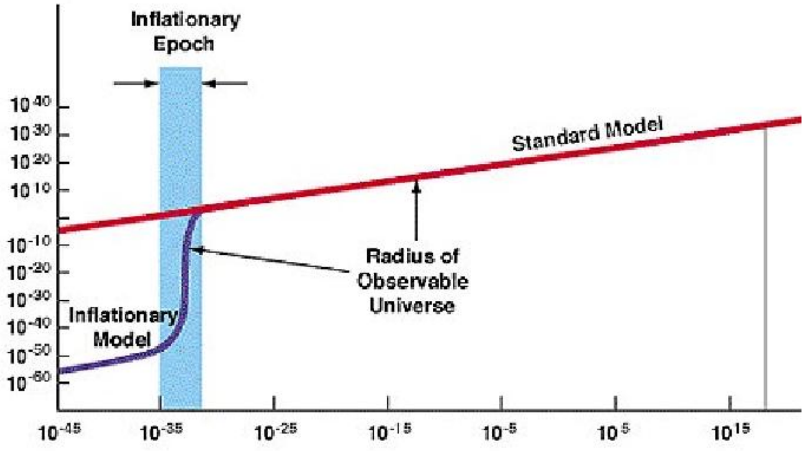


Figure 8.23: Time evolution of the size of the inflationary universe

One tantalizing conclusion derived from the inflationary hypothesis is that there are regions in the universe which we have not yet seen and which might look *very* different. Since no light has reached us from those regions we are currently unaware of their existence, only our inheritors will see the light coming from these distant reaches of the universe.

It is a challenge for current researchers to produce models that generate the intergalactic voids, yet with the *same* amount of dark matter required to understand the rotation of stars (Sect. 8.5.1) and using the inflation hypothesis such models actually exist. The corresponding computer simulations produce results such as the one shown in Fig. 8.24 which should be compared to the observations (Fig. 8.22).

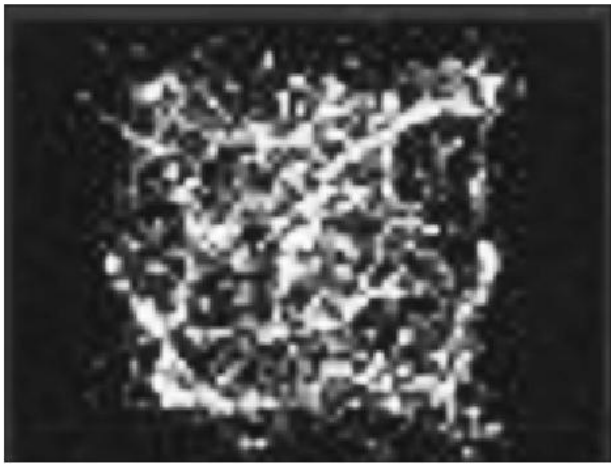


Figure 8.24: Simulation of the generation of structures in the universe assuming the presence of dark matter and an early epoch of inflation.

8.5.4 Summary

Though the General Theory of Relativity has produced a generous amount of verified predictions, its application to the universe at large has also generated a set of puzzles which, coupled to recent observations, are the topics of intense research. Whether there is a cosmological constant, whether the universe is filled with dark matter and the nature of this stuff and whether our current models of the universe are accurate enough to understand physics to the very earliest of times are issues currently addressed by researchers. The near future will provide more puzzles and some answers leading us, we hope, to a better understanding of the universe, our home.

Chapter 9

The lifes of a star

9.1 Introduction.

When stars are plotted in the H-R diagram, the number of stars in and out of the main sequence, together with models of stellar evolution provides a description of the possible ways in which stars are born, evolve and, eventually, die. During this process the star “move about” in the HR diagram (see Fig. 9.1). Since most stars *are* in the main sequence it is reasonable to suppose that during their life most stars *stay* in the main sequence, evolving into it when they are born and out of it when they are about to die. Models of stellar evolution confirm this.

For large objects (such as stars, galaxies, etc) the one ever-present force is gravity. This is always an attractive force which tends to condense stars and such into smaller and smaller objects. There are (fortunately) other effects which, at least temporarily, can balance gravity and stop this contraction. These effects are generated by the material which makes up the star and are always associated with various kinds of pressure (which tends to enlarge objects); a familiar example is the usual gas pressure

For large objects the one ever-present force is gravity

A less known type of pressure is produced by electrons [1] when they are brought in very close contact. Under these circumstances there is a very strong repulsion between the electrons, not only because they have equal charges (and hence repel each other), but because electrons, by their very nature, detest being close to each other: they require a relatively large breathing space. This repulsion between electrons is called *degenerate electron pressure* [2]. This effect has a quantum origin and has many interesting

Due to their dislike of being in close contact electrons produce a *degenerate electron pressure*

[1] Everything is made up of atoms. Atoms consist of a very dense and small nucleus and a bunch of electrons surrounding it.

