.

CONVERSATIONAL COSMOLOGY

An Overview of Our Current Understanding of the Large Scale Structure of the Universe

Henry E. Kandrup

Department of Astronomy and Department of Physics

and Institute for Fundamental Theory

University of Florida

Revised edition: August 1999

Note that, owing to copyright laws, the figures mentioned in the text are not included in this postscript file.

Introduction

This is a collection of notes for an undergraduate course on *cosmology*, the study of the large scale structure and history of the Universe. Most of the students in this course are focusing on the liberal arts and/or other "nonscientific" disciplines, but there are a few students majoring in the sciences as well. As I teach it, the aim of this course is threefold: (1) to present an overview of the history of our changing understanding of what the Universe as a whole is like; (2) to provide a relatively comprehensive description of our current understanding of our place in the Universe; and (3) to sumarise some of the major ideas of twentieth century science, including relativity and quantum mechanics, which play a role in understanding modern cosmology.

I have assembled these notes because, to my knowledge, there is no appropriate textbook which can be used for a course of this kind. The text aims to be self-contained. I assume no scientific sophistication and no mathematics more complicated than squares, cube roots, and the like. The Pythagorean Theorem comes up briefly in the lectures on relativity. The one significant way in which these notes *are* incomplete is that they contain no photographs of astronomical objects and relatively few diagrams. If and when a version of these notes ever sees publication as a book, it will be augmented heavily by a large number of such visual aids. In the interim, students using these notes will have to rely on their classroom experience and explorations of the web to fill this gaping lacuna.

I view my role in writing these notes as that of an old-fashioned story teller, one who aims to educate, enlighten, and entertain. As a physicist working actively in cosmology and a number of related fields, I myself find cosmology to be an interesting and exciting subject. That should hardly be surprising. However, it has been my experience that many other people, including nonscientists, share my enthusiasm. I consider it natural for people to be curious about what the Universe as a whole is like and how we fit into it. My aim here is to provide some answers to these general questions in what is hopefully a relatively nonmathematical, and user-friendly, way.

I like talking about cosmology to more or less anyone who is interested in listening and, over the years, I have had conversations about this sort of stuff with a large number of people, ranging from active scientists and graduate students to gardeners, truck drivers, construction workers, and engineers, as well as undergraduates. Each one of these conversations has, to a greater or lesser extent, helped me refine my story-telling skills and, as such, each of the people with whom I have discussed this material has helped to make these notes better than they might otherwise have been.

Sizes and Distances

the nucleus of a typical atom: 2×10^{-13} cm
the radius of a typical atom: 1×10^{-8} cm

the approximate wavelength of ordinary red light: 5×10^{-5} cm

the length of a largish ant: 1×10^{0} cm $= 1$ cm $= 0.01$ meter

the height of a typical adult: 175 cm $= 1.75$ meter

the approximate distance a person can walk comfortably in 8 to 10 minutes:
1000 meters $= 1$ km

the approximate length of Manhattan Island: 22 km

the approximate distance across Texas from El Paso to Texarkana: 1500 km

the approximate equatorial radius of the Earth: 6378 km

the approximate distance light can travel in 1 second in empty space:
$300,000$ km $= 3 \times 10^{5}$ km $= 3 \times 10^{10}$ cm

the average Earth-Moon distance: $\approx 3.84 \times 10^{5}$ km

the approximate equatorial radius of the Sun: 6.96×10^{5} km

the average Earth-Sun distance:
≈ 150 million km $= 1.5 \times 10^{8}$ km $= 1$ astronomical unit (1 AU)
≈ 500 light seconds $= 8\frac{1}{3}$ light minutes

the average Sun-Pluto distance $\approx$ 40 AU $\approx$ 6 billion km $= 6 \times 10^{9}$ km
$\approx 20,000$ light seconds ≈ 5.5 light hours ≈ 0.23 light days

the approximate distance from the Sun to the next nearest star:
4.3 light years

the diameter of a typical globular cluster: 100 light years

the approximate diameter of our Galaxy, the Milky Way:
$100,000$ light years $= 1 \times 10^{5}$ light years

the approximate diameter of our Local Group of galaxies:
3 million light years $= 3 \times 10^{6}$ light years

the largest scales in the Universe on which we detect clustering of galaxies:
$\approx 0.5 - 1.0$ billion light years $= 0.5 - 1.0 \times 10^{9}$ light years

the distance to the highest redshift quasars, the most distant objects which we can detect:
$\approx 10 - 15$ billion light years $= 1.0 - 1.5 \times 10^{10}$ light years

Sources and References

Much of the material presented in these lectures about archeoastronomy and the history of astronomy and physics before the twentieth century was derived from the following sources:

Boorstin, D. J. *The Discoverers* (Vintage, New York, 1984).

Campbell, J. *The Masks of God* (Penguin, New York, 1968).

Haskins, C. H. *The Renaissance of the Twelfth Century* (Meridian Books, New York, 1957).

Koestler, A. *The Sleepwalkers* (Hutchinson, London, 1959).

Kuhn, T. *The Structure of Scientific Revolution* (University of Chicago Press, Chicago, 1970).

Pannekoek, A. *A History of Astronomy* (Allen and Unwin, London, 1961).

Good college-level surveys of astronomy in general include

Abell, G. *Exploration of the Universe* (Holt, Reinhart, and Wilson, New York) and
Kaufmann, W. J. *Universe* (Freeman, New York),
both two-semester textbooks which have gone through multiple editions.
There is also an abridged one-semester version of the latter, namely
Kaufmann, W. J. *Exploring the Universe* (Freeman, New York).
The books by Kaufmann contain less information than Abell, but make somewhat easier reading.

Pretty – and instructive – pictures of astronomical objects have become *de rigeur* in all surveys textbooks on astronomy, including those listed above. Alternatively, large numbers of interesting images can be found on the web, at one or more of the following addresses:

http://www.stsci.edu/
(Space Telescope Science Institute: where Hubble Telescope data is analysed)

http://www.ctio.noao.edu/ctio.html
(Cerro Tololo Interamerican Observatory in Chile)

http://www2.keck.hawaii.edu:3636/index.html
(a list of sites provided by the Keck Telescope in Hawaii)

For images of Solar System objects – a subject not discussed in these lectures – try:
http://nssdc.gsfc.nasa.gov/imgcat
(from Goddard Space Flight Center in Greenbelt, Maryland)

Table of Contents

Lectures 1 and 2

OUR PLACE IN THE UNIVERSE

Astronomy and cosmology are sciences and, like all sciences, typically use the metric system of measurements. The basic unit of length is the meter, which corresponds to a length of about 39.36 inches. A kilometer equals 1000 meters, and corresponds to a distance of approximately 0.6214 miles. The basic unit of mass is the gram which corresponds in the English system to a weight of about 0.035 ounces. A kilogram equals 100 grams and corresponds to a weight of about 2.205 pounds.

In astronomy, one often encounters distances that are much larger than a kilometer and masses that are much larger than a kilogram. For this reason, other units of length and mass are also used. The *astronomical unit* (or AU) is defined as the average distance from the Earth to the Sun over the course of a year, and corresponds to about 150 million kilometers. The *lightyear* corresponds to the distance that light can travel in the course of a year when moving through empty space. Given that the velocity of light is about 300,000 kilometers, or, more precisely, about 299,793 kilometers per second, this works out to a distance of about 9.46 trillion kilometers. Equivalently, it corresponds to a distance of about 63,240 AU. Another measure of length that is often used (but which will be largely avoided in these lectures) is the *parsec*, which equals about 3.26 lightyears.

In astronomy and cosmology, one often has reason to use numbers that are much larger than, or much smaller than, one, involving lots of zeros. To avoid having to write these zeros, scientists usually use what is called *exponential notation* or *scientific notation*. For numbers bigger than one, this involves the identifications $10^0 = 1$, $10^1 = 10$, $10^2 = 100$, etc., so that, for a general $n > 0$, 10^n corresponds to a 1 followed by n zeros. Similarly, for numbers smaller than one, $10^{-1} = 1/10^1 = 0.1$, $10^{-2} = 1/10^2 = 0.01$, $10^{-3} = 1/10^3 = 0.001$, etc., so that 10^{-n} means that there is a 1 in the n^{th} slot to the right of the decimal point. With this convention,

$$63,240 \text{ AU} = 6.324 \times 10,000 \text{ AU} = 6.324 \times 10^4 \text{ AU}.$$

Similarly, 9.46 trillion kilometers corresponds to 9.46×10^{12} km.

At least three different temperature scales are in use throughout the world. In the United States, temperature is usually measured in degrees *Fahrenheit* whereas, in most of Europe, degrees *Centigrade* (or *Celsius*) is the norm. The Fahrenheit scale is so defined that water freezes at 32 F and boils at 212 F. In the Centigrade system, water freezes at 0 C and boils at 100 C. A body temperature of 98.6 F corresponds to a temperature of 37 C.

The Centrigrade scale is based on the convenient, but not fundamental, choice of a zero point at the temperature at which water freezes. A more fundamental choice is to pick the zero point as being the lowest possible temperature that can arise in nature, absolute zero, about -273 C or -459 F. This is done by measuring temperatures in degrees *Kelvin*. To convert from degrees Centigrade to degrees Kelvin, all that one needs to do is add 273 (or, more precisely, a value of about 273.1). A typical temperature on the Earth is about 300 K. The surface temperature of the Sun is about 10,000 F or 5,800 K.

One final bit of book-keeping. When we look at, and compare, objects in the sky it is useful to identify angular distances. Dating back to the ancient Babylonians, such measures of distance are based on dividing a circle into 360 degrees, one degree into 60 minutes of arc, and one minute of arc into 60 seconds of arc. The apparent sizes of the Sun and the Moon in the sky are both about 0.5 degree. (Interestingly, this system of time and angles, based on combinations of 6 and 60, was only used by the Babylonians in the context of religion and astrology, the arithmetic of the everyday world of business – with a few notable exceptions like beer – being conducted in base ten, i.e., using combinations of 10, 100, and so forth. It is also worth noting that the words "minute" and "second" were applied to geometry millenia before the late Middle Ages, when they came to represent the precise units of time that they do today.)

The Earth is one of nine planets in the Solar System, the third closest ot the Sun, about 150 million km away. Like the Sun and all the planets, it is approximately spherical, with an equatorial radius of about 6378 km, a value conveniently denoted $R_\oplus$ The Earth has a mass of approximately 5.976×10^{24} kg, usually denoted $M_\oplus$. The core of the Earth is comprised primarily of iron and nickel. The crust is comprised mostly of carbon, silicon, nitrogen, oxygen, and hydrogen. The atmosphere is mostly nitrogen and oxygen, with small amounts of other substances, including argon and carbon dioxide. The Earth revolves around the Sun once a year along an orbit that is almost, but not quite, circular: like all the planetary orbits, to a high degree of accuracy its orbit is given as an *ellipse*.

The Moon, which revolves around the Earth roughly once a month, again along a nearly elliptical orbit, is the closest natural extra-terrestrial object. Its distance from the Earth varies between about 356,000 and 407,000 km. It is slightly more than one quarter as large as the Earth in radius, about 1738 km or $0.272R_\oplus$, and has a mass of about 7.35×10^{22} kg or $0.0123M_\oplus$. The composition of the crust of the Moon is relatively similar to that of the Earth, except for the (near-)complete absence of water. The Moon has essentially no atmosphere since its gravity is not strong enough to hang onto any air.

The Sun, located at the center of the Solar System, is by far the largest material object therein. It has a mass of approximately 1.989×10^{30} kg or about $330,000M_\oplus$. This value is often denoted $M_\odot$. The equatorial radius of the Sun is about $696,000$ km or $109R_\oplus$, a value often denoted $R_\odot$. Like most of the Universe, and unlike the Earth and the Moon, the Sun is comprised almost entirely of hydrogen and helium.

Mercury is the first planet out from the Sun, separated from the Sun by an average distance $a \approx 0.38$ AU. (Here, and henceforth, a squiggly equal sign $\approx$ means "approximately equaling," indicating that the number is not exact. If the number is an even cruder estimate, this will often be represented by a single squiggle, i.e., $\sim$.) It has an equatorial radius $R \approx 0.38R_\oplus$ and a mass $M \approx 0.055M_\oplus$. Because it is close to the Sun, the side facing the Sun is typically very hot, with a temperature of about 350 C or 623 K. The side away from the Sun is very cold, with a temperature of about -170 C or 103 K. The fact that the side of Mercury facing the Sun is so hot is responsible for the fact that Mercury has essentially no atmosphere. This in turn is responsible for the fact that the side facing away from the Sun is extremely cold, much colder than any temperatures on the Earth. If the Earth had no atmosphere to "hang onto" heat at night, nights on the Earth would be much cooler than they actually are.

Venus is the second planet out from the Sun, separated by an average distance $a \approx 0.72$

AU. It has an equatorial radius $R \approx 0.949R_\oplus$ and a mass $M \approx 0.815M_\oplus$, which means that, of the other planets, it is the closest to the Earth in size. It is also the most similar to the Earth in terms of its location in the Solar System. However, despite these similarities it is a very different sort of world. Venus has a hot, dense atmosphere, comprised primarily of carbon dioxide, but also containing traces of various acids baked out of the rocks on the surface. The surface temperature on the planet is about 480 C or 750 K, i.e., 900 F, a temperature substantially higher than anything encountered on Mercury and hot enough to melt lead. The pressure of air weighing down on the surface of Venus is about 100 times that of air on the surface of the Earth. The air is in fact so dense that you could swim in it, if you had not first melted and/or corroded.

This noxious atmosphere was caused by a *greenhouse effect*, the onset of which was related to Venus' position in the Solar System: Because the planet was slightly closer to the Sun than is the Earth, it was a bit hotter than the Earth early on, which implied that water vapor in the atmosphere was unable to condense to form oceans. On the Earth, the oceans absorbed the carbon dioxide out of the primordial atmosphere, after which it was deposited into carbonaceous rocks on the ocean floors. On Venus it remained in the atmosphere, trapping heat on the surface of the planet, causing the planet to become even hotter. If the Earth had been located as little as ten percent closer to the Sun, it is believed that it would have undergone a similar fate.

Mars is the fourth planet out from the Sun, separated by an average distance $a \approx 1.52$ AU. It has an equatorial radius $R \approx 0.532R_\oplus$ and a mass $M \approx 0.107M_\oplus$. Of all other objects in the Solar System, the surface of Mars is most similar to the surface of the Earth. Mars shows clear signs of seasons, with changes in its polar icecaps and alterations in the terrain triggered by strong windstorms. It has an atmosphere, comprised primarily of carbon dioxide and argon, but the air is exceedingly thin, with an atmospheric pressure on the surface only about one percent the pressure of air at sea level on the Earth's surface. It is evident that, at earlier times, Mars had substantially more air than it does at the present, and that there was flowing water on the surface. It is believed that, because Mars' gravity is relatively weak, the air slowly escaped into space over the course of billions of years. This caused most of the water to evaporate, at which point it also escaped into space. Even today, there is at least a little water in the crust of Mars, both in the icecaps and stored as permafrost. Because it has little air, the temperature of Mars varies significantly, between -220 F or 133 K and 70 F or 293 K. During its daylight hours, Mars is the only place in the Solar System, aside from the Earth, where you could stand unprotected for any period of time without being killed by the environment.

The first four planets, Mercury, Venus, Earth, and Mars, known collectively as the *terrestrial* (or Earthlike) *planets*, are relatively similar in size and composition. Thus, in particular, all four have a solid rocky crust although, in the case of the Earth, about two thirds of that crust is covered by water. They are also similar in that they are *not* orbited by large numbers of moons. Neither Mercury nor Venus has any moons. The Earth, of course, has its own Moon but, from the viewpoint of astronomy, it is rather special: because the Moon is not that much smaller than the Earth, most astronomers and planetary scientists view the Moon as a sort of sibling planet of the Earth, rather than a typical moon. Mars has two tiny moons, Phobus and Deimos, each of which is only a few

kilometers across. They may be asteroids captured by Mars' gravity.
The next four planets, Jupiter, Saturn, Uranus, and Neptune, are relatively similar to each other but very different from the inner terrestrials. Each of these so-called *Jovian* (or Jupiter-like) *planets* is much larger than the Earth. Like the Sun, and unlike the terrestrial planets, they are made primarily of gaseous and liquid hydrogen and helium, although, unlike the Sun, each of them probably has a small, rocky earthlike core in its center. The Jovian planets are also distinguished by the fact that each of them has a large number of moons, all much smaller than the planets which they orbit. It has long been known that Saturn is surrounded by rings. In the past several decades, it has been discovered that the three other Jovians are also surrounded by rings, although their rings are much less conspicuous than those around Saturn.

Jupiter, the fifth planet out from the Sun, is separated by an average distance $a \approx 5.20$ AU. It has an equatorial radius $R \approx 11.3R_{\oplus}$ and a mass $M \approx 318M_{\oplus}$. Its mass is larger than all the remaining planets combined, although it is still less than one thousandth as massive as the Sun. If one defines the "surface" of Jupiter as corresponding to the tops of the clouds of ammonia floating in the hydrogen and helium, one infers that it has a surface temperature of about -110 C or 163 K. As of 1994, Jupiter had sixteen known moons, ranging from tiny Leda, about 4 km in radius, to Ganymede which, with a radius of about 2640 km, is larger than Mercury. None of these moons is visible without a telescope. However, the four largest of these moons – Io, Europa, Callisto, and Ganymede – are easily visible with even a small telescope, and were discovered by Galileo in the first decade of the 1600's.
Saturn, the sixth planet out from the Sun, is separated by an average distance $a \approx 9.54$ AU. It has an equatorial radius $R \approx 9.44R_{\oplus}$ and a mass $M \approx 95.2M_{\oplus}$ The mean surface temperature, again at the cloudtops, is about -180 C or 93 K. It has eighteen known moons, the smallest of which is Pan, about 10 km in radius. The largest of Saturn's moons, Titan, about 2575 km in radius, is interesting because it has a dense atmosphere of argon and methane, containing significant abundances of various hydrocarbons ("smog").
Uranus, the seventh planet out from the Sun, is separated by an average distance $a \approx 19.18$ AU. It has an equatorial radius $R \approx 3.98R_{\oplus}$ and a mass $M \approx 14.5M_{\oplus}$ The mean surface temperature is about -216 C or 57 K. It has fifteen known moons, the smallest of which – Cordelia, Ophelia, and Bianca – are about 25 km in radius.
Neptune, the eighth planet out from the Sun, is separated by an average distance $a \approx 30.06$ AU. It has an equatorial radius $R \approx 3.88R_{\oplus}$ and a mass $M \approx 17.2M_{\oplus}$ The mean surface temperature is again about -216 C or 57 K. It has eight known moons, the smallest of which is Naiad, about 30 km in radius.

The Earth's Moon and all of these planets have been studied by unmanned probes which flew past the planets, and two of the planets have actually been visited by unmanned probes. On several occasions during the period 1969 - 1972, American astronauts landed on and explored the surface of the Moon. Mercury was studied by several different Mariner probes, which flew by the planet starting in 1974. Venus was investigated more directly by a series of unmanned Venera probes, launched by the Soviet Union between 1975 and 1982. Each of the Veneras actually landed on the surface of the planet and sent back radio signals for a period of about an hour, before melting and corroding in Venus' atmosphere.

Mars was studied by two Viking probes that landed on the surface of the planet in 1976. This probe included a robot which made some simple experiments looking for signs of life: none were found. A more recent probe also failed to find any unambiguous evidence for life. The Jovian planets were all probed by Voyager I and/or Voyager II between 1979 and 1989. In particular, Voyager II reached Jupiter in 1979, Saturn in 1981, Uranus in 1986, and Neptune in 1989. Pluto, the ninth planet in the Solar System, has not yet been probed up close, and there are no plans to launch any probes in the next decade or so.

Pluto, at 39.44 AU the most distant planet from the Sun, is an oddity. It is by far the tiniest planet, with an equatorial radius $R \approx 0.18 R_\oplus$ and a mass $M \approx 0.002 M_\oplus$, which makes it substantially smaller than the Earth's Moon. Quite different from both the terrestrial and the Jovian planets, it is relatively similar to some of the moons orbiting the Jovian planets. However, Pluto is orbited by its own moon, Charon, which was first discovered in 1978. It is the coldest of the planets, with a temperature of about -223 C or 50 K.

On a dark night, Mercury, Venus, Mars, Jupiter, and Saturn are easily visible to the naked eye and, because of their motions in the sky, they have been recognised as objects distinct from the stars since ancient times. If you have somewhat better than average vision and are looking at the sky on a particularly dark and clear night, you may also be able to see Uranus. However, it is difficult to track the motion of Uranus in the sky without a telescope and, for that reason, Uranus was not recognised as a planet until 1781. Using an early telescope, in 1612 Galileo was apparently the first person to record the existence of Neptune. However, he did not see it move in the sky and, as such, did not recognise it as a planet.

The actual discovery of Neptune was an important event in the history of astronomy. After Uranus was recognised as a planet in 1781, astronomers regularly tracked its motion in the sky to verify that it was in accord with the laws of physics, as had been formulated by Newton in the late 1600's. However, after several decades they realised that its trajectory was not quite right, as if some other unknown object in the sky, presumably an undiscovered planet, were influencing Uranus by its gravitational pull. In 1845 John Couch Adams, a young student in England, predicted where this other planet should be, and Neptune was discovered in its predicted location in 1846.

The planet Pluto was only discovered in 1930.

The Solar System contains various other flotsam and jetsam in addition to the Sun, the planets, and their moons. *Asteroids* are chunks of rock, believed to be leftover debris from the original formation of the Solar System nearly five billion years ago. There are roughly 100,000 asteroids bright enough to appear on an earth-based photograph, but most of these are exceedingly small, less than a kilometer across. The largest of the asteroids is Ceres, with a radius of nearly 500 km, which accounts for about 30 percent of the total mass in asteroids. There are three asteroids with diameters greater than 300 km, another thirty with diameters between 200 and 300 km, and about 200 more between 100 and 200 km. Ceres was first discovered on 1 January 1801, the first day of the nineteenth century. It, like most of the asteroids, is located between Mars and Jupiter. However, there are other asteroids with orbits outside that of Jupiter and inside that of the Mars. Indeed, some of the asteroids are closer to the Sun than is the Earth.

Comets are ice-covered chunks of rock that typically spend most of their time in the outer reaches of the Solar System, much further away from the Sun than is Pluto. At those large distances, they are cold, dark objects, difficult to see. However, when a comet moves in towards the Sun, parts of the icy surface are vaporised, leading to the bright coma and tail that we observe.
Meteroids are bits of debris that derive both from asteroids and comets.

The orbits of the planets, moons, and asteroids all exhibit considerable regularities. All the planets and all the asteroids revolve around the Sun along orbits that are approximately planar and nearly circular, moving in the same direction, and the planes associated with the different orbits all rather nearly coincide. All the planets rotate around their axes, and, with two exceptions, Venus and Uranus, the axis of rotation is nearly perpendicular to the plane in which the planet orbits the Sun. Moreover, the direction of rotation is the same as that associated with the orbital motion. In addition, most of the moons revolve around their planets along planar, near-circular orbits, again nearly in the same plane and again revolving in the same direction. In this sense, most of the Solar System corresponds to a flat, nearly planar entity with a net sense of rotation. Comets provide a striking exception to these regularities, in that they have elongated, highly non-circular orbits which are not concentrated in or near the plane associated with the planets, the moons, and the asteroids.

Everything in the Solar System is held together by gravity, a force that acts between all objects in the Universe, with a strength that depends on their mass. Because the Sun is by far the most massive object in the Solar System, it dominates the motions of the other objects. This implies that, even though every object is, to a greater or lesser extent, moving around every other object, it is convenient and natural to speak of the planets, the comets, and the asteroids as orbiting around the Sun.

The Sun is a fairly typical star. It generates energy via nuclear fusion of hydrogen into helium, and it is this energy, radiated into space and subsequently absorbed by the planets, that is responsible for heating the Solar System as a whole. Accordingly to our best understanding, the Sun is about 5 billion years old; it is believed that it will continue to exist in its current form for another 5 billion years.

Aside from the Sun, the nearest stars to us are Alpha Centauri A and B, which are about 4.3 light years away, i.e., it takes light emitted by these objects about 4.3 years to travel through space and reach the Earth. To appreciate how large a distance that is, and to get a sense of just how much empty space is out there, note that it only takes about 500 seconds – a little over eight minutes – for light to travel from the Sun to the Earth and only about 330 minutes – less than six hours – for light to travel from the Sun to Pluto.
The Sun is one of roughly 100 - 200 billion stars contained within our galaxy, the so-called *Milky Way*. The Milky Way is a typical spiral galaxy, comprised of two more or less distinct components, namely a *disc component* and a *spheroidal component*. To get a rough sense of its shape, envision the disc component as a medium pizza and the spheroidal component as a beach ball which has been sliced in half, with one half centered above the pizza and the other half below.
In addition to the stars, the Milky Way galaxy contains a large amount of gas and dust, the whole system being held together by gravity. On the basis of the observed motions of

the stars and gas, one can infer how much gravity is acting and, given this information, one can try to estimate how much mass must be in the galaxy. By so doing, astronomers arrive on a galactic mass somewhere between 100 billion and 1 trillion times the mass of the Sun, i.e., $1 - 10 \times 10^{11} M_{\odot}$, with $2 - 5 \times 10^{11} M_{\odot}$ being perhaps the most likely range. This total mass is substantially larger than what can be accounted for in terms of what is actually seen, which suggests that there must be a fair amount of *dark matter* which has not yet been detected. This is one manifestation of the so-called *missing mass problem.*

Some of the stars, like the Sun, orbit the galaxy independently, not bound in any other structure. However, many of the stars are clumped together into smaller units, called open clusters and globular clusters. Globular clusters are striking, nearly spherical collections of up to a million stars, with most of the stars concentrated in a central region with a diameter between 30 and 60 lightyears. Open clusters are less conspicuous collections of tens to hundreds of stars, with a diameter of about 15 to 30 lightyears. In addition to this relatively large scale clumping, it is observed that many stars are in binary or triplet systems, in which two or three stars are bound closely together because of their mutual gravity. The two stars Alpha Centauri A and B are members of such a binary system.

The disc and spheroidal components of our galaxy are very different in composition. Overall stars in the disc are much younger than stars in the spheroidal component. The open clusters are all contained in the disc, along with almost all the dust and gas. One striking feature in the disc is the *spiral arms*, places where, even today, gas is being converted into new stars. The spheroidal component contains almost no dust and gas and no young stars. The globular clusters, distributed throughout the spheroidal component, contain the oldest stars in the galaxy, which are more than 10 billion years old.

The Milky Way is one of about two dozen galaxies clumped together in the *Local Group*, which is about 3 million lightyears across. The Milky Way and Andromeda, the two largest galaxies in the Local Group, are both spiral galaxies. However, there exist other types of galaxies as well.

Overall, the most common type of galaxy is probably an *elliptical galaxy*, which appears in the sky as a nearly featureless blob of stars. In a first approximation, an elliptical galaxy looks like the spheroidal component of an elliptical galaxy, from which the disc component has been detached. Like the spheroidal components of spirals, elliptical galaxies contain little dust or gas and almost no young stars.

In addition to spirals and ellipticals, there are the so-called *irregular galaxies*, so named because they look more irregular in appearance. Like spirals, these galaxies tend to have lots of gas and to contain both young and old stars. However, they do not evidence any conspicuous structures like spiral arms.

Just like stars, galaxies exhibit a strong tendency to clump into collections of objects, again in response to their mutual gravitational attraction. In certain cases, like the Local Group, the clumps are relatively small, containing only a few galaxies. However, sometimes they are much larger and more conspicuous. The *Virgo Cluster*, the nearest rich cluster, is about 50 million lightyears away and contains about 2500 galaxies that are bright enough to be seen from the Earth, within a diameter of about 7 million lightyears. Even more striking is the *Coma Cluster*, a nearly spherical collection of more than 10,000 galaxies

about 350 million lightyears away.
It is an interesting, and not completely understood, fact that different types of galaxies tend to be found preferentially in different types of environments. Rich, regular clusters containing large numbers of galaxies tend to be populated primarily by elliptical galaxies, especially in the central regions. Less densely populated clusters tend to contain higher abundances of spiral galaxies. It is believed that this is somehow related to the fact that most galaxies in rich clusters have been involved in one or more galaxy-galaxy collisions during the history of the Universe.

Using the best telescopes, astronomers can detect galaxies that are billions of lightyears away and use them as probes of the large scale structure of the Universe. Thus, e.g., cosmologists have already charted over 2700 different galaxy clusters within 4 billion light years of the Earth. However, one important thing to recognise is that, as one looks at increasingly distant objects, he or she is also looking into the past, when the Universe was younger and, apparently, different than it is today. Thus, e.g., when one sees the Coma cluster today, he or she is really seeing it the way it was some 350 million lightyears ago.

One of the most striking facts to have been observed about galaxies is that, apparently like almost all objects outside the Local Group, they tend to be moving away from us in a fashion described by the so-called *Hubble Law*. Specifically, it has been found that, on the average, a galaxy that is located a distance r from the Milky Way will be moving away from the Milky Way at a speed v which satisfies a simple linear relation, $v = Hr$. Here H denotes the so-called *Hubble constant*. It is to this fact which cosmologists refer when they state loosely that *the Universe is expanding*. Determining the precise value of this Hubble constant is currently one of the hottest – and most contentious – areas of research in cosmology.

Although one can see galaxies that are over a billion lightyears away, the most distant galaxies are simply too faint to be observed with current telescopes. The most distant objects which we see are *quasars*, which can be detected at distances greater than 10 billion lightyears. These quasars, which can shine hundreds of times brighter than current galaxies, are generally recognised as the precursors of the galaxies that we see today or, more precisely, as bright nuclei located in the centers of newly born galaxies The fact that we see no nearby quasars is interpreted as evidence that several billion years ago the Universe was substantially different than it is today.

One important ingredient in constructing an overall picture of the Universe and our place within it is the determination of how old various objects are. Here three different tacts play an important role:
1. To analyse the ages of solid rock and other similar materials one can use so-called *radioactive dating*. This technique implies that the oldest rocks on the surface of the Earth are about 3.9 billion years old. Because the crust of the Earth is a dynamic, changing entity, this does not imply that the Earth itself, or even its crust, is itself 3.9 billion years old. Rather, what it implies is that the Earth must be *at least 3.9 billion years old*. Examination of rocks brought back to the Earth from the Moon shows that some of these are older, as much as 4.48 billion years in age. This is interpreted as reflecting the fact that the Moon's crust is less dynamic than the Earth's. The oldest rocks to have been

analysed come from meteorite debris and give ages as large as 4.6 billion years.
2. The ages of stars can be estimated theoretically, using models that have been constructed to explain their current structure. These models allow one to do two different, and important, things. The first is to compute the overall brightness and total lifetime of a star as a function of its mass, the basic conclusion being that more massive stars burn much brighter and have substantially shorter lives. The second is to determine how the visual appearance of a star changes over the course of its lifetime. It is on the basis of the second of these that astronomers infer that the Sun is about 5 billion years old. The first of these is used to estimate the ages of globular clusters and other collections of stars. Globular clusters are distinguished by the absence of bright, massive stars, and a determination of the brightest stars that *do* exist is used to provide an estimate of the age of the cluster. Estimates for their age range between 10 and 16 billion years. In the early to mid 1990's, 12 - 14 billion years or seemed to be the mostly likely range, but the interpretation of more recent observations suggests that ages of 10 - 12 billion years may be more reasonable.
It is important that radioactive dating and stellar modeling provide two completely distinct ways of estimating ages. The fact that the oldest rocks in the Solar System appear to be comparable in age to, but slightly younger than, the Sun is an important confirmation of the currently accepted theory that the entire Solar System was created about the same time, roughly 5 billion years ago.

3. The fact that galaxies in the Universe are moving apart implies that, at earlier times, they were closer together. If one retrodicts backwards from what is observed today, one infers that, at a finite time in the past, everything in the Universe was piled up on top of each other, and that the Universe has expanded from an original much denser state. Estimates as to the time since the beginning of this expansion range between 10 and 20 billion years, depending on the precise value of the Hubble constant and other observable quantities.

According to the currently accepted cosmological model, the Universe began some 10 to 20 billion years ago in a hot, dense Big Bang, and everything has been blowing apart and cooling off since then. If one looks in otherwise empty space, one observes that the Universe is innundated in a bath of cold microwave radiation, with a temperature about 2.7 degrees Kelvin above absolute zero. This radiation is interpreted as a relict of the primordial fireball. At present it is not clear whether the Universe will continue to expand forever or whether instead it will eventually recontract back towards a Big Crunch. However, within the context of the standard Hot Big Bsng model, this is a question that can in principle be answered, given sufficiently accurate observations.

Mathematical Appendix: Temperature Scales

To convert from T_F, degrees Fahrenheit, to T_C, degrees Centigrade: first subtract 32 and then multiply by 5/9, i.e., $T_C = \frac{5}{9}(T_F - 32)$
To convert from T_C, degrees Centigrade, to T_F, degrees Fahrenheit: first multiply by 9/5 and then add 32, i.e., $T_F = \frac{9}{5}T_C + 32$.
To convert from T_C, degrees Centigrade, to T_K, degrees, Kelvin, add a number approximately equalling 273.1, i.e., $T_K \approx T_C + 273.1$.

Lectures 3 and 4

ARCHAEOASTRONOMY AND GREEK ASTRONOMY AND THE GREEK COSMOLOGICAL WORLD VIEW

An analysis of historical records indicates that astronomy originally developed in response to two human needs. The first, and more pragmatic, was the need by agrarian societies to obtain accurate predictions of the seasons and other yearly events, e.g., the flooding of the Nile in ancient Egypt. The second, more esoteric, was the apparent human need to identify humanity's place in the Universe. Because of this second motivation, the oldest astronomy/cosmology was closely allied with religion, sharing the same imagery. In particular, most if not all ancient cultures appear to have developed cosmological models which place the Earth at the center of a relatively small Universe.

A good example thereof is provided by the following description of the Egyptian Cosmos, due to Koestler:

> "The universe of the Egyptians was a more rectangular oyster or box; the earth was its floor, the sky was either a cow whose feet rested on the four corners of the earth, or a woman supporting herself on her elbows and knees; later, a vaulted metal lid. Around the inner walls of the box, on a kind of elevated gallery, flowed a river on which the sun and moon gods sailed their barques, entering and vanishing through various stage doors. The fixed stars were lamps, suspended from the vault, or carried by other gods. The planets sailed their own boats along canals originating in the Milky Way, the celestial twin of the Nile. Towards the fifteenth of each month, the moon god was attacked by a ferocious sow, and devoured in a fortnight of agony; then he was re-born again. Sometimes the sow swallowed him whole causing a lunar eclipse; sometimes a serpent swallowed the sun, causing a solar eclipse. But these tragedies were, like those in a dream, both real and not; inside his box or womb, the dreamer felt fairly safe."

Historical records indicate that the ancient Babylonians and Hebrews both had fundamentally similar cosmologies. Other cultures had somewhat different cosmologies, but even these all shared the common thread of being closely allied with religion and theology. This is, e.g., consistent with the ideas of Joseph Campbell, who would argue that a major (if not *the* major) aim of ancient religion and mythology was to connect the microcosm of the individual and the macrocosm of the Universe as a whole via the mesocosm of an intervening social structure.

In many different cultures, including the Middle East and the Far East, astronomical record keeping went back a very long time, in some cases thousands of years; and, over the course of such long periods of time, led to extremely accurate predictions. For example, Chaldean clay tablets dating from 3800 BC already show a well-established *old* astronomical tradition. By 750 BC or so, the Babylonians knew the length of the *sidereal year* (i.e., how long it takes the earth to revolve around the sun relative to distant stars) to much better than one part in a thousand. Indeed, the errors of the Babylonian astronomers were only three times as large as the errors of western astronomers working during the nineteenth century, who were already equipped with relatively sophisticated telescopes.

In this sense, ancient astronomy was an accurate, predictive discipline, even though it was motivated by, and couched in, the language of religion and mysticism.

There are seven particularly conspicuous objects in the sky, visible to the naked eye, which move relative to the remaining stars, that appear frozen in the sky. These are: (1) the Sun, (2) the Moon, (3) Mercury (Nebo in the ancient Middle East), (4) Venus (Istar), (5) Mars (Nergal), (6) Jupiter (Marduk), and (7) Saturn (Nonib). These objects do not wander all over the sky but, instead, are confined to a relatively narrow band in the sky, the so-called belt of the Zodiac. Today we know that this is because the Earth and all the other planets orbit the Sun in approximately the same plane, and that the plane in which the Moon orbits the Earth is again almost the same. Much/most of astronomy and cosmology in ancient times was devoted to the problem of characterising and predicting the motions of these seven objects.

Early Greek astronomy, e.g., at the time of *Homer* (the author of *The Iliad* and *The Odyssey*) was not very different from the astronomy of the Egyptians and the Babylonians, i.e., an admixture of mysticism and astrology, rather than anything approaching a science. However, during the sixth century B. C., the same century that saw Confucius, Buddha, and Lao-Tse, Greek culture underwent a significant transformation that led to the earliest known semblence of what might reasonably be termed science. Stated simply, the key new ingredient was the idea that the Universe is susceptible to rational thought, and, specifically, that there are immutable physical laws, accessible to human thought, that govern how the Universe works. It is this transformation of viewpoint which is usually interpreted as corresponding to the birth of astronomy as a *science*, as opposed to a collection of ideas combining mysticism, religion, and astrology.

The most influential of the early Greek astronomers were members of the Ionian school of philosophy:

The first of these was *Thales of Miletos* (624 - 547 BC or 640 - 546 BC), best known as a mathematician, who is credited with introducing the concept of abstract geometry, which led inexorably to Euclid's geometry, as is still taught in high school today. Thales is best known in astronomy textbooks as the first recorded individual to have predicted a total solar eclipse before it happened. However, from the viewpoint of fundamental science he is more important as the first influential Greek to have clearly articulated the ideas (1) that the Universe is governed by fixed laws and (2) that humans can, at least in principle, ascertain what these laws are.

Thales himself had a cosmology very similar to that advocated by Homer, in which the Earth was viewed as a disc floating on water. In this sense, he fits very much in the Egyptian/Babylonian tradition. However, he took a significant step beyond that tradition when he went on to ask: What is the primal stuff of which the Universe is made? Most of his ideas seem arcane from a modern perspective. However, his interest in causes and antecedents was an important influence on subsequent Greek thinkers, and hence played a crucial role in the evolution of Greek astronomy and cosmology.

A second Ionian, *Anaximander* (c. 611 - 546 BC), was, to the best of our knowledge, the

first thinker (at least in the West) to articulate concretely the idea that the Universe is infinite in both extent and duration. In responding to Thales' question, he argued that the raw stuff of the Universe is something more primal than anything that we recognise around us. Out of this stuff, he believed, all things developed and into it all things will return. He believed further that, before our current world, infinitely many other Universes existed; and that, in the future, after this Universe ends, infinitely many other Universes will exist. It is interesting to note that such a Phoenix Cosmology has been resurrected in the twentieth century, in the context of the Hot Big Bang model.
Anaximander is also credited with having constructed the first concrete, *mechanistic* model of the Universe, in which, e.g., the boat of the sun god Ra is replaced by the wheels of a clockwork. A good description of this model is provided by Koestler:

> "The earth is a cylindrical column, surrounded by air; it floats upright in the centre of the universe without support or anything to stand on, yet it does not fall because, being in the centre, it has no preferred direction towards which to lean; if it did, this would disturb the symmetry and balance of the whole. The spherical heavens enclose the atmosphere 'like the bark of a tree,' and there are several layers of this enclosure to accomodate the various stellar objects. But these are not what they seem, and are not 'objects' at all. The sun is merely a hole in the rim of a huge wheel. The rim is filled with fire, and as it turns round the earth, so does the hole in it – a puncture in a gigantic tyre filled with flames. For the moon we are given a similar explanation; its phases are due to recurring partial stoppages of the puncture, and so are the eclipses. The stars are pin-holes in a dark fabric through which we glimpse the cosmic fire filling the space between two layers of 'bark.' "

Anaximenes, an associate of Anaximander, is responsible for another important contribution, namely the idea that the stars are attached like nails to a transparent sphere in the sky which rotates around the earth. Although bizarre from a modern perspective, this was an extremely influential idea that dominated cosmology from the time of the Greeks until the beginning of modern times.

Other competing cosmological models were proffered by different schools of Greek philosophy, notably the Eleatic philosophers. Like the Ionian cosmologies, each of these seems bizarre from a modern perspective. However, the one important feature unifying them all is the fact that they involve no "sun-eating serpents" or other mystical interventions. Rather, there is the underlying belief that there exist *natural causes* which can explain why the Universe is the way that it is.

The next major innovations were associated with *Pythagoras of Samos*, and his students:

Pythagoras (c. 570 - 500 BC), originally from Samos, moved to Kroton, a Greek town in southern Italy, and founded the so-called Pythagorean Brotherhood. At one point, this Brotherhood had considerable political power but, by the end of his life, Pythagoras had been banished from the town, his students and disciples had been killed or exiled, and his meeting houses had all been burnt to the ground. It is not hard to see why he might have annoyed the locals: Pythagoras was believed by his followers to be semi-divine, son of the Hyperborean Apollo. According to Herodutus, and to other ancient Greek historians, Pythagoras was reputed to have worked miracles, conversed with demons from heaven, and, upon one occasion, to have descended into Hades before returning to Earth. It is also known that he wielded complete and absolute authority amongst his disciples. Aside from these eccentricities, the Brotherhood probably ran afoul of authority because of its controversial political philosophy, which included a belief in socialism/communism (rather akin to the Utopian Socialists of the nineteenth century AD), vocal advocacy of women's equality, and a quasi-monotheistic religious doctrine.

The Pythagorean world view was an all-encompassing one, which aimed to unify all aspects of reality. Several important words and concepts which we use today trace back to the Pythagorean school. Thus, e.g., the word "philosopher" (Greek for lover of knowledge) is Pythagorean in origin, as is the use of the word "harmony" in its broadest sense (transcending music). Refering to numbers as "figures" is again Pythagorean in origin.

The all-embracing Pythagorean philosphy aimed to unify what seemed disparate, if not contradictory, ideals: religion and science; mathematics and music; medicine and cosmology; body, mind, and spirit. One concrete example, of fundamental importance to the Pythagoreans, was the connection between music and mathematics. Specifically, they recognised that, when blowing into a pipe or plucking a string, the pitch of the resulting note depends on the length of the pipe or string; and that concordant intervals are produced by simple ratios of length, e.g., the octave (2:1), the fifth (3:2), the fourth (4:3), and so on. (This notion is still important today, although the the intervals were modified slightly with the advent of complicated harmonies on keyboards and the development of well-tempered scales.)

From the viewpoint of astronomy and cosmology, and indeed all subsequent science, the key ingredient of the Pythagorean viewpoint was the idea that mathematics constitutes the fundamental language of the Universe. The Pythagoreans did not feel that translating the Universe into mathematical language, and hence identifying its regularities, would cheapen it in any way. Quite the contrary: it was key to their philosophy that expressing other things in the Universe – astronomy, music, etc. – in the language of mathematics would only serve to ennoble them. The basic tact of their philosophy was summarised by Koestler as follows: "All things have form, all things *are* form, and all forms can be defined by numbers."

The Pythagoreans also advocated the concept of "The Harmony of the Spheres." Basically, they believed that the orbits of the planets, the moon, and the sun travelling around the earth represent the strings of a gigantic lyre curved into gigantic circles, with the intervals between the strings at concordant musical intervals. Thus, e.g., the Pythagorean scale in music – C, D, E-flat, G, A, B-flat, B, C – is based on the presumed distances between the orbits of these different heavenly objects. Pythagoras himself was able to *hear* the music of the heavenly spheres, although his disciples apparently did not share this unique ability. All this sounds nonsensical from a modern perspective, but, nevertheless, this viewpoint served as the motivation and framework in terms of which Johannes Kepler eventually constructed the solid edifice of modern astronomy in the early 1600's.

It should perhaps be noted that the idea of heavenly spheres did *not* originate with the Pythagoreans. Rather, it was an essential ingredient of many early mythologies, including those of the Near East, India, and China. Even the connection with music seems more universal, also arising, for example, in ancient Chinese culture, where music was viewed as a harbinger of both good and bad fortune. (It should also be noted that Pythagoras was *not* the first person to come up with the so-called Pythagorean Theorem: $a^2 + b^2 = c^2$. It would appear that the Babylonians were aware of this and other similar geometric formulae as early as 1800 BC.)

The Pythagoreans were very much enamoured with the notion of perfect shapes, e.g., circles and spheres. This played a considerable role in the subsequent ideas (1) that the Earth is a sphere and (2) that the orbits of the heavenly bodies must all be circles, ideas that not even Copernicus was able to shake off.

The Pythagoreans saw the study of "pure science" *both* as an intellectual delight *and* as a spiritual release, a purification of the soul, the latter in much the same sense that the Buddhist practises nonattachment. (The word the Pythagoreans used was *katharsis*, which translates into English as *catharsis.*) Koestler would argue that "The Pythagorean concept of harnessing science to the contemplation of the eternal entered, via Plato and Aristotle, into the spirit of Christianity and became a decisive factor in the making of the Western world."

Up to this point, all the Greek philosophers had envisioned *geocentric* cosmologies, in which the Earth was the center of the Universe and everything else in the Universe orbits about the Earth. *Philolaus* (5th century BC), a pupil of Pythagoras, is credited as the first to advocate the idea that the Earth may in fact move in space. Philolaus did *not* introduce the idea that the Earth rotates around its axis once a day, or that the Earth revolves around the Sun once a year. Rather, he had a more complex model, whereby the Earth revolves around an extraneous fixed point in space once every 24 hours. This has almost the same effect observationally as having the earth rotating around its axis: by completing a circular orbit in one day, an observer on the Earth has the illusion that everything else orbits around us once a day. Philolaus' model therefore provided for the first time a clear distinction between *diurnal* (once a day) motions and *annual* (once a year) motions.

Herakleides of Pontus (c. 375 - 310 BC), a student of Plato, built upon the ideas of Philolaus and is the first known philosopher/astronomer to articulate the idea that (at least some of) the planets orbit the Sun, rather than the Earth. Herakleides believed correctly that the earth rotates around its axis once a day, and that this explains the observed diurnal motions of the heavenly bodies. In his cosmology, the stars were stuck to what was essentially a "cosmic pin-cushion" which is frozen in the sky, presumably on a celestial sphere, but which appears to move around the Earth once a day because of the Earth's rotation.

The remaining question was then to account for the motions of the moon, the sun, and the other planets, as they move through the sky relative to the fixed stars. In particular, the most complicated task was to explain the so-called phenomenon of *retrograde motion*.

Today we know that all the planets orbit the Sun along trajectories which are nearly, but not exactly, circular, and that the *orbital period* (i.e., how long it takes to make one revolution around the Sun) increases as one moves further away from the Sun. This implies that an observer looking at the Solar System from the perspective of the Sun would see something quite simple. However, an observer on the Earth will see more complicated motions whereby, in particular, for part of its orbit a planet can appear to switch directions in the sky and move in an opposite, retrograde sense. Retrograde motion is easy to understand within the framework of a sun-centred *heliocentric* model of the Solar System. However, within the context of a geocentric model, explaining retrograde motion is a real problem.