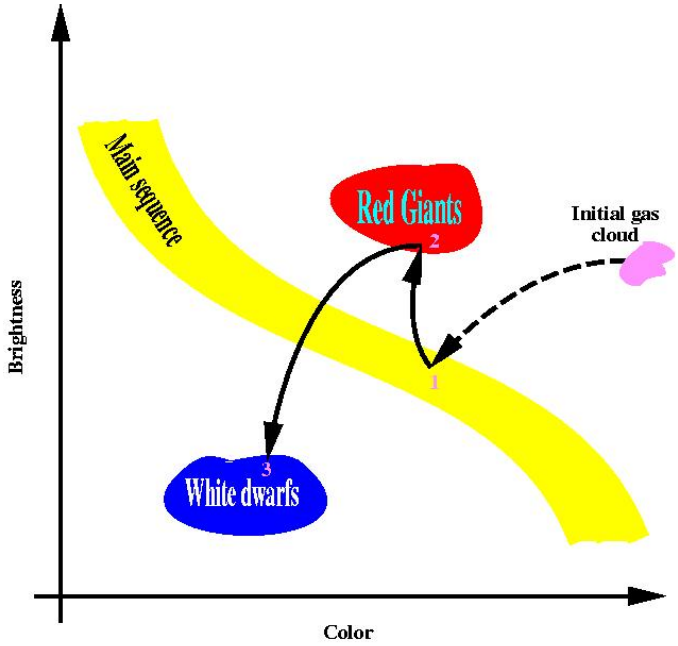


Figure 9.1: Diagram illustrating the evolution of a sun-like star. Born from a gas cloud it moves towards the main sequence (1) where it spends most of its life. After all Hydrogen is consumed in its core, the star burns Helium and becomes a red giant (2). Finally, when the Helium is consumed nuclear reactions subside and the star becomes a white dwarf (3) where it will spend its remaining (billions of) years.

consequences, to mention two, thanks to this strong dislike of electrons for occupying near-by locations, the floor supports your weight, and atoms have different chemical properties.

Electrons are not the only kids of particles that dislike being in close contact with one another. For example, the nucleus of a Hydrogen atom, called a *proton* also exhibits this property. Finally, and this is important for stellar evolution, other particles called *neutrons* also dislike being close to each others. Neutrons have no electric charge and are slightly heavier than protons; they are also found in atomic nuclei and are, in fact, a common sight in nature. All the atomic nuclei (except for Hydrogen) are made of protons and neutrons, with the neutrons serving as buffers, for otherwise the

[2] This is just a peculiar name and should not be interpreted as a judgment on the moral character of the electrons.

electric repulsion of the protons would split the nuclei instantly. When close to each other neutrons produce a *degenerate neutron pressure* and protons a *degenerate proton pressure* (see Fig 9.2).

Neutrons in close contact also produce a *degenerate neutron pressure*

Protons in close contact produce a *degenerate proton pressure*

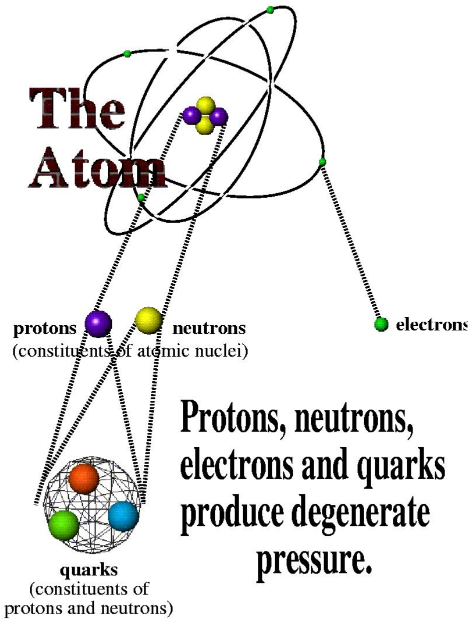

The Atom
protons
neutrons
(constituents of atomic nuclei)
electrons
Protons, neutrons, electrons and quarks produce degenerate pressure.
quarks
(constituents of protons and neutrons)

Figure 9.2: List of the most important particles which generate a degenerate pressure when in close contact. Also in the picture, the places where these particles are most commonly found.

The various stages of stellar evolution are classified according to the origin of the pressure which counterbalances gravity's pull. For most stars a balance is reached in the final stages of the star's life; there are some objects, however, for which gravity's pull overwhelms all repulsion in the stellar material, such objects are called *black holes*.

The various stages of stellar evolution are classified according to the origin of the pressure which counterbalances gravity's pull

The mass of the star largely determines its history, light stars (such as our Sun) will end in a rather benign configuration called a *White Dwarf*; heavier stars (with masses below 3-4 solar masses but larger than one solar mass) end as *neutron stars* after some spectacular pyrotechnics. Very massive stars end their lifes as black holes. This will detailed below, but before we need to understand what makes stars tick.