In Herakleides' early cosmology, the Moon and the Sun are both taken to orbit the Earth. However, the remaining planets are *not* all assumed to orbit the Earth. Rather, at least some of them are assumed to orbit around the Sun. Thus, in the earliest versions of his cosmology, the inner planets – Mercury and Venus – are visualised as satellites orbiting the Sun in much the same sense that the Moon is a satellite orbiting the Earth, although the outer planets – Mars, Jupiter, and Saturn – are assumed to orbit around the Earth. The idea that Mercury and Venus are especially tied to the Sun is very natural from an observational perspective: because they are physically proximate to the Sun, they are only seen in the sky near the Sun, just after dusk or shortly before dawn (hence their designation in several different cultures, including Babylonian, Chinese, and Aztec, as *morning* and *evening stars*).

In the later versions of his cosmology, Herakleides seems to have proceeded further by assuming that the outer planets also orbit the Sun. The net result is a world model which has the Earth at the center, with the Moon and Sun as satellites orbiting the Earth, and with the remaining planets as other satellites orbiting the Sun.

Aristarchus of Samos (c. 310 - 230 BC), the last Pythagorean astronomer, was apparently the first to take the final step towards a true heliocentric cosmology by articulating the idea that the Earth, like the remaining planets, actually orbits the Sun. Unfortunately, most of Aristarchus' writings have disappeared. However, a number of ancients have recorded descriptions of his cosmology which still exist today. Thus, e.g., the ancient Greek biographer Plutarch (AD 46? - 120?) wrote that, according to Aristarchus, "the heaven is at rest, but ... the earth revolves in an oblique orbit, while it also rotates about its own axis." The obliquity to which Plutarch refers is the fact that the axis about which the Earth rotates once a day is tilted by about 23 degrees relative to the plane in which the Earth orbits the Sun, a fact which we now understand as explaining the Earth's *seasons*.

Aristarchus was apparently one of the most respected Greek thinkers of his time and, as such, one might have supposed that his heliocentric model would have gained widespread acceptance. This, however, was not the case. Rather, his system was rejected emphatically by his successors in favor of a more complicated, and arcane, geocentric model of the Universe. Precisely why this was so is something about which historians of science still argue. However, what *does* seem clear is that, even before the time of Aristarchus, there had been a fundamental change in the intellectual climate of the times, constituting a rejection of the empirical attitude of such ancient scientists as Aristarchus, Hippocrates (considered the father of medicine), and Archimedes (a great mathematician, physicist, and engineer), in favour of an anti-empirical attitude, typified by the philosopher Plato and his school of thought.

The empirical attitude is typified by the following comments by a Hippocratic physician writing about the disease of epilepsy, which was a great mystery to the ancient Greeks:

"It seems to me that the disease [epilepsy] is no more 'divine' than any other. It has a natural cause, just as other diseases have. Men think it divine merely because they do not understand it. But if they called everything divine which they did not understand, why, there would be no end of divine things!" By contrast, Koestler summarises the attitude of Plato as follows: "The stars, he explains, however beautiful, are merely part of the *visible* world which is but a dim and distorted shadow or copy of the *real* world of ideas; the endeavor to determine exactly the motions of the imperfect bodies is therefore absurd." Plato himself wrote: "let us concentrate on (abstract) problems, said I, in astronomy as in geometry, and dismiss the heavenly bodies, if we intend truly to appreciate astronomy." In other words, Plato believed that looking at the heavenly bodies would provide no useful insights into astronomy! By contrast, Plato and his followers were *very* interested in geometry. Indeed, the entrance to Plato's Academy was inscribed with the words: "let no one destitute of geometry enter my doors."

It should perhaps be stressed that *Plato* (428 - 347 BC) had essentially no knowledge of astronomy, as it was understood at the time of his birth, and that he appeared to find the subject completely boring. However, proceeding purely on philosophical grounds, he articulated the guiding principles that dominated astronomy for almost two thousand years, namely that "*the world must be a perfect sphere, and that all motions must be in perfect circles at uniform speed.*"

Plato's ideas were further elaborated by his student *Aristotle* (384 - 322 BC) who, although he did not completely share Plato's anti-empirical bent, did accept his precept regarding spheres, circles, and uniform motion. In a first approximation, Aristotle's geocentric Universe was based on nine concentric spheres in the sky, centred on the Earth with increasingly large radii. The Moon was attached to the innermost sphere. The eighth sphere was that upon which the fixed stars were attached. The nineth, and most distant, sphere was that associated with the "Prime Mover," presumably Aristotle's version of a deity. The intervening spheres, two through seven, corresponded to the Sun and the other five known planets.

It is significant that, within this cosmology, perfection, i.e., God, is relegated to the most distant sphere, and that the Earth and its associated human imperfection is placed at the center of the Universe. This idea, key to Aristotle's philosophy, was elevated in medieval Christian cosmology/theology into a broad distinction between the imperfect, transitory Earth, and the perfect, immutable Heavens. It is very much counter to the Pythagorean world view, in which God, realised as the primal fire, is placed at the center of the Universe, i.e., in the middle of a spherical Earth.

Aristotle's cosmology is closely related to his physics. Aristotle believed that everything in our environs is constructed as combinations of four basic elements, namely earth, water, fire, and air, which in turn are combinations of hot/cold and wet/dry. Each of these elements inherits a unique *natural motion*: earth to sink towards the center of the Universe, i.e., towards the Earth; fire to float upwards away from the Earth; and air and water to float sideways. These natural motions can be countered by so-called *violent motions*, as caused, e.g., by people. All changes on or near the Earth are interpreted as involving transformations amongst the four elements. By contrast, the perfect, unchanging world of the heavens is populated by an immutable fifth element, not found in our environs. The natural motion of this fifth element is circular with constant speed since, as had been articulated by Plato, the only perfect motion is motion in a perfect circle at a constant speed. (Parenthetically, the notion of a fundamental, perfect meaning, transcending ephemeral appearances, persists in the English adjective "quintessential," which derives from the Latin words for "fifth essence" or "fifth element.")

To the ancients, the number 5 also correlated with the five known planets – Mercury, Venus, Mars, Jupiter, and Saturn –, it being well recognised from the earliest antiquity that the other two of the seven moving objects – the Sun and the Moon – are somehow different from the planets. Indeed, the presumed connection between the five elements and the five planets, and the mystical significance of the number 5, dates back at least to the ancient culture of Sumer, and the city of Ur, from whence the Biblical Abraham originated.

The big problem with this, or any simple geocentric model, is that it cannot account for the retrograde motion of the planets. This was not a problem for Plato, who was not interested in what we see, but it *did* concern Aristotle and his successors, who were interested in constructing a model which, even if unphysical, correctly describes and predicts what is actually observed. To account for retrograde motion, Aristotle modified his simple model of nine concentric spheres by allowing for a complicated, arcane, and confusing collection of spheres within spheres, whereby planets behave essentially like balls trapped between rolling spheres which, by themselves executing circular motion, drag the planets along complicated trajectories. This notion of "spheres within spheres," further amplified by *Eudoxus* (409 - 356 BC), another student of Plato, was so bizarre that even Aristotle's enormous personal prestige could not gain it acceptability. However, "spheres within spheres" was eventually replaced by another monstrosity, "wheels within wheels," which constituted the accepted cosmology until the 1600's.

This alternative approach was developed slowly, over a period of more than 300 years, in both Greece and Rome, the principal contributors being *Apollonius of Perga* (flourished 250 - 220 BC), *Hipparchus of Rhodes* (flourished c. 125 BC), and *Claudius Ptolomy* (AD 90 - 168), the last of whom, a follower of the Stoic philosophers, was responsible for codifying the theory in a highly influential book now called *The Alamagest* (a bastardisation of *Al Magistri*, an Arabic translation of the original Greek title, *Megisty Syntaxis* or *Great Composition*). The basic mathematical idea underlying this approach is that any periodic motion can be described as a combination of (a sufficiently large number of) circular motions. Each planet is assumed in a first approximation to move at constant speed along a principal circular orbit, the *deferent*. However, this motion is then perturbed by adding an additional circle, an *epicycle*, which rolls along the deferent, and assuming that the planet is attached to the epicycle. With judicious choices for (1) the ratio of the radius of the deferent to the radius of the epicycle and (2) the circular velocities of motion along each, it is possible to get a reasonable approximation to the observed motions of the planets, including a prediction as to when each planet will execute retrograde motion. The price to be paid for this success is that one is forced to invoke epicycles which orbit about empty points in space.

This sounds crazy enough. However, an additional complication arises: The true orbits of

the planets around the Sun are not exactly circular, describing instead slightly squashed *elliptical* orbits. This implies, however, that to get a reasonable description of what is seen, the deferents must be centred not exactly around the Earth but, instead, about a point in empty space slightly displaced from the Earth (the so-called equant). Moreover, +even this is not adequate to describe the observational data available to the ancient Greeks and Romans, which had been acquired over a period of nearly one thousand years. Instead, they were forced to add additional epicycles, rolling on the primary epicycles, and (according to some historians of science) to cheat with some of the oldest data, in order to get a satisfactory fit.
The final Ptolomaic model involved a total of 39 "wheels" to explain the motions of the seven known moving heavenly objects, plus one wheel to which the immutable stars were attached, this totalling 40 different wheels. By any stretch of the imagination, this would seem an arcane description, perhaps best described by *Alphonso X* of Castille, a Moorish King of the 1200's AD: "If the Lord Almighty had consulted me before embarking upon the Creation, I should have recommended something simpler." Even Ptolomy was modest about what his cosmology achieved: "I do not profess to be able thus to account for all the motions at the same time, but I shall show that each by itself is well explained by its proper hypothesis."

Several comments about this geocentric Universe are in order:
Astronomers working at or before the time of Ptolomy had more precise instruments available for measuring the motions of the heavenly bodies than any other astronomer before Tycho Brahe in the late 1500's, and they had a thousand years of accumulated data with which to compare their models.
The model incorporated in *The Almagest* was sufficiently accurate in predicting the motions of the heavenly bodies that it did not require significant revision for nearly a thousand years. However, by the beginning of the thirteenth century, small but non-negligible errors had begun to accumulate. This led Alphonso X to assemble a collection of medieval astronomers who, over a period of ten years, tweaked the Ptolomaic model to remove the errors. The resulting predictions, assembled in *The Alphonsine Tables*, were sufficiently accurate that they were used for navigation by Chistopher Columbus and Vasco da Gama over two hundred years later. (Parenthentically, Ptolomy had another important influence on Columbus: In his *Geography*, he (1) gives the circumference of the Earth as about $18,000$ miles, much smaller than the correct value of about $25,000$ miles or $40,000$ km, and (2) significantly overstimates the east-west extent of Asia, thereby suggesting to Columbus that sailing west to Asia would entail a much shorter journey than is actually the case.)
It should also be stressed that this geocentric cosmology was accepted in ancient Greece and Rome even though many educated people knew that the heliocentric Universe provided a neater explanation of what one sees and despite an explicit recognition, articulated, e.g., by Plutarch, Pliny, and Cicero, that the Sun "dominates" the motions of the planets.
Finally, it should be stressed that, despite wedding themselves to a geocentric cosmology, the Greeks and Romans had reached some significant, nontrivial conclusions about the size of the Earth and the distance to the Moon. Thus, e.g., Hipparchus had used geometrical arguments to measure the distance to the Moon to better than 1 percent accuracy, concluding that this distance equals approximately $30\frac{1}{4}$ earth diameters. Similarly, *Eratosthenes*

(276 - 195 BC) had determined the diameter of the Earth to within 1 percent or so of the currently known value of $12,756$ km.

Eratosthenes' logic is simple and compelling, even today: He knew that, at high noon on the first day of summer, the Sun rises directly overhead in the town of Syene (now called Aswan, Egypt), which is located almost exactly on the Tropic of Cancer. However, further north in Alexandria, where he lived, at the same time the Sun misses being overhead by some 7 degrees, approximately $1/50$ of a circle of 360 degrees (more precisely, about $1/51.4$). This he was able to determine directly by measuring the length of a shadow cast by a vertical stick stuck in the ground. Eratosthenes then interpreted this difference as resulting from the fact that the surface of the Earth is not flat. Knowing that Syene is a distance of approximately 5000 stadia due south of Alexandria, and assuming that this distance corresponds to $1/51.4$th of the circumference of the Earth, he was led to the conclusion that the Earth's circumference should be about $51.4 \times 5000 = 257,000$ stadia. Assuming that one stadium corresponds to a length of 157 meters or 515 feet (this unit of length comes from the size of an ancient Greek athletic stadium), this translates into a circumference of $40,350$ km or a diameter of $12,840$ km.

It should be noted that, even though the planets and stars are not adfixed to spheres in the sky, the concept of a *celestial sphere* is a useful one in describing the observed locations of objects in the sky. Specifically if, standing on the Earth, one looks out into the sky, it is not unreasonable to visualise various objects in the heavens as if they were attached to the surface of a transparent globe centered on the Earth. Extending into space from the terrestrial North and South Poles one can then identify *celestial North* and *South Poles*, and, extrapolating outwards, one can identify celestial analogues of lattitude and longitude which, respectively, the astronomer terms *right ascension* and *declination*. The apparent location of an object in the sky can then be characterised by specifying its right ascension and declination in the same way that, on the Earth, locations can be given in terms of lattitude and longitude. (Parenthetically, in his *Geography* Ptolomy was apparently one of the first authors to list systematically the locations of places on the Earth in terms of (often erroneous estimates of) lattitude and longitude.)

Finally, it should be stressed that, although arguably the founders of science, the ancient Greeks embraced various ideas which today would be considered decidedly unscientific. For example, both the Greeks and the Romans accepted completely the validity of astrology. Indeed, the strongest attack on astrology as practiced in Greece and Rome appears to have come from *Seneca* (c. 4 BC - AD 65), who argued that the principal reason that many astrological predictions are in error is probably because they are based on an analysis of only a small fraction of the stars in the heavens. Ptolomy himself was responsible for the *Tetrabiblios*, a four volume work which provides an overview of astrology as practiced well into the Middle Ages. Arguably, however, the Greek motivation for considering astrology was not completely unscientific: Given that two of the heavenly bodies – the Sun and the Moon – obviously impact phenomena on the Earth, it was not unreasonable, albeit erroneous, to suppose that the others might as well.

To conclude, one can say that, despite the promising beginnings of the Pythagorean astronomers, Greek and Roman astronomy ended with the development of what Koestler calls a "new mythology," based on three key precepts: "the dualism of celestial and sub-luminary worlds; the immobility of the earth in the centre; and the circularity of all heavenly motion." Exactly why this rigid orthodoxy evolved his hard to say. However, many historians of science, noting that it began to develop at a time when Greece was racked with internal wars (notably Athens versus Sparta), attribute it to a desperate fear of change, "the craving for permanence in a disintegrating culture."

In any event, with the collapse of Rome at the end of the fifth century AD, and the concommitant decay of civilisation as Europe entered into the Dark Ages, even the high points of Greek astronomy were eventually lost. In the sixth century BC, it was well known that the Earth is round; by the sixth century AD, it was believed once again that the Earth is flat!

Lecture 5
THE COPERNICAN REVOLUTION

The quotation by *Thabit* presented in this lecture comes from Chapter 10 of the book *The Renaissance of the Twelfth Century* by Charles Homer Haskins (Meridian Books, New York, 1957).

After the fall of Rome, progress in astronomy and cosmology more or less ground to a halt. There are several reasons for this: In Europe, this period corresponded to the Dark Ages, a time in which, aside from the unifying influence of the Catholic Church, society was fragmented into small pieces and there was little room for intellectual advancement in any field. Further to the East, the tradition of Rome extended through the Eastern Roman Empire (Byzantium), but this Empire was extremely conservative in its structure, jealously maintaining the religious and political orthodoxy of bygone times. Indeed, one can argue that, from the viewpoint of intellectual history, the Dark Ages really began with the anti-pagan edicts of the Christian Emperor Theodosius of Byzantium (ruled AD 379 - 395), which led to a mass exodus of intellectuals – philosophers, scientists, and artists – eastwards to Persia and even India. (The city of Rome itself was only sacked by the Goths some years later on 24 August 410.) Particularly in the West, intellectual thought in general, and astronomy/cosmology in particular, was dominated by the theology of the Catholic Church which, over the course of time, reflected a number of different interpretations of Plato and Aristotle, all of which evolved into rigid dogmas, leaving little room for evolution or growth.

In the early part of the Middle Ages, roughly from AD 400 to AD 1000, some theologians, like *St. Augustine* (354 - 430), born Aurelius Augustinus, embraced various (arguably distorted) interpretations of Plato (Augustine was originally a Neoplatonist). Others, however, like *St. Jerome* (340? - 420), rejected Plato and other Greek thinkers completely because they were pagan. Thus, e.g., Jerome referred upon one occasion to "the stupid wisdom of the philosophers," by which he meant all pre-Christian thinkers.

The cosmology of this period is closely related to that of the Babylonians and Hebrews which, in turn, is similar to that of the ancient Egyptians. Thus, in particular, most thinkers, including St. Augustine, took literally the description of the Universe provided in the Bible, assuming (1) that the Earth is flat and shaped like the Holy Tabernacle and (2) that the firmament is surrounded by waters. The idea that the Earth is roughly spherical was completely rejected.

A good description of the prevailing theories of cosmology at the time is provided by *Cosmas*, a sixth century profligate turned monk originally from Alexandria (Egypt) who, from his cloister on Mt. Sinai, wrote a book entitled *Topographica Christiana*, the first part of which is directed explicitly "Against those who, while willing to profess Christianity, think and imagine like the pagans, that the heaven is spherical."

According to Cosmas' cosmology, the Earth is exactly twice as long as it is wide, stretching lengthwise from East to West, this particular shape corresponding to the shape of the Holy Tabernacle reported in Exodus. The Earth is surrounded by an ocean, and this ocean is in turn surrounded by a second earth, which corresponds to the seat of Paradise. The

edges of the outer earth have four vertical planes, the four walls of the Universe. The roof of the Universe is a half-cylinder resting on the north and south walls, like "a Victorian travelling trunk with a curved lid." The floor of the Earth is not flat, but rather slants downwards from northeast towards southwest, this explaining the fact that, as recounted in Ecclesiastes, "the sun goes down, and hasteth to his place where he rose." For this reason, southward flowing rivers like the Tigris and Euphrates were presumed to have a faster current than northwards flowing rivers like the Nile; and ships were able to sail more quickly towards the southeast than towards the northwest. The stars in the sky were borne through space below the roof of the Universe by assemblages of angels, and hidden behind high mountains during the day. Correspondingly, the same mountains hid the sun at night, the presumption being that the Sun is much smaller in physical size than the Earth.
Cosmas was not a high eccesiastical authority, and his cosmology was never adopted as dogma. However, his cosmology is representative of the type of world view common in the early Middle Ages. Others represented the earth as a disc with Jerusalem in the center since Isaiah spoke of the "circuit of the earth" and Ezekiel stated that "God had set Jerusalem in the midst of the nations and countries."

The Earth began to become round again about AD 700, when the English monk *Bede* (c. 672 - 735) rediscovered the astronomical writings of the Roman naturalist *Pliny*. However, this viewpoint was not generally accepted by the Church until the end of the millenium when, in 999, a talented astronomer/geometer/musician named Gerbert was elected *Pope Sylvester II.* Although Sylvester died in 1003, he had as Pope already succeeded in reestablishing the Greek notion of a spherical Earth, floating in empty space. Coincidentally, around the same time the earlier cosmology of Herakleides had been rediscovered, this corresponding to the world model in which Mercury and Venus orbit the Sun. However, Herakleides' writings did not have a major impact on the thinking of the times.
It is significant that, although the Church had a well defined dogma about astronomy in the early Middle Ages, it did *not* make any serious effort to force its orthodoxy on the freethinkers of the time, provided that they did not disagree on questions of theology. Rather, dissent was typically ignored and orthodoxy was enforced primarily by inner compulsion.

To summarise, one can say that, by about AD 1000, the West had recovered astronomy and cosmology at the level of sophistication of Greece in the fifth century BC. However, there were no immediate further advances since that Greek cosmology subsequently became transformed into a new orthodoxy.

In the early Middle Ages, the work of Aristotle and many other Greek thinkers were not known in the West, and, to the extent that it embraced non-Christians, philosophical thinking was dominated by Plato. However, starting in the 1100's the work of these other Greeks made it the West in the form of Arabic translations, which were conveyed by the Spanish Moors to England and Ireland, from whence they eventually reached Northern and Central Europe. Thus, e.g., Euclid's *Elements*, the foundation of plane geometry, made it to England around 1120, and *The Alamagest* was translated from Arabic to Latin, the universal language of the educated, in 1175.
It is interesting that the Arabs, who had been acquainted with Greek learning for centuries, made few if any advances in theoretical science, although they *did* make significant practical

use of various aspects of what they had learnt, e.g., through the construction of *The Alphonsine Tables*. Indeed, it would appear that most of the scholars writing in Arabic were actually Jewish or Persian. In fact, a strongly intellectual Muslim culture only began to evolve around AD 750, nearly 125 years after Mohammad's death, when the Islamic world came to be dominated by the Persian Abbasids, who made Baghdad the world center for classical learning for several hundred years. However, once the Greek learning reached the West it had a very strong impact, triggering what it sometimes called the first Renaissance of the twelfth century.

From the viewpoint of cosmology and of science in general, perhaps the most important influence of this Renaissance was the increased importance that came to be associated with rational thought. To St. Augustine and other early Christian thinkers, influenced largely by one particular (mis-)interpretation of Plato, reasoning *per se* was a minor thing, the "handmaid of faith." However, to *St. Thomas Aquinas* (1225 - 1274) and other later Christian thinkers, more heavily influenced by another (mis-)interpretation of Aristotle, reasoning was elevated to the more lofty status of "bride of faith." Within the rigid hierarchy of medieval society, this meant that reason was still subservient to faith, its spouse, but it had least acquired the status of an independent being in its own right. The philosophy of Aristotle was also positive from a scientific viewpoint in that, unlike Platonic philosophy, it embraced empiricism through the belief that it is important to describe and characterise what is actually seen.

Unfortunately, however, both these positive features were squelched within a new orthodoxy that evolved through a perversion of Aristotelean thinking, which viewed the writings of Aristotle himself as the ultimate authority and arbiter of truth. Presumably in response for the love of order and hierarchy which characterised all of the Middle Ages, Aristotle was enshrined as an absolute and complete authority, and it was believed (or at least asserted) that every statement about nature to have come from his pen must in fact be true, and that there was no need to make any further investigations of nature since the answers were already known.

The orthodoxy of the times, and the unwillingness of Western intellectuals to accept the evidence of their own eyes, is well illustrated by the twelfth century Syrian physician *Thabit* in his description of an encounter with European medicine:

> They brought to me a knight with an abscess in his leg, and a woman troubled with fever. I applied to the knight a little cataplasm; his abscess opened and took a favorable turn. As for the woman, I forbade here to eat certain foods, and I lowered her temperature. I was there when a Frankish doctor arrived, who said, "This man can't cure them!" Then, addressing the knight, he asked, "Which do you prefer, to live with a single leg, or to die with both of your legs?" "I prefer," replied the knight, "to live with a single leg." "Then bring," said the doctor, "a strong knight with a sharp axe." The knight and the axe were not slow in coming. I was present. The doctor stretched the leg of the patient on a block of wood, and then said to the knight, "Cut off his leg with the axe, detach it with a single blow." Under my eyes, the knight gave a violent blow, but it did not cut the leg off. He gave the unfortunate man a second blow, which caused the marrow to flow from the bone, and the patient died

immediately.

As for the woman, the doctor examined her and said, "She is a woman with a devil in her head, by which she is possessed. Shave her hair." They did so, and she began to eat again, like her compatriots, garlic and mustard. Her fever grew worse. The doctor then said, "The devil has gone into her head." Seizing the razor he cut into her head in the form of a cross and excoriated the skin in the middle so deeply that the bones were uncovered. Then he rubbed her head with salt. The woman, in her turn, expired immediately. After asking them if my services were still needed, and after receiving a negative answer, I returned, having learned from their medicine matters of which I had previously been ignorant.

An earlier example of this innate conservatism, provided by Boorstin, refers to Bede (673 - 735), the English monk, who, although living within sight of Hadrian's Wall, which was constructed by the Romans after their conquest of England, preferred to quote its description by an ancient Roman author rather than describing what he himself had actually seen.
From a modern perspective, the big problem with an uncritical acceptance of Aristotle's ideas was that, regardless of any opinions one may have about Aristotle's purely philosophical writings, it must be recognised that his theories of physics and astronomy are simply wrong. His theory of natural motions, to be discussed in more detail later on in connection with Galileo, involves arcane reasoning which, albeit useful to Thomas Aquinas in providing a proof of the existence of God, contradicts everyday reality. Thus, for example, Aristotle asserted that heavy objects invariably fall more quickly than light objects, and that, in the absence of perturbing influences, all massive objects tend to rest, both of which assertions contradict our everyday experience. As more than one historian of science has emphasised, Aristotle said things that people knew from everyday experience were wrong. However, despite empirical evidence to the contrary, in at least one compartment of their mind they continued to believe what they knew to be false.

As stressed, e.g., by Koestler, by around 1500 there had evolved an elaborate scientific orthodoxy based on several critical tenets, each of which worked against the development of a satisfactory understanding of the Universe:

1. an anti-empirical bias, based on citing authority and ignoring concrete physical evidence;
2. an unquestioning belief in a geocentric Universe, despite substantial evidence to the contrary;
3. an equally unquestioning belief in uniform circular motion;
4. a rejection of the Pythagorean notion of mathematics as the language of science; and
5. the Aristotelean idea that moving bodies tend to sink towards rest, rather than remaining in motion.

It is within the framework of this network of beliefs that the Copernican revolution arose.

Nicholaus Copernicus was born in 1473 in Torun, in what is now Poland, the son of an affluent, patrician merchant. A quiet and secretive man throughout his life, essentially nothing is known about his childhood and youth except that, upon the death of his father, he became a ward of his Uncle Lucas, the future Bishop of Ermland, who eventually arranged for him finanically by having him made Canon of Frauenberg in 1499. Starting in 1491, Copernicus spent fourteen years at Universities in Cracow, Bologna, and Padua, where he received a broad Renaissance education, studying a bit of everything: philosphy, law, mathematics, medicine, astronomy, and the classics. He became a Doctor of Canon Law at the age of 30 in 1503. Despite drawing a salary as Canon, he was not required to do anything or even live in Frauenberg. Rather, upon the completion of his University education he moved to Heilsberg, where he served as physician and private secretary to his uncle, who had by that time been appointed Bishop. Only in 1512, upon the death of his uncle, did he move to Frauenberg and assume his responsibilities as Canon, which basically entailed the financial and bureaucratic management of Frauenberg Cathedral. At no time did Copernicus ever become a monk or a priest: his was a completely secular position.

Copernicus first became interested in the heliocentric theory in Italy where, while studying under several Neoplatonists, he was exposed to the writings of Aristarchus. He apparently began the development of his own cosmological model shortly thereafter, although his great work, *De Revolutionibus Orbium Celestium*, or *On the Revolutions of the Celestial Spheres*, was not published until 1543, the year of his death. According to his own testimony, he had begun work on his theory more than 36 years earlier.

It would appear that, in virtually every respect, Copernicus was an exceedingly cautious, conservative, and colourless man. Born at a very interesting time in his history, after the invention of the printing press but before Columbus had reached the Americas, he lived to see the Protestant Reformation and overlapped with such figures as Luther, Henry VIII, Charles V, Michaelangelo, Leonardo da Vinci, Machiavelli, and the Borgias. In terms of his personality he seems to have been an anachronism, a man of the Middle Ages living in the Renaissance. He apparently believed completely in Aristotle's theories of physics and, despite believing in a heliocentric Universe, constructed a world model based on deferents and epicycles.

Aside from *De Revolutionibus*, he only published one other book in his life, namely a 1509 translation from Greek to Latin of a collection of excruciatingly trite and boring aphorisms by an obscure seventh century AD Byzantine historian named Theophylactus Simocatta. In many respects this choice of text is quite revealing. In Copernicus' youth a knowledge of Greek was considered somewhat subversive, at least in Germany and Poland. This was probably in response to the Catholic theologian *Erasmus*' recent translation of the Greek New Testament into Latin, which had demonstrated explicitly that the canonical Latin Bible, *The Vulgate*, was fraught with numerous translational errors. By 1509, a knowledge of Greek had become accepted as a suitable intellectual accomplishment, so that Copernicus' translation from Greek to Latin would have shown simply that he was on the winning side of one historical controversy. However, to avoid any potential criticism he had selected a relatively modern text, not written by the ancients, which was completely uncontroversial and of which, therefore, no one could conceivably complain.

Around 1512, Copernicus prepared a short, handwritten manuscript, outlining the general ideas underlying his new cosmology, and several copies were quietly circulated amongst his friends and colleagues. Shortly thereafter, he began work on *De Revolutionibus*. He essentially completed the text of this book well before 1530, but then locked it away. He himself made absolutely no effort to disseminate or defend his ideas. Nevertheless, word about his ideas and his interest in astronomy slowly spread through Europe. In 1514 he was invited to participate in a Lateran Council, organised by the Vatican to address the problem of calendar reform, although he declined to attend. In 1532 a lecture on the Copernican theory was presented at the Vatican and, by all accounts, favourably received. Three years later, Copernicus was *asked* explicitly by one Cardinal Schoenberg, a confidant of the Pope, to publish his ideas. Still he hesitated.

Why Copernicus hesitated is not clear, and several different explanations have been proffered, more than one of which may be at least partially correct. Copernicus knew that he could not *prove* that his model was correct and, further, recognised that big pieces of it were contradictory and wrong. It is not even completely clear that he more than half believed in a heliocentric Universe. He was in every sense of the word a Middle Ages man. He himself did not like to do observations, and instead relied almost completely on ancient observations provided by the Chaldeans, the Greeks, and the Arabs, rather than doing his own: he saw himself as a mathematician and philosopher, who left the star-gazing to others. Perhaps the most compelling hint as to why he hesitated to publish is to be found in the introduction to *De Revolutionibus*, where he expressed his fear that people might not like his theory and that he might be "hissed off the stage." In fact, the only seriously unflattering criticism which he ever received during his lifetime was from *Martin Luther*, the Protestant reformer, who called him "a fool who went against Holy Writ."

In any event, he was finally convinced by a close friend, Bishop Tiedermann Giese, and a young disciple, the Lutheran Professor of Astronomy *Rheticus*, to allow Rheticus to publish a summary of the Copernican theory, albeit on the condition that Copernicus' name never be used: On the title page, the author of the theory that Rheticus aims to describe is given as "the learned Dr. Nicholas of Torun." Rheticus' pamphlet was published in 1539, causing a considerable stir, which led to more requests that Copernicus publish his book. Finally, in 1540 Copernicus agreed and Rheticus, then a visitor of Copernicus in Frauenberg, began writing out by hand the final draft of the manuscript to be given to the printer. By August 1541, Rheticus had completed transcribing and editing 424 pages of tiny scrawl, and returned home to Wittenberg to attend to other professional duties. In May 1542 the manuscript was presented to the printer in Nuremberg and the laborious process of typesetting began. Copernicus received a copy of the printed book on 24 May 1543, the day that he died.

(Parenthetically, Luther's reservations actually figured in the unsigned introduction that appeared in the first edition of *De Revolutionibus*, which would suggest strongly that Copernicus associated little, if any, fundamental "reality" with the heliocentric model. Only after his death was it learnt that that unsigned introduction had been written by the Lutheran theologian Andreas Osiander (1498 - 1552) who, when completing the editing process begun by Rheticus, aimed to reconcile Copernicus' thinking with his own personal orthodoxy.)

The key ideas of the Copernican world view are all contained in the handwritten pamphlet prepared around 1512, where he articulated what he considered his seven basic axioms:

1. The heavenly bodies do not all move around the same center.
2. The Earth is not the center of the Universe, although the Moon does orbit around it.
3. The Sun is the center of the planetary system and hence of the Universe.

4. Compared with distances to the stars, the distance from the Earth to the Sun is extremely small.
5. The apparent diurnal motions are associated with the Earth rotating around its axis.
6. The apparent annual motions are associated with the Earth revolving around the Sun.
7. Retrograde motion and similar irregularities are all caused by the fact that the Earth revolves around the Sun.

Axioms (1) - (3) indicate that Copernicus recognised clearly the distinction between planets, which orbit the Sun, and the Moon, which is a satellite that orbits the Earth. He apparently did not conceive of the possibility that other planets might also have moons. However, there was no evidence for such Moons until, using the newly invented telescope, Galileo discovered the four large moons of Jupiter in the first decade of the 1600's. Axioms (5) and (6) state correctly that the rotation of the Earth about its axis and the Earth's revolution about the Sun account respectively for the systematic diurnal and annual motions of objects in the sky. Axiom (7) indicates that Copernicus also recognised that a heliocentric model can provide a relatively simple explanation for the apparent retrograde motion of planets and other apparent irregularities.

However, Axiom (3) also indicates that the Copernican cosmology is woefully incomplete from a modern perspective: The Earth has been displaced by the Sun as the center of the Solar System, but there is no hint that the Sun, and the Solar System, are not situated at the center of the Universe. In particular, there is as yet no inkling of the idea that stars could be large, significant bodies in their own right. Rather, as was the case in all of the Greek cosmologies, the stars are still viewed as secondary minor players surrounding the Solar System, attached to a fixed sphere in the heavens. As to whether the Universe was finite or infinite in extent, Copernicus refused to say: when asked his opinion on this issue, he preferred to leave that question "to the discussion of the natural philosophers."

Axiom (4) requires more comment, both as to *why* Copernicus was led to believe it and as to *what* it implies about his world view. Stated suscinctly, the observational motivation for this Axiom is the fact that the sizes and relative orientation of the stars in the sky do not change over the course of a year. Medieval astronomers had reasoned that, if the stars are attached to a heavenly sphere centered about the Sun and if the Earth orbits around the Sun, the stars should look different to us at different times of the year. Thus, in particular, when we are closer to one side of the sphere of the stars, the constellations on that side should look larger and, when we are closer to the other side, the constellations on that other side should look larger. This, however, did not seem to be the case. The obvious resolution to this apparent paradox, the one which Copernicus recognised, is that the stars are so far away that, even though this effect exists in principle, it is much too small to be seen.

(Actually, using telescopes and the like it *is* possible to detect small changes in the relative locations of nearby stars over the course of a year as the Earth orbits the Sun. The first detection of this so-called *parallax* was by Friedrich Wilhelm Bessel in 1838. However, even for these nearby stars, the observed changes in apparent location are very small, a second of arc or less, this implying that even the closest stars [aside from the Sun] are several lightyears away.)

The notion that the Universe must be very large is perhaps the most striking innovation in Copernicus' cosmology. Everything else in these Axioms could have been, and in fact was, said by Aristarchus almost 2000 years earlier. (In the original manuscript of *De Revolutionibus*, Copernicus credited Aristarchus and his predecessors for the idea of a heliocentric Universe. However, he crossed out Aristarchus' name in ink, so that it does not appear in the final version, although the book *does* state that the heliocentric model had been known to the ancients.) However, although the ancient Greeks were aware that the Universe is much larger than the Earth (cf. Hipparchus's estimate of the distance to the Moon), they had not contemplated explicitly the idea that the observable Universe is really *that big*. Even Anaximander, who believed the Universe to be infinite in extent, seems to have thought that the Solar System and the stars are embedded in the center of a Universe which was otherwise largely empty space.

Aside from the misinterpretation of the relationship between the Solar System and the stars, all the preceeding Axioms are correct from a modern perspective. However, they do not in themselves constitute a specific concrete model. The construction of such a model was Copernicus' other major aim, the task that occupies so much space in *De Revolutionibus*. Unfortunately, however, unlike the Axioms his model is almost completely wrong. Copernicus was a man of the Middle Ages, enamoured of the teachings of the ancient Greeks, so that, when it came to constructing his model, he used the same incorrect building blocks which had been used by Ptolomy.

Although Copernicus had rejected the geocentric orthodoxy, he was still wedded to circular orbits and constant speeds, which made it difficult for him to describe the elliptical orbits along which the Moon and the planets move with varying speeds. He was therefore led to introduce his own version of a "wheels within wheels" model, replete with multiple sets of epicycles and the like, which is just as complicated, arcane, and unphysical as the Ptolomaic world model. In fact, the Copernican model requires more "wheels" than entered into the Ptolomaic model, 48 as opposed to 40. Moreover, just as for the Ptolomaic model the true center of the Universe is a point in empty space: for Ptolomy a point slightly displaced from the Earth, for Copernicus a point in space approximately three times the diameter of the Sun away from the center of the Sun. (Interestingly, though, Copernicus is deliberately vague as to whether his celestial spheres correspond only abstractly to the orbits of the planets or whether they have tangible reality as physical objects.)

Given all of this, it is natural to ask: why was it that Copernicus and his *De Revolutionibus* were so influential? Large parts of the book are wrong and/or unreadable, and Copernicus certainly does not come across as a forward-looking visionary. Moreover, alternatives to the Ptolomaic cosmology had been discussed freely in the West long before him. Indeed, several hundred years earlier Herakleides' cosmology had been the cosmology of choice for many astronomers. The biggest reason, apparently, that Herakleides had not become the standard for pragmatic astronomers was *not* intellectual orthodoxy but, instead, the fact that his cosmology had not been implemented in a concrete model. Copernicus was the first to construct a concrete model based on the heliocentric hypothesis, a model just as systematic, aracane, and confusing as the geocentric model constructed in *The Almagest* (and slightly less accurate: *The Prutenic Tables*, based on the Copernican world model and named for its patron, the Duke of Prussia, are arguably slightly less successful than *The*

Alphonsine Tables). Indeed, the fact that Copernicus worked within the same framework as Ptolomy, sticking to circles, spheres, and uniform motion and arguing in the language of Aristotle's natural motions, may have played an important role in his success. His book was anti-Ptolomaic in the sense that it was heliocentric but it was very much in the tradition of Ptolomy in every other way.

Lecture 6

THE NEW EMPIRICISM OF BRAHE AND GALILEO

The quotations from Kepler presented here come from Part IV of the book *The Sleepwalkers*, by Arthur Koestler (Hutchinson, London, 1959). The quotations from Galileo come from Chapter 41 of the book *The Discoverers*, by Daniel Boorstin (Vintage, New York, 1984).

By around 1500 natural philosophy in general, and astronomy in particular, was dominated by an elaborate scientific orthodoxy, based on five different tenets, each of which had to be overcome in the construction of a modern science of astronomy/cosmology:

1. an anti-empirical bias, based on citing authority, ignoring concrete physical evidence;
2. an unquestioning belief in a geocentric Universe, despite substantial evidence to the contrary;
3. an equally unquestioning belief in uniform circular motion;
4. a rejection of the Pythagorean notion of mathematics as the language of science; and
5. the Aristotelean idea of natural motion, in particular the idea that moving bodies tend to sink towards rest, rather than remaining in motion.

Copernicus played an important role in the ultimate rejection of the entrenched idea of a geocentric Universe, and he embraced the notion, dating back to Pythagoras and Plato, that mathematics should be the language of science. However, he was definitely of an anti-empirical bent: As stated already, he himself did not like to do observations and his overall approach towards astronomy and science in general involved citing, and accepting, authority. Moreover, like that of Ptolomy his entire world model was constructed in terms of uniform circular motion, and it was motivated completely within the context of Aristotle's notion of natural motion.

Two figures subsequently played particularly significant roles in overthrowing the anti-empirical bias. The first of these was *Tycho Brahe* (1546 - 1601), who, albeit working before the introduction of telescopes into astronomy, was responsible for the best systematic observations of the heavenly bodies since the ancient Greeks. The other was *Galileo Galilei* (1564 - 1642), who played a major role in the development of the refracting telescope and is credited with being the first astronomer to have used a telescope systematically to observe the heavens.

Galileo also played a major role in the ultimate repudiation of Aristotelean physics, although he was *not* the first scientist to reject explicitly Aristotle's theory of natural motion: That honour belongs apparently to *William of Ockham* (d. 1349?), a fourteenth century Parisian, and his followers.

The explicit rejection of the orthodoxy of circular motions was due to *Johannes Kepler* (1571 - 1630), the first to deduce from an analysis of observations that the orbits of the planets are in fact ellipses, and to articulate three empirical laws which characterise completely the motions of the planets around the Sun.

With the contributions of Galileo and Kepler, astronomy was presented for the first time with (1) an accurate mathematical characterisation of the motions of the planets around the Sun and (2) the underpinings of a theory of physics which should ultimately be able to *explain* the origins of these motions from first principles. It was *Isaac Newton* (1642 -

1727), born the same year that Galileo died, who provided that final synthesis, erecting the foundations of physics and astronomy, as they were accepted uniformly until the middle of the nineteenth century.

Tycho Brahe was born the son of a rich Danish nobleman, heir to a tradition of absolute power over his peasants (*jus primae noctis*, etc.), who were viewed as little more than property. However, even in the context of the haughty nobility of his times, he acquired a reputation for being especially disgusting, a fact which impacted on the overall course of his life upon more than one occasion. Like any typical Danish nobleman, Brahe was sent to a University, in his case the University of Rostock, to study law and the classic liberal arts and sciences (these classical studies divided into the so-called *trivium*, comprised of grammar, rhetoric, and logic/dialectic, which were examined for the Batchelor of Arts Degree, and the *quadrivium*, comprised of geometry, astronomy, arithmetic, and music, which were examined for the Master of Arts degree), with the aim of training him for a life at court. However, for reasons largely unknown, Brahe was more interested in astronomy. Always a fiery-tempered man, he managed somehow to lose his nose during his student days. The legends about this vary: some say that it happened during a duel, others in a tavern brawl; some say the argument was over astronomy, others over the favours of some young woman. The most likely version is that, at age twenty, he and another student dueled over the issue of who was the better mathematician. In any event, all through his later life Brahe was forced to wear an artifical gold or silver nose which was attached to his face with wax. (When his tomb was opened in 1901 on the $300th$ anniversay of his death, a green stain was found on his skull in the nasal opening which would indicate that his gold and silver noses were adulterated with copper.)

It is known that, by the age of 17, Brahe had discovered significant errors in both *The Alphonsine Tables* and *The Prutenic Tables*, and that this motivated him to devote his professional life to constructing an improved cosmology which would avoid these sorts of errors. Of particular importance was his recognition that, in order to construct such an improved alternative, he would need highly accurate and reliable observational data, much better than anything that was available in Europe at the time. Copernicus himself had also recognised the importance of improved observations. Indeed, the reason he rejected an invitation to a Vatican conference on calendar reform was his belief that satisfactory reform was impossible without much better observations. However, unlike Copernicus, who did not like observing, Brahe actually enjoyed it. His only other serious occupations were apparently eating, drinking, and wenching.

In 1572, Brahe and other astronomers were treated to the spectacle of a "new star" that appeared to turn on in the sky and slowly dimmed away after about a year. Today we know that this was a *supernova*, the last burst of a supergiant star before it dies as a neutron star. At the time, the nature and location of this object were questions of substantial interest: was it a new moving object in the heavens, or was it fixed in the sky like the stars? Brahe made repeated observations of this object over several months and, given the absence of any conspicuous motions vis-a-vis the other stars, came to the conclusion that it must be quite far way. His observations and conclusions were published in 1573 in a book entitled *De Stella Nova* or *The New Star*.

On the basis of the talents displayed in this book, Brahe convinced the King of Denmark, Frederick II, to fund an observatory, Uraniborg, to be constructed on the island of Hveen off the coast of Denmark. (The fact that Brahe's father had once saved the King's life probably did not hurt him in his efforts to secure funding from the King.) By fleecing his peasants, Brahe also raised additional funds, enabling him to build a truly impressive castle/palace/observatory, which looked roughly like a cross between the Kremlin and a Renaissance palace in Italy. Uraniborg was equipped with four different observatories – one on each corner of the palace –, an impressive, working torture chamber, and a court jester – Jepp – to entertain Brahe during his meals.

Brahe and his servants spent the next twenty years charting out the motions of the Moon, the Sun, and the planets as they move relative to the fixed stars and, in addition, measured the positions of 777 stars to an accuracy of better than 1/15 of a degree. The latter achievement is particularly impressive, since it was performed without the use of telescopes, which only came into use about the time that Brahe died. Instead of telescopes, Brahe resorted to the same tools as the ancients, including huge *diopters* (or cross-staffs, first invented by Hipparchus) and *quadrants* (or backstaffs), and other gadgets and devices. The resulting catalogue, published in 1602 as *Progymnasmata*, was the first to supercede the star catalogues of Ptolomy.

For at least three different reasons, Brahe himself did not believe in the Copernican world view: (1) He recognised that, since the relative sizes and orientations of the constellations do not change over the course of a year, either the Earth must be at rest relative to the stars or the stars must be very far away. The latter alternative was inconsistent with his view of a relatively small Universe. (2) Like many others before and after him, he found it difficult to accept the idea that an object as "big and sluggish" as the Earth can be hurdling through space at high speeds. (3) Despite his personal excesses, Brahe considered himself a devout Lutheran, and he was well acquainted with Luther's rejection of Copernicus.

However, Brahe was too much of a pragmatist to accept the Ptolomaic model, with its arcane explanation of retrograde motion. Instead, he concocted his own, Tychonic, world model, which (to his mind) incorporated the best ingredients of both. Specifically, like Herakleides he constructed a hybrid model, in this case allowing for the Sun and the Moon to orbit the Earth, but arranging for the rest of the planets to orbit the Sun.

Eventually King Frederick II drank himself to death (as was cheerfully noted by his biographer), and was replaced by his son, a somewhat more moderate Christian IV. The son was apparently willing to overlook Brahe's arrogant attitude towards the King and everyone else (partly because he had enjoyed playing at Uraniborg as a child). However, he drew the line at various irregularities that had occured on Hveen, including open defiance of orders from the High Court which had forbidden Brahe from holding the family of one of his tenants in chains. As a result, Brahe's sinecure was decreased substantially (albeit not withdrawn). Brahe responded by packing up his belongings and moving down towards Prague, where he was to be employed as the Imperial Mathematicus for Rudolph II, the Emperor of the Holy Roman Empire. In 1599, after a two year caravan through Europe, he reached his new residence, Benatek Castle, located 22 miles northeast of Prague. Here Brahe assumed his new responsibilities which included mathematics, astrology, and alchemy, as well as astronomy. The specific task for which he had been hired was to use his cosmological model to construct *The Rudophine Tables* as a new alternative to *The*

Alphonsine and *Prutenic Tables.* To aid him in his endeavors, he hired a large number of assistants, including Kepler, his eventual successor.

Finally, in 1601, he met his end, in a fashion recorded in Kepler's diary:

> "On October 13, Tycho Brahe, in the company of Master Minkowitz, had dinner at the illustrious Rosenberg's table, and held back his water beyond the demands of courtesy. When he drank more, he felt the tension in his bladder increase, but he put politeness before health. When he got home, he was scarcely able to urinate.
>
> At the beginning of his illness, the moon was in opposition to Saturn...
>
> After five sleepless nights, he could still only pass his water with the greatest pain, and even so the passage was impeded. The insomnia continued, with internal fever gradually leading to delirium; and the food he ate, and from which he could not be kept, exacerbating the evil. On October 24, his delirium ceased for several hours; nature conquered and he expired peacefully among the consolations, prayers and tears of his people.
>
> So from this date the series of celestial observations was interrupted, and his own observations of thirty-eight years have come to an end.