9.2 Stellar Power

The main power source for all stars is furnished by nuclear reactions. Possibly the most familiar of these reactions are the ones used in nuclear power plants; these, however, are *not* the ones or relevance in stellar processes. The relevant reactions present inside stars go under the name of *nuclear fusion.*

Recall that all atoms are made of a very dense and small nucleus which is positively charged and a bunch of electrons, which are negatively charged, and which surround the nucleus. At the center of the stars temperatures as very high (at least a few million degrees Celsius); pressures are also high [3] Under these circumstances the electrons are stripped off the nuclei and float around. In this large-temperature environment both electrons and nuclei very high speeds, so high that when two nuclei collide they often overcome the the repulsion produced by the fact that they both have positive charge. But when the nuclei come in such close contact with each other they will "stick". The result is a *new* nucleus and also *energy is released.* For example one can imagine slamming Hydrogen nuclei to produce the nucleus of a new element, Helium (see Fig. 9.3).

The result of the nuclear reaction in Fig. 9.3 is the depletion of Hydrogen in the star, the creation of Helium, and the release of energy in the form of radiation. Some of the radiation will heat the environment encouraging more nuclear reactions of the same type, but a small fraction of this energy will make its way to the star's surface and escape into space. Knowing the equivalence of mass and energy this implies that the stars become slightly lighter through this process. For our sun the loss is of "only" 1.35×10^{14} (135 trillion) tons per year (which is only about 7×10^{-12} – 7 trillionths – of a percent of the total solar mass). The jargon is that this reaction "burns" Hydrogen and that resulting "ashes" are mainly Helium.

The main nuclear reaction in stars "burns" Hydrogen and that resulting "ashes" are mainly Helium

The above is just one of a very large number of fusion reactions but it is the most common, and is present in all stars at some stage of their lifes. Other reactions are also important, I will talk about them later.

As time goes on the amount of Hydrogen drops and, eventually, there is not enough left to generate appreciable amounts of energy. There are nuclear reactions involving Helium (which is now quite abundant), but they require higher temperatures. So, when the Hydrogen is used up, the nuclear reactions turn off and the star continues to contract due to the gravitational pull. But, just as before, as the contraction proceeds, the temperature at the

[3] Remember that the pressure must balance gravity's pull. A star is a very massive body, hence gravity's pull will be very large; the pressure must then be also very large to cancel it.

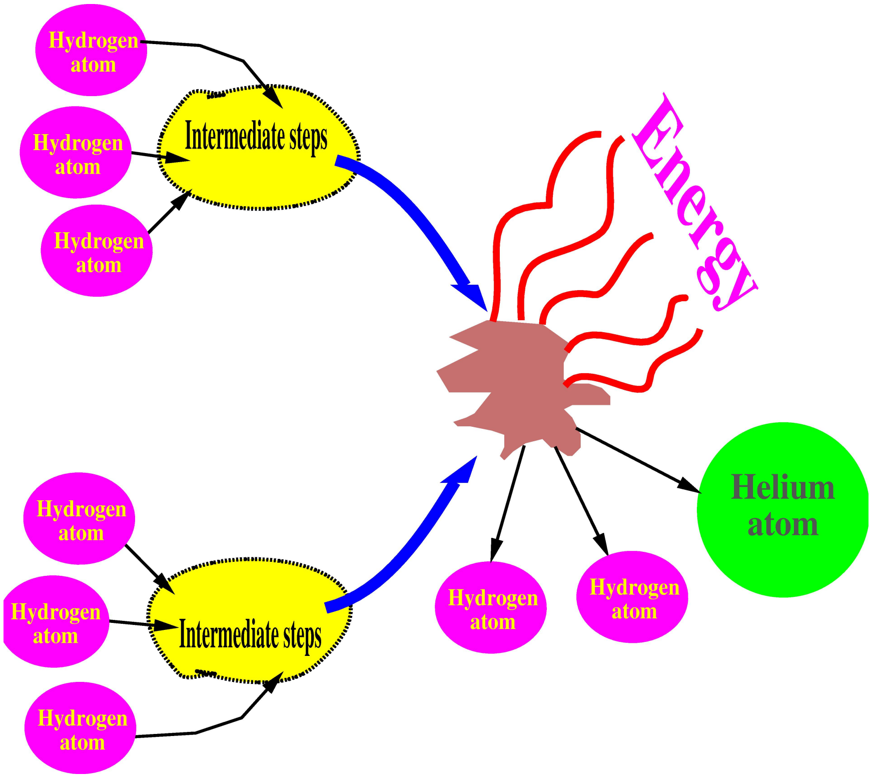


Figure 9.3: Illustration of the nuclear reactions which create Helium from Hydrogen. At very high temperatures Hydrogen atoms are slammed together, as a result a new element, Helium is created, the amount of Hydrogen is slightly depleted and energy, in the form of radiation, is released (in the "intermediate steps" some unstable nuclei are created).

core raises, eventually reaching the threshold of nuclear reactions involving Helium.

9.3 The lifes of a star

9.3.1 In the beginning

It all begins with a swirling cloud of dust and debris

It all begins with a swirling cloud of dust and debris (perhaps some old-star remnants). Gravitational attraction causes this cloud to slowly contract. As it contracts the cloud speeds up its rotation (much as an ice-skater turns faster when he/she draws her hands towards his/her body), and it heats up. The cloud becomes unstable and separates into blobs, some might be ejected due to centrifugal force, others condense into planets. The center of the cloud condenses into a big blob of matter (mainly Hydrogen since this is the most abundant element in the universe). This process takes about one billion years to complete and produces a primitive planetary system: a protostar (which is very big but too cold to produce nuclear reactions) circled by protoplanets. As time goes on, the protoplanets in their orbits will "sweep-out" the remaining debris from the cloud.

The center of the cloud condenses into a big blob of matter

9.3.2 A rising star

Through the evolution of the star the only force opposing the gravitational collapse is the pressure of the stuff the central blob is made of; this pressure is initially very small compared to the pull of gravity. This means that the blob will contract until pressures and temperatures at the center are so high that nuclear reactions turn on. At this point the energy release from the fusion reactions heats up the stellar material, this in its turn increases the pressure and the contraction stops. As mentioned above, the main reaction occurring at this stage consume Hydrogen and produce Helium: the star "burns" Hydrogen into Helium. This goes on for a long time: if the star is light (as our sun) it proceeds for about 10 billion years, much heavier stars use up Hydrogen much faster (for the heaviest ones it takes 'only' 1 million years).

Pressures and temperatures at the center are so high that nuclear reactions turn on

The main reaction occurring at this stage consume Hydrogen and produce Helium

9.3.3 A Giant appears

After the supply of Hydrogen in the core is depleted the corresponding nuclear reactions stop (there are other fusion reactions, but they can occur only at higher temperatures than the ones present at the center of the star at this stage). Then the pressure drops and the gravitational collapse proceeds. During this process the center of the star is compressed more and more, increasing the central temperature until, finally, it becomes so hot that nuclear reactions involving Helium start up: Helium atoms slam together, and, after a complicated reaction produce Carbon. When these reactions turn on the energy output is enormous, the core becomes extremely hot and radiates a very large amount of energy. This radiation pushes out the outer layers of the star, and as they are pushed out they become a bit cooler and thus look redder. The star then becomes a *red giant* a bloated result of the burning of Helium Our sun will eventually go through this process and will grow to the point that it will engulf the orbits of Mercury, Venus and,

When burning Helium the star becomes a *red giant* possibly, the Earth.

9.3.4 And so it goes

What happens when the supply of Helium is used up? The story is repeated: gravitational contraction takes over and the star collapses further. Eventually other nuclear reactions become viable, power increases until the various nuclei are depleted, then contraction takes over again. In this manner the star produces, Oxygen, Silicon and, finally, Iron. this is, in fact, the way in which these elements are manufactured in nature. Every bit of Carbon in a flower's DNA, every bit of Oxygen we take in every breath, every bit of Silicon in a sandy beach was created in a star.

Stars create Oxygen, Silicon and, finally, Iron

When the core of the star turns into Iron all nuclear reactions stop, permanently. The reason is that Iron is a very stable nucleus so that if two Iron nuclei are slammed together they will only stick if energy is *supplied* (in contrast, two Hydrogen atoms stick and also release energy). When nuclear reactions stop gravitational contraction continues again and will proceed until the electrons in it are closely squashed together. As mentioned above electrons dislike being in close contact with each other and when squashed will generate a pressure which opposes gravity; whether this pressure is sufficient to stop collapse depends on how heavy the star is.

Light stars

For stars lighter than 1.4 solar masses the electron degenerate pressure will balance gravity. The star has by now contracted from its red-giant size to the size of a small planet (like Earth). The material of this star is so dense a teaspoon of it would weight 1 ton on Earth.

When this final contraction occurs there is a certain amount of overshoot and bouncing back and forth before stability is achieved; in this process all the outer layers of the star are ejected. The end result is a beautiful ring of stellar material which spreads out, at the center of which a small star, called a **white dwarf**, remains (see Fig. 9.4). White dwarfs are is stable and their racy days of nuclear reactions are forever gone; they slowly radiate their remaining heat little by little and eventually become dark cinders. This is the end of a star whose mass is smaller than 1.4 times the mass of the Sun; this process is summarized in Fig. 9.5

It is interesting to note that the theory *predicts* that these objects will always be lighter than 1.4 solar masses. Observations have confirmed this. This theory is a combination of quantum mechanics and gravitation and, in

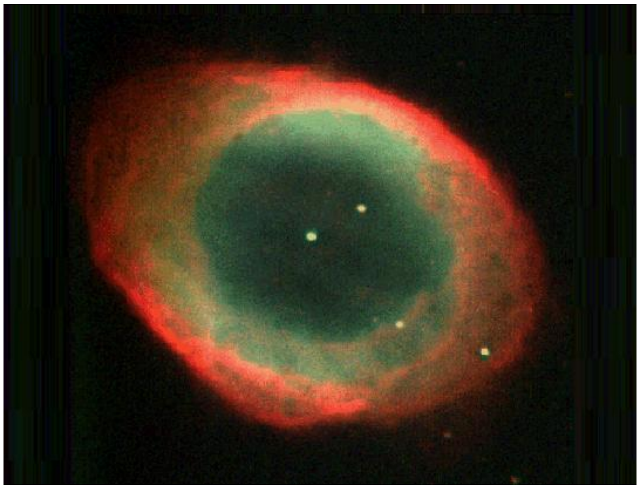


Figure 9.4: Photograph of a ring nebula. The central white dwarf has, in its last throes, expelled its outer layers appearing here as a ring surrounding the small remnant.

fact, it provided the first application of quantum physics to stellar objects.