Before Brahe died, he made sure that Kepler was appointed as his successor, albeit at a much reduced salary since, as he informed the Court, Kepler was far less talented.

Galileo Galilei was an Italian, born in Pisa the son of an impoverished, but well educated and intellectual, nobleman. He was educated for a time at a school in a Jesuit monastary, and then sent to the University of Pisa to study medicine, although he was apparently more interested in physics and astronomy. Despite acquiring an immediate reputation for brilliance, he was unable to secure a scholarship – probably because of his coldly sarcastic personality – and was forced to leave the University without a degree. Four years later, he was appointed a lecturer at the same University that had refused him a degree, a feat which Galileo claimed was especially impressive because he had not given anyone a bribe. Three years later, in 1592, he was appointed to the Chair of Mathematics at the University of Padua, where he remained until 1611. While in Padua, he was apparently involved actively in investigations of dynamics and physics. He developed a number of important gadgets, including a forerunner of the thermometer, which he sold to the community, but he did not publish anything about astronomy until 1610. When he became a believer in the Copernican world view is unclear: his first public endorsement came only in 1613. It is known that Galileo lectured at the University of Padua on the Ptolomaic world view in 1597, without suggesting that there were any major problems.

In 1609, Galileo learnt about the telescope from Dutch sailors visiting Padua, quickly improved on their primitive model, and even more quickly "sold" the improved model to the city, apparently for a tidy profit. (Actually he donated the telescope to the city and was promptly rewarded by an extension of his professorial appointment, which was about to expire, with a hefty salary increase from 520 to 1000 florins *per annum.*) Immediately thereafter, he started to use the telescope to investigate the heavens, quickly finding concrete evidence that the Ptolomaic model must be wrong. His observations and their interpretation were published in 1610 in a book entitled *Siderial Nuncius* or *The Siderial Messenger.*

Basically, Galileo made four concrete observations.

1. He found that Venus has phases, just like the Moon. The reason for this is easily understood in the context of a heliocentric model: the side of Venus facing the Sun will be illuminated, but that does not in general coincide with the side of Venus facing the Earth. According to the Ptolomaic model Venus cannot have seasons.
2. He also discovered that Jupiter has at least four moons. Aside from such seemingly transient objects as comets and meteors, these were the first moving objects in the heavens to have been discovered since antiquity. Galileo knew that they were in fact moons orbiting Jupiter because he saw them pass behind Jupiter and then reappear in front again.
3. Sunspots are intrinsic disturbances on the surface of the Sun; and
4. the surface of the Moon is not smooth. Both these latter observations contradict the doctrine of a perfect and immutable heaven, dating back to Aristotle and crucial to the medieval world view. Without the improved resolution provided by a telescope, it was easy for the ancients to argue that sunspots are really clouds passing between the earth and the sun, and that the irregularities on the surface of the Moon are either more clouds or, perhaps, a reflection of lights from the Earth. The improved observations made that impossible.

When published this book made quite a stir. The following year, Galileo went to Rome to lecture about his new ideas, teaching several different important and influential people, including Cardinal Barberini, who was later to become Pope. Over the next several years he stirred up a fair amount of controversy, writing numerous letters and pamphlets in support of the Copernican world view, that began to get him in trouble. As long as he taught the Copernican cosmology as a *hypothesis* that can lead to a useful description of the heavens the Catholic Church had no problem with him. However, once he began to assert that he had *proven* the validity of the Copernican cosmology and, particularly, once he began to discuss it in the context of Biblical scripture, he was bound to get in trouble.

Basically, Galileo's religious/theological justification for his findings was encapsulated in the argument that, although there is only one Truth, that Truth is communicated in two forms: one in the language of the Bible and the other in the language of Nature:

> ...both the Holy Scriptures and nature proceed from the Divine Word, the former as the saying of the Holy Spirit and the latter as the most observant executrix of God's orders. And since it is needful in the Scriptures, in order to accomodate these to the understanding of ordinary people, to say many things which appear different (as to the meaning of the words) from absolute truth, while on the other hand nature is inexorable and immutable and cares nothing about having her hidden reasons explained to the understanding of man, so long as she never oversteps the bounds of the laws imposed on her, it seems that none of the physical effects that are either placed before our eyes by sensible experience or are the conclusions of necessary demonstrations should ever on any account be placed in doubt by passages of the Scriptures which seem to have a different import ... two truths can never contradict one another ....

In 1615 Galileo was denounced to the Church and, although nothing came of the charges, the Inquisition opened a file on him. Finally in 1616, 73 years after the death of Copernicus, almost completely in response to Galileo's aggressive attitude, the Catholic Church felt compelled to take a stand on the teachings of Copernicus. The so-called *Qualifiers*, i.e., the theological experts of the Church, originally ruled that the Earth is at rest at the center of the Universe and that to say otherwise is heresy. However, under pressure from several powerful Cardinals they softened their opinion. The final report termed Copernicus' teaching "altogether opposed to Holy Scripture," but did not deem it heresy. Moreover, although *De Revolutionibus* was banned temporarily, it was made clear that the book could and would be unbanned if only a few particularly objectionable passages were removed, e.g., the argument (possibly inserted by Osiander) that Copernicus' theory does not contradict Scripture. At the same time, Pope Paul V had Galileo ordered, under threat of imprisonment, to stop believing in the Copernican theory; whether he was officially prohibited from discussing it is not completely clear from Vatican records. In any event, Galileo backed down, at least for a while.

In 1623, Galileo's erstwhile student, Cardinal Barberini, was elevated as Pope Urban VIII. Galileo appears to have thought that Urban would overturn the earlier ruling of the Inquisition, for he sought and received a private meeting with the Pope. Urban was apparently conciliatory but vague, making no promises. However, Galileo was sufficiently encouraged that, the following year, he began work on a new book, ostensibly an objective analysis intended to contrast the Ptolomaic and Copernican world views. The book was completed in 1629 and ready for the Church censors who, in seventeenth century Italy, had to approve any book before it could be printed. With more than a bit of finagling, finding particularly unknowledgeable censors and carefully assuring them that the book was an objective tome, Galileo eventually got the book past the censors. The *imprimatur* was granted, and publication occured in 1632.

The book, entitled *Dialog Dei Due Massimi Sistimi* or *A Dialogue Concerning Two Chief World Systems*, is constructed as a discussion amongst three characters: Salviati, a brilliant, erudite defender of Copernicus; Sagredo, a reasonably competent but uninformed believer in Ptolomy, who is eventually converted to the new theory; and Simplicio, a diehard Ptolomaist who, as his name might suggest, is not portrayed as being terribly bright.

For at least three reasons, its publication resulted in instant trouble. The first obvious point is that the book is not unbiased, as Galileo had claimed, and that it makes clear the fact that Galileo himself believed in the Copernican theory. A second point relates to style. Unlike his earlier *Siderial Nuncius*, the text is written in Italian, rather than Latin, a language that even nonscholars could understand, and, rather than being a scholarly work, it is a blatantly one-sided. That much was bad enough. The worst trouble, however, reflected the real-life identities of the three principles. Salviati and Sagredo were clearly modeled after two of Galileo's old friends with the same names, both of whom were safely dead, so that it was natural to ask who Simplicio might represent. Galileo himself insisted that he was modeled after Simplicius, a sixth century commentator on Aristotle. That may in part be true. However, at more than one point in the manuscript he had Simplicio uttering several of the less intelligent things that then-Cardinal Barberini had said when Galileo was teaching him the Copernican theory. The end of the book is especially offensive. Here Simplicio presents an absolutely stupid argument, describing it as coming "from a most eminent and learned person, and before whom one must fall silent" whereupon Salviati and Sagredo declare themselves silenced by "this admirable and angelic doctrine," and decide to "go and enjoy an hour of refreshment in the gondola that awaits us."

The resulting rumour that Galileo was insulting the Pope caused enough of a scandal that Urban VIII himself had Galileo ordered before the Inquisition. Galileo procrastinated because of illness but, when the Pope threatened to have him brought to the Vatican in chains, delivered himself to Rome in 1633 to face the charges against him. Basically, there were three: (1) that he had treated the Copernican theory as being more than a hypothesis (of this he was obviously guilty); (2) a more technical objection that he had stated falsely that Copernicus, and only Copernicus, had a correct explanation of the tides (again guilty); and (3) that he had lied in stating that he no longer believed in the Copernican theory (guilty yet again).

His original defense was that he had in fact treated the Copernican world view as a hypothesis and that he really did not believe it himself, claims which must have seemed completely ludicrous to the highly educated Inquisitors who examined him. As indicated by the records, the exchanges between Galileo and his Inquisitors were confusing, in large part because, as was standard in Inquisitorial proceedings, the accused was never informed of the charges against him. What *is* clear from the records is (1) that the Inquisition was not ready to execute Galileo but (2) that they wanted to make an example of him. He apparently was a scared, seventy year old man, particularly when it was hinted that he might be accused of *atomism*, the belief that all matter is composed of atoms, for which *Giordano Bruno* had been burnt in 1600. (In point of fact, the principal crime of Bruno, who was anathematized by both Catholics and Protestants, was a belief in the imminence of God in all material objects, which placed him much closer to the teaching of the Gnostics than to orthodox Catholicism.)

On June 22, 1633, Galileo knelt before the Inquisition and read a recantation in which he admitted his errors, stating in particular that he was guilty of believing in the Copernican heresy (even though the teachings of Copernicus had never been declared heretical):

> I, Galileo, sone of the late Vincenzio Galilei, Florentine, aged seventy years, arraigned personally before this tribunal and kneeling before you, Most eminent and Lord Cardinals Inquisitors-General against heretical pravity throughout the entire Christian Commonwealth, having before my eyes and touching with my hands the Holy Gospels, swear that I have always believed, do believe, and with God's help will in the future believe all that is held, preached and taught by the Holy Catholic and Apostolic Church. But, whereas, after an injunction had been lawfully intimated to me by this Holy Office to the effect that I must altogether abandon the false opinion that the sun is the center of the world and immobile, and that the earth is not the center of the world and moves, and that I must not hold, defend, or teach, in any way, verbally or in writing, the said false doctrine, and after it had been notified to me that said doctrine was contrary to Holy Scripture, I wrote and printed a book in which I treated this new doctrine already condemned and brought forth arguments in its favor without presenting any solution for them, I have been judged to be vehemently suspected of heresy, that is, of having held and believed that the sun is the center of the world and immobile and that the earth is not the center and moves.
>
> Therefore desiring to remove from the minds of Your Eminences, and of all faithful Christians, this vehement suspicion rightly conceived against me, with sincere heart and unpretended faith I abjure, curse, and detest the aforesaid errors and heresies and

> also every other error and sect whatever, contrary to the Holy Church, and I swear that in the future I will never again say or assert verbally or in writing, anything that might cause a similar suspicion toward me; further, should I know any heretic or person suspected of heresy, I will denouce him to this Holy office or to the Inquisitor or Ordinary of the place where I may be ....

Galileo was subsequently sentenced to life imprisonment which, through the direct intervention of the Pope, was commuted to house arrest. He returned to his villa where he lived until his death in 1642. Despite being enjoined from doing any more science, he spent the remainder of his life working on more experiments in dynamics, with a final book, *Dialogues Concerning Two New Sciences*, completed in 1636. The manuscript was smuggled out to Protestant Leiden and published by Elsevier Press. During the last years of his life, he had numerous visitors, including Milton in 1638. He was buried in the Church of Santa Croce, near Michaelangelo and Machiavelli.

In 1979, Pope John Paul II reopened the case against Galileo. In October 1992, the commision appointed to reexamine the case concluded of Galileo's Inquisitors that "This subjective error of judgment, so clear to us today, led them to a disciplinary error from which Galileo had much to suffer." John Paul's comment was that "A tragic mutual incomprehension has been interpreted as the reflection of a fundamental opposition between science and faith ... this sad misunderstanding now belongs to the past." Maybe, but Galileo is not the only scientist to have been persecuted for his beliefs.

Lecture 7

SETTING THE STAGE FOR NEWTON: GALILEO'S PHYSICS AND KEPLER'S LAWS

The quotations from Kepler presented here come from Part IV of the book *The Sleepwalkers* by Arthur Koestler (Hutchinson, London, 1959) and Chapter 39 of the book *The Discoverers*, by Daniel Boorstin (Vintage, New York, 1984).

Recall that, at the time of Copernicus, the prevailing theory of physics in the intellectual community was based on Aristotle's notions of *natural* and *violent motions*. According to Aristotle, everything on the Earth and its environs are constructed from four basic elements – earth, water, air, and fire – and each of these elements is endowed with a natural motion which it will execute in the absence of any disturbing influences. In particular, Aristotle believed that the natural motion of material objects is to fall towards the center of the Universe, i.e., the Earth, and once there to remain at rest. The only way to counter this tendency, causing a violent motion, is the continuous application of a perturbing *force*. If a body is not being subjected to a force, it will instantaneously cease any violent motion and simply fall to the Earth.

As will be discussed below, Aristotle's idea that material objects always tend to come to rest is incorrect. Moreover, his idea of forces, and their role in inducing violent motions, also diverges significantly from our current understanding. Physics today is predicated on the notion of forces that act between objects that need not be directly touching each other. Thus, e.g., two electrons, just as any pair of identically charged particles, will experience a repulsive electric force even if they are not in direct physical contact. This is understood as resulting from the fact that each electron generates a surrounding electric field, and that the field of each electron influences the other electron. This concept of a field was completely foreign to the thinking of Aristotle.

The ideas (1) that a moving object must be subjected continually to some force in order to remain moving and (2) that forces only act through direct physical contact led Aristotle to some bizarre interpretations of nature. A classic of example thereof is the Aristotelean description of what happens when an arrow is shot from a bow: It would seem naively that, according to Aristotle's physics, as soon as it ceases to have contact with the bowstring the arrow should immediately stop moving horizontally and fall to the ground. This, however, obviously does not happen. To explain this, the Aristotleans countered that, as the arrow first starts moving, when it is still in contact with the bowstring, it creates an air current or vortex, which subsequently moves through space, dragging the arrow along.

In addition to these erroneous theories, Aristotle made two other concrete predictions, both of which are false. The first is that a falling object like a ball will plummet downwards at a constant speed. The second is that more massive objects will fall more rapidly than less massive objects. In point of fact, neglecting air currents, friction, and the like, two balls of unequal mass will, when dropped, fall towards the Earth identically. Moreover, as they fall they will both be accelerated. Starting from rest, each will have acquired a speed of approximately 9.8 m/sec (or about 32 ft/sec) after one second, 19.6 m/sec after two seconds, and so forth.

Today it is understood that the two balls will be accelerated as they fall towards the Earth because of the Earth's gravitational pull. Because it is more massive, the heavier ball will feel a stronger pull; however, again because it is more massive, the heavier ball will be harder to pull. These two tendencies exactly cancel, with the net result that the accelerations of the balls are exactly equal, independent of their masses. This fact, called the so-called *Equivalence Principle*, lies at the heart of Einstein's theory of general relativity, as developed in the twentieth century.

Beginning with William of Ockham and his students in the fourteenth century, it began to be realised in the West that Aristotle's theories of physics are seriously wrong. Galileo, whose earliest work was in experimental dynamics, fits very much into that tradition. He played an important role in developing better clocks to facilitate his experiments and, using these clocks, made several important discoveries. In his studies of the pendulum, he was, e.g., apparently the first person to recognise that, for a small amplitude oscillation, the period of oscillation (how long each swing takes) is essentially independent of the amplitude.

Part of Galileo's studies involved an analysis of the motions of balls as they roll up and down inclined planes. He recognised that a ball accelerates as it rolls downwards and decelerates (i.e., experiences a negative acceleration) when it rolls upwards, and, given these observations, he reasoned that, in the absence of friction, a rolling ball on a horizontal plane would neither speed up nor slow down. Rather, it should experience no acceleration and, instead, simply move along in the same direction at a constant speed.

Galileo's crucial extrapolation from these experiments (a conclusion almost, but apparently not quite, reached by the Ockhamists) was that it is the absence of acceleration, rather than the absence of motion *per se*, that corresponds to what Aristotle would have termed the true natural motion. If an object is initially at rest, it will remain at rest unless it it disturbed somehow by a force. Alternatively, if it is already moving in some direction at a given speed, it will continue to move in the same direction with the same speed until something constrains it to do otherwise.

Galileo codified this basic idea in what is called his *Law of Inertia:* Any material body will move along a constant velocity trajectory, i.e., in the same direction with constant speed, forever unless there is some force that causes it to accelerate. In other words, the true natural motion is constant velocity motion, not necessarily zero velocity motion.

Galileo also recognised another point: If, as is manifest, the planets do not follow rectilinear trajectories, this must be because there is some force that is pulling them around the sun. In other words, he inferred that some force like gravity must exist.

Galileo also introduced another, perhaps even more insidious, concept, which has come to be known as *Galilean relativity.* Basically, the idea here is that if it is changes in velocity that are fundamental, rather velocities themselves, there is no way in which to determine fundamentally who or what in the Universe is really at rest.

Suppose, for example, that you are sitting inside a sealed train that is moving along at constant velocity. Given that you cannot see the scenery outside, there is no way in which you can tell whether the train is at rest, crawling along at 10 miles an hour, or speeding along at 150 miles per hour. Indeed, many of you have probably noticed that, while sitting

in an airplane moving along at 500 miles per hour or more, you feel no effects from the motion unless the airplane accelerates relative to the Earth, i.e., unless it speeds up, slows down, or changes direction.
Accepting Galilean relativity, it is not possible to say that, in a fundamental sense, the Sun is at rest and the Earth and the remaining planets are moving around it. Indeed, short of measuring an acceleration or the absence thereof, one cannot even say conclusively that the Sun and the rest of the Solar System are not travelling through space. In fact, it is known today that the Solar System is actually orbiting around the center of the Milky Way galaxy with a speed of more than 200 km/sec, and that the Milky Way galaxy is moving towards the galaxy Andromeda, again with a speed of hundreds of km/sec.
Galileo's ideas about dynamics and his Law of Inertia were formulated in his last book, *Dialogues Concerning Two New Sciences*, which was completed in 1638.

According to the horoscope which he cast for himself, *Johannes Kepler*, Keppler, Khepler, Kheppler, or Keplerus (he used all five of these different spellings during his life) was conceived on 16 May AD 1571 at 4:37 AM and born on 27 December at 2:30 PM after a pregnancy lasting 224 days, 9 hours, and 53 minutes. How he determined the time of conception so accurately is unknown, but it was very important to him since he, like many educated people of his time, was deeply enamoured of astrology.
Kepler was born in Weil-der-Stadt, a tiny hamlet in Swabia, in what is now southwest Germany, the son of a mercenary soldier. His father was apparently an unsavoury character who, during Kepler's childhood, was forced for reasons unknown to disappear from Weil-der-Stadt and was subsequently hanged *in absentia*. Kepler would appear to have had the classic miserable childhood. His family was impoverished; he grew up a scrawny child who succumbed to essentially every childhood illness, including smallpox at age four and recurring mange; he was clumsy and accident-prone, very unpopular with all the other children. In adult life, he was a neurotic hypochondriac, incredibly unsure of himself, with a permanently scabby face, the result of chronic skin disease. By all accounts, his mother was an unpopular shrew. At one point in his later life, Kepler was forced to devote considerable attention to providing her with a legal defense to keep her from being burnt as a witch.
Kepler's diary and notes on his childhood, which still exist, are amongst the more degressing documents which one is ever likely to read. A self-portrait of himself at age 26, written in Latin and reprinted from Koestler's book, is typical:

> "That man [i.e., Kepler] has in every way a dog-like nature. His appearance is that of a lap-dog. His body is agile, wiry and well-proportioned. Even his appetites were alike: he liked gnawing bones and dry crusts of bread, and was so greedy that whatever his eyes chanced on he grabbed; yet, like a dog, he drinks little and is content with the simplest food. His habits were similar. He continually sought the goodwill of others, was dependent on others for everything, ministered to their wishes, never got angry when they reproved him and was anxious to get back into their favour. He was constantly on the move, ferreting among the sciences, politics and private affairs, including the lowest kind; always following someone else, and imitating his thoughts and actions. He is bored with conversation, but greets visitors just like a little dog;

yet when the least thing is snatched away from him, he flares up and growls. He tenaciously persecutes wrong-doers – that is, he barks at them. He is malicious and bites people with his sarcasms. He hates many people exceedingly and they avoid him, but his masters are fond of him. He has a dog-like horror of baths, tinctures and lotions. His recklessness knows no limits, which is surely due to Mars in quadrature with Mercury, and in trine with the moon; yet he takes good care of his life.... [He has] a vast appetite for the greatest things. His teachers praised him for his good dispositions, though morally he was the worst among his contemporaries.... He was religious to the point of superstition. As a boy of ten years when he first read Holy Scriptures ... he was grieved that on account of the impurity of his life, the honour to be a prophet was denied him. When he committed a wrong, he performed an expiatory rite, hoping it would save him from punishment: this consisted in reciting his faults in public....

"In this man there are two opposite tendencies: always to regret any wasted time, and always to waste it willingly. For Mercury makes one inclined to amusements, games, and other light pleasures.... Since his caution with money kept him away from play, he often played by himself. [The word for 'play', *lusu*, may refer here either to gambling or to sex.] It must be noted that his miserliness did not aim at acquiring riches, but at removing his fear of poverty – although, perhaps avarice results from an excess of this fear...."

An extremely devout Lutheran throughout his life, Kepler went to the University of Tübingen on a scholarship, with the aim of studying to become a Lutheran pastor. There as an undergraduate he was taught by *Michael Maestlin*, one of the first thinkers to teach the Copernican world view in a European university. He received a batchelor degree and then entered the theological school at Tübingen. However, he was forced to leave the university to support his mother and younger siblings, accepting a position in Graz, a town in the Austrian province of Styria. He was hired as a teacher of mathematics and astronomy at a Protestant school attached to the Catholic University of Graz. A fiery, enthusiastic, and incoherent speaker, he was apparently a terrible teacher. His first year he only had a few students; his second year none at all. This meant that he had lots of free time, which he devoted to his other assigned activities, writing a yearly set of astrological predictions and doing research in mathematics and astronomy.

Kepler's first book, *Mysterium Cosmographicum* or *The Cosmic Mystery*, was published in 1596. Almost completely rubbish from a modern perspective, it is a brilliant but crazy admixture of astrology, astronomy, mysticism, and mathematics. From Kepler's viewpoint, the most important objective of the book was to outline a new model of the Solar System, wherein the spacings of the planets in the heavens are related to the so-called five *perfect solids*. As had first been noted by Euclid, these five three-dimensional figures are the only solid figures that can be constructed with faces of identical sizes: the tetrahedron (four identical faces), the cube (six), the octahedron (eight), the dodecahedron (twelve), and the icosahedron (twenty). The Pythagoreans had noted that, because these objects are symmetric in shape, they can be inscribed in a sphere is such a fashion that all the corners touch the sphere. Similarly, they can be circumscribed around a sphere, with a sphere touching the center of each face. Kepler's basic idea in this book – completely wrong – was that the five planets all revolve around the Sun, attached to heavenly spheres which are nested between five different perfect solids:

> And then ... it struck me: why have plane figures among three-dimensional orbits? Behold, reader, the invention and whole substance of this little book! In memory of that event, I am writing down for you the sentence in the words from the moment of conception: The earth's orbit is the measure of all things; circumscribe around it a dodecahedron, and the circle containing this will be Mars; circumscribe around Mars a tetrahedron, and the circle containing this will be Jupiter; circumscribe around Jupiter a cube, and this will be Saturn. Now inscribe within the earth an icosohedron, and the circle contained in it will be Venus; inscribe within Venus an octahedron, and the circle contained in it will be Mercury. You now have the reason for the number of planets....
>
> This was the occasion and success of my labors. And how intense was my pleasure from this discovery can never be expressed in words. I no longer regretted the time wasted. Day and night I was consumed by the computing, to see whether this idea would agree with the Copernican orbits, or if my joy would be carried away by the wind. Within a few days everything worked, and I watched as one body after another fit precisely into its place among the planets.

All this nonsense aside, *Mysterium Cosmographicum* did articulate one important idea, namely that the motions of the planets around the Sun are caused by a "mystical force" emanating from the Sun. In other words, Kepler, like Galileo, had arrived upon the concept that the Sun holds the Solar System together via some force that acts through space.

It should be noted that this sort of weird mysticism was an important characteristic of Kepler's persona throughout his life. Kepler believed in the eventual possibility of interplanetary travel and, in fact, started to write what would have been the first modern science fiction book, *Somnium*, about a trip to the Moon. Apparently, however, the book was never completed. Moreover, like Pythagoras, who was also fascinated by the perfect solids, in later life he claimed to be able to hear the music of the heavenly spheres. However, he also did serious historical and religious scholarship, determining that Christ was born about 4 or 5 BC, in agreement with the currently accepted date.

Kepler sent copies of his book to the two biggest astronomers of his day, namely Brahe and Galileo, both of whom seem to have been both impressed (by the mathematics) and amused (by much of the rest). This triggered a rather interesting exchange of letters with Galileo, many of which still exist, wherein Galileo seems to be picking Kepler's brains without offering a whole lot in return. (For example, in this correspondence Gallleo never admitted that he believed in a heliocentric Universe.) In any event, unlike almost everyone in Italy, Kepler came away with a positive impression of Galileo as a person, perhaps because they never met. Brahe had a more positive and immediate response, offering Kepler a job in Prague, where he had just moved his astronomical equipment.

For Kepler the offer from Brahe was timely. He and his family were Lutherans in a largely Catholic country and, at this period just before the Thirty Years War, religious persecution was picking up. Prague was a city known for its religious tolerance. More importantly, Kepler had apparently come to recognise that the motions of the planets around the Sun might not be strictly circular and that they might not involve motion at a constant speed, and he realised that, in order to characterise them correctly, he would need access to Brahe's data.

When Kepler arrived at Benetek Castle, he encountered what must have struck him as a circus: Some of the equipment was still being assembled, other pieces had been lost; his life was dominated by interactions with Brahe, a loud, aggressive hedonist, who was Kepler's antithesis in almost every respect. He quickly became embroiled in a collection of intrigues involving Brahe's jester Jepp, his son Tycho Junior, and one of his assistants, Junker Tengnagel, who was busy impregnating Brahe's daughter Elisabeth during an illicit affair.

In any event, Kepler was put in charge of the analysis of the motions of Mars, in part because he convinced Brahe by his eagerness and enthusiasm, in part because his direct predeccesor, Longomontaneus, had screwed it up. Of the orbits of the planets known to antiquity, Mars's orbit is the most interesting from an observational perspective. Because it orbits the Sun with a relatively short period of about 1.88 years, it moves across the sky very quickly. However, because it is further from the Sun than is the Earth, Mars is visible throughout much of the night, unlike Mercury and Venus, which are only visible just after dusk and just before dawn. Moreover, the distance between the Earth and Mars varies enormously, between about 80 and 380 million km.

When Brahe died, Kepler was hired to complete *The Rudophine Tables* using Brahe's theory. However, Kepler was a Copernican and had no intention of using a theory in which he did not believe. This enraged Brahe's heirs, who confiscated all of Brahe's equipment. However, Kepler absconded with all of Brahe's data, which he subsequently used to complete his investigations.

Based on his analysis of Mars' orbit, Kepler came to his first important conclusion, namely that the orbits of the planets are *not* exactly circular. In addition, he recognised that the speed with which a planet travels varies along its orbit: a planet moves fastest when it is closest to the sun, a point called *perehelion*, and slowest when it is farthest from the Sun, at so-called *aphelion*. More precisely, he discovered the law of "equal areas in equal times," a simple relation which, for any given planet, relates the distance of the planet from the Sun to the speed with which it is orbiting the Sun.

What this law means is illustrated in Figure 7.2, which represents schematically the planar orbit of the planet around around the Sun (grossly exaggerating the deviation from a circular orbit: none of the planets has an orbit nearly that elongated!). When the planet is far from the Sun, for example at point A, it is moving relatively slowly, so that it will take (say) 10 days to move the relatively short distance to point B. When it is close to the Sun, for example at point C, it is moving faster, so that in an identical period of 10 days it can move all the way to point D. The equal areas in equal times law says that, even thoush the distance along the orbit from A to B is much smaller than the distance from C to D, the areas swept out during the orbit moving from A to B and from C to D, as indicated by the darkened regions in the Figure, are identical.

This law, now known as Kepler's *Second Law*, is a major achievement. However, Kepler had

not yet discovered a precise characterisation of the *shape* of the orbit. He subsequently spend a huge amount of time working on that problem, trying all sorts of alternatives, including ovals and the like, before finally hitting upon the correct answer: *an ellipse*.
The construction of an ellipse is illustrated schematically in Figure 7.1. As is exhibited more carefully in Figure 7.3, an ellipse is characterised by two *foci*, a *semi-major axis*, defined as half the length of the ellipse along the long axis, and a *semi-minor axis*, defined as half the width of the ellipse along the short axis. Note that, for the special case in which the two foci coincide, an ellipse degenerates into a circle.
Kepler's *First Law* states that the orbit of each of the planets around the Sun is an ellipse, with the Sun located at one of the two *foci*. It is a relatively straightforward exercise in geometry to show that the average distance from the planet to the Sun is equal to the length of the semi-major axis. Thus, e.g., one can state that, by definition, the semi-major axis of the Earth's orbit around the Sun is 1 AU in length.
In 1609 Kepler published these results in *Astronomia Nova* or *The New Astronomy*. This was an extremely unpopular book, especially in that it contains a large amount of hard – and wrong – mathematics. Indeed, reading this book it is hard to understand how Kepler actually got to the right answer: he makes numerous obscure, tortuous arguments, augmented by a substantial number of mathematics mistakes which, seemingly miraculously, all cancel out in the end.

In 1611, Rudolph II abdicated as Holy Roman Emperor, and Kepler was sacked as the Imperial Mathematicus. He was forced thereafter to rely primarily on astrology as a source of income, but still continued his research on the motions of the planets. Having discovered how to characterise the motion of each planet separately, the remaining task was to establish a connection between the orbits of the different planets. More precisely, what Kepler sought to determine was how the *size* of the orbit is related to the *orbital period* of the planet, i.e., how long it takes for the planet to orbit the Sun.
Kepler's answer to this question, his so-called *Third Law*, was published in *Harmonice Mundi* or *The Harmony of the World*, which appeared in 1619, the same year that the English poet John Donne visited him in the remote town of Linz. This Third Law provides a direct one-to-one connection between the size of a planet's orbit, as characterised by the semi-major axis, and the duration of the orbit, as characterised by the orbital period. Specifically, what it says is that the square of the period is proportional to the cube of the semi-major axis. Recognising that the Earth has a period of one year and a semi-major axis of one AU, this can be summarised by saying that, for each of the planets,

$$P_{year}^2 = a_{AU}^3,$$

where P_{year} is the period of the planet expressed in years and a_{AU} is the semi-major axis expressed in units of AU.
The meaning of this Law can be illustrated by noting the following pattern:

Mercury has $a \approx 0.39$ AU and $P \approx 0.24$ yr.
The Earth has $a = 1.00$ AU and $P = 1.00$ yr.
Jupiter has $a \approx 5.20$ AU and $P \approx 11.9$ yr.

As the semi-major axis grows, so does the period and, in fact, the period grows *faster* than the semi-major axis. This can be understood by noting that, as one goes further away from the Sun, planets have further to travel *and* they are moving more slowly along their orbits.

Although important because of its articulation of Kepler's Third Law, *Harmonice Mundi* again contains a large number of weird ideas. In particular, in this book Kepler returns to one of his favorite themes, the five regular solids and their supposed role in determining the size of the plamets' orbits and, for the first time, articulates in print the idea that associated with each of the planets is a musical theme which he (at least) can decipher. The general tone of the book can be inferred from its principal theses, as formulated by the American historian of science I. Bernard Cohen in his *The Birth of a New Physics*:

> "Concerning the five regular solid figures.
>
> "On the kinship between them and the harmonic ratios.
>
> "Summary of astronomical doctrine necessary for contemplation of the celestial harmonies.
>
> "In what things pertaining to the planetary movements the simple harmonies have been expressed and that all those harmonies which are present in song are found in the heavens.
>
> "That the clefs of the musical scale, or pitches of the system, and the kinds of harmonies, the major and the minor, are expressed by certain movements."

Figure 7.4 indicates the shapes of the five regular solid and provides a line drawing of a gadget constructed by Kepler to elucidate his vision of the Solar System. Figure 7.5 presents the musical themes which Kepler associated with the heavenly spheres. (The musically inclined who might be interested in reproducing these themes on a piano should note that they were based on the Pythagorean scales of absolute pitch, which dominated European music until the eighteenth century, when it was replaced by the well-tempered scales pioneered by such composers as Johann Sebestian Bach. In particular, the notes that Kepler "heard" could have differed by almost a half-tone from the notes that would be heard if the themes are played on a modern keyboard.)

In 1628, Kepler finally published *The Rudolphine Tables*, based on his Three Laws, and dedicated them to Tycho Brahe. Kepler died in 1630.

Mathematical Appendix: The Equation of an Ellipse

Consider an ellipse oriented in the (x-y) plane, with the major and minor (i.e., long and short) axes aligned respectively along the x- and y-axes. The curve representing the ellipse can then be written as

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

where a and b represent, respectively, the lengths of the semi-major and semi-minor axes. The degenerate case of a circle corresponds to $a = b = R$, so that $x^2 + y^2 = R^2$, where R is the radius.

.

Lectures 8 and 9

NEWTON'S GRAND SYNTHESIS

The quotations from Newton presented here come from Motte's translation of *Principia*, modified by Florian Cajori and published as *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World* (Berkeley, University of California, 1946).

Kepler's Three Laws provided for the first time an accurate mathematical characterisation of the motions of the planets as they orbit the Sun. However, they do not provide a fundamental explanation of *why* it is that the planets execute the orbits that they do. This explanation was finally provided by Isaac Newton, building upon Galileo's theories of dynamics.

Isaac Newton (1642 - 1727) was born on Christmas day 1642, the son of an illiterate yeoman, in the English hamlet of Woolsthorpe in Lincolnshire. A sickly child, he almost died at birth. His father had died three months before Isaac was born, and young Newton was apparently deeply embittered when, at the age of three, his mother married an affluent clergyman and abandoned him to the care of his maternal grandmother. When he was eleven, his mother's second husband died and she returned to Isaac's household with three younger children. She subsequently withdrew him from school with the aim of making him a farmer, but he proved inept at farm chores. At the urging of a local schoolteacher and of his uncle, a clergyman, young Newton returned to school where he learned much Latin but little mathematics. He eventually went to University, but only as a "subsizar," a poor scholar forced to work his way through, a fact which apparently contributed to a deep-seated sense of insecurity. However, despite all these personal tragedies, Newton was lucky in one important respect: The society into which he was born was one very supportive of science and practical intellectual thought, witnessing the growth of a strong merchant class which played an important role in developing the technology that led to the Industrial Revolution.
In 1661 Newton matriculated at Trinity College in Cambridge, where he studied with *Isaac Barrow*, one of the premier English dynamicists and natural philosophers of the day. In 1665 the University was closed because of an outbreak of plague, and Newton, who had just received his B. A., returned home to Woolsthorpe. Over the next eighteen months he apparently made most the great scientific contributions which have come to be associated with his name:

1. He synthesized and extended the thinking of earlier French mathematicians to invent differential and integral calculus and the theory of infinite series.
2. He formulated what are known as Newton's *Three Laws of Physics* and used them to begin developing so-called *Newtonian gravity*.
3. He discovered that so-called white light is really an admixture of all the different colours of the rainbow.

In 1667, he returned to Trinity College as a Fellow (i.e., a faculty member). He remained there for the rest of his life, succeeding Isaac Barrow as the Lucasian Chair of Natural Philosophy, a position which he retained for many years. In 1669, he began teaching his theories of optics and certain aspects of his dynamics. However, he was apparently very reluctant to publish any of his results about dynamics and gravity. In part, this was probably a reflection of the fact that, at least in his youth, he was an intensely reticent, insecure, and private man. However, it would also appear that several important pieces of his theory of gravity were not completely understood until well after 1665. In any event, by the middle of the 1670's several other scientists had begun to reach similar conclusions, including *Christiaan Huygens* (1629 - 1695) from Holland, who is best known for his work in optics and various improvements of the telescope, and *Robert Hooke* (1635 - 1703), an English dynamicist.

In response to the competition, Newton finally decided to publish his results, and began writing a book on mechanics and gravity in 1684. The manuscript was completed and published in 1687, the publication costs being borne by a young Englishman named *William Halley*, for whom Halley's Comet is named. In terms of its broad impact on human intellectual thought, especially in Europe, *Philosophiae Naturalis Principia Mathematica* or *Mathematical Principles of Natural Philosophy* is arguably the single most influential scientific book ever written, before which even the works of Darwin and Freud pale. (The nineteenth century naturalist, Thomas Henry Huxley, recognised implicitly the importance of *Principia* when he called Darwin's *Origin of Species* "the most potent instrument for the extension of human knowledge which has come into men's hands, since the publication of Newton's Principia.") The theories articulated therein gained speedy acceptance throughout Europe by around 1700, largely because they facilitate accurate, numerical predictions. Moreover, once accepted Newton's theories were really not questioned seriously by anyone until the latter part of the nineteenth century. Today the physicist's world view has been altered through the discovery of Einstein's theories of relativity and through the development of quantum mechanics. However, despite these alterations, even now Newtonian mechanics lies at the foundations of physics and is the first topic taught to prospective physicists. One can argue that, in many respects, Newton plays a role in modern natural philosophy similar to that played by Aristotle in the natural philosophy of the Middle Ages.

In later life, Newton combined his interest in physics, mathematics, and astronomy with other interests, including politics, religion, and alchemy (he was also an arsenic eater). In 1689 he was elected to the English Parliament as the representative from Cambridge – at that time Oxford and Cambridge Universities were each assigned their own representative – and he was subsequently reelected. In 1693 he was stricken with mercury poisoning, apparently a result of his alchemic experiments, which caused him to lose all his hair. In 1696 he was made Warden of the Mint of England, where he helped to oversee the production of English coins; and, in 1699, he was appointed Master of the Mint, which brought him the then-extraordinary salary (for a commoner) of £4000 *per annum*. In 1703, he was elected President of the Royal Society of London, chartered in 1662 by Charles II as the first national scientific organisation in Europe, a position to which he was reelected yearly until he died. In 1705, he was knighted, the first commoner to be so honoured for his scientific achievements. In 1727 he died and was buried in Westminster Abbey, the traditional place of internment for English royalty and statesmen.

Newton's seemingly catholic interests were very much in accord with the spirit of the times,

a spirit which was well served by organisations like the Royal Society, whose organisation and everyday business were dominated by the Second Secretary, Henry Oldenburg (1617? - 1677) during the first fifteen years of its existence. Weekly meetings in London, attended by such luminaries as Robert Hooke and the chemist Robert Boyle (1627 - 1691), provided a venue for professionals and amateurs alike to describe their newly discovered results, both in person and through letters directed to the Society. Oldenburg, a polyglot who was fluent in modern Dutch, English, French, German, and Italian, as well as classical Greek, Hebrew, and Latin, played an especially important role in an era where Latin was no longer the universal language of scholarship by translating letters from one language to another, thus facilitating cross-fertilisation between "natural philosophers" from different countries. The *Philosphical Transactions*, a periodical founded by Oldenburg for the Royal Society and which, with only a few breaks, has appeared regularly since 1665, became the model for the modern scientific journal. In particular, the plain and precise style of expression mandated by the Royal Society, and the emphasis on repeatable experiments as opposed to anecdotal description, helped to define the modern form of scientific writing. Also important was the notion that science was an incremental undertaking: Advances in science could, and generally would, involve large numbers of smaller contributions from numerous individuals rather than huge leaps provided by a single dominant thinker. In the words of the physician Edward Tyson (1651 - 1708), "far better a little with accurateness, than a heap of rubbish carelessly thrown together."
(Parenthetically, Tyson himself played several important roles in the development of modern science and medicine. His dissection of a chimpanzee played a major role in convincing naturalists that, at least physiologically, humans are not all that different from some of the anthropoids; and, as director of the Bethlehem Hospital [Bedlam], the second oldest insane asylum in Europe, he proved that, with appropriate treatment, many so-called insane individuals could be cured to the point of being able to function in the real world.)

Legends dating to the eighteenth century would suggest that Newton was a pleasant and affable man. In point of fact, however, he was apparently quite unpleasant, especially in his later years. One of his students wrote that he only observed Newton to laugh upon one occasion, when someone was fool enough to ask him whether there was any benefit in studying Euclidean geometry. He also became notorious in demanding priority for his inventions and discoveries, and upon more than one occasion used his enormous power and prestige to impose his wishes. For example, when his former assistant, William Whiston, his successor to the Lucasian Chair but a man of unorthodox theology, was proposed for membership in the Royal Society in 1720, Newton threatened to resign as President if he were elected as a member.
Perhaps the most blatant example of Newton's unsavory character is manifested by his use of his position as President of the Royal Society in a lengthy priority battle which he had with *Baron Gottfried Wilhelm von Leibnitz* (1646 - 1716), himself a member of the Royal Society since 1673, over who was the true inventor of calculus. (Modern historians of science tend to argue that the two discovered calculus at about the time essentially independently.) In 1699, Newton sent the Royal Society a communication that accused Leibnitz of plagarism. Several years later, as President of the Royal Society, Newton assembled an allegedly impartial committee of seven, including five confirmed Newtonians,

to investigate the charge. In 1712, this committee submitted to the Society a report, authored by Newton himself although this fact was not made public, which, not surprisingly, confirmed Newton's original accusations. The committee's verdict was then read publicly at a special meeting of the Royal Society to which, in an attempt to further embarrass Leibnitz, the entire diplomatic community had been invited. In 1715, almost an entire issue of the *Philosophical Transactions* was devoted to a polemic against Leibnitz that was written by Newton himself. At the time, the strongest piece of alleged evidence to support Newton's case of plagarism was a letter that one John Collins (1625 - 1683) had sent Leibnitz in 1672 which described some of Newton's pioneering work on differential calculus. Only much later in 1852 was it established unambiguously that Leibnitz never received this letter, but only a copy from which, somehow, the suggestive passages were removed.
While Newton's defense of his priority was extreme, it was hardly unprecedented. Galileo, for example, had been involved in various disputes regarding his development of telescopes and their use as astronomical tools. In the Middle Ages such bickering was unusual. However, with the growth of the scientific community and the concommitant increase in the rate of scientific advancement, with the easy dissemination of ideas by printing. and with increased possibilities of financial remuneration for one's discoveries, such priority battles became progressively more common, and are hardly unusual even today. One avowed purpose for the reading of letters and papers at meetings of the Royal Society was to allow scientists to "register" their findings, thus enabling them to claim priority.

Logically speaking, Newtonian physics divides into two distinct pieces. The first of these is *kinematics*, which is concerned with characterising the properties and motions of various objects. The second is *dynamics*, which is concerned with explaining and predicting changes in these properties and motions. One might naively think that kinematics is relatively trivial and, in fact, describing a physical system at a given instant is typically much simpler than predicting its subsequent evolution. However, kinematics is important in that it provides the language in terms of which the laws of dynamics can be expressed.

The fundamental building block of Newtonian kinematics is the notion of an *event*. Characterising an event corresponds to the specification of the position of some object in three-dimensional space, say coordinates x, y, and z, at some instant of time, t. Changes in the position of an object are characterised by its *velocity*. Thus, for example, if the object moves a distance Δx in the x-direction during some short time interval Δt, one says that it has a velocity

$$v = \frac{\Delta x}{\Delta t}$$

directed in the x-direction. Similarly, changes in the velocity of an object are characterised by its *acceleration*. For example, if the object experiences a change in velocity Δv in the x-direction, one says that it has an acceleration

$$a = \frac{\Delta v}{\Delta t}$$

directed in the x-direction. A velocity has units of distance per time, e.g., 5 m/sec. An acceleration has units of velocity per time, e.g., 20 m/sec/sec.

Couched in the language of velocities and accelerations, Aristotle's notion of natural motion corresponds to zero velocity, whereas Galileo's notion thereof corresponds to zero acceleration.

It is important to note that the term *velocity* is *not* equivalent to *speed*: To specify the velocity of an object, one must state both (1) how fast it is going (which *is* the same as its speed) and (2) what direction it is moving. Similarly, *acceleration* refers to *any* change in velocity: speeding up, slowing down, and/or changing direction. For example, a moon of Jupiter moving along a (almost exactly) circular orbit at a constant speed still has a changing velocity and hence a nonzero acceleration.

In Newtonian physics, another important characteristic of an object is its *mass*, a measure of the amount of stuff that it contains. It is important to recognise that mass is *not* the same as weight. An astronaut walking on the Moon only feels about 1/6 as much gravity as he or she would when walking on the Earth and, as such, only weighs about 1/6 as much as he or she would on the Earth. However, the mass of the astronaut is the same.

The notions of time, position, velocity, acceleration, and mass provide the language in terms of which Newton's laws of dynamics are expressed.

Stated in his own words in *The Principia*, Newton's First Law says:

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

If a body is at rest, it will continue to remain at rest in the absence of forces. However, if it is already moving with some velocity, it will continue to move with that velocity, rather than coming to rest. This is a clear and unambiguous statement of the Law of Inertia, as formulated by Galileo. Indeed, even the language is reminiscent of Aristotle and Galileo, with *uniform motion* corresponding to a constant velocity. (Strictly speaking, Newton uses the word "motion" to refer to what is now called the momentum mv of an object, given as the product of its mass and velocity. However, for an object of fixed mass m, constant mv means constant velocity v.) By way of further explication, Newton continues with what seems almost a direct attack against Aristotle and his theory of how an arrow flies through the air:

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

Stated in modern language, Newton's Second Law says that, in order to change the velocity of some object and induce an acceleration in a given direction, a force must be imparted in that direction. Moreover, there is a precise connection between F, the value of the imparted force, and a, the values of the induced acceleration:

$$F = ma \qquad \text{or} \qquad a = \frac{F}{m},$$

where m is the mass of the object that is accelerated.

This Law, and the equation $F = ma$, have three concrete implications: (1) For fixed mass m, a larger force induces a larger acceleration. Thus, for example, if a force twice as large is imparted to an object of fixed mass, an acceleration twice as large will be induced. (2) For a fixed force F, a larger mass means a smaller acceleration. If one object is twice as

large as another, it will be twice as hard to acccelerate, so that the same force will only an induce an acceleration that is one half as large. (3) The acceleration experienced by the object is oriented in the same direction as the force to which it was subjected.
This law is really quite intuitive: If you push gently on a shopping cart filled with bricks you will only induce a small acceleration, but if you push very hard you will induce a large acceleration. Similarly, if you remove the bricks from the cart and replace them by a tiny newborn baby, the cart will be less massive, and it will be much easier to push it towards the pile of bricks with a large acceleration. In each of these cases, when you push the cart it is accelerated in the direction that you push it.
All of this may sound terribly terrestrial. However, the First and Second Laws have a direct, and important, astronomical implication: Given that the orbits of the planets and the other heavenly bodies are not rectilinear, it must be true that some force is acting to accelerate them.

Again stated in his own language, Newton's Third Law says that
"To every action there is always opposed an equal reaction; or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts."
This Law, oftentimes abbreviated by the statement "action equals reaction," implies a fundamental sense of reciprocity within the physical world, a balance between cause and effect. Newton himself provided some simple examples of precisely what it was that he meant: "Whatever draws or presses another is as much drawn or pressed by that other. If you press a stone with your finger, the finger is also pressed by the stone. If a horse draws a stone tied to a rope, the horse (if I may so say) will be equally drawn back towards the stone; for the distended rope, by the same endeavor to relax or unbend itself, will draw the horse as much towards the stone as it does the stone towards the horse, and will obstruct the progress of the one as much as it advances that of the other. If a body impinge upon another, and by its force change the motion of the other, that body also (because of the equality of the mutual pressure) will undergo an equal change, in its own motion, towards the contrary part."

In a more cosmic context, Newton used this notion of reciprocity to argue that, if the Sun exerts an attractive force on the Earth, the Earth must in turn exert an attractive force on the Earth, equal in magnitude but in the opposite direction. In other words, the Earth must pull on the Sun just as hard as the Sun pulls on the Earth. This implies in turn that each exerts an acceleration on the other although, since the Sun is much more massive, the acceleration which it feels is much smaller: $F_{S \to E}$, the force that the Sun exerts on the Earth, induces for the Earth an acceleration a_E satisfying $m_E a_E$, where m_E is the mass of the Earth. Similarly, $F_{E \to S}$, the force that the Earth exerts on the Sun, induces for the Sun an acceleration a_S satisfying $m_S a_S$, where m_S is the mass of the Earth. Given, however, that the two forces are equal, it follows that

$$m_E a_E = m_S a_S,$$

so that

$$a_S = \left(\frac{m_E}{m_S} \right) a_E.$$

The fact that the Sun is about $330,000$ times as massive as the Earth then implies that the Sun will be accelerated only $1/330,000$ times as much.

Newton's Laws talk about the effects of arbitrary forces. They do not, however, explain what force it is that is responsible for the motions of the planets and moons in the Solar System. This Newton did through the identification of gravity as a force. In arriving upon the notion of gravity, Newton was clearly motivated by the idea of *universality*, i.e., the idea that the same law of nature explains why dropped balls fall towards the Earth, why the Moon orbits the Earth, or why the Earth orbits the Sun.

That the same physics can explain both why a ball falls to the Earth and why the Moon orbits the Earth is not hard to understand. Suppose, as in Figure 8.1, that the Earth is approximated as a perfect sphere with nothing on its surface except a cannon. Suppose further that the cannon is given only a small charge of powder, and that a cannonball is fired horizontally. Because the charge is weak, the ball only experiences a tiny initial acceleration in the horizontal direction, whereas it experiences a substantial ongoing acceleration downwards because of the continuing influence of gravity. The net result is that the ball falls almost directly downwards. Suppose now that a cannonball is fired with a somewhat larger charge. In this case, it will experience a larger initial acceleration and, consequently, travel further in the horizontal direction before hitting the ground. If the Earth were flat, this would be the whole story. However, the Earth is round, in this case, approximated as exactly spherical; and, because of its curvature, the Earth is, so to say, falling away from the cannonball as the ball drops towards the Earth. The crucial point then is that, if the ball is given a sufficiently large initial acceleration, it will actually be able to circle all the way around the Earth and come back to where it began. In other words, it would have been placed in orbit as an artificial satellite. Newton interpreted the motion of the Moon as a natural manifestation of the same phenomenon.

Gravity *per se* Newton viewed as a "mystical emanation" spreading out from an object in the same way that light emitted from a bulb spreads out in space. Viewing the light from a bulb as spreading out in an expanding sphere of area $A = 4\pi R^2$, where R is the radius of the sphere, one infers that the amount of light that is seen from the bulb falls off as the square of the distance that he or she is away from the ball. Thus, for example, someone standing fifty feet away from the bulb will see only $\frac{1}{10}\times\frac{1}{10} = \frac{1}{100}$ times as much light as someone standing 10 times as close, only five feet away. By complete analogy, Newton argued that every massive object in nature exerts a gravitational force on every other object, with a strength that varies inversely as the square of the distance between the two objects. Given the additional assumption that the force that (say) object 1 exerts on object 2 depends linearly on the separate masses m_1 and m_2, Newton was led to infer that $F_{1\to2}$, the magnitude of the force that object 1 exerts on object 2, which is equal to $F_{2\to1}$, the force that object 2 exerts on object 1, satisfies

$$F_{1\to2} = F_{2\to1} = \frac{Gm_1m_2}{R^2}.$$

Here R denotes the distance between the two objects and G is simply a constant, the so-called *gravitational constant*, which determines how strong gravity is overall.

The two ideas (1) that gravity is some mystical force emanating from the Sun and (2) that the strength of the Sun's gravity decreases with increasing distance from the Sun had been considered by a number of earlier astronomers, including Kepler. Indeed, following a line of reasoning rather similar to Newton's, Kepler had made the concrete, albeit erroneous, prediction that the strength of the Sun's gravity varies inversely as distance R, rather than inversely as R^2. The only flaw in Kepler's argument was that, visualising the Solar System as a nearly planar entity, he thought of the gravitational emanations as spreading out along a circle of circumference $2\pi R$, rather than a sphere of area $4\pi R^2$.

In his thinking about gravity as a force that acts though empty space, Newton was probably most significantly impacted by the English physician and natural philosopher *William Gilbert* (1540 - 1603) who, in *De Magnete*, his 1600 treatise on magnets, had for the first time articulated clearly the idea that the entire Earth is a gigantic magnet, emanating a magnetic field into space. (Parenthetically, magnets had long fascinated many in the West. Already in use as navigational tools in China by about AD 1000, they were first mentioned in Western writings by the English monk Alexander Neckham (1157 - 1217). However, long after their introduction as a navigational tool they were treated with considerable suspicion because of the idea that they were magical entities. Indeed, Roger Bacon (1219? - 1294), the English scientist and necromancer, was at one point detained in Paris by Church officials because of concern about his experiments with "lodestones." As late as the time of Columbus, many ship captains were inclined to hide their magnetic compasses from their crew because of the misplaced notion that magnets were tools of the Devil.)

Kepler, whose 1596 *Mysterium Cosmographicum* had suggested that the Sun exerts something like a force acting through empty space, was also very much aware of – and impressed by – Gilbert's treatise on magnets. Indeed, in response to this work Kepler was led to a vision of the Universe, more commonly accepted in the eighteenth century, where one could "show that the celestial machine is not so much a divine organism but rather a clockwork ... inasmuch as all the variety of motions are carried out by means of a single very simple magnetic force of the body, just as in a clock all motions arise from a very simple weight." (Kepler is referring to pendulum clocks, and the weight he mentions is the weight that compensates for friction and prevents the pendulum from grinding to a halt.)

Although Newton may have been the first person to identify explicitly the $1/R^2$ gravitational force law, presumably in 1665 or 1666, he was not the first to publish. That honor goes to Christiaan Huygens, who published in 1673, fourteen years before *Principia*. Moreover, Robert Hooke claimed that Newton got the $1/R^2$ law from him, which may or may not be true. (Parenthetically, Hooke's claim of priority on this point so enraged Newton that he removed references to Hooke's work from the manuscript of *Principia*. He also refused to publish his *Opticks* or accept the Presidency of the Royal Society until Hooke's death in 1703.) However, the overall synthesis of physics and gravity into a coherent picture of the Universe is undoubtedly Newton's accomplishment.

The $1/R^2$ force law implies that gravity acts between every two objects in the Universe, even if they are extremely far apart: there is no sense in which gravity "turns off" if one gets far enough away. This implies that, strictly speaking, one will never escape completely from the Earth's gravitational field, even if one is billions of lightyears away in a distant galaxy. The only thing that one can really say is that, if one is sufficiently far from the

Earth, its gravitational influences will be negligible compared with the gravity of other, more proximate, objects.
Despite this fact, it is, nevertheless, useful to introduce the concept of an *escape velocity*, the initial velocity that a body needs to escape from the surface of the Earth or any other celestial body, neglecting the presence of any other objects. If, standing on the surface of the Earth, one tosses an object up into the air with a small initial velocity, it will travel upwards for a short time but then, having reached the uppermost part of its orbit, quickly fall back to the Earth because of the gravity which the Earth exerts on that object. If the object be thrown upwards with a somewhat larger initial velocity, it will reach a somewhat higher point, but still fall back to the Earth. However, if it be given a sufficiently large initial velocity, it will continue to travel upwards and away from the Earth indefinitely, eventually moving infinitely far away unless it is impacted significantly by the gravity of some other object. The minimum upwards speed required for this "escape to infinitely far away" is the escape velocity from the Earth.

Newtonian gravity, as codified in Newton's Three Laws and his Law of Gravitation, provides a coherent physical explanation of *why* Kepler's Laws arise. The fact that the planetary orbits are ellipses is a direct consequence of the fact that they are interacting via an attractive force law that varies as $1/R^2$. If, e.g., the strength of gravity varied as $1/R$, as Kepler had conjectured, the planetary orbits would not be ellipses.
Similarly, the law of equal areas in equal times is a consequence of what the physicist would call *conservation of angular momentum*, which holds for any force that depends only on the distance between two objects. Despite its fancy-sounding name, conservation of angular momentum is something with which many people, and certainly most athletes, are acquainted. If, e.g., a figure skater wants to end a routine spinning as fast as possible, he or she will draw his or her arms and legs very close together, which causes him or her to rotate faster. Alternatively, if a gymnast wants to make a two and a half flip somersault, he or she will tuck as tightly as possible, again in order to rotate faster. As will be seen later on in these lectures, conservation of angular momentum plays an important role in explaining the behaviour of objects like neutron stars and black holes.
The correlation between the semi-major axis, a, of a planetary orbit and the length, P, of its orbital period, i.e., the fact that P^2 is proportional to a^3, is again a consequence of the $1/R^2$ force law. The overall normalisation, e.g., the fact that, for a planet 1 AU out, it takes 1 year to orbit the Sun, rather than 2 or 3, depends explicitly on the mass of the Sun. If, for example, the Sun were only one quarter as massive but the Earth were still 1 AU out, it would take the Earth about two years to orbit around the Sun.
It should be stressed that, with the appropriate normalisation, Kepler's Laws also describe with very high accuracy other motions in the Solar System and the Universe as a whole, e.g., the orbit of the moons around the planets or of artifical satelllites around the Earth. All these different problems involve the same gravitational force and, as such, lead to the same sort of physical motions.

One final point needs to be stressed again: Every object in the Solar System exerts a gravitational influence on every other object and hence causes it to accelerate, to a lesser or greater degree. This implies that, strictly speaking, one cannot say that the Sun is exactly at rest in the middle of the Solar System. Moreover, it is not exactly legitimate to say that the Earth is moving in the Sun's gravitational field, neglecting completely the effects of everything else in the Solar System. Indeed, even supposing that the entire Universe were comprised only of the Earth and the Moon, it would not be entirely correct to say that the Earth orbits arond the Moon. Rather, what is true is that the Earth and the Moon orbit around a common *center of mass*.

The concept of a center of mass is one which almost everyone must understand from playing on a see-saw as a child. If two children of the same mass sitting on the see-saw want to rock around a stable equilibrium, they should both sit the same distance from the pivot point. However, if a small, skinny kid rides the see-saw along with a big, fat kid, they will quickly discover that, to get a stable balance, the fat kid must sit much closer to the pivot. When the two children have seated themselves correctly to achieve the desired balance, the pivot resides at the center of mass for the two children.

Because the Earth is much more massive than the Moon, the center of mass for the Earth-Moon system is much closer to the middle of the Earth than to the middle of the Moon. Indeed, it is actually located inside the Earth. Similarly, because the Sun is much more massive than the Earth, the center of mass for the Sun-Earth system is much closer to the Sun than to the Earth.

Because the Sun is so much larger than anything else in the Solar System, its gravitational influences are much more important overall than that of any other object in the Solar System, and it is reasonable to approximate the Sun as being at rest in the center with the other objects orbiting it. It is in this approximation that Kepler's Three Laws are correct.

Principia was of fundamental importance in that it introduced the Three Laws of Motion and the Law of Gravitation, and used them to provide an explanation of the motions of the heavenly bodies in the Solar System However, *Principia* also played another important role scientifically by codifying what has come to be known as the *scientific method*. Specifically, in a three page interlude about two thirds through the book of over 600 pages, Newton articulates his four *Rules of Reasoning in Philosophy* which, taken together, constitute perhaps the best coherent statement of what science is supposed to be all about and how, ideally, science ought to be done:

"Rules of Reasoning in Philosophy

1. *We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearance.*

   To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.

2. *Therefore to the same natural effects we must, as far as possible, assign the same causes.*

   As to respiration in a man and a beast; the descent of stones in *Europe* and *America*; the light of our culinary fire and of the sun; the reflection of light in the earth, and in the planets.

3. *The qualities of bodies, which admit neither intensification nor remission of degrees, and which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever.*
   For since the qualities of bodies are only known to us by experiments, we are to hold for universal all such as universally agree with experiments; and such as are not liable to diminution can never be quite taken away. We are certainly not to relinquish the evidence of experiments for the sake of dreams and vain fictions of our own devising; nor are we to recede from the analogy of Nature, which is wont to be simple, and always consonant to itself ....

4. *In experimental philosophy we are to look upon propositions inferred by general induction from phenomena as accurately, or very nearly true, not withstanding any contrary hypotheses that may be imagined, till such time as other phenomena occur, by which they may either be made more accurate,or liable to exceptions.*
   This rule we must follow, that the argument of deduction may not be evaded by hypothesis."

The first of these guiding principles emphasises the minimalist approach to science: The aim of science is to explain as much as possible as simply and with as few assumptions as possible. This means in particular that, if several different possible interpretations present themselves, one should select the simplest of these as being the most satisfactory.

The second of these rules, really an implication of the first, emphasises the idea that, to the extent possible, science aims to see various different phenomena as manifestations of the same basic processes and, consequently, to explain them identically. The same laws should explain how both animals and people breathe; the same laws should explain the motion of a ball falling to the ground, the Moon orbiting the Earth, or the Earth orbiting the Sun.

The wording of the third of these guiding principles is a bit more obscure. However, it seems clear from Newton's writings elsewhere that what he is trying to do here is *define* the aim of science as an intellectual discipline, as distinguished, e.g., from other "pure" philosophical speculations. The bottom line is that science is intrinsicially an empirical discipline, which aims to explain what will happen if one acts in a particular way or performs a particular experiment. In this sense, its aim and *raison d'etre* are fundamentally distinct from those of religion and theology. There is no implicit value judgment about the relative importance of science and other mental pursuits, just the recognition that their aims are different.
This particular point was extremely important to Newton, who was a devoutly religious – and mystical – man (although it would be a mistake to consider him a true mystic). Indeed, especially in his later years much of Newton's intellectual focus was on questions related to the Bible and Prophecy, as well as alchemy, as is reflected by the 650,000 words he wrote about alchemy and the 1,300,000 words that he wrote about religion and theology.

The last of these principles articulates perhaps the most fundamental and important concept of all, namely that, from the viewpoint of science, truth is only a working hypothesis. When we have constructed some basic theory in terms of which to explain a class of physical phenomena, we should, as minimalists, stick with that model until or unless we are forced to reject it, rather than introducing other alternative theories. However, we should not believe that the model is fundamentally true in some deep and ultimate sense. The aim of science is to explain the physical world, and all that we can really say about a model is that it is good or bad at explaining what is actually seen.

.

Lecture 10

EIGHTEENTH AND NINETEENTH CENTURY COSMOLOGY

Newton's physics provided a coherent and compelling description of the motions of the pieces of the Solar System known in the early 1700's, and was also extremely successful in explaining the motions of other planets and moons that were discovered in later years. However, neither Newton nor any of his contemporaries had managed to fit the stars into their overall cosmology. Given the dynamical world view described in *Principia*, it seemed ludicrous to view the stars as being attached to a celestial sphere frozen somehow in the sky, but it was not yet clear either what the stars are or how distant from us they are.

Already before Newton there were clear observational indications that, contrary to what had been believed in ancient times, the stars are not completely unchanging entities. *Supernovae* in our galaxy which, on the average, are visible on the Earth every couple of centuries, had been observed by both Brahe and Kepler, who saw them as distant objects that appeared to turn on abruptly and then fade away over the course of about a year. Other so-called *variable stars* had also been discovered. In 1596, *David Fabricius* had discovered a relatively bright star, visible to the naked eye, that appeared to fade away. However, in 1638 *Holwerda* had seen the star reappear and, by tracking it over the course of several years, discovered that it actually oscillated in brightness with a well-defined period of about 332 days. Numerous other variable stars were discovered subsequently.

Further advances resulted from the development of increasingly powerful telescopes, which rapidly improved from the primitive pair of lenses used by seamen at the time of Galileo. Newton himself had played an important role in the development of the so-called *reflecting telescope*, which focuses light with a mirror, unlike Galileo's *refracting telescope*, which relies instead on a lens. (Indeed, it was Newton's working model of a reflecting telescope that secured his membership in the Royal Society in 1672, fifteen years before the publication of *Principia*.) Huygens had also played an important role in the development of the telescope, using one to discover Titan, the large moon of Saturn, in 1655. The Italian astronomer *Giovanni Cassini* (1625 - 1712) working in Paris, the first to discover structure in the rings of Saturn, also discovered four of Saturn's moons between 1671 and 1684.

One individual who played an especially important role in recognising the existence of complicated structures outside the Solar System was *William Herschel* (1738 - 1822). Born in Hanover, Germany, and a musician and music teacher by profession, he and his sister, *Caroline Herschel* (1750 - 1848), immigrated to England, where they began to indulge their amateur interests in astronomy. William Herschel was responsible for the construction of several big telescopes, by far the largest that had ever been constructed, including one that was 40 feet long and had a mirror 58 inches in diameter. Unfortunately, this particular telescope was too big and unwieldy to manipulate easily, but the Herschels made numerous significant discoveries using a smaller 20 foot telescope with a 19 inch mirror. Equipped with a suitable eyepiece, even this much smaller telescope was able to achieve enormous magnifications, as much as the unheard of factor of 6,450, along with a very high resolution, much better than anything else that had ever been achieved.

By this time, astronomers had begun to believe that stars must be objects that actually move through space, but they also recognised that they must be extremely distant since they do not exhibit large, conspicuous motions over the course of a year as the Earth orbits around the Sun. The real question was: how far away? To answer this question it was realised that what one should look for was tiny apparent motions of nearby stars over the course of the year, relative to the more distant stars, a phenomenon now known as *parallax*.

William Herschel set out to look for evidence of such parallax and, although he never found it, he *did* make several other crucial discoveries. In 1781, he actually discovered an object that appeared to move relative to the remaining stars. However, it quickly became apparent that it was relatively close to the Sun – only about 19 AU away – so that it must be a planet, rather than a star: Herschel had discovered Uranus. Subsequently, in 1787 he found two of Uranus' moons and in 1789 he discovered two more moons of Saturn. The discovery of Uranus, originally named Georgium Sidus (or George's star) after King George III, secured Herschel an annual stipend from the British Crown which let him stop teaching music students and work full time as an astronomer.

Although he did not detect a yearly motion of stars resulting from the Earth's motion around the Sun, he *did* discover longer term, systematic motions of some stars relative to each other. Specifically, he discovered the existence of *binaries*, pairs of stars which orbit around each other with periods of tens to hundreds of years. Over the course of his lifetime, he tracked over 700 such objects. The motions which he observed were tiny, only a few seconds of arc, so that is was difficult for him to measure angular distances very accurately. However, he was able to determine unambiguously that the members were moving relative to each other since their mutual orientation relative to other stars was changing over the course of time. This discovery of binary stars was important in changing astronomers' view of the Universe because it demonstrated the existence of other systems, comprised of more than one object, outside the Solar System.

In addition to these motions of stars in binaries, Herschel observed systematic motions for several individual stars over the course of a number of years. These so-called *proper motions* were important in that they reinforced the idea that the stars must be moving through space.

Herschel also played an important role in developing the recognition that there are objects other than individual stars outside the Solar System. Improved telescopes had led to the discovery of numerous so-called *nebulae*, seemingly amorphous blobs of light that were named for their overall nebulous, or cloudy, appearance. Some of these are now known to be galaxies, collections of millions or billions of individual stars that could not be resolved until the development of yet better telescopes. Others, however, like the Orion nebula, are relatively nearby and hence much easier to see. In his observations of the Orion Nebula, Herschel detected both individual stars and an ambient milky glow which he interpreted as evidence that something other than stars must be there. Today it is known that this glow is associated with interstellar gas that is heated to a glow by energy emitted from the visible stars.

Despite these advances, Herschel had not measured regular annual motions of the stars and, until this was done, it was not possible to determine how far away they are. Such measurements only became possible in the mid to late 1830's, with the development of improved technology to measure the locations of the stars in the sky. One major advance was the development of improved measuring apparatuses, also important in surveying, which permited more accurate determinations of angles in the sky. The other major advance, due largely to *Joseph Fraunhofer* (1787 - 1826), was the development of improved eyepieces that helped minimise the distortion of an image as it is focused in a telescope.

During the mid 1830's, several different astronomers found indications of the elusive stellar parallax. However, the individual who is generally credited with the first conclusive measurements is the astronomer *Friedrich Wilhelm Bessel* (1784 - 1846) originally from Königsberg, Germany, who moved to St. Petersburg and became one of the founders of Russian astronomy. Originally employed as the clerk for a German merchant, Bessel developed an interest in navigation and, subsequently, in astronomy. He was both an extremely good mathematician (after whom so-called Bessel functions are named) and a talented observer. In particular, he was a pragmatist who played an important role in convincing other astronomers to make careful estimates of the intrinsic errors in their measurements, particularly those reflecting human error.

As illustrated in Figure 10.1, an astronomer measures distances to nearby stars in the same way that a surveyor might, for example, measure the distance to some relatively nearby object like a tree. The surveyor proceeds first by setting up a *baseline* of fixed length L, and, standing at each end of the baseline, lines up the tree with some much more distant object, e.g., a mountain that is far away. Given the angles, α and β, between the baseline and the tree, and indulging in some high school geometry, he or she can then determine the so-called *parallax angle* p and hence the distance d to the tree. When surveying the distances to the stars, the astronomer proceeds identically, using the motion of the Earth around the Sun to define a baseline of length 2 AU.

Between about 1837 and 1840, Bessel measured parallaxes for several different stars, measuring angles p corresponding to distances of around $500,000$ AU. The largest parallax angle which he found, corresponding to the closest star, was about 0.91 seconds of arc for the star Alpha Centauri. Subsequent more accurate measurements gave 0.76 seconds of arc, this corresponding to a distance of about 270,000 AU away, i.e., about 4.3 lightyears. In determining which stars to study, Bessel was guided by two ideas: (1) He should look at bright stars, both because they are easier to see and (so he thought) because they should be closer to the Earth. (2) He should look for stars that exhibited large peculiar motions since again they should probably be closer to the Earth.

When astronomers began trying to measure parallaxes for other stars, they rapidly came to two interesting, and significant, conclusions. The first of these was that many stars, particularly dim ones, do not have measureable parallaxes, this indicating that they are substantially further away. Indeed, only about a thousand stars have parallaxes sufficiently large to permit a direct distance determination, bigger than about one hundredth of a second of arc, the best that was possible as late as the 1980's. The second point was that although, as had been expected, brighter stars tend overall to have larger parallaxes, there is not a very tight correlation between the observed brightness of a star and how far away it is.

This latter fact follows as a consequence of an important, and fundamental, point: When we look at some object, it may look bright *either* because it is intrinsically luminous *or* because it is relatively nearby. If the object looks dim, this may be either because it is intrinsically dim or because it is very far away. The inference drawn from the weak correlation between distance and apparent brightness is that stars come with a broad range of intrinsic brightnesses.

It was also important to be able to determine more accurately how bright a star actually looks. During the middle of the nineteenth century, this was typically done using a *photometer*. The idea underlying this tool was to compare the light enitted by some star with a standard light signal of variable intensity, and to determine the intensity of the variable signal to exactly match the observed, or apparent, brightness of the star. Towards the end of the century, with the development of photography, it became more common to determine the apparent brightness by analysing photographic plates, a technique that was widespread until comparatively recently, when it was replaced by the use of so-called *CCD*'s or *charge-couple devices*.

Given a knowledge of how bright some star appears and how far away it is, one can work backwards to find out how bright it really is intrinsically. In doing this, one uses the idea that light emitted by a star travels outwards on a sphere of increasing area $4\pi R^2$, so that, e.g., an object of fixed intrinsic brightness that is twice as far away will only look one quarter as bright. By proceeding in this fashion, astronomers came to the conclusion that the stars that they saw were in fact extremely bright, typically comparable in brightness to the Sun. This suggested that stars might be objects relatively similar to the Sun.

This idea was corroborated by another result, to be discussed more carefully in a later lecture. Specifically, in the middle of the nineteenth century it was discovered that, for opaque objects like stars, there is a direct correlation between the colour of the object and the amount of light being emitted by each little patch on its surface. More precisely, the surface temperature of the object determines completely both the colour of the light and how much energy each portion of the surface emits. Thus, if one knows both the colour of the star and its intrinsic brightness, and assumes that it is spherical in shape, one can determine what its radius must be. This exercise leads to conclusion that, overall, most stars have radii comparable to that of the Sun.

These various discoveries led finally to the conclusion that the Sun must be a star and, consequently, the idea that there are many other worlds aside from our Solar System. Thus, in particular, there was a growing belief that other stars might also be surrounded by planets. However, such planets (presuming that they existed) were much too faint to be seen with nineteenth century telescopes. Only in the 1990's has relatively direct evidence emerged for the existence of planets orbiting other stars.

Having established the existence of stars as entities comparable to the Sun, it remained to address the question of how they all fit together: How large is the Universe overall and what is its structure?

Even the telescopes of the eighteenth century were able to reveal that the blur across the sky known as the Milky Way contains large numbers of stars. Motivated by this observation, a number of different individuals, notably *Johann Heinrich Lambert* (1728 -

1777) in 1761, articulated the idea that thousands of stars orbit around the Sun, forming a single gigantic system. Indeed, Lambert and the German metaphysician *Immanuel Kant* (1724 - 1804) went further to argue on philosophical grounds that there might also exist other similar gigantic systems as well. However, at the time this was rank speculation, with no observational motivation at all.

The Herschels introduced a simple argument to determine somewhat more precisely the position of the Sun in the surrounding *galaxy*. Their key assumptions were (1) that the stars around the Sun are distributed more or less homogeneously within a region of finite size and (2) that nothing obscures our vision, so that it is possible to see all the way to the edge of this region. Given these assumptions, one infers that if one sees many stars in one direction the Sun must be relatively far from the edge of the galaxy in that direction, but that if one sees only a few stars in that direction the Sun must be close to the edge. Based on this reasoning, the Herschels were led to what has come to be known as the *grindstone model*, whereby the Sun is assumed to live near the center of a conglomeration of stars that is shaped roughly like a giant grindstone or wheel.

Using the typical distance between nearby stars to set the overall scale, and again assuming a roughly homogeneous distribution, in the late nineteenth century the Dutch astronomer *Jacobus Kapteyn* (1851 - 1922) went further to estimate the linear dimensions of the wheel. His conclusion was that it is about 30,000 lightyears across and 6,000 lightyears thick. Today it is known that these numbers grossly underestimate the true size of the galaxy.

Another major lacuna in nineteenth century scientists' understanding of the Universe was a clear sense as to how old the Earth and other objects in the Universe are, and when the Universe was first created. The Irish clergyman James Ussher (1581 - 1656) had used an analysis of scripture and other theological writings to conclude that the Universe was created on 26 October, 4004 BC at 9:00 AM. However, in the eighteenth and nineteenth centuries this caused a serious problem for the rapidly evolving discipline of geology. It was recognised by that time that processes like erosion and volcanism slowly change the surface of the Earth, but it was also clear that, without invoking phenomena very different from what is observed at the present time, it is impossible to explain all the changes inferred from geological strata in a time of less than 6000 years. In the eighteenth century this led to the postulated existence of special epochs called "catastrophes" when these changes occurred with particular rapidity.

The first attempt to prove scientifically that the Earth must be significantly older than 6000 years was provided through an ingenious argument presented by *George-Louis Leclerc, Comte de Buffon* (1707 - 1788). Basically, Buffon assumed that the Earth was created somehow from debris ejected from the Sun when that object first formed. However, supposing that much of the Earth is comprised of iron and that, at the time of its creation, the Earth must be at least white hot, Buffon extrapolated from experiments performed on iron balls in his laboratory to conclude that the Earth must be at least 74,832 years old.

By the late nineteenth century, geologists like *Charles Lyell* (1797 - 1875) who embraced the doctrine of Uniformitarianism, which advocated a slow and continuous evolution of the Earth's surface, were convinced that the Earth must be millions of years in age. However, a clean estimate as to the age of the Earth – more than four billion years – was impossible until the advent of radioactive dating, a technique developed by the physicist *Ernest*

Rutherford in 1929.

By 1900, there had emerged a cosmology with the Sun as a typical star living in the middle of a galaxy of comparable stars, roughly 30,000 by 6,000 lightyears in size. Whether there also existed other galaxies was unclear: Astronomers had discovered so-called *spiral nebulae*, cloudy pin-wheels in the sky, and a number of them wanted to interpret these spiral nebulae as other galaxies. Thus, in particular, *William Parsons*, who had constructed the largest telescope in the world in 1845 – 60 feet long with a 6 foot diameter mirror – was convinced that he could detect individual stars in the so-called *Whirlpool Galaxy*. However, at the time that was still a minority opinion. Enlarging the Milky Way Galaxy to its proper dimensions and then establishing conclusively the existence of other galaxies were the most significant achievements of astronomy during the first quarter of the twentieth century.

Mathematical Appendix: Half-Lives

Modern theories of particle physics, based on so-called quantum mechanics implies that certain particles (like neutrons) and certain combinations thereof (including various nuclei like carbon-14 and strontium-90) are unstable in the sense that they will eventually decay into smaller constituents. The obvious question then is: when will such unstable particles (or particle combinations) decay? According to the same laws of quantum mechanics, this question is impossible to answer definitively for any given particle. However, quantum mechanics *does* allow one to provide simple *statistical characterisations* of how a *collection of particles* (or particle combinations) will eventually decay. Specifically, given an initial sample of N_0 particles, one finds that the average number $N(t)$ at later times will decrease exponentially in time t, in a fashion characterised completely by the *halflife*, $T_{1/2}$, for the particular particle to decay:

$$N(t) = N_0 2^{-(t/T_{1/2})}.$$

On the average, after a time $1T_{1/2}$ half ($\frac{1}{2}$) of the particles will still remain, after a time $2T_{1/2}$ one quarter ($\frac{1}{2}\times\frac{1}{2}$) will remain, after $3T_{1/2}$ one eighth ($\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}$), and so forth.

Lectures 11, 12, and 13

THE BASIC IDEAS OF SPECIAL RELATIVITY

The basic framework of Newtonian mechanics, as outlined several lectures ago, really involved two ingredients, namely (1) the Three Laws of Motion, which are intended to describe how an object will react in response to *any* force that one might envision, and (2) the Law of Universal Gravitation, which identifies the particular force, gravity, which is responsible for the bulk motions of the objects in the Solar System and throughout the Universe. However, this theory did not identify explicitly any of the other fundamental forces that exist in nature, such as electricity and magnetism, and, moreover, it was relatively disjoint from Newton's theories about light, what he himself called *opticks*.

In Newton's day, the nature of light was hotly debated by natural philosophers: was light particulate in nature, as Newton himself assumed, or was it instead fundamentally wavelike in character, as argued, e.g., by Huygens? Using a model which views light as a collection of small particles, Newton found it easy to explain how it is that light reflects off a mirror. However, a particulate model has a difficult time explaining such phenomena as *diffraction*, the blurring of light as it passes close to an opaque object or through a thin aperature, which results in a regular pattern of light and dark regions. For this reason, the interpretation of light as a wave became increasing accepted during the eighteenth and early nineteenth centuries.

At the same time, during the late eighteenth and early nineteenth centuries, numerous experiments performed by various different natural philosophers were also beginning to yield an understanding of electricity and magnetism, phenomena which had seemed completely mysterious as late as the seventeenth century. In particular, there emerged a recognition that electric effects are associated with charged particles. More precisely, it was recognised that some particles contain specific amounts of *charge* that generate a surrounding electric field in the same sense that all particles with mass generate a surrounding gravitational field. Moreover, it was realised that magnetism is associated with *currents*, i.e., motions of charged particles.

For a long time, it was not clear how the different pieces of the puzzle that were emerging could fit together into a single coherent whole. However, it was eventually realised that, fundamentally, electricity and magnetism are in fact two different manifestations of a single phenomenon. This realisation served as the basis for the unified description of electromagnetism finally provided in 1865 by the Scottish physicist *James Clerk Maxwell* (1831 - 1879), who formulated what are now known as Maxwell's equations.

In a real sense, Maxwell played the same role for the evolving science of electromagnetism in the nineteenth century as Newton had played for mechanics and gravity in the seventeenth century. In particular, Maxwell consolidated and systematised the work of earlier experimentalists like the American *Michael Faraday* (1791 - 1867) in much the same way that Newton systematised the work of Galileo and Kepler. (It is interesting that, in his later years, it was a photograph of Faraday that hung in Einstein's office, and that it was the experimentalist Faraday that the theorist Einstein felt was the predecessor on whose work he most relied.)

Subsequently, it was recognised mathematicially that Maxwell's equations make the phys-

ical prediction that charged particles can emit *electromagnetic radiation*, i.e., waves of energy, into space. This was confirmed experimentally in 1887 by the German physicist *Heinrich Hertz* (1857 - 1894), who was apparently the first to recognise that the phenomenon we know as light is in fact electromagnetic radiation. Thus, in particular, he understood that light emitted from the surface of the Sun carries energy away from it, and that this energy, when deposited on the earth in the form of heat, is responsible for the fact that the Earth is as warm as it is.
Actually, the idea that light is somehow related to electricity and magnetism predates both Hertz and Maxwell. The third edition of *Encyclopaedia Britannica*, published in the 1790's, contains a weird (and largely incorrect) article on electricity by one James Tytler that asserts that electricity is a vibration like light, rather than a material flow. More concretely, in 1845 both Faraday and the young *William Thompson* (1824 - 1907), later known as Lord Kelvin, performed simple experiments which demonstrated that light is affected by electric and magnetic fields.

What, however, is a wave? A wave is in fact something with which everyone is acquainted in real life, corresponding to a periodic pattern that travels through some medium. Examples include gentle undulations of the water in the bathtub when you wiggle your toe; the pattern of compression that is generated when you play with a slinky©; the wave in which you participate as a football spectator; and seismic waves associated with earthquakes, which travel through the mantle and core of the Earth. The phenomenon which we perceive as sound is in fact a wave: When you talk to someone, your vocal chord generates waves in the air around you which propagate through the air until they reach your audience's ears, where they cause oscillations in the eardrums which are interpreted by the brain as sound.

How should one characterise a wave? By definition, a wave is a repeating pattern, so that one can obviously ask the distance required before the pattern repeats itself. This distance is termed the *wavelength* of the wave, and is customarily denoted by the Greek letter *lambda*, i.e., λ. Similarly, one can ask how big overall the wave gets, i.e., whether the wave is strong or weak. This quantity is characterised by its *amplitude*, A. The definitions of λ and A are exhibited explicitly in Figure 11.1.
If, e.g., one generates a water wave in the bathtub, the water in the wave moves vertically, as is obvious from the behaviour of a rubber duckie floating in the water. However, the *pattern* of the wave travels horizontally, and it is easy to characterise the rate at which the wave moves. For example, one could make a mark on the side of the tub and then use a stopwatch to measure the time between one crest passing the mark and another passing. Having done this, one could then compute the number of crests passing per time interval, which is known as the *frequency*, customarily denoted by the Greek letter *nu*, i.e., ν.
One final point: If one knows the distance between successive crests, i.e., the *wavelength*, and how many crests pass by per unit time, i.e., the *frequency*, he or she can multiply them together to obtain the speed c with which the wave is travelling by, i.e.,

$$c = \lambda\,\nu.$$

For example, if the length of the wave is 20 cm and 5 crests pass by each second, i.e.,

$\lambda = 20$ cm and $\nu = 5/\text{sec}$, the speed of the wave is

$$c = \lambda \times \nu = 20 \text{ cm} \times 5/\text{sec} = 100 \text{ cm/sec} = 1 \text{ m/sec}.$$

All of this is true for any wave. What makes light special is that, at least in empty space, all wavelengths travel at the same speed c, about $300,000$ km/sec. This implies that there is a direct one-to-one correlation between the wavelength of light and its frequency: the longer the wavelength the lower the frequency.
Another important fact is that the visible light with which we are all familiar is only one very special type of light, characterised by a very narrow range of wavelengths. The colours of the rainbow extending from red through green through violet correspond to wavelengths between about 700 and 400 nanometers, where one nanometer equals 10^{-9} meters. The totality of light, or electromagnetic radiation, extends in decreasing wavelengths through radio waves, microwaves, infrared radiation, visible radiation, ultraviolet radiation, x-rays, and gamma rays. Radio waves can have wavelengths as long as a meter or even more; gamma rays have very short frequencies, down to 10^{-13} meters and less. Because they have the longest wavelengths, radio waves have the lowest frequencies; and similarly, because they have the shortest wavelengths, gamma rays have the highest frequencies.

The statement that the speed of light is a constant, independent of wavelength, only refers to motion through a vacuum. Light does in fact slow down somewhat as it passes through various materials – e.g., when passing through glass or water or, by a very small amount, when passing through air – and how much it slows down depends somewhat on the wavelength. Moreover, various types of material are nearly opaque to various wavelengths of light. Thus, e.g., the Earth's atmosphere blocks gamma rays, x-rays, ultraviolet radiation, and some infrared wavelengths impinging from outer space. It follows that, in order to see extraterrestrial signals at these wavelengths, astronomers must place telescopes above the Earth's atmosphere.

But what does all of this have to do with astronomy? Basically, there are two obvious connections. The first is the fact that when we detect the presence of some object in outer space, this almost invariably involves observing some of the light which that object has emitted. The second point is that, in varying degrees, every object in the Universe emits both light and heat, albeit quite possibly at wavelengths which are invisible to the naked eye. (As will be discussed later, classical general relativity implies that black holes do not emit any light. However, any material near a black hole will be impacted by the hole and, consequently, emit various light signals. Moreover, incorporating general relativity and quantum mechanics leads to the prediction that black holes *will* in fact emit various forms of radiation.) For example, we see the Sun because the Sun emits light, which radiates to the Earth, interacts to a certain extent with the Earth's atmosphere, and eventually enters our eyes. Moreover, this light carries energy and, as will be discussed in a later lecture, the phenomenon which we perceive as heat is a form of energy. It follows that it is the light emitted from the Sun which is responsible for heating the Earth to the temperature which it has.

All the other examples of waves denumerated above – water waves in a bathtub, compression of a slinky, seismic waves, and sound – travel *through* something, be it water, earth, or air, and, by analogy, scientists in the nineteenth century wanted to visualise electromagnetic waves as propagating through some medium. Thus, for example, Maxwell and many others could not imagine a field as a self-supporting entity propagating in free space. They therefore embraced the concept of an *ether*, a strange undetected substance that was supposed to represent the material through which light moves. In particular, it was with respect to a frame of reference in which this ether is at rest that light was supposed to travel with speed c. Today, it is known that no such ether exists, i.e., that light really travels through empty space, i.e., a *vacuum*. However, the concept of an ether was of fundamental importance to the thinking of the late nineteenth century, and played a direct role in the discovery of Einstein's relativity. (The idea that so-called empty space is really filled with an ether through which everything, including particles and light, moves arose well before the nineteenth century, dating back at least to the French mathematician and philosopher *Rene Descartes* (1596 - 1650).)

Another key idea in nineteenth century thought was the belief that Newton's Laws in general, and Galilean relativity in particular, should hold for light as well as any other substance in nature. Galilean relativity makes a specific, concrete prediction about how velocities add: Suppose that, when some baseball pitcher (or cricket bowler) throws a ball as hard as he can, it leaves his hand at 90 mph, regardless of whether he is standing still or in motion. If he stands still and throws the ball towards a stationary batter (or batsman) as hard as he can, that batter will obviously see and feel the ball coming towards him at 90 mph. Suppose, however, that the pitcher is running forward at 10 mph and then throws the ball as hard as he can. In this case, Galilean relativity (and common sense) both imply that the batter will perceive the ball coming towards him at $90 + 10 = 100$ mph. The fancy way of saying this obvious fact is that, according to Galilean relativity, as incorporated in Newton's Laws of Motion, velocities should add linearly.

The nineteenth century scientist would have said that if this argument works for balls travelling through the air, it should also work for light travelling through the ether. Suppose, first of all, that you are standing still relative to the motion of the ether and that, performing a careful experiment, you measure that the velocity of some light signal is c. Then suppose that you start moving relative to the ether with a velocity v and that, while moving, you again measure the velocity of the same light signal. In this case, the nineteenth century scientist would say that you should *not* measure the same velocity: instead of c, you should instead measure a value $c + v$.

Light travels so fast that this effect, even assuming it exists, would be unobservable in everyday life. Nevertheless, nineteenth century scientists thought that there *was* at least one way in which the effect could be seen: They knew that the Earth travels around the Sun with a speed of around 30 km/sec, about 10^{-4} times the speed of light. Assuming, however, that the Earth is not stationary with respect to the ether, an observer looking at some light signal at different times of the year or travelling in different directions should measure it as having slightly different speeds. For example, the measured velocity should be slightly different in magnitude when the Earth is moving downstream from the ether

than when the Earth is moving upstream. The effect to be measured is very small, only about one part in ten thousand, but, with the improved technology of the late nineteenth century, it was big enough to be detected, provided of course that it existed.
In the 1880's the American physicist *Albert A. Michelson* (1852 - 1931) and his collaborator *Edward W. Morley* (1838 - 1923) began trying to measure this effect, their principal motivation being to determine the motion of the Earth relative to the ether. However, after a series of extremely careful experiments they were led to the conclusion, published in 1887, that *there is no measurable effect!* It appeared that the speed of light does not depend on the direction in which the light travels, and that the Earth's presumed upstream or downstream motion relative to the elusive ether does not matter.
Michelson and Morley repeated their experiments over the years as the requisite technology improved, but no effect was ever detected. These experiments have continued even to the present, leading to the conclusion that, irrespective of the direction of motion, the measured speed of light on the Earth is constant to within 1 m/sec.

This conclusion came as a nasty shock to physicists and astronomers in the nineteenth century, since it forced them to a distinctly unpleasant conclusion: either the Earth is at rest relative to the ether or else Galilean relativity, the starting point of all of Newtonian physics, must be wrong. The idea that the Earth was at rest relative to the ether seemed unacceptable, given that the Earth was orbiting around the Sun and that the Sun was presumably moving through space (although some scientists tried to concoct models whereby the Earth dragged a surrounding shroud of ether along with it). However, rejecting Newtonian mechanics seemed close to heresy.
The problem became even more acute when, in the 1890's, several scientists, notably the Dutch physicist *Hendrick Lorentz* (1853 - 1928), recognised that Maxwell's theory of electromagnetism implied, in agreement with the findings of Michelson and Morley, that the observed velocity of light *should not change* during the course of the year. Rather, Maxwell's equations make the specific unambiguous prediction that, contrary to the predictions of Galilean relativity, *every observer should measure the same value for the velocity of light*. Maxwell's theory of electromagnetism also disagreed with Newtonian mechanics!
It is worth stressing exactly how strange and counter-intuitive this result is using the earlier baseball analogy: If the pitcher stands on the mound and aims a flashlight at a stationary batter, that batter will detect the light coming towards him with a velocity c. So far so good. However, if the pitcher runs towards him with velocity v and again shines the flashlight towards him, the batter will *not* see the light coming at velocity $c + v$. Rather, he still sees it coming with the same velocity c.

Given these observations, there seemed two alternatives available for the scientist around 1900. One involved trying to modify Maxwell's theory of electromagnetism to avoid the prediction of a constant velocity and to hope that – somehow, anyhow – the Michelson-Morley result would disappear. The other involved accepting that the velocity of light is the same for all observers and investigating what that implies.
Given that Maxwell's theory had already withstood many other experimental tests and that it actually seemed to explain what was seen, the former alternative might seem relatively contrived from a modern perspective. The fact that many scientists clung to it is therefore

evidence of the extraordinary influence that Newton and his thinking still exerted on the scientific community. In any event, this tact was not successful. It seems difficult, if not imposible, to construct an alternative theory consistent with experiments involving light that does not violate Galilean relativity.

Scientists were thus forced to entertain the idea that the speed of light is in fact constant, independent of observer. However, this idea is inconsistent with Galilean relativity, which follows as a logical consequence of two, and only two, basic assumptions:

1. *Our ordinary notions of space and time are correct.* In particular, time is an attribute of nature that moves forward uniformly, independent of anything that happens in the Universe. The duration of any physical process has objective meaning, independent of how anyone measures it; and, presuming that measurements are done correctly, everyone will agree on how long something takes. Similarly, at any given instant of time the distance in space between two different objects has objective meaning, again independent of how anyone measures it; and presuming that the measurments are done correctly, everyone will agree on how far apart the two objects are.

2) The natural motion of the Universe corresponds to motion with a constant velocity. In particular, the laws of dynamics and physics imply that, in the absence of perturbing forces, objects move along rectilinear trajectories at constant speed. It is changes in velocity that are fundamental, rather than their absolute values. Moreover, it is impossible to establish absolutely whether or not an object is at rest. *Provided that an object is not being accelerated, the best that one can do is ascertain whether or not it is moving relative to some other object.*

Special relativity is an alternative to Newton's world view, predicated on the notion that it is changes in velocity that are fundamental but surrendering the ordinary notions of space and time. In particular, the original 1905 formulation of special relativity developed by the German, and later American, *Albert Einstein* (1879 - 1955) was based on assuming (1) that changes in velocity are fundamental and (2) that the velocity of light is the same for all observers, and then working out the logical implications of these assumptions.

Because it involves surrendering the ordinary notions of space and time, special relativity leads to physical predictions which, albeit verified experimentally, directly contradict our ordinary intuition. However, it *does* imply that, at least when dealing with objects travelling at low velocities, much less than c, Galilean relativity and Newton's world view are approximately correct. This is essential since, over the past three hundred years, Newton's dynamics have been found to provide a highly accurate description of many observable phenomena, including the motions of the planets and other celestial bodies.