For heavier stars the pull of gravity overcomes the degenerate electron pressure and collapse continues.

Electrons, protons and neutrons. Matter in most situations is composed of electrons, protons and neutrons. Electrons are negatively charged and weigh 9×10^{-31}kg, protons weight 1.8×10^{-27}kg and have positive charge, exactly opposite to that of the electrons. Neutrons weigh as much as protons and have no charge. Usually protons and neutrons are bound together in atomic nuclei and are surrounded by a cloud of electrons so that the whole systems is neutral. If, however, matter is subjected to higher and higher pressures, eventually the atoms are crushed together to the point that the electrons can jump around from the vicinity of one nucleus to another.

If the pressure is increased still further the nuclei themselves are brought into close contact and lose their identities. At this point the protons undergo a reaction in which they absorb an electron and turn into protons while emitting a neutrino (yet another subatomic particle). Because of this process most of the matter turns into neutrons and neutrinos. The latter interact very seldom and just leave the system; because of this what remains is essentially an enormous number of neutrons

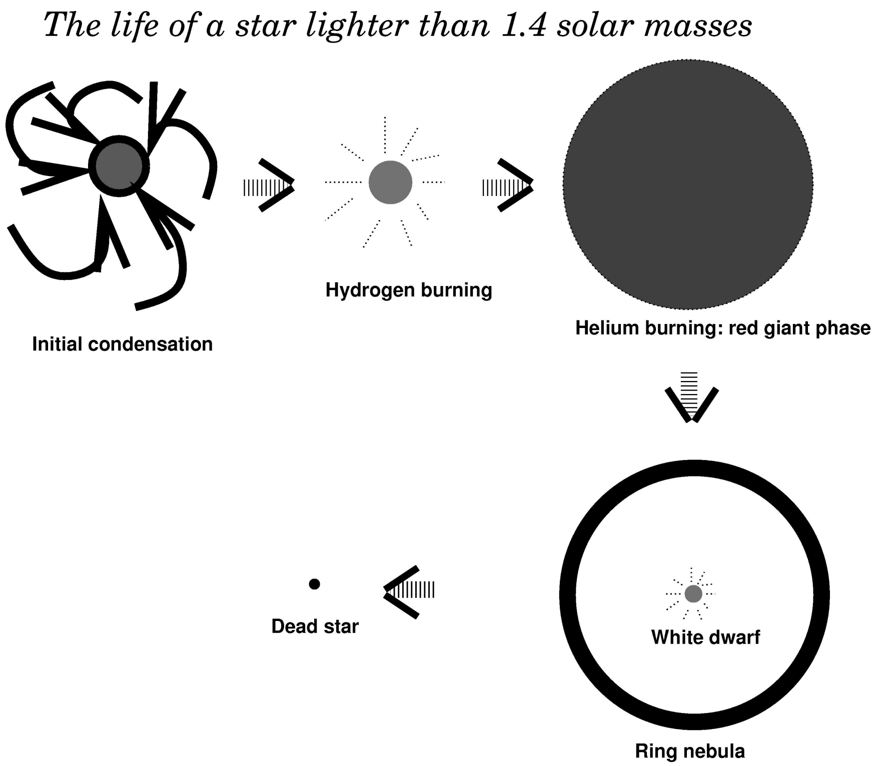

The life of a star lighter than 1.4 solar masses
Initial condensation
Hydrogen burning
Helium burning: red giant phase
White dwarf
Ring nebula
Dead star

Figure 9.5: Time and life of a star of mass below 1.4 times the solar mass (less than about 3×10^{27}tons).

Medium-size stars

For stars heavier than 1.4 solar masses but lighter than about 3–4 solar masses (the calculations are still a bit uncertain), the electron pressure is not strong enough to balance gravity. The contraction then goes crushing the electrons together and braking apart the Iron nuclei into their constituents. These constituents, neutrons and protons, also detest being close to each other and, as mentioned above, produce a (degenerate) pressure which opposes gravity. For a star in the present mass range this pressure is sufficient to stop further collapse, but is effective only when the material is extremely dense which occurs only when the star has contracted to an object a few kilometers in diameter.