The first, and most obvious, prediction of special relativity is that *all objects with nonzero mass must travel with speeds smaller than the speed of light.* Light, which is a massless entity, travels at c, the speed of light, but any object that does have mass must travel more slowly. According to Newton, the bigger the force the bigger the acceleration; and, if an object is given a sufficiently large acceleration, it can achieve an arbitrarily large velocity. Not so in special relativity. Special relativity agrees that larger forces induce larger accelerations. However, as an object is accelerated up to larger velocities approaching c it becomes progressively more difficult to accelerate. One way of thinking about this is that, as the object moves faster, it acts as if it has a larger mass. In any event, it would require an infinite amount of energy to accelerate the object up to a velocity c. This prediction is corroborated by physicists every day in accelerators and cyclotrons, which can accelerate subatomic particles up to speeds in excess of $0.999999c$, but not all the way to c.

One particularly striking, and counter-intuitive, physical prediction of special relativity is *time dilation*, i.e., the fact that *time slows down for a moving object*. Thus, in particular, one infers that a clock moving with a nonzero velocity will tick more slowly than a clock that is sitting at rest. Here one is *not* saying that there is some force or anything like that which perverts the clockwork to make it tick more slowly. Rather, one is saying that, for the moving clock, time itself has slowed down. Special relativity mixes the notions of space and time in such a way that, for an object moving through space, time moves more slowly. This fact gives rise to what is sometimes called *the twin paradox*. Suppose that there are two identical twins, born at essentially the same time, each of whom is presented with a spaceship capable of traveling at a speed very close to the speed of light. One of the twins chooses to remain on the Earth. However, the other twin takes advantage of the spaceship to take a high speed trip through space, which she thinks lasts for five years and during which her body ages five years biologically. When she returns to the Earth, she might naively expect that she would see a twin who, like her, has aged by five years. However, she finds instead that two hundred years have elapsed on the Earth, and that her twin (and all her other acquaintances) are long dead.

If you think about this scenario for a couple of minutes, you might come to the conclusion that, at least superficially, it contradicts the idea that only relative motion is fundamental. The twin sitting on the Earth thinks that the spaceborne twin is moving, but couldn't the twin sitting in the spaceship argue equally well that she is at rest and that it is the twin on the Earth that is moving at a high velocity? But, if that is the case, it should be the twin in the spaceship, not the twin on the Earth, that first grows old and dies.

As long as neither twin is accelerated, they would both be equally right – or equally wrong – in asserting that the other twin, moving at a high velocity, is aging less. However, the twin that is in the spaceship must undergo huge accelerations, first to enable her to move away from the Earth at a very high velocity, then to turn the spaceship around, and finally to slow down again once she has returned to the Earth. This breaks the symmetry and implies that it is the twin remaining on the Earth that has aged more.

The important point here is that, because time is an observer-dependent notion in special relativity, one cannot uniquely answer the question of which twin is older without bringing them back in contact with each other. Each of them might think the other older but, short of bring them together to effect a direct comparison, neither can really say.

Another prediction of special relativity is so-called *length contraction*, the fact that *moving objects are shorter*. Suppose, in the context of the twin paradox, that, when both spaceships were on the Earth, their lengths were measured and found to be exactly equal. Then suppose that, while the spaceborne twin is flying her spaceship off into space, the twin on the Earth measures its length, e.g., via radar, which entails reflecting radio waves off the ship. What she will find is that, regardless of how she does the measurement, the spaceborne ship is shorter than the spaceship that has remained on the Earth. Analogously, the twin on the spaceborne ship could measure the spaceship remaining on the Earth, and she would conclude that it was the ship on the Earth, and not hers, that is shorter. Again both – and neither – are right. In order to effect a direct comparison between the spaceships, either the spaceborne ship must return to the Earth or the ship on the Earth must go into space and catch up with the spacebound one. But, if the two ships are brought together, and their lengths compared, both twins will agree (1) that both ships are equal in length and (2) that they have the same length that they had when measured originally.

At this juncture, you might object that, given this explanation, one should say that moving objects *appear* shorter rather than that moving objects *are* shorter. Maybe. The key realisation, however, is that space and time are the fundamental building blocks in terms of which we define our notions of reality; and, if every measurement that an observer can conceivably make implies that something has become shorter, one can reasonably say that, from the perspective of the individual performing the measurement, the object *is* shorter.

Both time dilation and length contraction illustrate a more general point about special relativity: Because different observers see space and time differently, they will in general disagree on such questions as how much time has elapsed or how big something is. To ask which of the observers is *right* fundamentally is not meaningful in any practical way. The only way to decide say which of two objects is older or bigger is to bring them together at a single point in space and time and effect an explicit comparison. A more precise way of saying this is that all observers will agree on the properties of an individual event, i.e., what happens at a given point in space at a given time, but that they will not agree on the properties of different events and how they are connected.

Because special relativity warps the notions of space and time, it implies that velocities add in a way that is different from the simple linear rule of Galilean relativity. To explain the new rule, it is useful to return again to the baseball analogy. Suppose that the pitcher is running towards the batter with a velocity u, and that he releases the pitch from his hand at a velocity v. Galilean relativity would then predict that the batter will see the pitch coming towards him at a velocity $w = u + v$. However, special relativity implies instead that the batter will see the pitch coming towards him with a slightly smaller velocity,

$$w = \frac{u+v}{1+uv/c^2},$$

where c is the speed of light. This law holds quite generally for the addition of velocities. Presuming that the velocities u and v are both much smaller than c, it follows that $uv/c^2 \ll 1$ (the symbol $\ll$ means "much less than") and the denominator on the right hand side is almost exactly equal to one. It follows that, to a very high accuracy, $w \approx u + v$. In this limit one recovers Galilean relativity. However, if the pitcher were to shine a light signal towards the batter, rather than throwing a pitch, in the preceding equation the velocity of the baseball, v, would be replaced by the velocity of light, c. In this case, the observed velocity

$$w = \frac{u+c}{1+uc/c^2} = \frac{u+c}{1+u/c} = c\left(\frac{u+c}{u+c}\right) = c.$$

In other words, even though the pitcher is running towards the batter when he shines his light, the batter will see the light approach him with velocity c.

Many people think that understanding how special relativity works requires a knowledge of rather sophisticiated mathematics. This, however, is not the case. All that one requires is a knowledge of high school algebra and geometry, in particular the Pythagorean theorem. The purpose of the remainder of this lecture is to show how, using a straightforward generalisation of the Pythagorean theorem, it is possible to derive an explicit formula that says exactly how much time slows down for a moving object. This will also provide an explicit demonstration of what is meant by the statement that special relativity views time as a fourth dimension, coequal with the other three spatial dimensions.
If you get a bit lost in the following details, do not worry; it is the general idea that is important.

Suppose that there is some object near you, located 4 meters to your right and 3 meters in front of you. What is the total distance in space that it is separated from you? As illustrated in Figure 11.2, to answer this question, you need only invoke the Pythagorean theorem. Suppose, for specificity, that you call the right-left direction the x-direction and the direction in front or back of you the y-direction. Now draw a right triangle, with one side of length $\Delta x = 4$ units towards the right and a second side extending ahead from that point for a distance $\Delta y = 3$ units. The Pythagorean theorem implies that the total distance Δr to the desired point satisfies

$$\Delta r^2 = \Delta x^2 + \Delta y^2.$$

Given, however, that $\Delta x = 4$ and $\Delta y = 3$, it follows that $\Delta r^2 = 3^2 + 4^2 = 9 + 16 = 25$, so that $\Delta r = 5$ units and the object is 5 meters away.
If you add a third direction, also allowing the object to be displaced upwards or downwards, the game is the same. If the object is located a distance Δx to the right or left, a distance Δy to the front or back, and a distance Δz up or down, the total distance to it satisfies

$$\Delta r^2 = \Delta x^2 + \Delta y^2 + \Delta z^2.$$

Thus, e.g., if the object is 4 meters to the right, 3 meters in front, and 12 meters overhead, one has $\Delta r^2 = 3^2 + 4^2 + 12^2 = 9 + 16 + 144 = 169$, so that $\Delta r = 13$.

According to the ordinary notions of space and time, as in Newtonian mechanics, if you determine (correctly) that the distance from you to the object is Δr, then everyone else in the Universe measuring the distance should find the same length Δr. Galilean relativity implies that distances between objects are something about which everyone can agree.

Similarly, if you determine correctly the duration Δt of some event, again everyone making the measurement has to agree. However, in special relativity this is not the case! If you measure the distance in space and the elapsed time between two different events, e.g., you throw a hand grenade high into the air and measure how far it travels and for how long before it explodes, then in general you will get a different answer than someone else will. Nevertheless, there *is* one thing about which you, and every other observer, will agree, namely the so-called *four-dimensional interval* between two events.
If you determine that two events are separated in time by an interval Δt and in the three spatial directions by distances Δx, Δy, and Δz, you can of course compute the quantity

$$\Delta s^2 = c^2 \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2.$$

So can anyone else, given his or her measured intervals in time and space. The key idea of special relativity, which explains precisely how our notions of space and time are mucked up, is that *all nonaccelerated observers will agree on the length of the four-dimensional interval* Δs. Separations in time and space do not have fundamental meaning individually. However, there is a combination of the two, the four-dimensional distance, or interval, Δs about which all observers can agree. In this sense, the fundamental notion of distance is four-, rather than three-dimensional, referring to the separation between events in both space and time.

The fact that this interval is independent of observer implies directly that moving clocks slow down. Suppose that an observer looking at a properly working clock, which is at rest relatively to him, sees that successive ticks are separated by a time interval ΔT, the nominal period between ticks intended by the manufacturer. Since for him the clock is at rest, it does not move in space between ticks, so that Δx, Δy, and Δz are all zero. This observer therefore infers a four-dimensional interval between successive ticks satisfying

$$\Delta s^2 = c^2 \Delta T^2.$$

Now consider another observer who sees the clock moving in the x-direction with some velocity v. For him Δy and Δz are both zero, but Δx is not. Rather, during the time interval Δt the clock will have moved a distance $\Delta x = v\Delta t$. This second observer therefore computes a four-dimensional interval between successive ticks of length

$$\Delta s^2 = c^2 \Delta t^2 - \Delta x^2 = c^2 \Delta t^2 \left[1 - \frac{1}{c^2}\left(\frac{\Delta x}{\Delta t}\right)^2\right] = c^2 \Delta t^2 \left[1 - \frac{v^2}{c^2}\right],$$

where Δt is the time between ticks, as measured by the moving observer. However, special relativity implies that both observers must agree on the length of the interval Δs, which means that

$$c^2 \Delta T^2 = c^2 \Delta t^2 \left[1 - \frac{v^2}{c^2}\right]$$

or

$$\Delta t = \frac{\Delta T}{\sqrt{1 - \frac{v^2}{c^2}}}$$

The denominator on the right hand side of this expression is less than one, so that Δt is bigger than ΔT. It follows that the time interval between ticks, as measured by the observer moving relative to the clock, is larger than the time between ticks, as measured by the observer who sees the clock at rest. In other words, as far as the moving observer is concerned, time as measured by the clock has slowed down.

One final point: The relation $\Delta s^2 = c^2 \Delta t^2 (1 - v^2/c^2)$ implies that, for velocities less than the speed of light, $\Delta s^2 > 0$, but that Δs actually gets smaller as v increases. In other words this four-dimensional interval has the curious property that a large velocity, which generates a large three-dimensional distance, yields a small four-dimensional distance. The limiting case $v = c$, satisfied by light and other massless particles, corresponds to $\Delta s = 0$. Light travels at high speeds through three-dimensional space but has a four-dimensional distance equaling zero.

.

STUDY GUIDE FOR LECTURES 1 - 5

Words and names that you should recognise and understand:

(Most of the objects mentioned in the first two lectures were only described in a very cursory way, so that you are only responsible for a very general notion as to what they are. The majority will be discussed more carefully later on in the course.)

meter	kilometer	gram
kilogram	astronomical unit (AU)	lightyear
parsec	exponential notation	Fahrenheit
Celsius	centigrade	Kelvin
degree of angle	minute of angle	second of angle
Mercury	Venus	Earth
Mars	Jupiter	Saturn
Uranus	Neptune	Pluto
planets	terrestrial planets	Jovian planets
moons	The Moon	planetary rings
asteroids	comets	meteroids
Mariner	Venera	Viking
Voyager I	Voyager II	gravity
stars	nuclear fusion	Milky Way
spiral galaxies	disc component	spheroidal component
dark matter	missing mass problem	open clusters
globular clusters	spiral arms	elliptical galaxies
irregular galaxies	galaxy clusters	Local Group
Virgo Cluster	Coma Cluster	Hubble Law $v = Hr$
Hubble constant H	quasars	radioactive dating
Hot Big Bang Model	sideral year	Thales of Miletos
Anaximander	Anaximenes	Pythagoras
Philolaus	diurnal motions	annual motions
Herakleides of Pontus	orbital period	retrograde motion
geocentric Universe	heliocentric Universe	morning/evening stars
Aristarchus of Samos	empiricism	Plato
Aristotle	Aristotle's elements	Aristotle's natural motion
Aristotle's violent motion	Eudoxas	Claudius Ptolomy
The Almagest	deferent	epicycle
The Alfonsine Tables	Eratosthenes	celestial sphere
celestial North Pole	celestial South Pole	right ascension
declination	Jerome	Augustine
Cosmas	Pope Sylvester II	Copernicus
De Revolutionibus	parallax	rotation
revolution	*Prutenic Tables*	

Recommended Questions:

1. What is the largest planet in the Solar System? the smallest?
2. What is the planet closest to the Sun? most distant from the Sun?
3. On which planets have Earth-based probes been landed? Which planets have been explored by passing spacecraft? About which planet is the least known?
4. What are the principal differences between terrestrial and Jovian planets?
5. What are comets? asteroids? meteroids?
6. How long does it take light to travel from the Sun to the Earth? from the Sun to the next nearest star?
7. About how large is the Milky Way galaxy? About how many stars does it contain? What is the distance to the next large galaxy?
8. How old do astronomers think the Sun and the Solar System to be? On what basis have they reached their conclusion?
9. How old do astronomers think the Universe as a whole to be? On what basis have they reached their conclusion?
10. Explain why the Earth has seasons. (A picture may help.)
11. How would the seasons change if the Earth's rotation axis were not tilted about 23.5^o relative to the Earth-Sun orbital plane? if it were tilted 90^o? (An explicit answer to this question was not presented in the text, but hopefully you can think through what the answer must be!)
12. What were the basic ideas of the Pythagorean world view, at least as it related to astronomy and cosmology?
13. What were the basic ideas of the Platonic and Aristotelean world views, at least as they related to astronomy and cosmology?
14. Explain what is meant by retrograde motion. What is the physical origin of this phenomenon? (A picture may help.)
15. Explain the way in which Ptolomy's geocentric cosmology accounted for the phenomenon of retrograde motion.
16. Explain the physical ideas underlying Eratosthenes' calculation of the circumference of the Earth.
17. What were the major ideas underlying Copernicus's cosmology? In what ways was his cosmology right? In what ways was it wrong?
18. Why was Copernicus forced to the logical conclusion that the stars must be much further away from the Earth than the Sun, the Moon, and the other planets are?

STUDY GUIDE FOR LECTURES 6 - 10

Words and names that you should recognise and understand:

Tycho Brahe	Galileo Galilei	William of Ockham
Johannes Kepler	Isaac Newton	supernova
Uraniborg	Emperor Rudolph II	*Rudolphine Tables*
Cardinal Barberini	Salviati	Sagredo
Simplicius	atomism	Aristotle's natural motion
Galileo's natural motion	Law of Inertia	force
force field	Equivalence Principle	Galilean relativity
perfect solids	ellipse	focus (foci) of an ellipse
semi-major axis	semi-minor axis	aphelion
perehelion	Kepler's First Law	Kepler's Second Law
Kepler's Third Law	Christiaan Huygens	Robert Hooke
Principia	William Halley	kinematics
dynamics	event	velocity vs. speed
acceleration	mass	weight
Newton's First Law	Newton's Second Law	Newton's Third Law
Newton's universality	gravity	Law of Universal Gravitation
escape velocity	angular momentum	center of mass
scientific method	Rules of Reasoning	variable stars
refracting telescope	reflecting telescope	William Herschel
Caroline Herschel	parallax	binaries
proper motions	nebulae	Frederich Wilhelm Bessel
photometer	J. H. Lambert	Immanuel Kant
Johannes Kapteyn	grindstone model	William Parsons

Recommended Questions:

1. What was Tycho Brahe's principal contribution to astronomy and cosmology?
2. As regards astronomy, what was the principal contribution provided by William of Ockham and the Ockhamists?
3. What were Galileo's principal contributions to astronomy and physics?
4. What were Kepler's principal contributions to astronomy and cosmology?
5. What were Newton's principal contributions to astronomy, mathematics, and physics?
6. Why did Tycho Brahe not accept the idea of a heliocentric cosmology?
7. What was the alternative cosmology advocated by Tycho Brahe?
8. What were the four early discoveries made by Galileo which demonstrated conclusively (at least to him) that Ptolomy's geocentric cosmology could not be correct?
9. How did Galileo manage to run afoul of the Catholic Church?
10. Contrast Aristole's and Galileo's notions of natural and violent motions.
11. Contrast Aristotle's notion of a force with the modern notion of a force acting via a field.
12. What is the Equivalence Principle, and how is it manifest in our everyday world?
13. What precisely is meant by Galileo's Law of Inertia?

14. What precisely is meant by Galilean relativity? Give an example of its implications in our everyday world.
15. Why did Kepler accept a position with Brahe in Prague?
16. Draw a picture of an ellipse, indicating the foci, the semi-major axis, and the semi-minor axis. In what sense is a circle a particularly simple kind of ellipse?
17. State in words the physical content of Kepler's First Law.
18. State in words the physical content of Kepler's Second Law.
19. State in words the physical content of Kepler's Third Law.
20. What is meant by kinematics? dynamics?
21. Distinguish between mass and weight.
22. What is the difference between velocity and speed?
23. Define precisely what is meant by an acceleration. Can an object have a constant speed and still be accelerating?
24. State in words the physical content of Newton's First Law.
25. State in words the physical content of Newton's Second Law.
26. State in words the physical content of Newton's Third Law.
27. How large is the force that the Sun exerts on the Earth compared with the force that the Earth exerts on the Sun? Why is it that the Sun is accelerated by the Earth much less than the Earth is accelerated by the Sun?
28. Describe the physical content of the Law of Universal Gravitation, $F = Gm_1m_2/R^2$.
29. Define what is meant by an escape velocity.
30. In what sense do Newton's Laws explain Kepler's Laws? In what sense do they transcend Kepler's Laws?
31. Explain what is meant by conservation of angular momentum.
32. Define what is meant by a center of mass. In what sense is it true that the planets in the Solar System orbit the Sun?
33. How is it that the same Law of Gravity can explain both how objects fall towards the Earth and how the Moon orbits the Earth?
34. Explain the physical content of Newton's four "Rules of Reasoning in Philosophy."
35. What is meant by parallax, and how is it measured by astronomers?
36. How is it possible, given the distance to some star and its apparent brightness, to figure out its true, or intrinsic, brightness?
37. How, roughly, is it possible to figure out the size of a star if one knows its color and its intrinsic brightness? (This question will be answered more carefully in a later lecture.)
38. How were astronomers in the nineteenth century led to the conclusion that stars must be objects comparable to the Sun?
39. Describe the so-called grindstone model of the Milky Way galaxy, and explain the logic that the Herschels used in arriving upon it.
40. Summarize what scientists knew about the large scale structure of the Universe, i.e., outside the Solar System, in about 1900.

STUDY GUIDE FOR LECTURES 11 - 13

Words and names that you should recognise and understand:

electric charge	electric current	J. C. Maxwell
electromagnetism	electromagnetic radiation	Heinrich Hertz
wave	wavelength λ	amplitude
frequency ν	speed of light c	$c = \lambda\nu$
nanometer	radio waves	microwaves
infrared radiation	visible radiation	ultraviolet radiation
x-rays	gamma rays	ether
vacuum	A. A. Michelson	E. W. Morley
Hendrick Lorentz	Albert Einstein	Galilean relativity
special relativity	time dilation	Twin Paradox
length contraction	addition of velocities	four-dimensional spacetime
Pythagorean Theorem	four-dimensional interval	

Recommended Questions:

1. Given several examples of wavelike phenomena arising in nature.
2. Define what is meant by the amplitude, frequency, and wavelength of a wave.
3. How are the frequency ν and the wavelength λ of a wave related to the speed c of the wave?
4. How is electromagnetic radiation connected with the fact that the Earth is warm?
5. Why is it necessary for astronomers to place telescopes above the Earth's atmosphere to detect gamma rays, x-rays, ultraviolet radiation, and infrared radiation from outer space?
6. What type of visible light has the longest wavelength? the lowest frequency? Why are the answers to these questions related?
7. What type of visible light has the shortest wavelength? the highest frequency?
8. Denumerate the different types of electromagnetic radiation in order of decreasing wavelength, in order of increasing frequency.
9. Assuming that the speed of light $c \approx 300,000$ km/sec and that Mars is about 225 million km from the Sun, about how long does it take light to travel from the Sun to Mars?
10. What did the nineteenth century scientist mean by *ether*? Does such an ether really exist?
11. Describe the Michelson-Morley experiment. What was its original aim?
12. What was the surprising outcome of the Michelson-Morley experiment?
13. Precisely what is meant by the statement that Maxwell's theory of electromagnetism contradicts Galilean relativity?
14. State the two assumptions from which Galilean relativity follows.
15. State the two assumptions from which Einstein's special relativity follows.
16. Why operationally is it impossible to take a massive object with a velocity less than c and accelerate it to a velocity of c or above?
17. What does the scientist mean by the term *time dilation*?

18. Discuss the so-called Twin Paradox and explain why it is that there really is no paradox.
19. What does the scientist mean by the term *length contraction*?
20. Discuss the difference between how velocities *add* in Galilean relativity and in special relativity.
21. In what sense is it true that Galilean relativity is *almost right*?
22. Define what is meant by the four-dimensional spacetime interval between two points.
23. Suppose that an object is 5 feet in front of you and 12 feet above you. What is the total spatial distance between you and the object? Will every observer agree with your answer? Is there a notion of distance between two objects about which all nonaccelerated observers *can* agree?
24. About how large is the speed of light c? How do you think your perception of the world around you would change if the speed of light were only one hundredth as large?

Lectures 14, 15, and 16

QUANTUM MECHANICS AND THE MODERN THEORY OF MATTER

So is light a particle or is it a wave? The unified theory of electromagnetism developed during the nineteenth century identifies light as a wavelike phenomenon, namely electromagnetic radiation, and this identification has proven extremely useful in understanding how light propagates through space. However, it does not constitute a completely satisfactory theory, in that it cannot explain how light is created, emitted, and/or absorbed. When describing the emission and absorption of light, one must instead invoke the modern atomic theory of matter, which treats light as a particulate entity, with the basic particle of light being the *photon*.
From a modern perspective, one can say that, strictly speaking, light – like ordinary matter – is neither a wave nor a particle. Rather, it must be described in the framework of a more sophisticated theory, so-called *quantum mechanics*, which says that light behaves like a wave in certain respects and like a particle in others.

Before turning to the issue of *why* a wavelike theory of light is not adequate, it is worth stressing one of the big successes that *does* arise from the wavelike interpretation of light, namely an explanation of the *Döppler effect.* This phenomenon, which is a general property of all waves, is something that you have no doubt encountered in real life, although you may not have thought very much about it. A good example of the Döppler effect is provided by the sound emitted by an ambulance siren as the ambulance drives past you. As the ambulance approaches, you hear a relatively high pitched sound, corresponding to a higher frequency. However, when the ambulance has passed you and is moving away, the pitch which you hear is lower, this corresponding to a lower frequency. In other words, the frequency of the sound waves that you hear depends on the motion of the emitter towards or away from you.
The origin of this effect is illustrated in Figure 14.1. Suppose in the first instance that you are standing at a fixed, unchanging distance from the emitter of some light signal, and that you look at the light and determine its wavelength and frequency. Then suppose that you start moving towards the emitter, or that the emitter starts moving towards you – special relativity says that these two possibilities are really equivalent. In this case, the distance between successive crests of the waves that reach you will be shorter than what was the case originally, so that the frequency with which they pass you will be higher, i.e., you will detect a larger number of crests per second. The net result is that you measure a higher frequency and a shorter wavelength. This implies, for example, that if the emitter is shining yellow light towards you, you will see that light shifted towards the shorter wavelength, blue end of the spectrum. One thus says that the light has been *blue-shifted.*
Similarly, if you and the observer are moving apart, you will measure a lower frequency amd a longer wavelength. In this case, the yellow light will be shifted towards the longer wavelength, red end of the spectrum, so that the light is said to be *red-shifted.*
The relative motion of you and the emitter triggers an overall dilation or compression in observed wavelengths, whereby different colours of light are all shifted by the same fractional amount. In the context of Newtonian mechanics, the formula for the change in wavelength is especially simple. If the light is emitted with wavelength λ, the observed

wavelength will be shifted by an amount $\Delta\lambda$ satisfying

$$\frac{\Delta\lambda}{\lambda} = \frac{v_r}{c},$$

where v_r is the velocity with which the light is moving towards you or away. Positive velocity v_r corresponds to motion away from you, with $\Delta\lambda$ positive. For the case of high velocities, Newtonian mechanics is no longer adequate and one must incorporate the effects of special relativity, which modify this equation.
The Döppler effect is of fundamental importance in astronomy and cosmology because it provides a direct physical way in which to determine whether an object is moving towards or away from the Earth and the speed of this relative motion. Thus, for example, it is by measuring such Döppler shifts that cosmologists know that most galaxies are moving away from us.

At the same time that some nineteenth century scientists exploring electromagnetism were discovering that the wavelike interpretation of light can explain many physical phenomena, other scientists were discovering a different class of phenomena which such a wavelike interpretation simply cannot explain. The key idea, which emerged from a different series of experiments, is that of a *black body*.
In the middle of the nineteenth century scientists discovered that every object in the Universe emits electromagnetic radiation, to a lesser or greater degree. In general, the details of what is emitted – how much light and what colours – turns out to be a complicated mess. However, it was discovered that, for the special case of opaque objects, i.e., objects that are not transparent to light, everything about the emitted radiation turns out to be determined completely by the temperature of the emitter. Good examples of near-perfect black bodies are stars, planets, and people, each of which emit electromagnetic radiation.

The first obvious question to ask is: *how much light does a black body emit*? The answer to this question is provided by the so-called Boltzmann Law, formulated by the Austrian physicist *Ludwig Boltzmann* (1844 - 1906), who made a number of different fundamental contributions towards the understanding of black body radiation and the modern theory of matter and heat. What this law says is that U, the total amount of electromagnetic radiation emitted from each spot on the surface of a black body, as measured by the amount of energy given off per square centimeter per second, is proportional to the fourth power of the surface temperature, as measured in degrees above absolute zero. This is written conveniently as

$$U \propto T^4,$$

where T is the temperature and the symbol $\propto$ means "proportional to."
This law implies that, as temperature increases, the amount of energy that is emitted increases very quickly. Thus, e.g., a star with a surface temperature twice as high the Sun has the property that each little spot on its surface emits $2^4 = 2 \times 2 \times 2 \times 2 = 16$ times as much radiation. Similarly, as the temperature decreases the amount of energy that is emitted decreases very quickly. For an object that is only one third as hot as the Sun, each spot on the surface will emit only $\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{81}$ as much radiation. As the temperature approaches absolute zero, i.e., zero degrees Kelvin, the amount continues to

decrease. One definition of *absolute zero* is that it corresponds to a temperature at which no blackbody radiation is emitted.

The next obvious question is: *what colour light does a black body emit*? The answer to this is slightly more complicated, and involves the concept of a *black body spectrum.* Basically, the light will be emitted with a distribution of wavelengths, all different colours being represented, but the relative amounts of energy coming out at different wavelengths depends on the temperature. One useful characterisation of the overall form of this distribution is provided by Wien's Law, which was formulated by *Wilhelm Wien* (1864 - 1928), a German physicist. What this law says is that the wavelength λ_{max} where the most radiation is emitted varies inversely with the temperature T above absolute zero. Symbolically, this can be written as

$$\lambda_{max} \propto \frac{1}{T}.$$

Objects at lower temperatures emit most of their radiation at longer wavelengths and lower frequencies. Thus, e.g., an object which is only half as hot as the Sun will emit the most radiation at a wavelength twice as long as the wavelength at which the Sun emits the most radiation.

Figure 14.2 exhibits a graph which plots the amount of energy emitted per square centimeter per second at different wavelengths for objects at three different temperatures. Wavelengths are expressed in microns or micrometers, μm, one of which equals 10^{-6} meters. It should be apparent that the three curves are rather similar, each being a compressed or dilated version of the others. The graph corresponding to the highest temperature peaks further towards the left, this representing the fact that, for this object, most of the light comes out at shorter wavelengths. The graph for the coolest object peaks further towards the right, this representing the fact that, for this object, the light comes out at longer wavelengths overall. The area under each curve represents the total amount of light emitted at all different wavelengths. That the curve for the highest temperature has the largest area beneath it represents the fact that the hottest object is, by far, emitting the most energy.

The bottom line is that every opaque object emits radiation with an intensity and colour that are determined completely by its temperature. The reason that the Sun looks yellow is that, having a surface temperature of about 5800 K, most of the electromagnetic radiation is emitted at wavelengths corresponding to yellow light. If the surface of the Sun were cooler most of the light would be emitted at longer wavelengths, and the Sun would look redder in appearance. So-called *red giant stars* look redder than the Sun precisely because they have a lower surface temperature.

Various solid objects on the Earth, including people, also emit black body radiation. However, since people are substantially cooler than the Sun, with a body temperature of about 310 K, most of the light which they emit comes out at much longer wavelengths, corresponding to infrared radiation. (The fact that people emit infrared radiation has long been used by the military to track people in the dark.) The Earth and the terrestrial planets all emit primarily infrared radiation. However, the most distant planets are so cold that a substantial portion of the radiation which they emit comes out at longer microwave wavelengths.

The astronomical significance of Wien's Law should be obvious: by looking at the colour of a star, or of any other opaque object in the sky, one can infer directly how hot the surface is. This is an extremely important first step in understanding what the object is like and how it works.
One final point is also worth making: In general, one is not really interested observationally in the amount of energy emitted by each little spot on the surface of an object. Rather, what is important is the *luminosity*, the total amount of energy that is being emitted into space. This is obtained by adding up the contributions from each little spot on the surface, i.e., multiplying the amount of energy from each little spot by the total surface area. Thus, in particular, for a spherical black body of constant surface temperature, the luminosity L depends on two, and only two, quantities, the surface temperature T and the radius R. This means that, if one knows both the luminosity L and the temperature T one can calculate the radius R. This was the technique first used by nineteenth century astronomers to demonstrate that distant stars are comparable in size to the Sun.

The modern theory of matter and light, quantum mechanics, the subject of the remainder of this lecture, was developed largely in response to the fact that the wavelike interpretation of light cannot account for black body radiation. In order to explain the basic ideas involved in this theory, it is useful to start with a simple digression, reviewing some very simple facts from chemistry and atomic theory.
Everything in the Universe can be broken down into various *chemical elements.* As of 1994, 105 such elements were known, 90 of which occur naturally on the Earth and the others of which have been created artificially. On the surface of the Earth, like the other terrestrial planets, the most abundant elements are hydrogen, nitrogen, carbon, oxygen, and silicon. In the core of the Earth, iron and nickel predominate. The Earth's atmosphere is about 78% nitrogen and 20% oxygen, with the remaining 2% in smaller traces of other elements. By contrast, the Sun, like the outer Jovian planets, is comprised primarily of hydrogen and helium.
In ancient times, many natural philosophers believed that chunks of various elements could be divided into successively smaller pieces *ad infinitum*, without the process every ending. Others, including *Democritus* (c. 460 - 362 BC) and *Epicurus* (c. 342 - 270 BC), believed that there is a smallest piece of an element that can exist. Today we know that these so-called *atomists* were correct. The smallest unit of an element is an *atom.* Atoms, the smallest pieces of chemical elements, combine together to make *molecules*, the smallest pieces of *chemical compounds.*
The first hard experimental evidence that matter is constituted of something like chemical elements, and not simply combinations of the four ancient elements – earth, air, fire, and water – was apparently provided by the English natural philosopher *Robert Boyle* (1627 - 1691). The concept of an element is usually attributed to the French chemist *Antoine Lavoisier* (1743 - 1794). The notion of a chemical compound, with fixed relative abundances of different elements, is usually attributed to the British chemist/physicist *John Dalton* (1766 - 1844). Chemical notation, with abbreviated symbols for the different elements that are combined into molecules (e.g., H_2O as the notation for the compound water, molecules of which contain two hydrogen and one oxygen atom), was pioneered by the Swedish *Brogelius* (1779 - 1848). (Since ancient times, alchemists had used symbols for

various chemicals, but in their ignorance they had implemented no distinctions between elements and compounds.)

Atoms are extremely tiny, typically about 10^{-8} cm across, and any macroscopic sample of material contains a very large number of atoms. For example, a single water molecule is comprised of three atoms, two hydrogens and an oxygen, and there about 10^{22} water molecules in a cubic centimeter (cc) of water, a volume comparable in the size to the smallest joint of your little finger. Another example of a molecule is carbon dioxide, which contains one carbon and two oxygens.

An atom is comprised of a *nucleus*, surrounded by one or more *electrons.* Even compared with a tiny atom, a nucleus is extremely small, with a typical size of about $10^{-13} - 10^{-12}$ cm. Crudely, an atom can be viewed as being like a tiny Solar System, with the electrons orbiting the nucleus. The Solar System is held together by gravity. An atom is held together by electric forces. (Many scientists in the seventeenth and eighteenth centuries had the idea that atoms are regulated completely by Newton's gravity, like all macroscopic phenomena understood at the time. However, Newton himself realised that the workings of atoms could not be explained without also allowing for the effects of some other force.)

There are two kinds of electric charge, commonly termed positive and negative. They have the property that like charges repel and unlike charges attract. Just as for Newtonian gravity, the strength of the attractive or repulsive force varies as $1/R^2$, where R is the distance between the charges. Electrons have negative charges, whereas the nucleus has a positive charge.

The simplest atom, ordinary hydrogen, consists of one electron, with charge $-e$, which is orbiting around a nucleus containing one *proton*, with charge $+e$. The charges of the proton and the electron are *exactly* equal in magnitude, so that the total charge of the atom, arising from both the electron and the proton, is *exactly* zero. So-called heavy hydrogen, or *deuterium*, is comprised of one electron orbiting a more complicated nucleus, containing one positively charged proton and another particle, which is electrically neutral. Not so creatively, this particle is called the *neutron.* A third type of hydrogen, *tritium*, is comprised of one electron orbiting a nucleus consisting of one proton and two neutrons.

All atoms contain a nucleus of protons and neutrons, surrounded by one or more electrons. Thus, e.g., ordinary helium contains two electrons, two protons, and two neutrons. Ordinary carbon contains six electrons, six protons, and six neutrons. Another type of carbon, so-called carbon-14, contains six electrons, six protons, and eight neutrons. For all atoms, the number of electrons equals the number of protons, so that the atom as a whole is exactly neutral electrically.

The interactions of atoms with each other, the subject of *chemistry*, depends primarily on the number of electrons which they have. For this reason, one identifies two different atoms with the same number of electrons and protons, but different numbers of neutrons, as being different *isotopes* of the same chemical element.

Electrons, protons, and neutrons are all extremely small. An electron has a mass of about 9.11×10^{-28} gm, a number often denotes m_e. The mass of a proton and a neutron are, respectively, about $1836\, m_e$ and $1839\, m_e$. As in the Solar System, most of the mass in an atom is located at the center.

Classical physics, according to either Newton or Einstein, says that, in the Solar System,

the planets, the moons, and any other object can assume essentially any orbit and have an arbitrary energy. This is *not* true for the orbits of electrons in an atom. Rather, quantum mechanics implies that only certain specific orbits are allowed, each being characterised by a specific value of the energy.
The allowed values of these orbits depend on the composition of the atom and, as such, are different for each kind of atom. If an electron is initially in some relatively high energy orbit and subsequently drops to a lower energy orbit, it can only do so by getting rid of the excess energy. This the electron does by *emitting a photon of light*.
The colour of the light to which this photon corresponds is determined completely by the energy in a fashion characterised by Planck's Law, named for *Max Planck* (1858 - 1947), the German physicist who first discovered it. What this Law says is that the energy E of the photon is proportional to its frequency ν, satisfying a law

$$E = h\nu,$$

where h is a constant, the so-called Planck constant. High frequency, short wavelength photons have high energy; low frequency, long wavelength photons have low energy.
The net result is that, if an electron moves from a high energy orbit to a lower energy orbit, it will emit a photon of light, with a well-defined energy and frequency which is unique to the specific atom in question. Determining the energy or colour of such a photon thus serves to provide a fingerprint, which can be used to determine what kind of atom the electron was in.

This atomic theory also provides a coherent explanation of the phenomenon that we perceive as temperature. The basic point here is that atoms and molecules do not simply sit there at rest. Rather, they execute complicated, irregular trajectories which, viewed macroscopically, appear to be essentially random. It is this random motion of objects that we perceive as heat, and the amount of motion determines the temperature which we would associate with the macroscopic object.
The typical speed with which molecules move depends on two things, namely (1) the temperature T and (2) their mass m. If the molecules have a high temperature they will move quickly, but if they are cold they will instead tend to move more slowly. Another way in which to characterise absolute zero is to say that it is the temperature at which there is no random motion. At a given temperature, the typical speed with which molecules move depends on their mass: light molecules tend to move quickly, whereas more massive molecules tend to move slowly. Both the light and heavy molecules will, on the average, have the same energy, but a heavy particle of given energy moves slower than a light particle with the same energy.

The notion of temperature as random motion, when combined with the concept of an escape velocity, provides a simple explanation of why it is that some planets and moons can keep an atmosphere whereas others cannot. Basically, it all devolves into a competition between two effects: High temperatures correspond to high speeds but, if a particle is moving too fast, it will have a speed in excess of the escape velocity, and leak out into space. It thus follows that, in order to keep an atmosphere, a planet must have enough gravity that the escape velocity is much bigger than the typical speed with which the molecules move

about. Any air on the surface of a hot planet close to the Sun will move very quickly, so that the planet needs lots of gravity to prevent its escape. Air on an object further out will be cooler, so that a planet or moon needs much less gravity to prevent the escape.
The other remaining point is that a planet or moon may have enough gravity to trap heavier, more slowly moving, molecules but not enough to prevent the lighter ones from escaping. Thus, e.g., the Earth has enough gravity to keep oxygen and nitrogen but not enough to keep hydrogen and helium (although it can retain water molecules comprised of both hydrogen and oxygen).

But what does all of this have to do with black body spectra? Here the key is an understanding of the role of *collisions* between atoms and molecules. As molecules move about in a near-random fashion, they will occasionally collide with one another; and, when this occurs, they will exchange energy. A good analogy to help visualise this process is a collection of balls rolling on a billiard table (neglecting friction, so that the balls never come to rest). As the balls roll about in a seemingly near-random way, they will occasionally collide with one another, and the net result thereof is that the directions and speeds of the different balls will change. Slow balls tend to be speeded up and faster balls tend to be slowed down.
Just as balls on a billiard table change velocities as a result of collisions, so also collisions between molecules will occasion changes in velocity. However, there is one additional complication, namely that some of the energy invested originally in the motions of the molecules can be used to *excite* an electron into a higher energy *excited state*. In fact, if the temperature is sufficiently high, and the molecules move with sufficiently high velocities, there may be enough energy to strip an electron completely away from an atom, so that it is no longer bound. In this case, one says that the atom has become *ionised*.
In any event, one possible result of a collision between molecules is that an electron can become excited. However, the electron does not "want" to be excited. Rather, its natural propensity is to have the lowest possible energy. The net result is that, either directly or in stages, the electron will drop back down towards the lowest energy *ground state.* To do this, the electron must rid itself of the excess energy, which it does by emitting one or more photons of light.
As noted already, these photons are emitted with very special energies and frequencies. However, once emitted these photons can in turn collide with the molecules in the material and thereby change their energies. After a few collisions, the net result is that a photon, emitted with an initial very special energy and frequency, will have "forgotten" its initial conditions and have a near-random energy. Again the analogy with balls on a billiard table is appropriate: If all the balls on the table were white, and one shot in the cue ball on a clean initial break, it would be very difficult immediately after the break to decide which was the special cue ball and which were the remaining balls. Viewed from on large, the motion of the cue ball would not look different from the typical motions of the other balls.
A *black body spectrum* characterises the energies of a collection of photons which have succeeded in randomising their energies in precisely this sense. It requires that the photons have been involved in a substantial number of collisions, so that it can only be generated in an opaque material through which photons cannot pass without collisions. The form of the spectrum depends on the temperature of the medium, since this temperature determines

how fast the molecules move overall and hence how hard they will collide with each other. If there are not enough collisions to randomise the energies of the photons, one observes instead emission at specific frequencies, the values of which are determined by the material that is doing the emitting.

The aforementioned heuristic description provides the basis for a precise mathematical explanation of three phenomenological laws first discovered experimentally by the German scientist *Gustav Robert Kirchoff* (1824 - 1887):

1. A sufficiently dense and opaque object – solid, liquid, or gas – will radiate light at all different wavelengths and hence produce a *continuous spectrum.*
2. By contrast, a sufficiently low density, nearly transparent, object will emit light with well-defined special wavelengths, producing what is called an *emission spectrum.*
3. If light characterised initially by a continuous spectrum passes through a cool, low density gas, it will be transformed into a modified *absorption spectrum* from which certain special frequencies have been removed.

This last phenomenon is caused by the fact that the intervening gas can absorb photons with just the right energies, using them to excite electrons to excited states. These excited electrons will eventually return to the ground state, but the photons which they subsequently emit will in general be aimed in a different direction than the undisturbed light passing through the gas.

To fully appreciate the preceeding discussion, it is important to characterise what is meant by the energy of the light and the molecules. As stated already, the energy of a photon emitted with frequency ν is $E = h\nu$, where h is the Planck constant. The notion of energy for the molecules requires a more careful discussion. According to Newton's laws of motion, any massive particle moving with a velocity v will, neglecting gravitational, electric, and all other interactions, have an energy $E = \frac{1}{2}mv^2$, where m is the particle mass. In particular, Newton would say that an object at rest has no energy.

Special relativity leads to a very different picture, whereby the energy is given instead as

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}},$$

where c denotes the speed of light. It is easy to see that, for $v = 0$, the denominator is equal to one and this reduces to $E = mc^2$, Einstein's celebrated formula. For a small but non-zero speed, $v \ll c$, it can be shown that the energy reduces to $E \approx mc^2 + \frac{1}{2}mv^2$, the second term agreeing with Newton's energy. Alternatively, for high speeds v approaching the speed of light, the denominator approaches zero and the energy becomes extremely large. E would in fact be infinite for $v = c$. This corresponds to the fact, described earlier, that it would require an infinite amount of energy to accelerate a massive particle up to the speed of light.

In the context of quantum mechanics and the atomic theory of matter, the important things about Einstein's energy are (1) that even a particle at rest has energy, what is often called its *rest mass energy*, and (2) that this rest mass energy, like any other energy, can be transformed into different forms. In particular, it is possible to take some of the rest mass energy associated with a particle and transform it into energy associated with particle motions and/or energy in the form of photons and other massless particles. This is, for example, the idea underlying the possibility of both *nuclear fission* and *nuclear fusion*.

Basically, the idea behind nuclear fission is to take a large nucleus like uranium and break it into two or more smaller pieces, emitting energy in the form of photons and/or so-called *neutrinos*, the latter to be described below. When fission occurs, one finds that the total mass of the by-products is smaller than the mass of the initial nucleus, the difference in mass having been converted into energy. This is the phenomenon underlying the atomic bomb of the Second World War and the nuclear reactors currently in use throughout the world.
Fusion, the mechanism responsible for the hydrogen bomb and the generation of energy inside the Sun, works the other way around, by sticking together smaller nuclei into larger pieces. Thus, e.g., most of the energy inside the Sun is manufactured by combining four protons to make a helium nucleus and two electrons, plus some additional energy emitted in the form of photons and neutrinos. In this case, again the total mass of the end products is smaller than that of the initial components, the difference having been transformed into energy.

As of today, physicists have identified four basic interactions which, in principle, are supposed to be able to explain everything we see in the physical Universe around us:
1. *Gravity* was first characterised usefully by Newton in terms of his Law of Universal Gravitation. As noted already, all forms of mass feel the attractive effects of gravity. For this reason, gravity dominates the dynamics of the Universe on the largest scales, even though this interaction is intrinsically much weaker than any other. The Law of Universal Gravitation, when combined with Newton's Laws of Motion, provides an accurate description of most of the observed features of gravity. However, the theory is *not* completely in accord with what we see in nature and, as will be discussed in the next lecture, has been replaced in the twentieth century by a more satisfactory theory, Einstein's *general relativity*.
2. *Electromagnetism*, first put on a completely sound footing by Maxwell, provides a unified description of all electric and magnetic effects involving charged particles. As noted already, it is the interaction responsible for holding an atom together.
The last two interactions were only discovered and understood much more recently.
3. The *strong interaction* is the interaction responsible for holding the nucleus of an atom together. Unlike gravity and electromagnetism, it only has appreciable effects between objects that are extremely close together, so that it is not observed directly in everyday life. However, without this interaction no complicated atoms could exist. The protons in the nucleus of an atom feel a strong electric repulsion since they all have the same charge, and it is only the strong interaction which manages to overcome this repulsion, providing the glue to hold the nucleus together. When *nucleons*, i.e., protons and neutrons, are close together, it gives rise to a very strong attraction, whence stems its name. However, when the nucleons are further apart the attraction becomes much weaker and the interactions between protons are dominated by their electric repulsion.
4. The *weak interaction* is the interaction responsible for radioactive decay, i.e., the fact

that pieces of a nucleus originally stuck together can eventually fall apart. As its name suggests, it is a relatively weak interaction, much weaker than either the strong interaction or electromagnetism, albeit much stronger than gravity.

Over the past fifty years, scientists have succeeded in constructing a description of nature, the so-called *Standard Model*, which incorporates the last three of these fundamental forces – electromagnetism, the weak interaction, and the strong interaction – into a unified theory which appears consistent with experiments involving the structure of matter on the smallest scales. Interestingly, however, it has not yet been possible to incorporate gravity into this description. It appears that general relativity provides a highly accurate description of the large scale effects of gravity, in the Solar System and beyond. However, it is clear that it must be subjected to some additional modifications if it is to be made completely consistent with the microscopic description of nature provided by the Standard Model. In particular, it is clear that general relativity must fail for sufficiently strong gravitational fields, as would arise in black holes or the very early Universe.