The contraction of these stars from their initial solar size to the size of a city is one of the most spectacular events in the heavens: a supernova. Imagine an object weighting 5×10^{27} tons (that is five thousand trillion-

trillion tons, or about 2.5 solar masses), which contracts from a size of 10^6 (one million) kilometers to about 10 kilometers, and it all happens in a fraction of a second. During collapse the amount of energy generated is fantastic, part of it goes into creating all elements heavier than Iron, part into creating neutrinos and part is transformed into light.

Radioactive elements are also created during the collapse. These elements rapidly decay, and the resulting radiation is so intense it produces a fantastic flash of light. At this point the supernova will out-shine a full galaxy of normal stars (several billion or up to a trillion of them!).

After the collapse there is a violent overshoot before equilibrium sets in, at this time all the outer layers of the star are ejected at speeds close to that of light. When this material goes trough any planets around the star (if any) it vaporizes them. In the middle of this cloud the core of the original star remains, a rapidly rotating remnant, protected against further collapse by it neutron degenerate pressure.

The overshoot is so violent that the elements created will be strewn all over the region surrounding the star, part of this material will end up in dust clouds which will become stellar systems (the shock produced by the supernova material colliding with a dust cloud may initiate the formation of a stellar system); this is how the Earth acquired all elements aside from Hydrogen and Helium. Every bit of tungsten used in our light bulbs came from a supernova explosion, as all the uranium, gold and silver. All the iron in your hemoglobin got there through a supernova explosion, otherwise it would have remained locked into the deep interior of some star.

The most famous supernova was observed by Chinese astronomers more than one thousand years ago (see Sect. **??**), its remnants are what we call the Crab nebula (Fig. **??**). We also met another important supernova (see Sect. **??**) observed by Tycho Brahe in 1572 (Fig. 9.6). In 1987 a star in our galaxy “went” supernova, since then we have observed the ejecta from the star and the remnant of the core (Fig 9.7. There are, of course, many known supernova remnants (see, for example, Fig. 9.9). The evolution of a middle-size star is illustrated in Fig. 9.8.

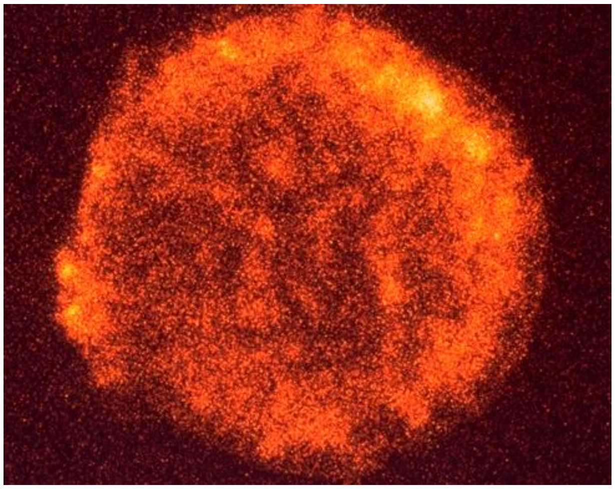


Figure 9.6: An X-ray photograph of the remnant of Tycho's supernova.

The Crab nebula. The Crab nebula in Fig. **??** is the remnant of a supernova explosion. The explosion was observed on July 4, 1054 A.D. by Chinese astronomers, and was perhaps about as bright as the Full Moon, and was visible in daylight for 23 days. It was probably also recorded by Anasazi Indian artists (in present-day New Mexico and Arizona), as findings in the Chaco Canyon National Park (NM) indicate

After gravity is balanced, and after the exterior shells are ejected the star stabilizes forever. But not without some fancy footwork: the remains of the star usually rotates very rapidly (up to 30 times per second!) and it also possesses a very large magnetic field. These two properties cause it to emit X-rays in a directional fashion, sort of an X-ray lighthouse. Whenever the X-ray beam goes through Earth we detect an X-ray pulse which is very regular since the star's rotation is regular. This is called a *pulsar*. As time goes on the rotation rate decreases and the star dies a boring *neutron star*. Neutron stars are very compact objects having radii of about 10 km (6 miles) so that their density is enormous, a teaspoon of neutron-star material would weigh about 10^{12} (one trillion) tons on the Earth's surface.

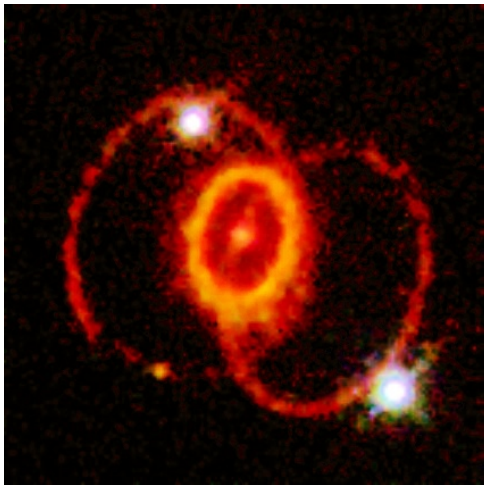


Figure 9.7: Left: a picture of the supernova 1987A remnant (the most recent supernova in our galaxy. Right: photograph of the core.).