The currently accepted theory of matter invokes two general classes of fundamental particles, *quarks* and *leptons*, known collectively as *fermions*. Interactions amongst these fundamental constitutents of nature are mediated by another class of particles, known as *bosons*. Quarks are particles that interact via all three non-gravitational interactions. Leptons are particles that do not feel the strong interaction and only interact via electromagnetism and the weak interaction. Electromagnetism was originally reconciled with the atomic theory of matter in the context of *quantum electrodynamics*, as developed during the 1940's and 1950's by two American physicists, *Richard Feynman* (1918 - 1988) and *Julian Schwinger* (1918 - 1994), and the Japanese *Shin-Ichiro Tomonaga* (1906 - 1979). A less fundamental theory of the weak interaction was developed by the Italian physicist *Enrico Fermi* (1901 - 1954) and his collaborators in the early 1950's, and that theory was subsequently combined with quantum electrodynamics in the *electro-weak theory* of the American *Steven Weinberg* (b. 1933) and *Abdus Salam* (b. 1926) from Pakistan. The general theory of the strong interaction is known as *quantum chromodynamics*. Various aspects of this theory date back to the 1960's, but it is still a comparatively active area of scientific investigation.
The Standard Model invokes twelve fundamental particles, six quarks and six leptons, divided into three *generations*, each comprised of two quarks and two leptons. The lowest generation consists of particles that (except, perhaps, for the neutrinos, which may all have zero mass) are less massive than the higher two generations. (As of 1999, it appears that neutrinos, albeit lighter than any other known massive particles, *do* have nonzero masses, and that the masses of the three different *species* are unequal. However, the evidence is not yet completely conclusive.) The more massive particles of the higher generations can decay into the lower mass particles of the first generation and, for that reason, only the lowest generation is readily observed in the everyday world.
As will be discussed more carefully below, to each of these *particles* there also corresponds an *anti-particle*, so that there also exist *anti-quarks* and *anti-leptons*.

The three generations contain the following particles:

Generation	Quarks		Leptons	
1	up	down	electron	electron neutrino
2	strange	charm	muon	muon neutrino
3	top	bottom	tau	tau neutrino

The cutesy names for the quarks have no fundamental meaning; they are just names with which to identify six different particles, each of which has specific, well-defined properties. (Back when they were being named, there was a movement to replace *top* and *bottom* by *truth* and *beauty* but, perhaps unfortunately, those alternative names never caught on.)
As noted already, it is the first generation that corresponds to ordinary matter as observed in the everyday world. The electron has already been described. Protons and neutrons are assembled in terms of appropriate combinations of *up*, *down*, *anti-up*, and *anti-down* quarks: a proton is made of two ups and an anti-down, and a neutron is made from two downs and an anti-up. Getting the total charges to work out correctly i.e., $+e$ for the proton and zero for the neutron, requires that quarks have fractional charge. The up and down quark have, respectively, charges $+\frac{2}{3}e$ and $+\frac{1}{3}e$. The anti-up and anti-down have opposite charges, $-\frac{2}{3}e$ and $-\frac{1}{3}e$.
Neutrinos are massless, or nearly massless, particles which do not interact via electromagnetism or the strong interaction and, as such, are only weakly coupled to other forms of matter. Because they are so weakly interacting, they are extremely difficult to detect. The Sun emits large numbers of neutrinos, a by-product of nuclear fusion, and these neutrinos flow out through the Solar System and beyond. Upwards of a trillion neutrinos pass through your body each second without having any appreciable effect. However, neutrinos are important astronomically, both because they play an important role in nuclear fusion within stars and because their presence had important consequences for the evolution of the Universe shortly after the Big Bang.
Interactions amongst quarks and leptons are mediated by bosons. The best known example of a boson is the photon, which is tied up with the electromagnetic interaction. The weak interaction is mediated via particles called *W-* and *Z-bosons*. The strong interaction is mediated by *gluons* which are named for the fact that they provide the glue that hold the quarks together into protons, neutrons, and the like.

The currently accepted theories of the elementary particles of nature, known collectively as *quantum physics,* view particles in a different way than do the ordinary laws of macroscopic dynamics. Thus, in particular, they incorporate the idea that one can never determine with absolute precision exactly where a particle is or exactly how fast it moves. Rather, the best that one can do is describe the particle probabilistically, for example by saying that, with overwhelming likelihood, it is located within some small three-dimensional region of space with velocities near some specified value. This general idea is often called the *Uncertainty Principle.*
The Uncertainty Principle admits to a simple physical interpretation, which dates back to the Danish scientist *Niels Bohr* (1885 - 1962). The idea is in fact quite simple: If you want

to determine where a particle is located, you have to perform some kind of experiment or measurement. For example, you might shine a light at the particle and watch what is reflected back towards you. However, the photons that are reflected off the object carry energy and, when they strike the object, they cannot help but impart a substantial impulse which will significantly change the velocity of the particle. The wavelike character of the light makes an accurate determination of position comparatively difficult and, to minimise any uncertainties in the determination, you need to use light with an extremely short wavelength. However, short wavelength light means high frequency, high energy photons, precisely the type of photons which will disturb the particle the most. The more precisely you pin down where the particle is, the more you change its velocity. Analogous reasoning shows that, if you try to perform an experiment to measure how fast the particle is moving, you cannot help but alter its position.

The fact that one cannot say with complete certainty both where a particle is and how fast it is moving, i.e., the fact that one cannot determine simultaneously both its position and its velocity, means that it is really not a particle in the primal sense that Newton envisioned. Rather it is a fuzzed-out entity which, in many respects, behaves more like a wave. It is in this sense that, like light, such objects as protons and neutrons can also be said to be both particle and wave or, if you prefer, neither particle nor wave.

One importance consequence of the Uncertainty Principle is the so-called *Exclusion Principle*, first formulated by the German, and later Swiss, scientist *Wolfgang Pauli* (1900 - 1958). Basically what the Exclusion Principle says is that if two fermions are brought too close to one another, and are therefore localised in a very small region of space, they will exert an effective repulsive force upon each other, which will tend to move them apart with a large velocity. Applied macroscopically to a large collection of particles this gives rise to so-called *degeneracy pressure*, a real, measureable source of pressure which resists attempts to compress the matter to very high densities. Degeneracy pressure plays an important role in explaining the working of *white dwarfs* and *neutron stars*, the corpses of dead stars. Some cosmologists believe that it may also play an important role in explaining the *missing mass problem.*

The basic ideas of quantum mechanics, as sketched out in this lecture, were first given a precise mathematical formulation in 1927 by two independent workers, the German physicist *Werner Heisenberg* (1901 - 1976) and the Austrian *Erwin Schrödinger* (1887 - 1961). However, that earliest version did not incorporate Einstein's 1905 theory of special relativity. The credit for the unification of special relativity and quantum mechanics is due primarily to *Paul Adrien Maurice Dirac* (1902 - 1984), an English electrical engineer. Proceeding purely on theoretical grounds with the aim of incorporating quantum mechanics and special relativity, and without any experimental motivation whatsoever, in 1928 Dirac was led to the prediction that, to each type of particle arising in nature, there corresponds an exact opposite, its anti-particle. Thus, e.g., to the electron there should correspond the anti-electron or *positron*, to the neutrino an anti-neutrino, to the proton an anti-proton, and to the neutron an anti-neutron. Positrons were first discovered in 1932 in *cosmic rays*, high energy charged particles impinging on the Earth from outer space.

Today antimatter is a well-established reality. For example, the Standard Model incorporates the fact that since a proton is made of two up quarks and an anti-down quark, an anti-proton must be made of two anti-up quarks and one down quark. Quarks and anti-quarks, like leptons and anti-leptons, are identical in that they have exactly the same mass, but are opposites in that they have opposite electric charges (and other opposite quantities associated with the strong and/or weak interactions). An up quark has charge $+\frac{2}{3}e$, so that the anti-up must have $-\frac{2}{3}e$. A down quark has charge $+\frac{1}{3}e$, so that the anti-down must have $-\frac{1}{3}e$. A proton has total charge $\frac{2}{3}e + \frac{2}{3}e - \frac{1}{3}e = e$; an anti-proton has total charge $-\frac{2}{3}e - \frac{2}{3}e + \frac{1}{3}e = -e$.

Dirac also introduced another, even weirder, idea, namely that there is a real sense in which so-called empty space, i.e., the *vacuum*, is not really empty. Rather, that entity which we ordinarily perceive as empty space is filled with so-called *virtual pairs* of particles and anti-particles. One way to visualise this is to think of empty space as containing infinitely many particles and infinitely many anti-particles, each of which "wants" to exist. However, under ordinary circumstances they cannot exist: the vacuum has zero energy, but it requires energy to convert these virtual particles into real particles.

One might therefore like to think that these virtual particles are safely inaccessible. However, this is not so since, given an appropriate source of energy, they can in fact be converted into real particles. For example, if so-called empty space is placed in a sufficiently strong variable electric field, energy from the field can be used to turn a virtual pair into a reality. This phenomenon, crazy as it may perhaps sound, has been demonstrated directly in laboratory experiments.

In a similar fashion, one expects that if space is subjected to a sufficiently strong variable gravitational field, some of the energy from that field could be used to create pairs of particles and anti-particles. This latter effect, first investigated systematically by the American *Leonard Parker* and the Russian *Yakov B. Zel'dovich* (1914 - 1989), and subsequently popularised by the English physicist *Stephen Hawking* (b. 1942), is responsible for the fact that a *black hole* should emit radiation into empty space.

When matter and antimatter come in close proximity, they can annihilate one another, converting the energy into photons and other bosons. Similarly, if photons of sufficiently large energy come into contact, they can convert themselves into a particle anti-particle pair. Thus, e.g., two photons, each with energy in excess of $m_e c^2$, with m_e the electron mass, can convert themselves into an electron-positron pair, which has total rest mass $m_e c^2 + m_e c^2 = 2m_e c^2$. Matter-antimatter annihilation is readily detected astronomically, for example in the form of high energy gamma rays emitted from the center of the Milky Way and other galaxies. This is interpreted as direct physical evidence that these regions in the galaxy contain significant amounts of antimatter, a fact which plays a role in the indirect argument that the centers of many galaxies must contain huge black holes of total mass 1 millon $M_\odot$ or more.

This leads to two empirical facts, the explanations of which are simply not known. It is apparent from laboratory experiments that the Universe as a whole, or at least the region which we can access, is almost exactly neutral electrically, i.e., that there are equal numbers of positive and negative charges. Because electricity is a much stronger force than gravity, it would completely dominate the world around us if it were not for the fact that the net charge of the Universe is essentially zero, so that electric effects tend to cancel out.

However, despite this fact it appears that, at least in our part of the Universe, there is currently much more matter than antimatter. If this were not the case, we would detect much larger numbers of high energy gamma rays resulting from the annihilation of particle anti-particle pairs. There can be no large antimatter galaxies at least within the nearest 200 million light years or so.
The idea that, overall, the Universe contains equal amounts of matter and antimatter, and that somewhere far away there are antimatter galaxies, is perhaps appealing aesthetically. However, it is rejected by most cosmologists as unlikely since there is no known plausible mechanism that could explain how matter and antimatter regions could come to be separated by such large distances. At one point earlier in the century it was conjectured that matter and antimatter repel each other gravitationally, and that this could have pushed matter and antimatter regions apart. However, subsequent experiments demonstrated unambiguously that this is false. We know experimentally that, to a very high degree of precision, matter and antimatter feel the same gravitational effects, a fact assumed implicitly in Einstein's *general relativity.* One is thus left with the unresolved puzzle of why it is that, seemingly, the Universe has much more matter than antimatter.

Mathematical Appendix: Döppler Shifts Done Right

The formula for the shift in wavelength, $\Delta\lambda$, presented near the beginning of this lecture,

$$\frac{\Delta\lambda}{\lambda} = \frac{v_r}{c},$$

is only correct for the case of objects moving relative to us with speeds small compared with the speed of light. For objects moving with larger velocities, one must allow for the effects of special relativity. A correct formula, valid for all velocities, is the following.
Let v denote the total velocity of the object relative to an observer, and let v_r denote the speed with which the object is moving away from the observer, i.e., v consists of both v_r, the so-called *radial velocity* and an additional component, the tangential velocity v_t, perpendicular to the line of sight. The emitted and observed frequencies, ν_{emit} and ν_{obs} then satisfy

$$\frac{\nu_{obs}}{\nu_{emit}} = \frac{\sqrt{1-(v/c)^2}}{1+v_r/c},$$

which implies that

$$\frac{\Delta\lambda}{\lambda} = \frac{1+v_r/c}{\sqrt{1-(v/c)^2}} - 1.$$

For speeds small compared with the speed of light $\sqrt{1-(v/c)^2} \approx 1$, so that

$$\frac{\Delta\lambda}{\lambda} \approx 1 + \frac{v_r}{c} - 1 = \frac{v_r}{c}.$$

.

Lectures 17, 18, and 19

THE BASIC IDEAS OF GENERAL RELATIVITY

As discussed already in an earlier lecture, one of the most striking features about gravity is the Equivalence Principle, which states that the gravitational acceleration experienced by an object is independent of its mass.

Consider three particles – a neutron, a proton, and an electron –, each inserted separately into an identical electric field, and suppose that no other influences are acting on them. The neutron is uncharged and, for this reason, will feel no electric force. The proton and electron have charges that are equal in magnitude and opposite in sign, and this implies that they will feel electric forces again equal in magnitude and opposite in sign. Because the neutron feels no force, it experiences no acceleration. The proton and electron, feeling nonzero forces of opposite sign, will be accelerated in opposite directions. However, because the proton has a much larger mass than the electron, about 1836 times as large, the acceleration which it experiences will be much smaller is magnitude, about $\frac{1}{1836}$ times as large. The net result is that each of the three particles experiences a different acceleration: the electron a relatively large one, the proton a relatively small one in the opposite direction, and the neutron none at all. Analogous results would be obtained if several different particles were subjected to the strong or the weak interaction.

Now suppose that the same three particles are inserted instead into a gravitational field. In this case, all three of the particles will experience a force acting in the same direction. The electron has the smallest mass and thus experiences the smallest force. The proton, being 1836 times as massive, experiences a force that is 1836 times as large as that acting on the electron. The neutron, with a mass about 1839 times as large as the electron, experiences a force about 1839 times as large. All three forces are different. However, the three different particles will each experience an identical acceleration. The electron, being the least massive particle, is the most easily accelerated of the three. The proton, 1836 times as massive, is 1836 times as hard to accelerate. However, the force acting on it is 1836 times as big. The net result is that the proton is accelerated precisely as much as the electron. Similarly, the neutron is 1839 times as massive as the electron, so that it is 1839 times as hard to accelerate. However, because it is subjected to a force 1839 times as large, it too feels the same acceleration.

For the electric interaction (and for the strong and weak interactions), two things determine the total acceleration experienced by some particle, namely (1) its electric charge (or the strong and weak analogues thereof) and (2) its mass. The charge determines how hard the particle is being pulled, i.e., the force acting on the particle. The mass determines how hard it is to pull. The total acceleration experienced by the particle depends on the ratio of the charge to the mass. Two particles with the same charge but different masses – e.g., an electron and a muon, both of which have charge $-e$ but which have different masses – are accelerated differently. Two particles with the same mass but opposite charges – e.g., an electron and a positron – are again accelerated differently.

The situation is very different for gravity. In gravity, the "charge" that determines how hard the particle is pulled, i.e., the total force to which it is subjected, is in fact the same thing as the mass that determines how easy or hard it is for the object to be accelerated.

The ratio of charge to mass is thus the same for each object (identically equal to one), and this implies that every object is accelerated identically. Some philosophers of science distinguish between the two roles of mass by identifying (1) the *gravitational mass*, which measures how hard the object is pulled, and (2) the *inertial mass*, which measures how hard it is to accelerate. In this language, the Equivalence Principle says that, for any object in the Universe, the gravitational and inertial masses are exactly equal.
The equivalence of gravitational and inertial mass, and the fact that the total gravitational acceleration is seemingly independent of the mass of an object, was recognised explicitly by both Galileo and Newton, who fully appreciated its profound significance. In the three centuries since then, an enormous amount of experimental work has been done to try to determine the extent to which it is exactly true, Some of the most important experiments were performed by the Hungarian *Lorand von Eötvös* (1848 - 1919) who, starting in 1899, made a series of high precision experiments demonstrating that, for a variety of different materials, the acceleration induced by the Earth's gravity is the same to better than one part in a billion. Today it has been established that the acceleration is the same to at least one part in a trillion. By the beginning of the twentieth century, the Equivalence Principle was accepted as a well-established scientific fact, and Einstein used it as the starting point for his theory of general relativity.

You have no doubt experienced at least one physical implication of the Equivalence Principle in your everyday life. Suppose, for example, that you get into an elevator on one of the upper floors of a building, carrying a heavy bag of groceries, and the elevator starts going down. What will you feel? As you initially accelerate, you and your bag will feel lighter, as if you were being subjected to a weaker gravitational attraction towards the Earth. However, when you stop accelerating you and your bag will feel the same as you did before the elevator started to move; and when the elevator again decelerates to a halt you and the bag will feel heavier, as if subjected to a stronger gravitational pull. The sensation of feeling lighter and then heavier is related entirely to the fact that you, and your bag, are both being accelerated identically.
This gives rise to the following hypothetical situation which, albeit crazy, makes an important point. Suppose that you passed out last night and then came to, hopefully the next morning, in an elevator. Depending on whether you felt lighter or heavier than expected, you could presumably infer whether or not the elevator and its contents were being accelerated and, if so, in what direction. If, for example, you felt lighter for a long time, you might think that you were in a skyscraper moving downwards for hundreds of feet. However, there is another, perhaps far-fetched, explanation which would explain equally well what you felt. Specifically, your sensations would be completely consistent with you and the elevator having been transported to Mars or some other extraterrestrial object and then placed at rest on the surface of that planet. The gravitational field on the surface of Mars is weaker than the gravitational field on the surface of the Earth and, for this reason, if you were to stand on the surface of Mars you would feel lighter than you normally do on the Earth, just as if you were in a falling elevator.
The important point then is the following: Sitting in a small elevator, *there is no way you could tell whether you were being subjected to a uniform acceleration or whether instead you had been placed in a uniform gravitational field that is different from that on the surface*

of the Earth. In this sense, a constant gravitational field and a constant acceleration are physically indistinguishable. This, perhaps, is the most physical way of stating the content of the Equivalence Principle.

The simplest notion of *relativity*, dating back to Galileo and incorporated by Einstein in his theory of special relativity, says that there is no way in which one can determine fundamentally whether or not an object is at rest. All that one can do determine is whether it is being accelerated and/or moving relative to other objects in nature. The Equivalence Principle goes one step further by saying that one cannot distinguish between a uniform acceleration and the effects of a uniform gravitational field.

It should be emphasised that the Equivalence Principle only refers to constant uniform accelerations. For this reason, it does *not* mean that, if one were prevented from looking out into space, it would be impossible to say that the Earth is experiencing the gravitational effects of the Sun and the Moon, rather than being subjected to some other mechanical acceleration. The reason for this is that the gravitational fields of the Sun, the Moon, and other objects in nature are not uniform. Rather, the strength of the gravitational field associated with any object varies with the distance R from the object, satisfying (at least approximately) the $1/R^2$ relation of Newton's Law of Universal Gravitation. This means, however, that the side of the Earth facing the Moon feels a bigger attractive force towards the Moon than does the side facing away from the Moon. This in turn implies that the innards of the Earth are being subjected to internal stresses that are trying to stretch the Earth, and it is only because of mechanical effects deep inside that the Earth is not pulled to pieces. Indeed, we see one tangible effect of this stretching in the ocean tides. Although the solid portions of the Earth cannot respond significantly to the stretching effects associated with the Moon (and, to a lesser extent, with the Sun and other objects), the liquid water flowing on the surface *can* respond by readjusting itself. As illustrated in Figure 17.1, it is this continual readjustment as the Earth revolves around its axis that is responsible for high and low tides.

The fact that we see tides implies that the Earth is not being subjected either to a uniform acceleration or to a uniform gravitation field. Rather, tidal forces associated with the fact that the strength of gravity changes with distance, demonstrate that the Earth is being subjected to a nonuniform gravitational field. The Equivalence Principle says that one cannot distinguish between a constant gravitational field and a constant acceleration, so that, in a real sense, constant gravitational forces are not fundamental. However, non-uniform gravitational fields, which give rise to so-called *tidal forces*, are fundamental in the sense that they cannot be mimicked by a single uniform acceleration.

It is the preceding interpretation of the roles of constant velocity and constant acceleration – as incorporated both in the principle of relativity and the Equivalence Principle –, combined with the empirical fact that the observed speed of light is the same for all observers, that leads inexorably to Einstein's theory of *general relativity.* Special relativity does not incorporate the effects of gravity. Indeed, Newton's theory of gravity is inconsistent with special relativity in that it allows for the possibility of arbitrarily high velocities. Reconciling special relativity with the Equivalence Principle leads to the replacement of Newton's theory of gravity by Einstein's general relativity.

Combining the principle of relative motion with the constancy of the speed of light led to a theory of special relativity that occasioned a serious change in our understanding of the notions of space and time. Incorporating the Equivalence Principle into a broader theory of general relativity perverts the notions of space and time even more.

Recall that Newton's world view treats space and time as distinct, well-defined entities about which everyone can agree. Thus, e.g., any two observers will (if they measure things properly) agree on how long some process takes or how far apart two objects are. In special relativity, this is no longer true: how long something takes and the distance between objects depends fundamentally on who is doing the looking. It is only by unifying space and time into a single four-dimensional entity, *spacetime*, that one can abstract a notion of four-dimensional distance about which different observers can agree.
The identification of a four-dimensional world as the arena of physical events entails a major change in perspective. However, the four-dimensional world of special relativity is still relatively conventional in the sense that it is "flat," so that, for example, the spacetime distance between two events is given by a simple generalisation of the Pythagorean Theorem. The key idea of general relativity, to be explained in the following, is that *the phenomenon that we perceive as gravity corresponds to the flat spacetime of special relativity becoming curved.*

It is difficult to visualise things in three dimensions, let alone four, and, for that reason, it is useful to describe what is meant by curved space and curved spacetime in the simpler setting of two dimensions (see Figure 17.2).
An ordinary piece of paper, unfolded and uncurled, is an example of a *flat* two-dimensional space where, for example, all the simple rules of Euclid's geometry hold true, including the Pythagorean theorem. In particular, three things about such a surface are well known:

1. Two straight lines, initially parallel, never intersect.
2. Every triangle has three angles adding up to 180 degrees.
3. The shortest distance between any two points is a straight line.

Now contrast this flat surface with the surface of a sphere. The surface of a sphere is again a two-dimensional entity, but in this case it is *curved* in on itself, and hence described as having *positive curvature*. As had been recognised already by the ancient Greeks, geometry on the surface of a sphere is very different from geometry on a flat two-dimensional surface. Thus, in particular, all the properties of flat space denumerated above fail in such a curved space.
On a sphere initially parallel lines *can* intersect. If, for example, one starts two parallel lines at the equator, each moving due north, and follows them forward, one will discover that they intersect at the north pole. Less obvious, perhaps, but also true, is the fact that triangles on the surface of a sphere always have three angles adding up to more than 180 degrees. A tiny triangle sees the surface of the sphere as looking almost completely flat – just as in our everyday life we do not usually recognise the curvature of the Earth –, so that its angles add up to a number only very slightly larger than 180 degrees. However, as the triangle is dilated the angles get bigger and the total gets much larger than 180 degrees. (As a practical matter, the Earth is sufficiently large that, as a practical matter, one can usually ignore the fact that it has positive curvature. However, the Earth's curvature *does*

prove important to architects when they design buildings with a very large surface area, such as the Pentagon.)
As known already to the ancient Greeks, on a sphere the shortest distance between two points is given by a trajectory along a so-called *great circle* passing through those two points. For a sphere of radius R, a great circle corresponds to a circle extending around the sphere which is centred about the center of the sphere, and thus also has a radius R. In any surface or space, flat or curved, the shortest distance between two points is called a *geodesic*. One can thus say that, on a flat surface, geodesics correspond to the usual straight lines, whereas on a sphere they correspond to great circles.
Another example of a curved two-dimensional surface is a saddle, or hyperboloid. Unlike a sphere, in this case the surface curves out on itself, so that it is said to have a *negative curvature*. For this curved two-dimensional surface, initially parallel lines *diverge*, rather than converging, and the sum of the angles of a triangle always adds up to less than 180 degrees.

Stated suscinctly, the basic idea of general relativity is the following: In the absence of any sources of matter and energy, there is no gravity and spacetime is flat. This means, in particular, that spacetime is characterised by the invariant measure of distance,

$$\Delta s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2,$$

appropriate for special relativity. Neglecting gravity and all other forces, a tiny particle moving in this flat space will follow a geodesic, i.e., it will move in a straight line at a constant speed.
Now suppose that some substantial sources of matter and/or energy are present. These sources will have the effect of making the spacetime curved, the details of precisely how the spacetime is curved depending on the overall distribution of the sources. Because of the curvature, the same tiny particle which had followed a linear trajectory in the flat spacetime will no longer move along the same trajectory as before. Rather, it will follow a new trajectory, corresponding to a geodesic in the curved spacetime.
Even more suscinctly: the presence of matter curves spacetime, but this curvature alters the trajectories along which matter moves.

This description treats gravity in a fundamentally different way from the other three interactions of nature. Unlike gravity, which is interpreted as a manifestation and consequence of the curvature of spacetime, electromagnetism and the strong and weak interactions are still viewed as honest forces which affect particle trajectories without altering the curvature of spacetime. A particle moving in flat spacetime will, when subjected to an electric force, still be moving in flat space, but its trajectory will no longer be a geodesic, i.e., it will no longer move in a straight line. Similarly, a particle moving in a curved spacetime will, when subjected to an electric force, be deflected from the curved space geodesic which it would otherwise follow. The Equivalence Principle says that all particles feel the same gravitational effects so that, viewed Newtonianly, they experience the same accelerations. This is why it is possible to view gravity as a manifestation of the curvature of spacetime, which treats everything identically. Other forces do not satisfy the Equivalence Principle and, for this reason, they cannot be described in the same way as a manifestation of spacetime curvature.

As a simple analogy (see Figure 17.3), suppose that, suppressing one dimension, three-dimensional space is visualised as a two-dimensional sheet of frictionless rubber that is stretched out horizontally. If no massive objects are placed on the rubber sheet, the sheet will be essentially flat. It follows that a tiny marble placed on the sheet and given some initial kick will simply move along in a straight line at constant speed. Now suppose, however, that a large stone is placed in the middle of the sheet, causing the sheet to sag. In this case, the sheet of rubber is no longer nearly flat and, if the marble rolls close to the rock, its trajectory will be altered significantly. The presence of the massive stone has curved the surface of the sheet, and this curvature has altered the trajectory of the marble.

At this juncture, it should be stressed that, although Newton's theory of gravity is not completely correct, it is almost correct in the sense that it provides a description of nature adequate to explain almost everything about gravity that we see. Indeed, under ordinary circumstances differences in the predictions of Newtonian gravity and general relativity are exceedingly small. In this sense, one cannot fault Newton or his theory. As inferred from his *Rules of Reasoning in Philosphy*, Newton himself would have said that his theory of gravity should be viewed simply as a working hypothesis which apparently does a very good job of describing and predicting what we see in nature. However, he would not associate any fundamental truth with his theory, any more than he would with Einstein's relativity. General relativity is a better theory in the sense that it permits more accurate predictions about nature, but this does not mean that it is "The Truth," any more than Newton's gravity.

One other important point also needs to be stressed. According to Newton, every massive object in the Universe has its own gravity and, as such, contributes to the total gravitational field acting anywhere. Special relativity says that matter and energy are really the same thing, in the sense that rest mass energy, mc^2, can be converted into other forms of energy. These two ideas are both incorporated into general relativity by the notion that *all sources of energy* contribute to that entity which we call gravity. Thus, in particular, the thing which we perceive as the gravitational field produced by some collection of objects reflects both the rest mass energy of the objects and the energy associated with their motions. Even massless particles like photons and neutrinos carry energy and, hence, contribute to the total gravity.
Similarly, every source of energy feels the effects of gravity. Ordinary particles have mass and therefore experience gravitational effects. Photons of light carry energy and they too experience gravitational effects. As a material object passes the Earth, the Sun, or any other celestial object, it will feel gravitational effects and its trajectory will be deflected by that object. In exactly the same sense, and for exactly the same reason, light passing by that object will also be deflected.

The prediction that light will be "bent" by a gravitational field is inconsistent with Newton's theory of gravity, which says that only massive objects experience gravitational effects. For this reason, the experimental confirmation that this effect really exists played an important role in convincing the scientific community that Newtonian gravity is not completely correct and that general relativity might actually be true. The story behind this is an interesting piece of culture:

Einstein, a vocal Jewish pacifist from Germany, first published the basic ideas underlying the theory of general relativity in 1916 and 1917, in the middle of the First World War, which pitted Germany against Great Britain. At that time, one of the two most prominent astronomers in Britain was *Arthur Stanley Eddington* (1882 - 1944), a Quaker and another vocal pacifist. When Britain started a wartime draft, Eddington was called up to join the military but, because of his religious convictions, he refused. The government was in an awkward position. They did not want to incarcerate Eddington, because of his national and international reputation. Neither, however, were they willing to excuse him from military service. Provisions did exist for alternatives to military service – work in hospital and the like – but Eddington refused to cooperate. Eventually a compromise was reached. As his alternative military service, Eddington, the Englishman, was to lead an expedition of scientists to South Africa to seek experimental confirmation of the German Einstein's prediction that light passing near the Sun should be deflected by the Sun's gravity.

How should one go about showing that light passing by the Sun is bent? Under ordinary circumstances, the Sun is so bright that it is impossible to distinguish anything that appears very close to the Sun in the sky. Thus, in particular, it impossible to see either stars or planets during daylight hours. However, during a total solar eclipse, the *photosphere*, the bright yellow surface of the Sun, is blocked off by the Moon. This means that, during an eclipse, one can in fact see the light emitted by a distant star passing very close to the Sun along its path to the Earth.

By 1917, astronomers were able to determine very accurately both the actual and expected locations of stars in the sky, i.e., where they appear to be and where they should appear to be in the absence of any disturbing influences. The crucial points, then, are (1) that during any eclipse there will be stars which appear in the sky to be very close to the Sun, and (2) that, because they appear to be close to the Sun, their light must be passing very close to the Sun and should therefore experience the Sun's gravity. This means, however, that the star should appear to be in a slightly different position than it would be if the light which is observed did not have to pass very close to the Sun (see Figure 17.4).

Total solar eclipses are a relatively infrequent occurence and, even when they happen, are not visible all over the Earth. However, in 1919 there was a solar eclipse visible in South Africa. The aim of Eddington's expedition was to look at stars near the Sun during the eclipse and to see whether they were slightly displaced from where they would otherwise have been expected to be. The predicted effect was not large, only about 1.75 seconds of arc for a ray of light just skirting the edge of the Sun. However, it was an effect which, if real, could be measured using available technology. The expedition confirmed explicitly that the light had in fact been bent and that the measured deflection was consistent with the predictions of general relativity. The results made the front page of *The New York Times* and other newspapers throughout the world, and, at age 40, the pacifist physicist Einstein was an international celebrity.

The same idea underlies the workings of a so-called *gravitational lens*. Suppose that some very distant object in the Universe, e.g., a quasar, is emitting light into space, some of which eventually reaches the Earth. This light will have to travel for millions or billions of years before reaching us, most of the time passing through near-empty space. However, it is possible that, as it travels through space, some of that light may come close to an intervening galaxy, or some other source of mass and energy, such as a black hole. In this case, the light will feel substantial gravitational effects from the galaxy, which will alter its visual appearance when it eventually reaches the Earth. Thus, in particular, two things can happen:

(1) As illustated in Figure 17.5, it is possible that there may be more than one image of the object in the sky. Given the perturbing influence of the galaxy, there can be more than one path, each curved, which the light can take to go from the quasar to the Earth. However, an observer on the Earth will be unaware of the curvature and, looking in the sky, will see several different images which would seem to be associated with different objects.

(2) The observed image or images can appear brighter than they would otherwise. The reason for this is that the galaxy can focus together light which, if the galaxy were not present, would have spread out in different directions in space. In this sense, the galaxy acts like the lens in a refracting telescope, which explains the origin of the name gravitational lens.

Gravitational lenses were first discovered accidentally in 1979, when astronomers found what seemed originally to be two different quasars, separated only by 6 seconds of arc in the sky, which, upon closer examination, appeared identical in terms of the colours of light which they emitted. It was quickly realised that they were seeing the effects of a gravitational lens, predicted by Einstein in 1936 but never before observed. By 1995, hundreds of gravitational lenses had been found.

Another prediction of general relativity, which was important historically in convincing scientists that general relativity is correct, was the predicted *precession of the perehelion of Mercury*. The idea here is the following. According to Newtonian gravity, the attraction between any two pieces of matter varies as $1/R^2$, where R is the distance between them. This means that, when calculating the net attraction between two objects of finite size, one must add up the $1/R^2$ contributions between each piece of the first object and each piece of the second. If both objects are exactly spherical, the net result of this calculation is a total $1/R^2$ force that depends only on the distance between the centers of the two objects. This fact leads in turn to the prediction that the orbits of the two bodies around their common center of mass are given exactly as ellipses. If the distribution of mass in one or the other of the objects is not exactly spherical, the orbits will manifest small but calculable deviations from a true ellipse. As illustrated in Figure 17.6, what one finds is that, rather than closing exactly on itself after one revolution, the orbit slowly precesses. For a planet orbiting the Sun, this means, for example, that the location in space where the planet is closest to the Sun, its *perehelion*, slowly changes in time. A similar effect also arises when one allows correctly for the fact that although the motion of any planet in the Solar System is dominated by the Sun, it also feels small effects associated with the gravity of the other planets.

By the end of the nineteenth century, astronomers were able to measure this effect for a number of different planets, including Mercury, which exhibits a precession of about 1.56 degrees per century. Most of the observed effects could be attributed to the oblateness of the Sun, i.e., the fact that it is slightly flattened, and to the perturbing influences of the other planets, predominantly big Jupiter. However, as demonstrated in 1882 by *Simon*

Newcomb (1835 - 1909), Newtonian gravity predicts a slightly larger precession for Mercury, about 43 seconds of arc per century, than what is actually seen. General relativity predicts that precisely this deviation should exist, the reason being that, according to Einstein, the force between two objects does not vary exactly as $1/R^2$. Rather, when two objects come very close together, the force between them grows slightly faster than what Newton would predict.

Yet another prediction of general relativity is the existence of *gravitational redshifts*, i.e., the fact that, as light moves away from a massive object with strong gravity, its colour and energy change. That this effect must exist is easy to understand. One knows on the one hand that photons of light carry an energy $E = h\nu$, the value of which is determined by their frequency. On the other hand, one knows that light feels gravity so that, for example, moving out of an attractive gravitational field requires the expenditure of energy. When light moves from a region with more gravity to a region with less gravity, its energy E decreases. This means, however, that its frequency ν also decreases, and that its wavelength increases. The net result is that originally yellow light will have its wavelength shifted towards the longer wavelength red end of the spectrum, whence derives the name gravitational redshift.

This effect was first confirmed experimentally on the Earth by *R. V. Pound* and *G. A. Rebka* in 1960, who showed that, if you shine light upwards from the ground towards the top of a tall building, an observer at the top of the building will find that the energy and frequency of the light have decreased.

Another, related, prediction of general relativity is that, near a massive object with strong gravity, time actually slows down. Indeed, as will be discussed more carefully later on in the context of black holes, if gravity is sufficiently strong time and space can be mucked up to the extent that, in a certain sense, they switch roles.

One of the best known predictions of general relativity is the possibility of *gravitational radiation*, which can carry energy away from a source. Maxwell's theory of electromagnetism implies that two charged particles orbiting each other along accelerated orbits will radiate electromagnetic radiation into space, thereby losing energy. This is a well understood phenomenon which, for example, underlies the generation and transmission of radio waves. In an analogous fashion, Einstein's theory of gravity implies that two massive objects orbiting each other will emit gravitational radiation into space, again losing energy.

This phenomenon has not yet been detected directly. However, strong indirect evidence for the existence of gravitational radiation derives from observations of the so-called *binary pulsar*, discovered in 1974 by *Joseph Taylor* and his student *Russell Hulse*. The binary pulsar consists of two tiny *neutron stars*, the corpses of ordinary stars, orbiting each other in a binary system. Because these neutron stars are extremely tiny, their physical structure is largely irrelevant and they can be visualised as essentially point mass objects (relatively easy to describe) orbiting around their common center of mass. Newton's gravity would say that, neglecting friction and any other external effects, these two objects could continue to orbit each other forever, following essentially unchanging trajectories. However, general relativity implies that, as they orbit each other, they will emit energy into space, with the net result that the binary system as a whole loses energy. How much energy is lost depends

in a calculable way on the masses of the two neutron stars and the sizes and shapes of their orbits, both of which quantities can be determined observationally.
The important point then is that as energy is lost from the system, the form of the orbits must slowly change. Specifically, what will happen is (1) that the two neutron stars will move closer together as the total energy of the orbits decreases, (2) that the shapes of the orbits become more nearly circular and, most easily measured, (3) that the periods of the orbits become shorter, i.e., it takes less time for the neutron stars to move around each other. The rate at which the period decreases is related directly to the amount of energy that is lost and, as such, the measurement of a decrease in period corresponds to an indirect measurement of how much energy has been lost. Data accumulated over a period of twenty years has demonstrated that the neutron stars do indeed lose energy, and that, to a high degree of precision, this loss corresponds exactly to the amount of energy that general relativity predicts should have been emitted into space.

A number of different groups around the world are currently working on the development of technology which should permit the direct measurement of gravitational radiation. Several different approaches are being developed, the most promising of which is probably an approach using something called *laser interferometry.* To explain how this experiment works, it is useful first to understand physically what a gravitational wave really is. Basically, a gravitational wave corresponds to a ripple in the curvature of space which, like an electromagnetic wave, propagates at the speed of light. As the wave moves through space, it has the effect of distorting the normal spacetime curvature, so that the measured distance between two nearby points will oscillate very slightly around the measured distance between them when the wave is not present.
The idea underlying the laser interferometry experiments is to look for these small changes in measureable distances. In its simplest guise, this entails setting up an experiment in which two laser beams of light are emitted from a single source in perpendicular directions, reflected off mirrors, and then reassembled near the source back into a single entity. It is hard to measure exactly the distance light travels in either direction. However, using the wavelike properties of light what one *can* do is determine very accurately the differences in the lengths travelled. The experiment is set up so that, in the absence of any gravitational waves, the distances travelled in the two orthogonal directions are the same. The idea then is that if a gravitational wave passes by it will change the distance in at least one of the directions, so that the two paths no longer have the same length. This mismatch in length is measured and interpreted as a direct manifestation of gravitational waves. (More precisely, it is very hard to determine if the two path lengths differ by an integer number of wavelengths, but, because of the wavelike properties of light, it is relatively easy to see when the path lengths differ by a fractional number of wavelengths. What is done as a practical matter is initialise the experimental apparatus so that, in the absence of any gravitational waves, the two path lengths differ by at most an integer number of wavelengths and then to monitor the system to see whether, at some future time, the difference in path lengths changes.)
Most workers in general relativity believe that, given ongoing advances in technology, such a direct detection will prove possible within the next decade. A number of prominent relativists (including, e.g., Kip Thorne at Caltech and Stephen Hawking at Cambridge)

have made friendly wagers as to whether such a direct detection will occur before 1 January 2001, the beginning of the next millenium.

The most famous prediction of general relativity is probably the existence of *black holes*. A rough sense as to what these objects are, and an explanation of why it is that they can exist, are both easily obtained by combining ideas from Newtonian gravity and special relativity. Newtonian gravity introduces the concept of an escape velocity, the initial velocity with which something must move away from a planet or any other object in order to succeed in getting arbitrarily far away. Special relativity makes the explicit prediction that nothing can travel faster than the speed of light: light travels at a speed c, but everything else travels at speeds $v < c$. The obvious implication is that if the gravity associated with some object is sufficiently strong, nothing will be able to escape, not even light. That objects like black holes could exist was first noted in 1783 by the Reverend John Mitchell, an amateur British astronomer. The first scientist to have contemplated the possibility of invisible objects with escape velocities larger than the speed of light was apparently the Frenchman *Pierre Simon de Laplace* (1749 - 1827), who discussed this eventuality in his 1795 treatise *Mécanique Céleste*, or *Celestial Mechanics*.

Dating well into the twentieth century there were heated arguments as to whether black holes actually exist in nature, or whether instead they are simply theoretical possibilities that are never realised. Even today, there is a fair amount of controversy about whether any specific object that we detect in nature is or is not a black hole. However, one definite conclusion that can be reached is that, given physically plausible assumptions regarding how matter is distributed in space, *general relativity implies that black holes must exist in nature.* This is the content of the so-called *Singularity Theorems* proved in the late 1960's by three mathematical physicists, the American *Robert Geroch* and the British *Stephen Hawking* and *Roger Penrose*.

Given that, according to general relativity, nothing from inside a black hole can ever escape, it would seem difficult to determine observationally precisely what is inside. This is in fact the case. Specifically, the so-called *Black Hole No-Hair Theorem*, proved in a number of different ways by several different physicists, implies that there are only three aspects of a black hole which can be inferred from outside: its mass, its net electric charge (if any), and its angular momentum or spin. At least in principle, these three quantities can all be measured from outside because they have specific effects on the trajectories of particles orbiting near a black hole.

The mass of a black hole can be determined in the same way that one determines the mass of the Sun, by measuring how long it takes for a planet or any other object to orbit around the object. A black hole has the property that, once inside, you can never escape. However, viewed from outside it behaves gravitationally very much like any other ordinary object. Thus, in particular, objects revolving around a spherical black hole satisfy a relativistic version of Kepler's Third Law, which relates the size and period of the orbit to the mass of the central black hole. (A derivation of Kepler's Law using Newtonian mechanics shows that the period of any planet's orbit depends on the masses of the Sun and the planet, as well as the size of the planet's orbit. Thus, assuming realistically that the Sun is far more massive than any planet, one predicts that the period P varies inversely with the square

root of $M_\odot$, the mass of the Sun. This means, e.g., that if the Sun were four times as massive as it is and the size of the Earth's orbit were the same, the period with which the Earth orbits the Sun would be only half as long.)
If the black hole has a net electric charge, this too can be inferred from outside since, like any other set of charges, the charges in the black hole will cause an electric field that influences the motion of charged particles outside the hole.
The angular momentum, or spin, of the black hole can be determined by the fact that, if the hole is spinning, the period of an object orbiting the hole will depend on the direction in which it orbits. If the object orbits the black hole in the same direction that the hole is rotating it will orbit with a shorter period than it would if orbiting in the opposite direction. Measuring the difference in periods between prograde and retrograde orbits permits a calculation of the total angular momentum.

The overall structure of a black hole is very simple. If it has no angular momentum, an isolated black hole must be exactly spherical (although external objects could exert gravitational effects that stretch and alter its shape). There is good reason to think that black holes arising in nature are probably neutral, or very nearly neutral, electrically, so that it is reasonable to approximate them as having no charge. An isolated black hole with zero charge and zero angular momentum is characterised completely by its mass, and only has two interesting "parts:"
1. The *event horizon* represents a one-way surface through which matter and energy may enter but through which, according to the laws of general relativity, they can never again exit. (As discussed in a later lecture, combining general relativity with the ideas of quantum mechanics leads to the prediction of *black hole evaporation*, a process related to particle creation in a gravitational field whereby, in at least some sense, material can again leave the black hole.)
2. The *singularity*, located at the center of the spherical black hole, represents a place where, according to general relativity, gravitational effects become infinitely strong.
It takes a very high density of matter to generate an escape velocity as large as the speed of light and, for this reason, black holes are very small considering the amount of mass that they contain. If the Earth were compacted somehow into a black hole, the event horizon would have a radius slightly less than 1 cm. If the Sun were compacted to a black hole, the radius would be about 3 km. The size of a black hole, as measured by the radius of the spherical event horizon, is known as the *Schwarzschild radius*, in honour of the German astrophysicist *Karl Schwarzschild* (1873 - 1916), who first used Einstein's theory to compute the spacetime curvature around a black hole while dying in a German military hospital from the effects of poisoned gas. This radius satisfies the simple relation

$$R_{sch} = \frac{2GM}{c^2} \approx 3\left(\frac{M}{M_\odot}\right)\ \mathrm{km},$$

where G is the gravitational constant, M is the mass, and c is the speed of light. The second equal sign says that the radius grows linearly with mass by about 3 km for each solar mass, $M_\odot$.

What would you see and feel if you were to fall into a black hole? That depends to a large extent on the size of the hole. However, one important thing to stress is that, in terms of what you actually feel, the event horizon is nothing special. You would not feel any special effects "turn on" as you pass through the horizon, and might not even notice anything was amiss, until you tried to turn around and escape. The only special significance of the event horizon is the fact that, once you have gone through, you can never get back out. The overall strength of gravity become stronger as you approach the singularity and, sooner or later, that *will* have an effect on you. The important thing here is not the overall strength of the gravity but, instead, the existence of enormous tidal forces, which would eventually tear you to shreds. The side of you closer to the singularity would feel a larger force than the side facing away from the singularity, so that you would feel stretched. As you moved closer to the singularity, the stretching would become more pronounced and, eventually, would be strong enough to tear you to bits.

(Parenthetically, this is the same effect which is believed responsible for the rings around Saturn and the other Jovian planets. If a would-be moon gets too close to the surface of a planet, inside what is called the *Roche limit*, the tidal forces acting on the moon will be strong enough to tear it to shreds, leaving only small pieces of debris behind.)

Mathematical Appendix:

Kepler's Second Law in Newtonian Mechanics and General Relativity

As first shown in *Principia*, the fundamental version of Kepler's Third Law assumes the form

$$P^2 = \frac{4\pi^2 a^3}{G(M_\odot + M_p)},$$

where $M_\odot$ and M_p denote respectively the mass of the Sun and the mass of the planet, a is the semi-major axis of the orbit, and G is the gravitational constant. If one replaces $M_\odot$ and M_p by the masses of individual stars, this formula also yields the period with which, according to Newtonian mechanics, two stars in a binary system orbit each other. In the limit that the Sun is much more massive than the planet, i.e., $M_p \ll M_\odot$, this expression reduces to

$$P^2 = \frac{4\pi^2 a^3}{GM_\odot}.$$

For the special case of spherical orbits, this simpler formula, appropriate for a small spherical object orbiting a much more massive spherical object, also holds in general relativity, provided that a is defined appropriately and that $M_\odot$ and P represent the mass of the larger object and the period of the orbit as measured by an observer far from the two-body system.

.

Lectures 20, 21, and 22

STARS AND HOW THEY WORK

In order to understand what stars are like and how they work, one first needs to determine some of their bulk properties, such as their surface temperatures, their luminosities, and their radii. This can be done more or less directly for nearby stars, where the distance to the star can be estimated using the method of parallax. In general, a direct determination of the mass of a star is impossible. However, it *is* possible to estimate the mass of a star if it is located in a binary system.

The surface temperature T of a star is relatively easy to determine. Stars emit a distribution of light that is extremely well approximated by a black body spectrum. However, by analysing that spectrum to determine the wavelength λ_{max} where the most light is emitted, and then invoking Wien's Law, $\lambda_{max} \propto 1/T$, one can infer the temperature T. Another distinct way in which to proceed is to examine the relative intensities of various spectral lines superimposed on the black body spectrum, the point being that these relative intensities are determined by the temperature of the star.

To determine L, the luminosity or intrinsic brightness of a star, one needs to measure two things, namely (1) how far away the star is and (2) how bright it looks. An object that looks dim may do so either because it is intrinsically not so bright or because it is very far away. Similarly, an object may appear very bright either because it is very bright intrinsically or because it is relatively nearby. If one knows how bright or dim the object looks and how far away it is, one can determine how bright or dim it must be intrinisically to look the way that it actually does.