9.3.5 The heavyweights

But what happens for stars heavier than about 3-4 solar masses? In this case the pressure from the squashed nuclei cannot stop the gravitational attraction and collapse continues. In fact no known effect can stop the collapse and it will go on and on until the star collapses to a point. This how a *black hole* is created (see Sect. **??**).

For this object the gravitational force is so big that even light cannot leave its vicinity: as mentioned in section **??**, if a light beam comes too close to the center of such an object, the bending effect is so severe that it spirals inwards. Light emitted from up to a certain distance will be bent back into the star. This distance defines a horizon: nothing inside the horizon can ever come out, nothing that crosses the horizon ever leaves the black hole. The more massive the black hole, the larger the horizon.

For a very massive black hole an astronaut may cross the horizon without feeling any personal discomfort, only later he realizes that he is inside a cosmic Venus fly-trap (or roach motel [4]) out of which there is no escape.

General relativity together with our knowledge of subatomic physics guarantees that a sufficiently large star will eventually collapse to the point where a horizon appears. The manner in which such a star evolves thereafter is impossible to know since no information from within the horizon can be

[4] You check in...but you never check out

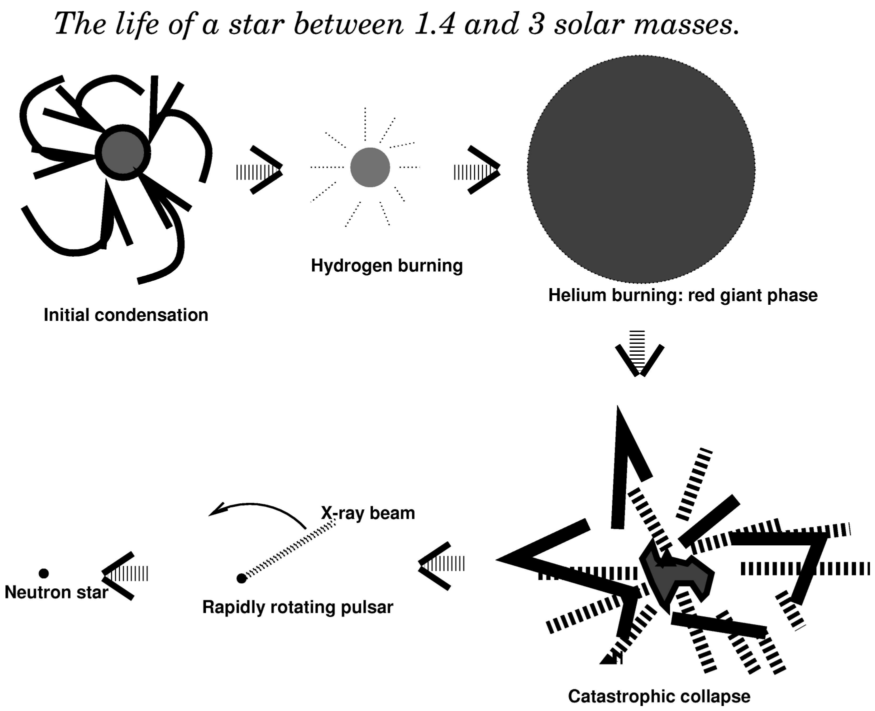


Figure 9.8: Time and life of a star of mass between 1.4 and 3–4 times the solar mass (between about 3×10^{27} and 7×10^{27}tons).

sent to the outside universe. There might be some new kind of effects which will stop the collapse of even the most massive stars, but even then the horizon *will* remain. The point is that our *present* knowledge of physics predicts the existence of black holes, even if we do not know *all* physical effects in Nature. The fact that we have several excellent black-hole candidates supports (albeit indirectly) our understanding of gravitation and physics in general.

The detection of black holes is difficult: one looks not for the object itself but for certain characteristics of the radiation emitted by matter falling into the black hole; see Fig. **??**. Anything coming near the black hole will be strongly attracted to it, it will swirl into the black hole, and in the process it will heat up through friction, this very hot matter emits electromagnetic radiation in a very characteristic way and it is this patter what the astronomers look for (see Sect. **??**).

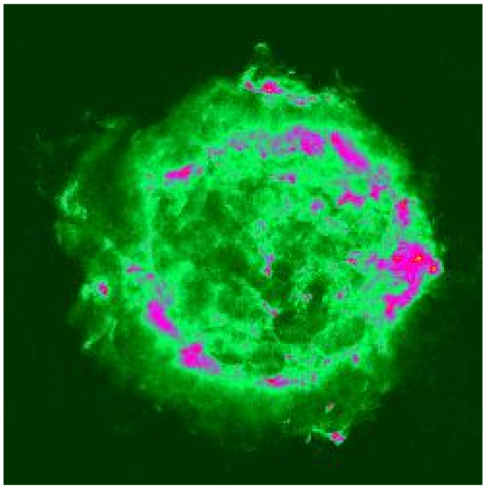


Figure 9.9: A radio picture of the Cassiopeia A nebula, a supernova remnant.