Recall that, for a black body, the amount of energy emitted by each little spot on the surface is determined by the surface temperature; and that the total luminosity is obtained by adding up the energy emitted from all the different little spots on the star. Assuming that the star is a spherical black body, its luminosity L is thus determined completely by its radius R and temperature T. This means, however, that if one knows the luminosity and temperature, using the techniques already described, one can immediately determine what the radius R must be. Again the physical idea is straightforward: An object can be extremely luminous either because it is very hot or because it is very big. If, for example, an object is known to be cool but it is still very bright intrinsically, the object must be very large.

The temperature T can be measured directly for any star that looks bright enough that we can measure the colour of the light which it emits (which means literally millions of stars). However, the luminosity L and radius R can only be determined for the relatively few nearby stars which have a measureable parallax. For this reason, until very recently a direct determination of L, T, and R was limited to about 1000 nearby stars. (Recently, the situation improved enormously with the launch of Hipparchus, a space telescope which can measure parallax angles *much* smaller than anything possible on the Earth, where resolution through a telescope is degraded by atmospheric distortion. Using various computer algorithms to "clean" images, by the late 1990's it has become possible in some cases to resolve details in visual wavelengths down to scales as small as 0.05 seconds of arc or less!)

Determining the mass of an individual isolated star is essentially impossible, since the mass can only be measured directly through its influence on other massive objects. However, if two stars are located in a binary system, it *is* possible (at least in principle) to compute their masses using an analogue of Kepler's Laws. If one knows the size of the two stellar orbits and the period with which the stars move around each other, it is possible to infer directly the sum of their two masses, $M_1 + M_2$. However, if one can also determine the location of the center of mass, i.e., the central pivot point of the orbits, he or she can infer the ratio M_1/M_2. Knowing the sum and the ratio allows one to formulate two algebraic equations which can be solved to yield M_1 and M_2 individually.

If one examines the values of L, T, and R for nearby stars, one observes a very striking and regular pattern, summarised in the so-called Hertzsprung-Russell diagram, which is named for the Danish *Ejnar Hertzsprung* and the American *Henry Norris Russell*, who, independently, first discovered these regularities between 1905 and 1913. An example of such a *Hertzsprung-Russell* (or *H-R*) *diagram* is provided in Figure 20.1. Here the abscissa, or x-axis, corresponds to surface temperature, with the unusual convention that T increases as one moves towards the left. The ordinate, or y-axis, corresponds to the intrinsic brightness or luminosity.

Basically one finds four distinct classes of stars. About 90% of the observed stars fit along the so-called *main sequence*, which extends from the upper left-hand corner towards the lower right-hand corner, i.e., from hot, very luminous stars towards cooler, less luminous stars. The Sun is a typical main sequence star, roughly average in surface temperature (about 5800 K) and radius (about $696,000$ km). The so-called *giants* or *red giants* are situated slightly above and to the right of the main sequence. These correspond to stars which are brighter than main sequence stars with the same temperature and, as such, they must have larger radii, which explains the origin of their name. The typical size of a red giant star ranges between about $3R_\odot$ and $200R_\odot$, i.e., as large as 100 million km or more. Above the giants are the *supergiants*, which are yet brighter and larger in radius. Supergiants can have radii of up to 1000 $R_\odot$, this corresponding to a size roughly equal to the distance from the center of the Sun to Jupiter. Finally, there are a few *white dwarfs* located below and to the left of the main sequence. These correspond to hot stars which, nevertheless, are very dim, so that they must have small radii. The typical size of a white dwarf is about $\frac{1}{100}R_\odot$, roughly equal to the radius of the Earth.

Aside from their differences in luminosity, radius, and black body temperature, these four classes of stars also exhibit differences in the forms of their spectral lines. This has the fortunate consequence that, even if one cannot measure the size of a star, he or she can infer from the spectrum whether it is a main sequence star, a giant, a supergiant, or a white dwarf.

One might expect that these four different classes of stars have very different masses, e.g., that the huge supergiants would be much more massive than the much smaller white dwarfs. This, however, is *not* true. It *is* true that no very light supergiants are seen and that no very massive white dwarfs are seen. However, main sequence stars come with a relatively broad range of masses, including values comparable to those observed for giants, supergiants, and white dwarfs.

The observed radii of stars vary considerably, from about $\frac{1}{100}R_\odot$ to $1000R_\odot$, a factor of 10^5, and the observed luminosities range from about $10^{-4}L_\odot$ to about $10^5L_\odot$ (where $L_\odot$ denotes the luminosity of the Sun), a factor of 10^9. However, the observed range of masses is considerable smaller: Few, if any, stars are detected with masses below about $0.05M_\odot$ or above $50M_\odot$. The overwhelming majority of stars have masses less than $3 - 5M_\odot$.
A plot of stellar luminosity versus stellar mass for main sequence stars exhibits a simple and regular trend, known as the mass-luminosity relation (see Figure 20.2). Basically what one observes is a relatively tight correlation between the mass of a star and its luminosity, the luminosity increasing very quickly with increasing mass: More massive stars burn much more brightly.

A direct determination of L and R is only possible for relatively nearby stars for which parallax can be measured. Mass determinations are even more limited because they can only be done for stars in binary systems. However, it is possible to extend the preceeding description to stars in general if one makes two exceedingly plausible assumptions:
The first of these is that the fundamental properties of stars are universal, and that distant stars for which parallax measurements are impossible have the same physical properties as nearby stars for which parallax measurements can be made. Granted this assumption, it is possible to estimate indirectly both the luminosity and the radius of a star, as well as its distance from the Earth, provided only that its spectrum can be analysed. By analysing the light received on the Earth, one can determine directly both the temperature of the star (from the colour) and what class of star it is (from the spectral lines). However, the H-R diagram reveals that, for nearby stars of a given class, there is a relatively sharp correlation between temperature and luminosity. It follows that, assuming distant stars share the same properties, one can estimate L, and hence R. However, by comparing the intrinsic brightness, as measured by L, with the apparent brightness, also easily measured, one can also determine how far away the distant star actually is. This method of estimating distances is sometimes called the method of *spectroscopic parallax.*
The second assumption is that the mass-luminosity relation observed for stars in binary systems also holds for individual stars not bound in a binary. This is believed true because the stars in a binary are usually – albeit not always! – separated by a distance sufficiently large that they should not exert a significant influence on each other's structure. Granted this assumption, an examination of the spectrum, which leads to a knowledge of L, can also be used to estimate the mass M.

The preceding explains how one can estimate various bulk properties about stars. However, it says nothing about how a star works. Why does it have the temperature, luminosity, and radius that it does?
Ultimately the regulator that determines what stars are like is gravity. Gravity is an attractive interaction operating on every source of mass and energy in the Universe, and its natural propensity is to compress everything down to the smallest possible dimensions. If something inside the Earth did not resist gravity, the Earth would be compressed down to essentially zero size. If something inside a star did not resist gravity, the star too would be compressed to zero size. On the Earth, what resists gravity is material stresses within the solid and liquid innards of the Earth, which derive ultimately from the electromagnetic

interaction. But what about a star?
By the end of the nineteenth century scientists knew that stars are balls of hot gas, and, from laboratory experiments, they were also well acquainted with the fact that a sample of hot gas is characterised by a large pressure that tries to blow the sample apart. They therefore inferred correctly that it is the fact that a star is hot that leads to a source of pressure which prevents gravity from compressing the star down to an extremely small size. However, they also recognised that heat from the star is continually being radiated out into space in the form of electromagnetic radiation. But this means that, without a continuing source of energy to produce the heat, the pressure in the star would decrease, thereby allowing gravity to compress the star to a smaller size. In particular, the Englishman *William Thomson, Lord Kelvin* (1824 - 1907), had demonstrated mathematically that, according to Newtonian gravity, if no source of pressure were present to resist gravity, the Sun would shrink down to about 37% of its current size (more precisely, it would shrink by a factor of $1/e$) in about 30 million years. Similarly, in the absence of some other source of pressure, the Sun must have been about 2.7 times larger 30 million years ago than it is today.
When this prediction was first made, there was no clear understanding as to how old the Solar System and the Earth really are, so that one could not rule out the possibility that the Sun is slowly contracting. However, by the early twentieth century geologists had accumulated evidence suggesting that the Earth must be at least a billion years old, so that it seemed hard to imagine a Sun which changes significantly over the course of a few tens of millions of years. A general consensus therefore emerged that there must be some energy source that continually replaces the energy that the Sun and other stars radiate into space.
There had also evolved the vague idea that this energy source must somehow be related to newly discovered nuclear reactions: any other known source of energy would be exhausted far too quickly. Thus, e.g., if the Sun were made completely of coal and its luminosity were produced by burning that coal, it would run out of fuel within about 3000 years, a much shorter period than recorded history. However, scientists of the time knew nothing about nuclear fusion and, consequently, had no clear idea as to the source of energy in a star.

Significantly, however, even though astronomers could not identify the source of energy in a star they were still able to construct highly accurate models as to what a star must be like inside: All that they needed to do was assume (1) that there is *some* source of energy to provide the needed pressure, and (2) that this energy source facilitates a stable balance between pressure, which is trying to blow the star apart, and gravity, which is trying to shrink it down to a point. Two of the leading figures in this enterprise were the German astronomer *Robert Emden* and the American mathematician *James Homer Lane*.
Crudely speaking, these astronomers were working like a crooked accountant dealing with dirty money. Even if the accountant has absolutely no idea where the money comes from, he or she can still balance the books for a stable, albeit illegal, business. The accountant balances income and expenditure, assuming that there is a source of income and a fixed set of expenditures. The astronomer balances pressure and gravity, assuming that there is some steady source of energy to produce the pressure and a fixed amount of gravity, determined by the total mass of the star.

Only in the 1930's and 1940's was it fully understood and appreciated by physicists like *Hans Bethe* (b. 1906) that the energy responsible for heating a star like the Sun and providing the necessary pressure derives from nuclear fusion. As noted already, the idea behind nuclear fusion is that lighter nuclei can be combined into heavier nuclei, with the net result that a tiny fraction of the original rest mass energy is transformed into energy in the form of photons and neutrinos. Deep inside a star, the photons generated via nuclear fusion heat the gas to extremely high temperatures, millions of degrees or more, thereby providing a source of pressure that enables the star to resist gravity. This energy slowly works its way out to the surface of the star and is eventually radiated into space as electromagnetic radiation.
How exactly does this fusion proceed? Basically, this involves a complicated network of nuclear reactions which has the net result of first combining hydrogen into helium, then helium into carbon, and so forth, generating progressively heavier nuclei. Deep in the core of a star, this process will continue indefinitely until one of two things happens:
(1) Some alternative source of pressure eventually comes into play, which obviates the need for nuclear fusion. One such source is provided by quantum mechanical *electron degeneracy pressure*, a pressure that arises in high density matter when electrons are scrunched very close together. As will be seen later, this electron degeneracy pressure is responsible for the fact that, at the end of its lifetime, a star like the Sun will have been converted into a white dwarf.
(2) Eventually the hydrogen in the core is converted completely into the isotope iron-56. This nucleus corresponds to the ultimate nuclear ash, i.e., the lowest energy state of nuclear material. Once the nucleons have been converted into iron, there is no way in which to extract any further energy via nuclear fusion. At this stage, there is no longer a source of pressure to resist gravity, so that the core of the burnt out star contracts down to form a tiny *neutron star* or *black hole*.

One other point needs to be explained. Fusion does not happen spontaneously on the Earth, or on the other planets, even when such basic constituents as hydrogen or helium are readily available. Nuclear fusion can only occur in a very high temperature environment. The reason for this is not hard to understand: Fusion involves sticking nuclei together, i.e., assemblages of nucleons containing protons. If one or more protons is brought sufficiently close together, they will experience a powerful attractive force associated with the strong interaction. However, if they are further apart their interaction will instead be dominated by the electric repulsion associated with the fact that they all have the same positive charge (like charges repel!). It follows that, in order to get two nuclei to stick together in a larger nucleus, they must somehow overcome the electric repulsion at larger separations
This is impossible at low temperatures, when nuclei move very slowly. However, at higher temperatures these same nuclei will move with higher velocities and higher energies. The important point then is that if the temperature is high eneough, when two nuclei collide into one another they may have enough energy to overcome the electric repulsion, so that they can get close enough together that the strong interaction comes into play. This explanation also accounts for another fact: Because larger nuclei contain more protons, they will experience a larger electric repulsion so that, if they are to stick together, a harder collision is required. This implies that to fuse progressively larger nuclei increaingly larger

temperatures are required. Fusion of hydrogen into helium requires a temperature of a few *million* degrees Kelvin. Fusion of silicon into iron requires a temperature well in excess of one *billion* degrees.

The key recognition in all of this is that the four observed classes of stars – main sequence, giant, supergiant, and white dwarfs – correspond to different stages in the life history of a single object, i.e., a star will evolve from class to class during the course of its lifetime.

Given this lengthy introduction, turn now to a brief description of the life history of a star, a never-ending battle between gravity and various different sources of pressure. Basically, the life of a star proceeds through four different stages:

The first stage begins with a cold, low density cloud of gas, comprised primarily of hydrogen and helium, floating in otherwise empty interstellar space. Because the gas is very cold, only some 5 to 15 degrees above absolute zero, it has relatively little pressure. However, this implies that, if the cloud is dense enough, attractive gravity will overcome the small repulsive pressure, and the cloud will begin to contract.
Why does a cloud begin to contract precisely when it does, rather than a million years earlier or a billion years later? This presumably happens because the cloud is subjected to some sort of disturbance that compresses the gas, thereby increasing the density, tipping the balance between pressure and gravity, and forcing the onset of contraction. In many cases, this is associated with *spiral waves*, which will be discussed later on in the context of galaxies. However, spiral waves are not the only thing that can trigger the contraction. Other possibilities include the explosive death of an older star and the radiation emitted from other, nearby stars, each of which can serve to squish a nearby cold gas cloud.
In any event, for one reason or another the gas cloud begins to contract. When this occurs, two things happen, more or less simultaneously. One is that the cloud can fragment into two or more separate pieces, each of which will eventually evolve into a different star. The other is that, as it contracts, the (pieces of the) cloud will become denser and, for this reason, increase in temperature. This latter phenomenon – that higher density means higher temperature – is something well known from terrestrial experience. For example, most of you have probably noticed that, when you pump air into a bicycle tire, inflating it to a very high pressure and density, the air in the tire will heat up. This explains the fact that the needle through which the air is pumped can feel quite warm to the touch.
When the gas is cold, fusion cannot occur since the gas molecules are not moving fast enough. Eventually, however, the temperature of the gas increases, first to the point that the molecules break apart into individual atoms, then to the point that the atoms are ionised into nuclei surrounded by unbound electrons, and, finally, to the point that protons, i.e., hydrogen nuclei, can stick together to form helium. How long this contraction takes depends sensitively on the mass of the star. For a $15M_{\odot}$ star, this may take about 100,000 years. For a star the size of the Sun, it would take longer than 10 million years. The difference reflects the fact that larger stars have more gravity, which can trigger a more rapid contraction.

The net result is that, once the central part of one of the fragments reaches a temperature of a few million degrees Kelvin, it is hot enough for hydrogen to fuse into helium, so that one sees the onset of so-called *hydrogen core burning.* This fusion serves a source of pressure which is finally strong enough to resist the gravity, so that the contraction of the gas cloud halts. At this stage, the gas cloud has been converted into a newborn main sequence star. This main sequence star exists as an equilibrium between two basic tendencies: gravity, which is still trying to compact all the gas into a tiny volume, and a repulsive pressure, generated by the fusion of hydrogen into helium in the core of the star, which is trying equally hard to blow the gas apart. At this stage, helium is not fused into anything heavier since the core of the main sequence star is not sufficiently hot to allow helium nuclei to stick into yet heavier nuclei.

The period that a star spends on the main sequence corresponds to the longest period of its life. Eventually, however, all the hydrogen in the hot central core is converted into helium, i.e., all the fuel is used up. How long this takes again depends sensitively on the mass of the star, which determines how much gravity the pressure has to overcome. For a $40M_{\odot}$ star, the fuel is depleted in about 1 million years; for a star like the Sun, the lifetime is about 10 billion years. A star of mass $0.5M_{\odot}$ will have a total lifetime of about 56 billion years, which means that, even if it were born very early in the history of the Universe, it would still be a main sequence star today.

How actually does the astronomer determine what the innards of a star must be like and/or how long it can live? Basically, as intimated already, this involves model-building, applying the known laws of physics and chemistry to predict what it is that a blob of gas of given mass will do. In constructing these models, the astronomer basically works like an accountant, working out the implications of four basic rules:

The first of these is the demand that the star satisfy the condition of *hydrostatic equilibrium*, i.e., that the fluid material within the star manifest a balance, or equilibrium, between the two competing influences that are acting on it, namely gravity and pressure. Near the center of a star gravity is especially strong, so that more pressure is required to resist the gravity. Out towards the surface there is less gravity and consequently less pressure as well.

The second is *conservation of mass*, which, in this context, entails keeping track of the location of the matter which, ultimately, determines the form of the gravitational field.

The third is *conservation of energy*, which entails keeping track of the generation of energy manufactured deep in the core of the star as a result of nuclear fusion.

The fourth, and final, ingredient involves *energy transport*, i.e., tracking how the energy, manufactured deep in the core of a star, eventually works its way to the surface and subsequently escapes into space.

Under ordinary circumstances, it is not necessary to worry about an analogous *mass transport* since typically there is very little net motion of mass in or out. The only significant exception is for the outer regions of a star in the earliest and latest stages of its evolution, when a substantial amount of mass can be ejected into interstellar space.

Constructing a model for some star involves specifying its mass and chemical composition and then invoking these basic rules to compute various quantities that characterise its bulk structure and internal composition. Thus, for example, one can compute how the density, often abbreviated ρ, or the temperature T varies as a function of distance r from the middle of the star. Similarly, one can compute $L(r)$, the total luminosity generated within a distance r of the center. Thus, e.g., for a star like the Sun one calculates that the central density is about 160 $\mathrm{gm/cm^3}$ and that the temperature at the center is about 15.5 million K. At the surface, the density has dropped off to essentially zero, and the temperature has dropped to about 5800 K. About 64% of the total mass resides, and almost all of the luminosity is generated, within the inner 30% in radius.

Such modeling also allows one to investigate how the luminosity, radius, and surface temperature vary if the mass of the star is changed. A large mass M means lots of gravity, which means that lots of pressure is required to resist the gravity. However, this requires higher temperatures, which means more energy must be generated, so that the star is forced to fuse faster. It follows that a large mass M corresponds to a large luminosity L, which explains the observed mass-luminosity relation. Even a small increase in mass corresponds to a large increase in luminosity. Thus, e.g., a $4M_\odot$ star will be about 128 times as bright as a $1M_\odot$ star with the same chemical composition. Massive stars burn much more brightly, exhausting their nuclear fuel much more quickly, which explains why it is that they have much shorter lifetimes.

As a mean sequence star grows old, its internal composition is altered as hydrogen is converted into helium, and this causes changes in such bulk properties as surface temperature and luminosity. The details of these changes depend on the mass of the star but, overall, the net result typically is that, as it ages, a star will become slightly cooler but slightly more luminous. The fact that L increases even though T decreases then implies that the radius R must also increase. The case of a 1 $M_\odot$ star like the Sun is more or less representative. According to our best understanding, the Sun was born as a main sequence star about 5 billion years ago with a luminosity of about $0.6L_\odot$, i.e., 60% of its current luminosity. By the end of its lifetime, about 5 billion years in the future, its luminosity will have increased to a value of about $2L_\odot$. This slow increase in luminosity will presumably cause the extinction of life on the Earth, at least as we know it. First the polar icecaps will melt, next the oceans will evaporate, and, finally, the Earth will lose its atmosphere.

Once the hydrogen in the core is finally exhausted and the source of pressure removed, gravity again wins and the core begins to contract. As the core contracts, its temperature will rise still more, and the heat from the very hot core will leak out to the remaining parts of the star where hydrogen is still abundant. However, once the regions outside the core are hot enough, they will be able to start fusing hydrogen into helium, which leads to the onset of so-called *hydrogen shell burning* in the regions immediately outside the core.

The sudden onset of hydrogen shell burning means that outside the core there is *too much*, rather than too little, pressure. For this reason, pressure beats gravity outside the core, and the outer portions of the star are blown away from the core until they assume very large physical dimensions. As the gas moves further away from the hot core, it cools again and, in the outer regions, a new balance is achieved between gravity and pressure. The main sequence star has been converted into a red giant: called giant because it is big, called red because its surface is cooler, and hence redder, than the surface of the original main sequence star from which it evolved.

Meanwhile, the core continues to contract and heat up until the temperature approaches 100 million degrees Kelvin, at which point it is hot enough for the helium to fuse into carbon via *helium core burning.* This new fusion serves as a source of pressure to resist gravitational contraction, so that, once again, the core stops contracting.

However, just as for the case of a main sequence star, the fuel is eventually depleted: all the helium in the core is eventually converted into carbon. At this stage, gravity wins yet again and the carbon core begins to contract and heat up. As before, heat from the core can leak out to the surrounding regions, where nuclear fuel, this time helium, is still plentiful. The net result is *helium shell burning*, which causes the outer portions of the star to expand yet more, converting the red giant into a supergiant.

The final fate of this supergiant depends on the mass of the star:
A relatively low mass star has less gravity overall and, for this reason, needs slightly less pressure, so that it does not get quite as hot as does a more massive star. What this means is that, before the carbon core gets hot enough to fuse into yet heavier elements, it becomes sufficiently small and compact that *electron degeneracy pressure* comes into play. This quantum mechanical source of pressure halts the contraction, preventing the need for any further source of fusion. The outer portions of the star, bloated out to huge physical dimensions, feel a relatively small gravitational attraction towards the core and slowly leaks out into space as a *planetary nebula*, so that the central core is revealed as a white dwarf, the corpse of a low mass star. A white dwarf is small and very compact, with a typical density of about 1 million gm/cm^3, this corresponding to about a ton per teaspoon.

For a higher mass star, with more gravity overall, the core will be somewhat hotter. This means in particular that, before electron degeneracy has a chance to halt the contraction, it becomes hot enough to trigger the fusion of carbon into yet heavier elements. Once this happens, the star undergoes a complicated series of nuclear reactions, whereby a number of different elements are generated, until finally the central core has been converted into iron. Working in from the outside, at this stage the supergiant will contain roughly spherical concentric regions with hydrogen shell burning, helium shell burning, carbon shell burning, neon shell burning, oxygen shell burning, and silicon shell burning, all surrounding the central iron core, which has a temperature of about 3 billion K.
The late stages in the lifetime of a star are important because it is only in giants and supergiants that the heavier elements are produced. According to the Hot Big Bang Model, the only elements present in any appreciable abundance when the Universe was young were hydrogen and helium (along with tiny amounts of lithium and beryllium). Carbon, nitrogen, oxygen, iron, and the like, the elements required for the existence of life as we know it, were all produced almost entirely inside high mass stars during the late stages of stellar evolution.
As the core begins fusing into progressively heavier nuclei, it evolves more and more rapidly. The reason for this is not hard to see: Because the core becomes denser and more compact, there is more gravity to overcome, so that an ever-increasing amount of pressure is required, which means that more and more fuel is expended at an ever-increasing rate. Thus, e.g., for a $25M_\odot$ star, hydrogen core burning will last for 7 million years, helium core burning for 500,000 years, carbon core burning for 600 years, neon burning for 1 year, oxygen burning for 6 months, and silicon burning for only about 1 day.

Once the core has been converted to iron, it is impossible to extract any more energy and pressure by fusing into yet larger nuclei. Gravity again wins, and the core begin to contract and heat up extremely quickly. Within about 0.2 *seconds*, the density is so high that the electrons in the core are literally squished into the nuclei, at which point they combine with the protons to form neutrons and anti-neutrinos, the latter of which stream out into space. The net result is that the core of the star is converted into what is essentially a single gigantic nucleus, comprised almost entirely of neutrons. After a total of about 0.25 seconds from the onset of the collapse, the density becomes comparable to the density of ordinary nuclear matter, about 4×10^{14} gm/cm^3 or 400 million tons/cm^3. At this stage a new source of pressure arises, namely *neutron degeneracy pressure*, which prevents the closely packed neutrons from being compacted any further.
Once neutron degeneracy pressure sets in, the matter is essentially incompressible, so that it cannot contract any more. However, material outside the neutron core is still falling down on the core with very high speeds, 15% the speed of light or more. When this material smacks into the core, its inward progress is halted, and, within milliseconds, the material bounces back out like a basketball rebounding off a hard surface. The net result is that the outer portions of the star blow out into space, and the star is seen to explode as a gigantic *supernova,* leaving behind a neutron star.

If the star is too massive, not even neutron degeneracy pressure can halt the collapse. In this case, contraction will proceed until the escape velocity is greater than the speed of light, at which point the core becomes a black hole.

Such supernovae only arise from rare high mass stars and, for this reason, are relatively infrequent occurences in our Galaxy, although they are readily detected in other galaxies. The most recent detection in the Milky Way was a 1604 supernova observed by Johannes Kepler. However, a supernova *was* observed as recently as 1987 in a nearby satellite galaxy called the Large Magellanic Cloud, less than 100,000 lightyears away.
Actually two different types of supernovae are observed, known creatively as Type I and Type II. At its brightest, a *Type I supernova* has a luminosity of about $10^{10} L_\odot$, 10 billion times the luminosity of the Sun, which means that it is as bright as an entire ordinary galaxy. The supernova "turns on" in about a day and then slowly fades away over the course of several hundred days. Type I supernovae are characterised by the fact that they show signs of only relatively small amounts of hydrogen gas.
At their brightest, *Type II supernova* are slightly less luminous, only about $10^9 L_\odot$. They too turn on in about a day and then fade away over the course of a year or so. They differ from Type I supernovae in that they shows signs of involving lots of hydrogen.

It is Type II supernovae that are believed caused by the death of a high mass star, the presence of hydrogen reflecting the fact that, even in the late stages of its lifetime, a star has lots of hydrogen in its outer parts.
Type I supernovae are apparently a different sort of beast, usually, if not always, arising in a binary system containing a white dwarf with a giant or supergiant companion separated by a relatively small distance. As a giant or supergiant, the companion star is very large, so large that, in certain cases, part of the gas in its outermost regions feels a stronger gravitational attraction towards the white dwarf than towards the center of the giant. The

net result is that gas originally part of the giant or supergiant can fall onto the white dwarf. The white dwarf, of course, is the relict carbon core from another star which was already a supergiant and had ejected its outer layers of hydrogen into space. As gas from the current giant or supergiant spills onto the dwarf, its gravity will compress the carbon on the surface of the white dwarf, causing an overall rise in temperature. The first result of this increase in temperature is the conversion of the infalling hydrogen into helium and then carbon. However, when the temperature is high enough something else will happen: the carbon will begin to fuse into heavier nuclei and the white dwarf itself will explode as a gigantic thermonuclear bomb.

The remnants of a supernova remain visible through a telescope for hundreds or thousands of years, corresponding to a milky glow of gas that eventually disperses through space. Thus, e.g., the Gum Nebula, with a radius of about 500 lightyears centred about 800 lightyears away, is the remnant of a star that exploded as a supernova about 4000 BC. At the time, it would have been visible to the ancient Chinese and Egyptians, appearing at its brightest to be about 10% as bright as a full moon, i.e., about 250 times brighter than Venus.

Lectures 23 and 24

WHITE DWARFS, NEUTRON STARS, AND BLACK HOLES

A white dwarf is a dense object, comparable in mass to the Sun, generated as the corpse of a relatively low mass star. Since low mass stars are much more common than high mass stars, there are no doubt many more white dwarfs than neutron stars or black holes. The Sun is expected to end its life as white dwarf. Exactly how many white dwarfs are out there is very difficult to say. When they first form, they are extremely hot, since they are the relicts of the high temperature carbon cores within supergiants. However, over the course of billions of years they slowly radiate this energy into space and slowly fade out of sight. Stellar astrophysicists predict that, even though white dwarfs constitute only a tiny fraction of the stars that we actually see in the sky in the neighbourhood of the Sun, there may be 60% as much mass in the form of essentially invisible white dwarfs as in the other types of stars that we can see.

Aside from the fact that they are small and compact, two other features about white dwarfs are well established: that they rotate very quickly and that they have very strong magnetic fields.

That white dwarfs rotate very quickly is a consequence of conservation of angular momentum, or spin. Every object that we know in the sky rotates about its axis, some rapidly and others much more slowly. For example, the Earth rotates around its axis with a period of about 24 hours. The Sun, a typical main sequence star, rotates with a period of about a month, a fact first established by Galileo, who watched sunspots appearing to sweep across the surface of the Sun as a result of the Sun's rotation (staring at the Sun through a telescope was probably responsible for Galileo's blindness in his later years). Suppose, however, that the Sun were to shrink down to much smaller dimensions, as will be the case for its central core as it evolves into a white dwarf. Conservation of angular momentum then implies that, because the object has become more compact, it would be forced to spin much more quickly. If, e.g., the Sun were to shrink to about one hundredth its current radius, one would anticipate that it would spin with a period about one ten thousandth as long, only about 45 minutes or so.

That white dwarfs have strong electric fields is also easily understood. Many astronomical objects, including the Earth and the Sun, have moderate magnetic fields. However, if the Sun were to shrink down to one hundredth its current size, dragging its magnetic field along, the same amount of magnetism would be compressed into a much smaller region of space, leading to a much stronger magnetic field. The magnetic field of a white dwarf can be determined observationally by the fact that it induces measureable deformations in various spectral lines of the light emitted by the dwarf, which can be detected from the Earth. Specifically, what one finds is that, in the presence of a magnetic field, what was originally a single spectral line can be split into several different lines, the size of the splitting depending on the strength of the magnetic field. This phenomenon known as the *Zeeman effect*, was first discovered experimentally by the Dutch physicist *Peter Zeeman* in 1896. By observing this line splitting in white dwarfs, one concludes that many of them have magnetic fields as large as a million times the magnetic field of the Sun, or even more.

Electron degeneracy pressure is an unusual source of pressure which leads to a striking, and significant, prediction about the structure of white dwarfs. Specifically, one can show that the laws of quantum mechanics, when combined with Newtonian gravity, imply that the radius of a white dwarf *decreases* with increasing mass. *Large mass M implies small radius R.* In particular, there is a maximum possible mass for a white dwarf, the so-called *Chandrasekhar limit*, which equals about $1.4M_\odot$. As first demonstrated in 1932 by the Indian astrophysicist *Subrahmanyan Chandrasekhar* (1910 - 1995), the radius of a white dwarf should actually become extremely small for masses above about $1.3M_\odot$, and it is in fact impossible to satisfy the condition of hydrostatic equilibrium for $M > 1.4M_\odot$ allowing only for electron degeneracy as a source of pressure. If the core of a star ever becomes larger than approximately $1.4M_\odot$, it cannot end its life as a white dwarf. It must collapse to a yet denser state.

The concept of degeneracy pressure was first appreciated in the early 1920's, with the advent of quantum mechanics. Subsequently, in 1926, the American physicist *Richard Fowler* suggested that electron degeneracy pressure is the source of pressure in a white dwarf star. The details of how white dwarfs work was developed in 1931 and 1932 by Chandrasekhar and the Russian physicist *Lev D. Landau* (1908 - 1968). In 1932, the neutron was discovered experimentally, and in 1933 and 1934, the Americans *Walter Baade* and *Fritz Zwicky* suggested that stellar cores with masses in excess of the Chandrasekhar limit would end up as neutron stars. In particular, they argued correctly that explosive supernovae should correspond to the death of high mass stars which could leave neutron stars behind. However, the scientific community as a whole was exceedingly sceptical. In part this was because the predicted size and density of a neutron star seemed absurd – an object as massive as the Sun compacted to a size $R \approx 10 - 20$ km, with a density $\rho \approx 4 \times 10^{14}$ gm/cm^3. In part this was also because Zwicky had acquired the reputation of being a bit of a flake, although most of his ostensibly flaky ideas have subsequently been shown to be correct.

The idea of neutron stars quickly disappeared from sight, gone albeit not completely forgotten. Then in 1967 a young graduate student in Cambridge, England named *Jocelyn Bell* made an extraordinary discovery. As her Ph. D. thesis topic, Bell had decided to observe and analyse the radio emissions from several *quasars*, recently discovered objects which at the time were perhaps the most glamorous objects in astronomy. However, when she started looking for quasars with a sophisticated radio telescope at Jodrell Bank, England, she discovered a very regular radio signal, corresponding to a nearly constant period of 1.3373011 seconds. What could this signal be? Originally, Bell and her supervisor thought that she was picking up some spurious signal, perhaps noise from the apparatus or some sort of disturbance on the Earth. However, Bell quickly became convinced that she was detecting a real signal from outer space and, after some exceedingly acrimonious interactions with her supervisor, succeeded in convincing him as well. They published their results and, in 1974, he was awarded the Nobel prize for the discovery of what had come to be known as *pulsars*.

The object which Bell discovered was named a pulsar because the signal seemed to pulse on and off, but there was originally no idea as to what it really might be. Thus, in particular, when the U. S. military learnt about the discovery they began studying the object, giving their progran the title *Project LGM* (*Little Green Men*) because of the belief that the signal might be coming from some extraterrestrial civilisation trying to communicate with other civilisations. Quickly, however, it was realised that that explanation would not do. Three other sources were soon discovered, each located very far from the others. Moreover, it was recognised that the amount of energy emitted by these objects is enormous: the observable emission extended through a broad range of wavelengths, much broader than would be necessary to send an easily detectable signal and involves much more energy than even a normal star can emit. Finally, theoretical astrophysicists brushed off the old ideas of Zwicky and Baade and realised that what they were seeing was probably a neutron star.

The chain of argument leading to that conclusion is instructive. At the time it was well known that many red giants pulsate, getting larger and smaller in physical dimension, with periodic changes in luminosity and surface temperature. It was therefore natural to ask whether these pulsars are variable stars. This, however, seemed unreasonable. The pulsations of red giants typically have a period of days or weeks, whereas pulsar signals have periods of about a second. There was no way in which any known object, even something as small as a white dwarf, could be made to pulse with a period that short.

Another possibility to be suggested was that one was seeing a "hot spot" on the surface of a rotating white dwarf which appears to "turn on" when it faces the Earth and to "turn off" when facing away. This, however, made no sense. There is no way that a white dwarf could ever rotate with a period as short as a second. When an object spins it feels a force that tries to flatten it out; and, if it spins sufficiently rapidly this force will be strong enough to pull it to pieces. Any object the size of a white dwarf that rotated with a period as short as 1 second would unquestionably blow apart.

However, it was quickly recognised that a neutron star *should* in fact rotate with a period of a second or so. As discussed already, if a main sequence star with a rotation period of about a month were shrunk down to the size of a white dwarf – 5,000 to 10,000 km in radius – it would be expected to rotate with a period of an hour or so. If it were shrunk yet further, down to about 15 km in radius, the typical size of neutron star, it would rotate with a period only about 2 billionths as long, only a fraction of a second.

Suppose, therefore, that a pulsar is a rapidly rotating neutron star. What then is to explain the origin of the "hot spot" that is beaming radiation out into space? This beaming is believed associated with the fact that the neutron star also has a very strong magnetic field. As discussed already, theory predicts that white dwarfs should have intense magnetic fields, and these fields have in fact been observed. However, the same theory also predicts that yet smaller neutron stars should have yet stronger magnetic fields, as strong as 1 to 100 billion times as strong as the Sun's magnetic field.

Another important point is that the magnetic North and South Pole of the rotating neutron star need not coincide precisely with the true North and South Pole associated with the star's rotation. Rather, like the Sun, the Earth, and the other planets in the Solar System, one might expect that the magnetic poles would be oriented in a somewhat different direction.

Given these preliminaries, one can now explain how it is that pulsars might be expected to work: A neutron star is of course comprised primarily of neutrons which, being uncharged particles, are relatively insensitive to magnetic effects. However, near the surface there are still a large number of charged protons and electrons. The important point, then, is that if these protons and electrons are in motion or if the magnetic field changes in time, the protons and electrons will feel an intense force that will accelerate them. This acceleration causes the generation of electromagnetic radiation, which is focused primarily along the magnetic poles and subsequently beamed into space. As illustrated in Figure 23.1, because the magnetic poles are not oriented along the true poles, the pulsar then behaves like a lighthouse beacon, sweeping past the Earth every period, emitting the light which we detect.

As of today, literally hundreds of pulsars are known, many associated with leftover debris from supernovae, and it has become apparent that, overall, the oldest pulsars typically rotate the most slowly. This is in fact easy to understand. A pulsar is beaming large amounts of energy into space, but this energy must come from somewhere. What is happening is that the enormous energy associated with the rapid rotation of the neutron star is slowly being converted into a different source of energy, namely the electromagnetic energy that is being radiated into space. The net result is that the period of a typical pulsar grows by a few billionths of a second per day. The amount of energy being beamed into space is in fact absolutely enormous, corresponding typically to a luminosity of about 75,000 $L_\odot$, i.e., 75,000 times the luminosity of the Sun.

Although *most* old pulsars spin more slowly, there are some interesting exceptions, associated invariably with pulsars located in a binary system, the companion being a giant or supergiant or, perhaps, a very young white dwarf or neutron star. As discussed previously, in close binary systems containing a giant or supergiant, there exists the possibility of mass transfer, whereby material from the larger star flows onto the smaller compact object. However, this mass transfer can also entail a transfer of angular momentum from the overall rotation of the binary system to the rotation of the compact object, the net result being that the old neutron star can be "spun up" to a higher rotation rate.

The fastest pulsars known at the present time, all located in close binaries, rotate with a period of 1 or 2 milliseconds. Their existence is important because it provides some useful insights into the behaviour of the matter of which the neutron star is comprised. If even an object as small as a neutron star were to spin too rapidly, it would blow apart, so that there is a maximum rotation rate at which a neutron star can spin. Exactly what that rate is depends theoretically on how dense the neutron star is overall, which in turn depends on the detailed workings of neutron degeneracy pressure in very high density matter. If the neutron star is slightly less dense, gravity is slightly weaker overall, and it is more difficult for the rotating star to stick together. Alternatively, a slightly denser star will have more gravity and thus be able to support a higher rotation rate. The net result is that the observation of very high rotation rates implies a constraint on how squishy the high density matter can be.

This information is in turn of crucial importance to the astrophysicist since it is only by knowing precisely the detailed behaviour of neutron degeneracy pressure in very high density matter that one can determine how large in mass a neutron star can possibly be. It *is* known that, as for white dwarfs, the radius of a neutron star decreases with increasing mass, and that there is an upper mass limit. However, the exact value of this maximum mass, which depends on the properties of the matter inside, is not yet known. The best that can be said definitively is that it must be between about $2M_\odot$ and $5M_\odot$.

At this stage it is useful to compare the overall strength of gravity on the surface of a white dwarf and a neutron star with gravity on a typical main sequence star like the Sun. This is perhaps best done by comparing escape velocities. The escape velocity from the Sun is about 619 km/sec. A white dwarf is only about one hundredth in radius, and thus has an escape velocity ten times as large, about 6000 km/sec. Neutron stars are smaller yet, with a typical escape velocity being as large as 150,000 km/sec. The obvious point then is that this last value is relatively close to the speed of light, about 300,000 km/sec. This means that, if a neutron star were too dense, its gravity would be so strong that, according to Newtonian gravity, its escape velocity should be larger than the speed of light. It follows that, to describe neutron stars correctly, one really needs to invoke general relativity. It also follows that, if the core of a high mass star is too large it cannot end its life as a neutron star and must instead become a black hole.

So what is a black hole? It is an entity predicted by the laws of general relativity which corresponds to an object where gravitational effects are so strong, and the curvature of spacetime is so large, that nothing inside can escape. Thus, in particular, if a hypothetical observer were to approach a black hole, he or she would find that the ordinary notions of space and time start to break down. Indeed, inside the event horizon of the black hole there is a precise mathematical sense in which the ordinary notions of space and time switch roles. The same law of nature which ensures that, under ordinary circumstances, time only moves forward although one can move both forwards and backwards in space, implies that, inside the event horizon, one's distance from the singularity can only *decrease*. What this means is that you, or any other observer, passing through the event horizon would necessarily fall to the singularity. Moreover, you would find that, according to your biological clock, this would happen within a finite period of time. Similarly, once the core of a contracting star becomes smaller than the size of the event horizon of a black hole with the same mass, it has no choice but to continue contracting indefinitely until it shrinks to zero size. No known force in the Universe can halt the contraction.
Interestingly, though, because of the mucked up character of spacetime, an observer outside the hole would never see this happen. As measured from far away, time slows down infinitely much at the event horizon, so that an external observer could never see anything pass through the event horizon. As far as he or she is concerned nothing ever gets there. However, as an infalling object approaches the event horizon very closely any light which it emits will be subjected to an enormous gravitational redshift, with the result that, viewed from far away, the signal becomes longer and longer in wavelength and progressively lower in energy. In this sense, a distant observer would see an infalling object slowly fade out of sight.

An earlier lecture discussed the fact that, viewed from afar, a black hole is characterised completely by three, and only three, numbers, namely its mass, its electric charge (if any), and its angular momentum. Most astronomers believe it extremely unlikely that a black hole arising in nature will have an appreciable amount of charge. If, originally, some black hole had a large net charge, it would exert a huge electric field that would attract oppositely charged particles into the hole, until the hole became very nearly neutral electrically. It thus seems reasonable to restrict attention to black holes for which only the mass and angular momentum are nonzero.

That earlier lecture also provided a simple physical characterisation of the structure of a spherical black hole, with zero angular momentum. However, the preceding discussion of neutron stars would suggest that black holes can spin very quickly. It therefore seems important also to consider the geometry of a rotating black hole with nonzero angular momentum. Basically, the most important new feature arising for such a hole is the existence of a so-called *ergosphere*, located directly outside the event horizon. Physically, this corresponds to a region where space and time have been distorted to the extent that, even though it *is* possible to get back out, it is *not* possible to remain at rest. The curvature of spacetime inside the ergosphere is so strong that it drags any object along a trajectory that orbits about the black hole.

The existence of an ergosphere also implies an interesting possibility, namely that, in principle, one can "mine" energy from a black hole. The basic idea is simple. If some rocket ship (or any other object) were to enter the ergosphere around a black hole, eject a portion of its cargo in an appropriate fashion, and then move far away from the black hole, one would find that, at the end of the day, the rocket will have more energy than the energy with which it began! Basically this happens because the cargo falls into the black hole in such a fashion that, although increasing the rest mass of material in the hole, it decreases its rotation rate. The net result is that energy associated originally with the rotation of the black hole is converted into the excess energy which the rocket carries away.

At this point, one might ask: what about magnetic fields? A neutron star presumably has a magnetic field billions of times stronger than that of the Sun; and one might expect that a black hole should have an even stronger field. Not so! A black hole *can* have an electric field, but its geometry is such that it cannot support a magnetic field. Similarly, a rotating neutron star will have a shape corresponding to a complicated gravitational field that depends on the details of the matter inside it in a fashion that cannot be characterised completely by the total mass and angular momentum. For a black hole where did all this complicated stuff disappear?

In short, as established by the *No-Hair Theorem*, the answer is that the magnetic field and any irregularities in the gravitational field are radiated into space in the form of electromagnetic and gravitational waves. Thus, in particular, one can demonstrate that, according to the laws of general relativity, a dense star core collapsing to form a black hole will, unless exactly axisymmetric, emit a burst of gravitational radiation. This is considered by many astronomers the most likely source for gravitational radiation that could be detected on the Earth. For this reason, most experimental searches for gravitational radiation have focused on the kind of radiation which one would expect to see emitted from a newly forming black hole.

A black hole is, by definition, invisible in the sense that not even light can escape, so how

can one ever hope to detect one? The answer is that, even though *it* is invisible, other visible objects near the hole will feel its gravity and react in an observable fashion that is indicative of the presence of a compact source of gravity. Viewed from afar, the observable effects induced by a neutron star and a black hole are not all that different, so that it may not be easy to distinguish between the two. If, however, one can demonstrate that the compact object is very massive, for example significantly bigger than $5M_{\odot}$, one can rule out the possibility of a neutron star and conclude that the object must be a black hole.
At present, the best environment in which to search for black holes resulting from the death of a high mass stars appears to be in close binary systems containing a giant or supergiant as well as some compact object. The basic idea again involves the potential implications of mass transfer from the large star onto the compact object. If the two stars were not rotating each other, mass from the large star could fall directly onto the small neutron star or black hole, achieving a direct, head-on collision and, in the case of a black hole, plummeting directly into the singularity. The net rotation of the binary prevents this from happening. Rather, because of this rotation gas spilling over onto the compact object will approach the hole with a sense of circulation, causing it instead to settle down into a so-called *accretion disc*, comprised of gas orbiting the compact object along circular, or near-circular, orbits. This is illustrated schematically in Figure 23.2.

A black hole is a tiny object, characterised by enormous gravity, so that an accretion disc involves a lot of material packed into a very small region of space following very high velocity orbits, with speeds approaching the speed of light. However, despite these complications, a relativistic version of Kepler's Third Law remains applicable. This Law, which relates the period of an orbit around the black hole to the mass of the hole and the size of the orbit, implies that material slightly closer to the hole will orbit with a period somewhat shorter than material slightly further out. However, this *differential rotation* means that bits of gas at slightly different radii will scrape against each other, which serves as a source of friction. Like any other source of friction, this causes the conversion of some of the energy of the orbiting material into heat, which is subsequently emitted into space in the form of high temperature electromagnetic radiation, typically x-rays. As the orbiting material loses energy and angular momentum, it slowly spirals in towards the hole. This is analogous to the behaviour of a satellite in the upper portions of the Earth's atmosphere which, because of friction associated with the surrounding air, slowly spirals down towards the Earth. The net result is that old material eventually falls onto the black hole, only to be replaced by new material being supplied from the companion giant or supergiant.
Accretion onto a neutron star or black hole is a very efficient way to generate an enormous amount of energy. The efficiency of nuclear fusion in a typical main sequence star is about 0.00717 in the sense that each little mc^2 of hydrogen converted into helium buys an energy of about $0.00717mc^2$. The efficiency of fusion of helium, carbon, etc., into yet heavier elements is substantially less. By contrast, accretion onto a spherical black hole has an efficiency of about 0.0572. What this means that each nucleon of mass m accreting onto the black hole must radiate into space approximately $0.0572mc^2$ worth of energy, nearly ten times as much as would have been available from fusing hydrogen into helium. For the case of a rotating black hole, the efficiency increases even more, to values approaching 0.2 or more.

X-rays emitted in this fashion are not visible directly from the Earth, because they cannot pass through the Earth's atmosphere unimpeded. However, they *are* easily seen from satellites placed above the atmosphere. The first probe to detect x-rays in outer space was *Uhuru*, which was launched in 1970 and sent back images for about three years. Several subsequent probes have discovered yet more sources. Amongst the x-ray emitters that have been found are a number of binary systems containing a compact object and a large companion. Some of these, notably Hercules X-1 (so named because it was the first x-ray emitted discovered in the constellation Hercules), almost certainly involve accretion onto a neutron star. However, in several cases it is much more likely that the compact object is in fact a black hole. Perhaps the strongest case for a black hole is provided by Cygnus X-1 (the first x-ray emitter discovered in the constellation Cygnus). A detailed analysis of the motions of the stars in this binary indicates that the compact object must have a mass of at least $7M_\odot$, a mass substantially too large to attribute to a neutron star.