The best candidate for a black hole was, for a long time an object in the constellation Cygnus and is called Cygnus X1. Very recently (May 1995) an object with the name GRO J1655-40 in the constellation of Sgittarius became an excellent black-hole candidate. In this object a star is accompanied by an object that emits no light, there is material falling into the companion and the X-rays from this material are unique to black-holes. Moreover, the mass of the companion can be determined to be heavier than 3.35 solar masses. The companion has then all the properties of a black hole.

Black holes are also supposed to be the engines at the center of active galactic nuclei and quasars (see Fig. 9.11). These are very distant objects which, by the mere fact of being detectable on Earth, must be immensely luminous. So much so that nuclear energy cannot be the source of that much radiation (you'd need more nuclear fuel than the amount of matter in the system). On the other hand, a black hole of several million and up to a billion solar masses can, by gulping down enough stellar material (a few suns a year) generate in the process enough energy.

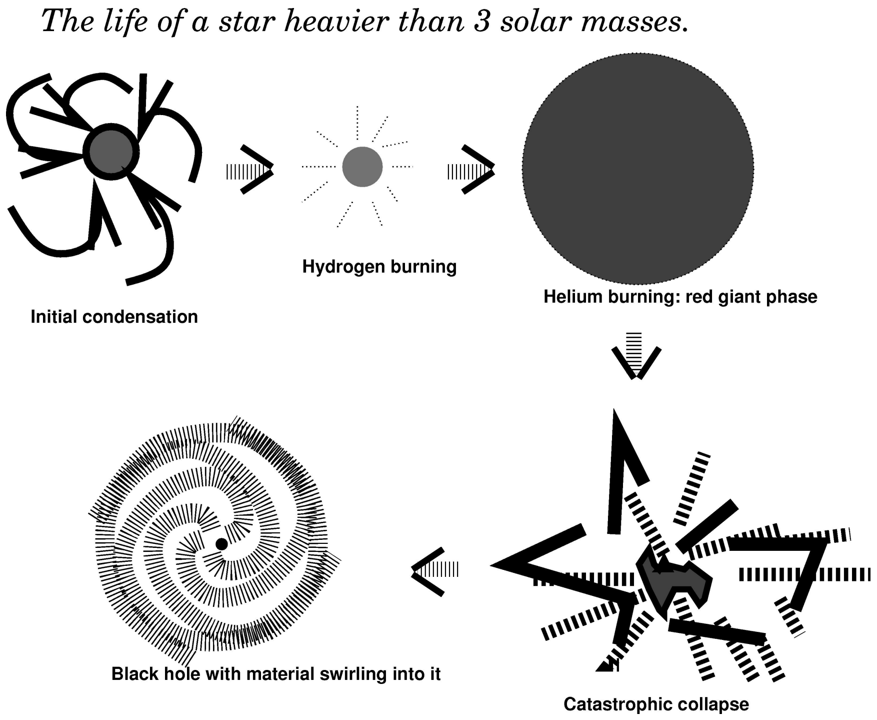

The life of a star heavier than 3 solar masses.
Initial condensation
Hydrogen burning
Helium burning: red giant phase
Catastrophic collapse
Black hole with material swirling into it

Figure 9.10: Time and life of a star of mass heavier than 4 solar masses (8×10^{27}tons).

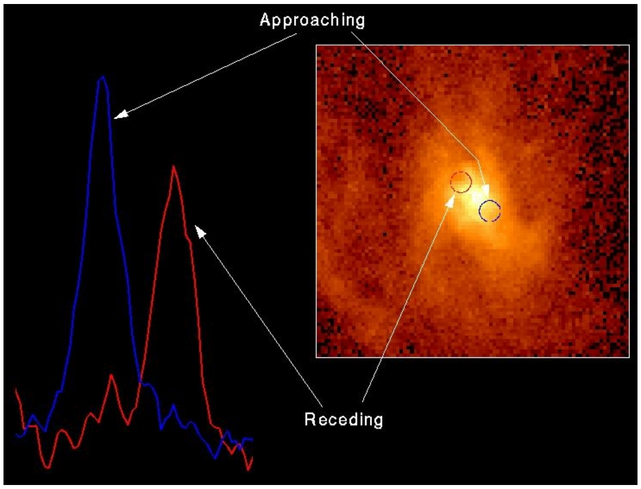


Figure 9.11: A disk of accreting matter onto a very compact object believed to be a black hole. The object at the center of M87 (located 50 million light-years away in the constellation Virgo) weights about three billion suns, but is concentrated into a space no larger than our solar system. The black hole is surrounded by a disk of matter which is being sucked into the center; as the matter falls in it radiates, and the emission from two regions are measured. Using the Doppler effect one can calculate the velocity of the material falling in; the region label;ed “approaching” emits blue-shifted light, while light from the “receding” region is red-shifted. The speed of the gas is enormous: 1.2 million miles per hour (550 kilometers per second).