In the middle of the 1970's the existence of black holes in nature was still a hotly debated issue. However, by the late 1980's most competent astronomers had come to agree that black holes do indeed exist. Indeed, there is evidence not just for relatively small black holes arising from the death of high mass stars, but also for much larger *supermassive black holes*, $10^6 - 10^9 M_\odot$ or more, that, as will be discussed later, appear to reside at the centers of many galaxies.

Mathematical Appendix: Space-Time Distances Near a Spherical Black Hole

Consider, for simplicity, a spherical black hole located at the spatial point $r = 0$ and restrict attention to spatial motion that is purely radial, i.e., only inwards or outwards without any rotations. Spacetime motion is thus restricted to two dimensions, one space and one time. If the black hole were not present, the notion of distance associated with this two-dimensional motion would satisfy

$$\Delta s^2 = -c^2 \Delta t^2 + \Delta r^2,$$

where r denotes radial distance, and the laws of special relativity would imply that any two observers would agree on the value of Δs^2. The presence of a black hole with mass M so distorts spacetime that the two-dimensional notion of distance becomes

$$\Delta s^2 = -(1 - 2GM/rc^2)c^2 \Delta t^2 + \Delta r^2/(1 - 2GM/rc^2),$$

where G denotes the gravitational constant and radial distance r is so defined that a sphere of radius r has surface area $4\pi r^2$. (Since space is curved, it is not completely obvious how to define distance: unlike the case in flat space, a sphere with surface area $4\pi r^2$ will *not* have volume $4\pi r^3/3$!) Note that, as the size of the black hole increases, i.e., as M gets larger, the quantity $(1 - 2GM/rc^2)c^2\Delta t^2$ gets smaller, this indicative of the fact that the presence of a black hole causes time to slow down. Note also that, for sufficiently small $r < 2GM/c^2$, the quantity $(1 - 2GM/rc^2)$ changes sign so that the coefficient of Δt^2 becomes positive and the coefficient of Δr^2 negative. In this sense, space and time have swapped their usual roles.

.

Lectures 25 and 26

GALAXIES AS LARGE SCALE BUILDING BLOCKS – THE MILKY WAY

One knows today that the Sun is a typical main sequence star, situated in a more or less typical location within a typical, albeit somewhat larger than average, spiral galaxy, the so-called Milky Way. The Milky Way contains roughly 100 to 200 million stars, along with an *interstellar medium* comprised of dust and gas, everything being bound together by gravity. Broadly speaking, as illustrated schematically in Figure 25.1, which views the galaxy both edge-on and from above, the Milky Way is comprised of two components, the disc component and the spheroidal component. The spheroidal component is comprised overall of older, so-called *Population II* stars, and contains almost no dust and gas. The disc component contains significantly younger *Population I* stars, as well as almost all of the dust and gas. All the globular clusters are located in the spheroidal component, whereas the disc contains all the open clusters. As measured across the disc from side to side, the galaxy has a diameter of some 80,000 to 100,000 lightyears. The Sun resides in the disc, about halfway out from the center.

Every fact stated in the preceding paragraph was unknown in 1900: Although scientists at the beginning of the twentieth century had a relatively good understanding of the Solar System, and although they were beginning to appreciate the fact that stars are clumped into various structures, including open and globular clusters, they did *not* know either how stars work or how, overall, they are distributed throughout the Universe. The Ptolomaic cosmology, which dominated Western thought from ancient Rome until the end of the Middle Ages, put the Earth at the center of the Universe, surrounded by the Sun, the planets, and the stars, the last affixed to a rotating sphere in the sky. Copernicus displaced the Earth from its preferred location, inserting the Sun in its stead, but he still failed to recognise the full enormity of the Universe or the role that stars play therein. The idea that we and the stars around us live in a galaxy of finite size was only introduced in the late eighteenth century, and only in the nineteenth century, with work of Kapteyn and the Herschels, was there a concrete, albeit incorrect, attempt to estimate the size and shape of that galaxy and our location therein. The existence of other galaxies had not yet been established: whether or not they exist was an extremely contentious issue.

The cosmology of 1900 was wrong for at least three reasons. (1) The accepted size of the galaxy – 30,000 by 6,000 lightyears – was much smaller than is now known to be true. (2) The Sun was believed, erroneously, to be located at or near the center of the Milky Way. (3) There was no understanding that galaxies outside the Milky Way also exist and, consequently, no sense as to how big the Universe really is.

The first astronomer to determine correctly the relative location of the Sun within the Milky Way, and to recognise that Kapteyn's estimated dimensions are much too small, was the American *Harlow Shapley* (1885 - 1972), a professor at Harvard University. The argument that Shapley used is simple and intuitive; and, when corrected for some technical details, actually yields the answers that we believe today. The starting point for Shapley was the observation that two types of clusters are observed – open and globular – and that they are distributed very differently in the sky. The open clusters are all concentrated in the broad band of the Milky Way, stretching from one side of the sky to the other. Most of the globular clusters are located on one side of the sky, and essentially none are found in the band of the Milky Way. Shapley's basic ideas were (1) that the two different types of clusters live in two different components of the galaxy; and (2) that the globular clusters are distributed in a more or less symmetric fashion about the center of the galaxy. The fact that globulars are not distributed isotropically around the Sun then implies that the Sun is *not* near the center of the galaxy. Mapping out the physical locations of the globular clusters should enable one to determine the Sun's distance from the center of the galaxy as well as estimate its overall size. The only question was: How could one determine the distances from the Earth to the different globular clusters?

To answer this requires a digression. By the early twentieth century it was well known that many red giant stars exhibit systematic variability, whereby their radii, luminosities, and surface temperatures oscillate with a very regular, well-defined period. Different types of variable stars were known, but it was possible to identify one class of objects, the so-called *Cepheid variables*, that seemed to exhibit particularly similar behaviour, albeit with different periods of oscillation. It was therefore natural to ask whether, for such objects, there are any regular patterns relating the period with which the stars oscillate to such properties as their average luminosity. This question was answered in the affirmative in 1912 by the American astronomer *Henrietta Leavitt* (1868 - 1921), a technician at Harvard University who, owing to the sexist attitudes of the time, was never permitted to become a member of the faculty.

What Leavitt did was study a large number of Cepheid variables located within globular clusters in a single distant "nebula," the Small Magellanic Cloud, which is now recognised as a small satellite galaxy of the Milky Way, situated about 100,000 lightyears away from the Earth. Because the overall size of the Magellanic Cloud is much smaller than its distance from the Earth, one can assume that the Cepheids therein are approximately equidistant from the Earth. It follows that, even if one does not know the distance to the Cloud, so that one cannot translate from apparent brightness to absolute brightness, one *can* at least assume that the apparent brightness of each Cepheid is directly proportional to its absolute brightness: The unknown distance to the Cloud should affect the visual appearance of each of the stars identically.

Leavitt demonstrated explicitly that, for the stars which she observed, there was a direct one-to-one correlation between the average apparent brightness and the period with which the brightness oscillates. The fact that the stars were all roughly equidistant then implied a direct relation between the intrinsic brightness, or luminosity, and the period, although the overall scale, or calibration, was not known. To fix the overall scale, one needed to determine the period of oscillation for one or more Cepheids for which the distance *could* be measured.

In 1918 Shapley proceeded to address this remaining problem by studying a few relatively nearby Cepheids whose distance could be determined directly. The net result was a calibrated period-luminosity relation that established a direct correlation between the average luminosity of a Cepheid and its pulsation period. Given this, estimating distances to globular clusters, or any other object containing Cepheid variables, is easy: all that one need measure is the period and apparent brightness of the Cepheids in the object. The

period-luminosity relation implies that, given a knowledge of the period, one can infer the intrinsic brightness, but by comparing the intrinsic and apparent brightnesses of an object one can estimate the distance to the object.
Using this approach, Shapley came to two principal conclusions, namely (1) that the galaxy overall is about 300,000 lightyears across and (2) that the Sun is located about halfway out from the center. The second conclusion is still accepted today, but it is now believed that Shapley's estimated size is about three times too large. Basically, there were two minor flaws in Shapley's chain of logic.

The first flaw was in his assumption that interstellar space is empty. In point of fact, interstellar space contains dust and gas, and that dust and gas acts like a "fog" to obscure objects, making them appear dimmer than they might otherwise be. This is an effect well known from everyday life. On a foggy night, a pair of headlights from an automobile will look much dimmer than they would in the absence of the fog, so that, without correcting for the effects of the fog, one would think the automobile much further away than it actually is.
The second flaw was that there are really two different types of Cepheid variables, which differ in chemical composition and, for this reason, satisfy different period-luminosity relations. The Cepheids which Leavitt analysed are so-called *Population II Cepheids*, located in globular clusters in the Small Magellanic Cloud and characterised by relatively low abundances of elements heavier than helium. The Cepheids which Shapley analysed are *Population I Cepheids*, located near the Sun which have substantially higher abundances of heavy elements. Basically, Shapley had calibrated Leavitt's apples using oranges.
Correcting for these two errors, one is led to a galaxy with a diameter between 80,000 and 100,000 lightyears, i.e., a radius between 40,000 and 50,000 lightyears. The Sun is located in the disc, at a distance estimated to be between 21,000 and 27,000 lightyears from the center of the galaxy. In other words, even neglecting the possibility of other galaxies there is no obvious sense in which the Sun is at or near the center of the Universe. Writing about his discovery in his later years (1969), Shapley had the following to say: "This is a rather nice idea because it means that man is not such a big chicken. He is incidental – my favorite term is 'peripheral'."

As stated already, the disc of the galaxy contains all the open clusters and most of the dust and gas, as well as most of the younger stars. But how thick is it? The answer to this actually depends on the type of star in question. The youngest stars tend to be located very close to the plane of the disc, within about 300 lightyears or so, but the older stars are more spread out, being found as far as 1500 lightyears away from the plane.
One of the more striking features of the disc is the presence of *stellar associations*, transitory collections of very young stars, born from the collapse of a single gas cloud, which will eventually disperse since, unlike open clusters, they are not held together by gravity.
The most striking features observed in other spiral galaxies are *spiral arms*, pin wheel shaped patterns consisting of bright stars, stellar associations, young open clusters, and particularly high concentrations of dust and gas, typically about 5000 lightyears thick. In our own galaxy, these arms are difficult to observe directly, because they are largely obscured by the dust and gas located in the disc. However, as will be described below, one

can get some indications as to the overall shape and structure of these spiral arms using radio waves, and it would seem that they are very much like the spiral arms observed more directly in other galaxies.

In the spheroidal component of the galaxy one can distinguish between two relatively distinct sub-components. One of these is the *galactic halo.* The halo corresponds to a thin sprinkling of old stars, located far from the center of the galaxy. It contains most of the globular clusters and almost no dust or gas. No young stars are observed, and there is no indication that there has been any star formation in the past ten billion years or so. Somewhat over 100 globular clusters have been observed directly. Allowing for the fact that many globulars are obscured by intervening dust and gas, particularly those located on the other side of the galaxy, it is estimated that the galaxy must contain about 200 globulars overall. These globular clusters are comprised of roughly 100,000 to one million stars, closely packed together, with the typical distance between stars being only one tenth as large as in the neighbourhood of the Sun.

The other sub-component of the spheroidal component is the *nuclear bulge*. The bulge corresponds overall to a very dense region of stars, most of them quite old. However, the bulge is not completely devoid of dust and gas, and there is evidence for some relatively young stars, including, perhaps, a few only forming at the present time. Again because of the intervening dust and gas, the bulge is essentially opaque at optical wavelengths. However, it can be probed directly in various other wavelengths, including radio waves, x-rays, gamma rays, and some infrared wavelengths.

The motion of the Sun relative to other stars in the galaxy can be determined directly by measuring the Döppler shifts of various spectral lines in these stars. A systematic analysis of these motions leads to the conclusion that the Sun is orbiting the galactic center along a roughly circular orbit with a speed of approximately 250 km/sec, i.e., upwards of half a million miles per hour. Assuming that the Sun is located about 25,000 lightyears from the center of the galaxy, this implies an orbital period of about 200 million years.

The speed with which the Sun, or any other star, moves through the galaxy depends on its location in the galaxy and on how much mass is pulling on that star. This is completely analogous to the fact that the speed with which the Earth orbits the Sun is determined by its distance from the Sun and the Sun's mass. It follows that, by mapping out the speeds with which different stars are orbiting the center of the galaxy at different distances from the center, one can estimate the total amount of mass that is contained within the galaxy. The motions of the Sun and other nearby stars indicates that there must be between 100 and 150 billion $M_\odot$ of material closer to the center of the galaxy than is the Sun.

Looking at stars further out from the center of the galaxy provides additional information about the distribution of mass in these outer regions. Unfortunately, as one moves further out distance determinations become progressively more difficult and fewer objects are bright enough to be detected, so that the uncertainties in the total mass increase significantly. However, there is a general consensus that the total mass of the galaxy must be between about 100 billion and 1 trillion $M_\odot$, i.e., $1-10\times10^{11}M_\odot$, with $4-8\times10^{11}M_\odot$ perhaps being the most likely range. (It is an interesting fact that the estimated mass of the Milky Way has increased monotonically over the past thirty years or so. Indeed, back in the 1960's estimated masses smaller than $10^{10} M_\odot$ were not uncommon. This large uncertainty reflects two facts, namely (1) that, as discussed below, the amount of matter that is directly visible is considerably smaller than the amount of matter inferred from an analysis of the motions of this visible material and (2) that the Galaxy contains gas clouds and other material that extends out to distances considerably larger than would be inferred by looking only at stars in the disc.)

Significantly, $4 - 8 \times 10^{11} M_\odot$ corresponds to a total mass much larger than what can be accounted for in terms of visible material. Some of the difference can be attributed to faint white dwarfs, diffuse gas clouds, and very low mass stars, which are simply too dim to be seen. However, particularly out towards the edges of the galaxy there seems to be more *missing mass* than can be attributed to such relatively conventional sources. This is one manifestation of the *missing mass problem*. As will be discussed in more detail later on, this missing mass problem has led many cosmologists to speculate that much of the Universe is inundated with some sort of exotic *dark matter*, comprised of something quite distinct from the protons, electrons, and neutrons which appear to be the dominant sources of massive stuff in the Solar System and its environs.

As noted already, the spiral arms within our own galaxy are difficult to detect directly because of the obscuring effects of dust and gas, which make the disc largely opaque at visible wavelengths. However, there is solid reason to believe that they are like the spiral arms in other galaxies, corresponding to a more or less coordinated spiral pattern that propagates through the galaxy at a speed substantially different from the typical speed with which the stars orbit the center of the galaxy.

The fact that these arms contain associations, young open clusters, and clouds of hydrogen gas heated by bright young stars suggests very strongly that they are regions associated with active star formation, i.e., where stars are being born even today. A bright $40 M_\odot$ star has a lifetime of only 1 million years or so, but this means that if it travels at a speed of 250 km/sec, comparable to the speed with which the Sun orbits the center of the galaxy, it will only move a distance of about 1500 lightyears in its entire lifetime. However this distance is considerably smaller than the thickness of a typical spiral arm, so that it follows that such a star will spend its entire lifetime within the spiral arm without ever escaping. Lighter stars born in the spiral arms will eventually exit the arms and, over the course of billions of years, orbit around the galaxy, intermittently passing through the spiral arms again.

Although the distant portions of the spiral arms of our galaxy are completely obscured in visible wavelengths, one *can*, as indicated in the left hand panel of Figure 25.2, observe nearby associations concentrated in three bands, blurring in the distance, which can be interpreted as pieces of more extended spiral arms. To probe the arms at larger distances, one can also exploit the fact that they contain a large amount of cold hydrogen gas that emits a strong radio signal with a wavelength of approximately 21 cm. Unlike visible light, radio waves at this wavelength can pass through the disc relatively unimpeded, which allows one to detect them halfway across the galaxy. On the far side of the galaxy, even the 21 cm line cannot be detected easily. However, the cold hydrogen gas in the disc is often accompanied by carbon monoxide gas which emits a more easily detected 2.6 mm line. Carbon monoxide *can* be detected directly even on the far side of the galaxy and, to the extent that this carbon monoxide traces the spiral arms, one can map out the arms to very large distances from the Sun.

The net conclusion derived by combining the results of these different probes is that our galaxy definitely has four distinct spiral arms which, as illustrated schematically in the right hand panel of Figure 25.2, are relatively intermittent or disjointed.

But what *are* these spirals arms? The first important point is that they do not correspond to some physically connected structure. Just like planets orbiting the Sun, objects orbiting the center of the galaxy have periods that vary significantly with their distance from the center, the more distant stars orbiting with a much longer period. However, this *differential rotation* implies that, if the spiral arms were originally connected physically, they would eventually be stretched out and torn to shreds.

Rather than connected entities, spiral arms appear to correspond to a wavelike phenomenon which propagates through the galaxy much like a sound wave generated when you speak propagates across a room. Indeed, spiral arms are generally interpreted as regions of compressed stars and gas that are circulating around the galaxy.

Another apt analogy is provided by a traffic jam, caused when a truck drives along a highway at 25 miles per hour on a road filled with cars that want to travel at 55 miles per hour (or more). The truck causes a high density of cars directly behind it but, hopefully, the same cars do not remain behind the truck indefinitely. Rather, cars directly behind the truck eventually pass the truck and speed away, only to be replaced by other cars, originally located further behind.

Perhaps the most important point about these *spiral density waves* is that they can trigger star formation. If a cold cloud of gas is compressed by a spiral arm, the balance between pressure and gravity can be upset, causing the cloud to begin to contract, thereby leading to fragmentation and the subsequent formation of stars. As noted already in an earlier lecture, the larger stars will form first, because they have more gravity driving the contraction. These large stars are typically very luminous, emitting large amounts of very hot, ultraviolet radiation; and, especially during the earliest portions of their lifetime, they eject large amounts of gas into interstellar space in the form of a *stellar wind*. However, the ultraviolet radiation and the ejected gas both serve as sources of additional pressure, which can further compress the interstellar medium and trigger further star formation. In addition, the most massive stars will actually explode as supernovae within a million years or so, well before the smaller stars have had a chance to form. This too can act as a source of compression, triggering yet further star formation. The net result is an epoch of so-called *self-sustaining star formation*.

Most astronomers believe that the Sun was made in one of the spiral arms of the galaxy about 5 billion years ago, possibly (as suggested by an examination of meteorite debris) as a direct result of a nearby supernova; and that the Sun has subsequently orbited the galaxy about 25 times, intermittently passing through various spiral arms again. A more controversial idea is that, when the Solar System passes through one of the gas clouds associated with a spiral arm, the amount of sunlight reaching the Earth will decrease, and that this decrease may be responsible for (at least some of) the periods of colder climate to which the Earth has been subjected.

It is completely impossible to probe the nuclear bulge at visible wavelengths, but there *are* various other wavelengths at which light from the bulge can reach the Earth relatively unimpeded.

Three types of radio waves are detectable. A 21 cm line serves as an indicator of cold hydrogen, and a 2.6 mm line indicates the presence of cold carbon monoxide. A third type of radio signal is provide by *synchrotron radiation*, which can only be produced by the acceleration of high energy electrons in a strong magnetic field.

Most infrared radiation is blocked by dust and gas in the disc, but there *are* two wavelength bands in the infrared for which light can travel through the galaxy relatively unimpeded. One band, centered near 2μm (1μm $= 10^{-6}$ meter), corresponds via Wien's law to a temperature of about 1500 K, and thus serves as a diagnostic for the presence of cool red giant stars. The other, centred near 10μm, corresponds to a temperature of about 300 K, and thus serves as an indicator of dust and/or planetary material heated by nearby starlight.

In addition, it is possible to detect x-rays and gamma rays, both of which are indicative of high energies and hence very high temperatures, millions of degrees Kelvin or more.

By piecing together signals detected at these wavelengths it is possible to obtain at least a rough sense as to the contents of the nuclear bulge. Moving from the Sun towards the center of the galaxy, about 17,000 to 12,000 lightyears out one detects strong emission at 2.6 mm which suggests the presence of a ring of cold gas clouds. Inside this ring, one observes infrared radiation indicative of a dense collection of old red giants, together with small quantities of dust and gas. Then, about 9,000 to 10,000 lightyears out one detects a 21 cm line, indicative of a ring of hydrogen gas. The portions of the ring closest to the Sun are significantly blueshifted, indicating that the material is approaching the Sun, whereas the portions on the opposite side of the galaxy are significantly redshifted, this indicating that they are moving away. The interpretation of the observed Döppler shifts is that one is seeing a rapidly expanding ring, which is moving out from the galactic center with a speed of approximately 100 km/sec. Given the observed expansion, and estimating the rate at which gravity would slow it down, one infers that this ring must have been produced by some gargantuan explosion near the center of the galaxy about 30 million years ago.

Inside this expanding ring, one observed synchrotron radiation of the form that one might expect to arise from a single gigantic object. Then between about 1600 and 1000 lightyears from the center one observes another ring of high temperature, ionised gas, tipped significantly with respect to the plane of the disc and rotating at a speed of about 200 km/sec. Inside this second ring, the density of stars is extremely high, over a million times as dense as the environment near the Sun, so that the typical distance between stars is only about 4 light*days* (recall that the nearest stars to the Sun are over 4 light*years* away!). Then, at or very near the center, in a tiny region no more than a fraction of an *AU* or so across, there is a strong source that is emitting copious amount of radiation with (at least) every wavelength that is possible to detect: radio, infrared, x-rays, and gamma rays.

The high speed of the rotating ring and, especially, the high speeds observed for the stars inside the ring, indicate that there must be an enormous amount of gravity acting in this region. If a strong source of gravity were not serving to attract the stars and the ring towards the center, their high velocities would allow them to escape to the outer parts of the galaxy. Some, perhaps all, of this gravity can be associated with the high density central concentration of stars. However, over the last decade or so there has emerged a growing belief on the part of many astrophysicists that the center of the galaxy contains a supermassive black hole. The only real question is: How massive is this black hole? In the late 1980's and early 1990's, many astronomers assumed that the hole would be as large as 3 to 5 million $M_\odot$, so that it would contribute significantly to the total mass of the galactic center. However, others argued that the black hole should be much smaller, with a mass of only a few thousand $M_\odot$ at most, claiming that otherwise the hole would be much more conspicuous than it is. Only in the late 1990's did high resolution imaging yield reasonably conclusive evidence that the black hole really is quite large. As of Summer 1999, the best estimate for the total mass was about $2.6 \times 10^6 M_\odot$, with an uncertainty of about 15%.

The presence of a supermassive black hole with a mass as large as 3 million $M_\odot$ or so would be useful in explaining the high velocities that are observed since it would provide a strong source of gravity. However, even more importantly the black hole would have the effect of attracting nearby material further in towards the center, thereby helping to account for the high density of stars that appear to be there. Short of postulating the presence of some central object it is difficult to explain just how the center of the galaxy came to be so dense.

The presence of a supermassive black hole, either big or small, would also lead naturally to an explanation of the strong emission of electromagnetic radiation which has been observed. Basically, the idea here is that one is again seeing the effects of accretion in a scaled-up version of what happens in a compact binary system. The presence of the black hole will serve to attract any gas or stars that stray nearby, and, once a star gets too close, the tidal forces associated with the hole will tear it to shreds, converting it back into unbound gas. However, just as for the case of a binary, the gas cannot fall directly onto the black hole. Because of the overall rotation of the galaxy, the infalling gas will have a net circulation, so that it will instead settle down into an accretion disc, before slowly spiralling into the hole.

The net result is that one would expect radiation emitted from a region of linear size maybe 5 to 10 times the Schwarzschild radius, R_{sch}. For a 3 million $M_\odot$ black hole, $R_{sch} \approx 9$ million km, or about $0.06AU$, so that most of the observed radiation should come from a region smaller than $1AU$ in extent. It is also important to stress once again that accretion onto a black hole is an extremely efficient way in which to generate a large amount of energy. Thus, e.g., a mass infall rate of $10^{-6} M_\odot$ per year, only one $M_\odot$ every million years, would, for the case of a spherical black hole, generate a luminosity of about 1 million $L_\odot$, a number that is independent of the mass of the putative black hole.

But how did this black hole get there in the first place? Back in the 1970's, many astrophysicists believed that an initially dense collection of stars would somehow evolve relatively quickly to such a high density that it would end up collapsing to form a large black hole with a mass of one million $M_\odot$ or more. However, detailed numerical modeling performed since then suggests that this may not be the case, although it *is* relatively easy to concoct formation scenaria for black holes as small as a few thousand $M_\odot$. In part for this reason, many astrophysicists now believe instead that any large supermassive black hole in this, or any other galaxy, must have formed very early in the history of the Universe, before galaxies even existed. Indeed, they would argue further that these primordial black holes could have played a role in the original formation of galaxies. This latter alternative is consistent with the idea that quasars are very young galaxies containing supermassive black holes with masses in the range $10^6 - 10^9 M_\odot$.

As is summarised in Figure 25.3, there is a simple progression in age manifested by stars in different parts of the galaxy. The oldest stars are in the halo, next oldest in the bulge, and then in the disc as a whole. The stars located in the spiral arms are all relatively young, 100 million years or less. The ages of these stars also correlate with the relative abundances of elements heavier than hydrogen and helium, what the astronomer sloppily terms *metals*: older stars have very low abundances of elements whereas younger stars have much higher abundances. This is understood as reflecting the fact that, early on, the Universe was comprised almost totally of hydrogen and helium, and that the heavier metals were only formed later on in the late stages of stellar evolution. The abundance of gas correlates with the typical age of the stars in that, without a ready supply of gas, it is imposssible to make any new stars.
Before the 1960's the fact that younger stars, particularly those in the disc, have nearly circular orbits, and that older stars have orbits more irregular in shape, was interpreted in a number of different ways. However, in the past thirty years or so there has emerged a simple and compelling picture which explains simultaneously both this and the aforementioned fact that older disc stars are more spread out above and below the disc than are younger stars. Specifically, it is now generally accepted that disc stars are originally manufactured at or near the plane of the disc in nearly circular orbits, but that, over the course of their lifetimes, they are subjected to occasional gravitational encounters with nearby gas clouds that make the orbits more irregular in shape and induce larger excursions from the plane.

These observed facts provide important clues as to how our galaxy formed and evolved, which will be considered later on in the context of galaxies in general.

.

	Population I		Population II	
	extreme	intermediate	intermediate	extreme
Location	spiral arms	disc overall	nuclear bulge	halo
Avg. Age (yr)	$<$ 100 million	0.2 - 10 billion	2 - 10 billion	10 - 16 billion
Avg. % Metals	3%	1.6%	0.8%	<0.8%
gas?	lots	some	a little	almost none
Orbit shape	circular	slightly irregular	moderately irregular	very irregular

Figure 25.3

STUDY GUIDE FOR LECTURES 14 - 16

Words and names that you should recognise and understand:

wave	photon	quantum mechanics
Döppler effect	redshift	blueshift
blackbody	Boltzmann's Law	Wien's Law
blackbody spectrum	luminosity	atomists
elements	chemical compounds	atoms
molecules	nucleus	electrons
protons	neutrons	isotopes
Planck's Law	collisions	excited state
ground state	ionized	Kirchoff's Laws
continuous spectrum	absorption spectrum	emission spectrum
energy of a photon	Newton's particle energy	Einstein's particle energy
rest mass energy	nuclear fission	nuclear fusion
neutrinos	Four Interactions	gravity
electromagnetism	strong interaction	weak interaction
Standard Model	fermions	quarks
leptons	bosons	quantum electrodynamics
electro-weak theory	quantum chromodynamics	three generations
anti-particle	anti-quark	anti-lepton
up quark	down quark	strange quark
charm quark	top quark	bottom quark
electron	electron neutrino	muon
muon neutrino	tau	tau neutrino
fractional charge	W and Z bosons	gluons
Niels Bohr	Uncertainty Principle	Pauli Exclusion Principle
degeneracy pressure	Werner Heisenberg	Erwin Schrödinger
P. A. M. Dirac	positrons	vacuum
virtual pairs	particle creation	matter-antimatter annihilation

Recommended Questions:

1. Explain physically what is meant by the Döppler effect, and how it works.
2. Why do you think that the Döppler effect is important in astronomy and cosmology?
3. What precisely is meant by a blackbody?
4. Explain the physical content of Boltzmann's Law.
5. Explain the physical content of Wien's Law.
6. Describe the content and implications of the graph presented in Figure 14.2.
7. Why does the Sun look yellow?
8. How would the Sun change in visual appearance if it were to be the same size but have a surface temperature of about 2900 K, about half its present value?
9. Upon what two quantities do the total luminosity of a spherical blackbody depend?
10. Describe the makeup of an atom, including the relative sizes and masses of the different constituents.

11. Under what circumstances will atoms emit photons of light?
12. Explain the physical content of Planck's Law.
13. In what sense can one say that: "heat is motion"?
14. Explain the physical ingredients that determine whether a planet or moon can hang onto an atmosphere.
15. Explain carefully how it is that atoms come to emit photons of light with special well-defined frequencies.
16. Explain qualitatively how photons emitted with well-defined frequencies are subsequently transformed into a (near-)blackbody spectrum.
17. State Kirchoff's Laws and explain why they are true.
18. How does Einstein's notion of energy differ from the notion of energy provided by Newton?
19. Explain roughly the idea underlying nuclear fission.
20. Explain roughly the idea underlying nuclear fusion.
21. Denumerate and describe the four fundamental interactions known to physicists.
22. Describe the particle content of the so-called Standard Model. Include the role of both fermions and bosons, both quarks and leptons, and the three known generations.
23. What are neutrinos?
24. State what is meant by the Exclusion Principle, and describe the physical interpretation of the Uncertainty Principle provided by Bohr.
25. State what is meant by the Exclusion Principle and degeneracy pressure.
26. What is meant by anti-particles?
27. In what sense is the vacuum not really empty space?
28. Explain what is meant by particle creation in an electromagnetic or gravitational field.
29. How do we know that, in our part of the Universe, the total electric charge is almost exactly zero?
30. How do we know that, in our part of the Universe, there are not equal amounts of matter and antimatter?

STUDY GUIDE FOR LECTURES 17 - 19

Words and names that you should recognise and understand:

Equivalence Principle	gravitational mass	inertial mass
Lorand von Eötvös	tidal forces	flat spacetime
curved spacetime	positive curvature	negative curvature
great circle	geodesic	gravitational mass-energy
light bending	Arthur Stanley Eddington	gravitational lens
precession of perehelion	Simon Newcomb	gravitational redshift
R. V. Pound	G. A. Rebka	gravitational radiation
binary pulsar	Joseph Taylor	Russell Hulse
laser interferometry	black holes	Singularity Theorems
Robert Geroch	Stephen Hawking	Roger Penrose
Black Hole No-Hair Theorem	event horizon	singularity
Schwarzschild radius	Roche limit	

Recommended Questions:

1. Explain why it is true that gravity satisfies the Equivalence Principle but that electricity does not.
2. What is meant by gravitational mass? by inertial mass? In terms of gravitational and inertial masses, what does the Equivalence Principle say?
3. In terms of uniform gravitational fields and uniform accelerations, what does the Equivalence Principle say?
4. In what sense can one say that uniform gravitational fields are not fundamental? In what sense are nonuniform gravitational fields, associated with tidal forces, fundamental?
5. Explain the origins of tides on the Earth.
6. What is meant by the statement that mass and energy curve otherwise flat spacetime?
7. Identify three attributes of a flat space or spacetime that are not true about curved spaces and spacetimes.
8. What precisely is meant by a geodesic?
9. How does an object with energy move in a spacetime that is curved?
10. What is meant by light bending in a gravitational field?
11. How did the expedition led by Eddington confirm the existence of light bending for starlight passing near the Sun?
12 What is a gravitational lens, and how does it work? Describe two different things that a gravitational lens can do.
13. What is meant by the precession of the perehelion of Mercury? What are two Newtonian sources that contribute to this precession? What is the origin of the remaining relativistic contribution?
14. What is meant by gravitational redshift and how does it work? Has it been measured directly on the Earth?

15. In what sense does the analysis of the binary pulsar discovered by Joseph Taylor and Russell Hulse provide an indirect confirmation of the existence of gravitational radiation?
16. How do some relativists intend to use laser interferometry to try to detect gravitational radiation directly?
17. How is it that combining Newtonian gravity and special relativity leads to the predicted existence of black holes?
18. What is meant by the Black Hole Singularity Theorems?
19. What is meant by the Black Hole No-Hair Theorem?
20. How in principle can one measure the mass, charge, and angular momentum of a black hole?
21. What is meant by the event horizon of a spherical black hole? The singularity of the black hole?
22. What physically is the Schwarzschild radius of a spherical black hole? How large is it in kilometers?
23. What would you feel as you fell towards a black hole?
24. What is meant by the Roche limit of a planet?

STUDY GUIDE FOR LECTURES 20 - 22

Words and names that you should recognise and understand:

luminosity	blackbody spectrum	method of parallax
Ejnar Hertzsprung	Henry Norris Russell	Hertzsprung-Russell (HR) diagram
main sequence star	(red) giant	supergiant
white dwarf	mass-luminosity relation	spectroscopic parallax
Lord Kelvin	Robert Emden	James Homer Lane
nuclear fusion	Hans Bethe	hydrogen core burning
hydrostatic equilibrium	conservation of mass	conservation of energy
energy transport	hydrogen shell burning	helium core burning
helium shell burning	electron degeneracy pressure	planetary nebula
carbon burning	neon burning	oxygen burning
silicon burning	iron burning	neutron degeneracy pressure
neutron star	Type I supernova	Type II supernova

Recommended Questions:

1. How can an astronomer measure directly the physical distance to about 1000 relatively nearby stars?
2. How can an astronomer determine the surface temperature T of a star?
3. How can an astronomer determine the luminosity L of a star that is sufficiently close that its distance d can be measured directly?
4. How can an astronomer infer the radius R of a star sufficiently close that its distance can be measured directly?
5. Under what circumstances is it possible to estimate directly the mass m of a star?
6. Draw a Hertzsprung-Russell (or *HR*) diagram. On this diagram, indicate the locations of the four different types of stars: main sequence, (red) giant, supergiant, and white dwarf. For the main sequence stars, where are the more massive stars located? the less massive stars?
7. What is meant by the mass-luminosity relation?
8. About how large, relative to the Sun, are typical main sequence stars? giants? supergiants? white dwarfs?
9. How much variability in radius is exhibited by different types of star? variability in luminosity? variability in mass?
10. Explain the logic whereby one can use the *H-R* diagram to estimate the distances to stars sufficiently far away that their distances cannot be measured directly.
11. In what sense can one say that a star exists as an equilibrium between gravity and pressure?
12. Why physically is it that nuclear fusion only happens at very high temperatures? Why is it that nuclear fusion into progressively heavier elements requires increasingly higher temperatures?
13. What two things happen as a low density cloud of gas contracts down to a smaller physical size?

14. Why does the time required for a fragment of a gas cloud to evolve into a main sequence star depend sensitively on the size of the cloud?
15. What is the source of pressure in the core of a main sequence star like the Sun?
16. How does an astronomer go about making predictions as to what the innards of a star are like? What are the basic rules which he or she uses?
17. How physically does stellar modeling account for the mass-luminosity relation?
18. How does a main sequence star change as it ages?
19. Explain the physical mechanism responsible for the onset of hydrogen shell burning? How does hydrogen shell burning transform a main sequence star into a red giant?
20. What is the principal source of pressure in the core of a red giant?
21. What is the physical mechanism causing a large red giant to bloat into a yet larger supergiant?
22. What is the principal source of pressure in a white dwarf? Why is it that a white dwarf is very small in physical size, but, typically, very hot?
23. Why is the formation of heavy elements in high mass stars important for humans?
24. Why is it that fusion never proceeds beyond iron?
25. Why is it that successive stages in the life of a high mass star are progressively shorter in duration?
26. Why is a *neutron star* aptly named?
27. What is the principal source of pressure in a neutron star?
28. What do we see when we look at a Type I supernova? a Type II supernova?
29. Explain how the death of a high mass star results in a Type II supernova.
30. Explain the workings of a Type I supernova. Why do they always arise in a binary system?

STUDY GUIDE FOR LECTURES 23 - 26

Words and names that you should recognise and understand:

white dwarf	angular momentum	Zeeman effect
Chandrasekhar limit	S. Chandrasehar	Lev D. Landau
Richard Fowler	Fritz Zwicky	Walter Baade
neutron star	Jocelyn Bell	pulsar
escape velocity	black hole	event horizon
ergosphere	No-Hair Theorem	accretion disc
differential rotation	Uhuru	Cygnus X-1
Hercules X-1	supermassive black holes	interstellar medium
Population I stars	Population II stars	disc component
spheroidal component	Harlow Shapley	Cepheid variables
Henrietta Leavitt	period-luminosity relation	stellar associations
open clusters	globular clusters	spiral arms
galactic halo	nuclear bulge	missing mass problem
dark matter	21 centimeter line	spiral density waves
stellar wind	self-sustaining star formation	synchrotron radiation

Recommended Questions:

1. How abundant do astronomers believe white dwarfs to be?
2. How dense are white dwarfs? How large in radius are they?
3. Why physically do white dwarfs rotate with a very short period, an hour or less?
4. Why physically do white dwarfs have very large magnetic fields?
5. What is the actual evidence that white dwarfs really do have large magnetic fields, a million times the magnetic field of the Sun, or even larger?
6. What is the source of pressure inside a white dwarf?
7. What physically is the Chandrasekhar limit? How large is it relative to the mass of the Sun?
8. How dense are neutron stars? How large in radius are they?
9. Describe, carefully and systematically, the line of reasoning leading to the conclusion that pulsars are rotating neutron stars.
10. Why is it that pulsars cannot be rotating white dwarfs?
11. What is responsible for the "hot spot" on a neutron star that beams radiation into space?
12. Where does the energy that a pulsar radiates into space come from?
13. What is the physical origin of so-called *millisecond pulsars*, which rotate with a period as short as about a millisecond, i.e., a thousandth of a second?
14. Why is the existence of millisecond pulsars important for scientists interested in understanding the workings of high density nuclear matter?
15. What is the upper mass limit for a neutron star?
16. In what sense do space and time switch roles inside a black hole?
17. In what sense is true that, as far as an external observer is concerned, black holes never form?

18. What physically is the ergosphere of a rotating black hole?
19. How in principle can one "mine" energy from a rotating black hole?
20. Why is it that black holes have no magnetic fields? Where have they gone?
21. Given that a black hole is by definition invisible, how can one detect its presence?
22. What is an accretion disc. How can it arise in a close binary system containing a giant/supergiant and some compact object, either a neutron star or a black hole?
23. What does the astronomer mean by saying that accretion onto a black hole is a more efficient source of energy generation than nuclear fusion?
24. What is Cygnus X-1? Hercules X-1?
25. What is the interstellar medium?
26. What are Population I and Population II stars?
27. In what senses was the cosmology of 1900 incorrect from a modern standpoint?
28. What are Cepheid variables? What is meant by the period-luminosity relation for Cepheid variables?
29. How did Henrietta Leavitt establish the existence of this period-luminosity relation?
30. How did Harlow Shapley calibrate Leavitt's period-luminosity relation?
31. What were Shapley's principal conclusions regarding the size and shape of the Milky Way galaxy?
32. What were the two mistakes made by Shapley that led him to an erroneous estimate of the true size of the Milky Way?
33. What today is believed to be the overall diameter of the Milky Way?
34. Where in the Milky Way is the Sun located?
35. How thick is the disc component of the Milky Way? How does this depend on what type of stars one looks at?
36. What are stellar associations?
37. What are spiral arms? Why is it easier to study spiral arms in other galaxies than the arms of the Milky Way?
38. What are the two principal components of the spheroidal component of the Milky Way?
39. Overall what is the galactic halo like?
40. Overall what is the nuclear bulge like?
41. How fast is the Sun rotating around the center of the Milky Way? How was this determined?
42. How massive do astronomers believe the Milky Way to be? How are their estimates obtained?
43. What is meant by the missing mass problem?
44. What is the evidence suggesting that spiral arms are regions of active star formation?
45. How are hydrogen gas and carbon monoxide used to map out the spiral arms of the Milky Way?
46. What physically are the spiral arms?
47. What is meant by the theory of self-sustaining star formation?
48. Why is it impossible to probe the center of the Milky Way at visible wavelengths?
49. What wavelengths of light can be used to probe the center of the Milky Way?

50. Describe the physical contents of the nuclear bulge, as one moves from the Sun towards the center of the Milky Way.
51. What do many astronomers believe that there is a supermassive black hole located at the center of the Milky Way?
52. What is the mechanism whereby such a supermassive black hole could generated the observed radiation emitted from the galactic nucleus?
53. Study the patterns summarised in Figure 25.3, which connect the ages and properties of stars in different parts of the Milky Way.

Lectures 27 and 28

GALAXIES AS LARGE SCALE BUILDING BLOCKS – OTHER GALAXIES

Although some astronomers like William Parsons had argued since 1845 that so-called *spiral nebulae* are distinct galaxies, outside the Milky Way, the existence of other galaxies was not proven definitively until 1924. By 1920 the debate as to whether or not other galaxies actually exist had become heated and acrimonious, two of the most prominent figures in the controversy being Harlow Shapley and *Edwin Hubble* (1889 - 1953), a Kentucky lawyer turned astronomer. Shapley, who had gained enormous prestige after his pioneering work in determining the size and shape of the Milky Way, was vociferously opposed to the idea that other galaxies exist, arguing that the spiral nebulae were in fact objects inside the Milky Way. Hubble, not yet as distinguished, argued equally vociferously for the existence of other "island Universes."

The earliest concrete, albeit indirect, evidence that these nebulae really are outside the galaxy was provided by the American astronomer *Vesto M. Slipher* who, starting in 1914, began using the Döppler effect to determine the velocities with which the brighter, and hence presumably more nearby, nebulae were moving relative to us. His surprising conclusion, established more and more definitively during the next ten years, was that the majority of these objects are in fact moving away at very high speeds. Thus, e.g., his original sample showed that 11 of 15 nebulae were moving away in completely different directions at velocities averaging about 300 km/sec.

This, however, seemed difficult to reconcile with the idea that these nebulae are objects within the Milky Way. An object moving with such a speed would exit the Milky Way in well under a billion years (recall that the Sun, halfway out from the center, has a speed of about 250 km/sec and orbits the galaxy once every 200 million years or so). However, by this time geologists were confident that the Earth must be at least a billion years old, so that, if these nebulae were in fact in the galaxy, they must somehow have been propelled outwards from the center well after the Earth had formed.

The first theoretical argument that the so-called Andromeda nebula is actually a distinct galaxy was provided by *E. Öpik* in 1922. The key to Öpik's argument was the recognition that the apparent size of this nebula in the sky and its apparent brightness both depend sensitively on how far away it is: a distant object looks both dimmer and smaller. His critical assumption was that Andromeda and the Milky Way are made of exactly the same stuff, in the sense that they have the same luminosity per unit mass, i.e., on the average each bit of mass in Andromeda and the Milky Way gives off the same amount of light. What Öpik showed was that, granted this assumption, one can estimate the distance r from the Sun to Andromeda in terms of (1) Andromeda's angular size in the sky, (2) its apparent brightness, and (3) the typical speed with which stars orbit in Andromeda, all of which were well known. His net conclusion was that Andromeda is about 1.4 million light years away, clearly outside the Milky Way. However, if it is that far away, Andromeda must be very large in linear dimension to look as big as it does, which implies in turn that it must be comparable in size to the Milky Way.

Both these arguments were unsettling, but neither was really conclusive. The conclusive proof that Andromeda is a galaxy in its own right, outside the Milky Way, was provided by the discovery of Cepheid variables in Andromeda. Interestingly, Cepheids in Andromeda were first discovered in 1918 by Shapley, who must certainly have recognised their fundamental significance. However, he did not publish his results (which would have proved conclusively that his objection to the existence of other galaxies was wrong). Five years later, in October 1923, Hubble rediscovered these Cepheids in Andromeda. Cepheid variables were known to be extremely bright intrinsically, about $15,000 L_\odot$. These, however, looked exceedingly dim, so that they must be very far away. Over the next several months Hubble tracked their oscillations in apparent brightness and, by December, using Leavitt's period-luminosity relation and Shapley's calibration, concluded that Andromeda must be approximately 900,000 to 1 million lightyears away from the Sun. When the results were reported at a meeting of the National Academy of Sciences in January 1924, the existence of other galaxies was generally accepted as an unavoidable reality.

In point of fact, both Öpik's and Hubble's numbers underestimate the distance to Andromeda. The best estimate today is that the distance from the Sun to Andromeda is approximately 2.4 million lightyears, with an uncertainty of about 0.1 million lightyears. Öpik's estimate was extremely good, considering the very simple argument that he used. Hubble's estimate was more seriously flawed, again because of problems with Shapley's calibration of the period-luminosity relation.

With the discovery that other galaxies exist, the Sun had once more been displaced, this time even more fundamentally than by being moved away from the center of our galaxy. It became recognised quickly that millions, if not billions, of galaxies must exist; and that, in terms of its bulk appearance, the Milky Way is not particularly special. In short order, this led much of the astronomical community to embrace an anti-anthropogenic principle, which holds that, overall, the general physical properties of our part of the Universe are not special in any way. As will be seen in a later lecture, this played an important role in the original construction of the Hot Big Bang Model, as it began to be developed in the 1920's.

Broadly speaking, three different types of galaxies are observed, namely *spiral galaxies* (and *barred spirals*), *elliptical galaxies*, and *irregular galaxies*. One also identifies a few so-called *lenticular galaxies* (sometimes called S0's), which seem intermediate in appearance between spirals and ellipticals.

Spiral galaxies are all like the Milky Way in that they contain both a disc and a spheroidal component. The former invariably includes two or more spiral arms – two or three being the norm –; the latter includes both a halo and a nuclear bulge. However, the detailed shape can vary significantly from one spiral galaxy to another. In some galaxies the spiral arms are tightly wound and extremely regular in appearance, while in others they appear to be more fragmented. In some galaxies the bulge is extremely thick and pronounced, while in others it is very thin. An interesting but not understood fact is that there is a strong correlation between the overall shape of the arms and the bulge. Tight arms correlate with a fat bulge, whereas flacid arms correlate with a wimpy bulge. For this reason, astronomers make a morphological distinction between three types of spiral galaxies: *Sa*'s, with tight arms and a fat bulge, *Sc*'s with loose arms and a wimpy bulge, and *Sb*'s, which are intermediate between these two extremes. The Milky Way and Andromeda are both of this intermediate type.

The disc component of a spiral galaxy typically has a net sense of rotation in that the vast majority of stars orbit the center in the same direction. Significantly, however, the spheroidal component does not have a strong net circulation. It appears that, in both the halo and the bulge, there are roughly the same number of stars orbiting in each direction. The origin of this distinction is not known.

Barred spiral galaxies are like ordinary spirals except for the fact that (some of) the spiral arms are attached to a bar extending out from the nucleus, rather than extending from the nucleus itself. They exhibit the same correlations between arms and bulges as do the ordinary spirals and, for this reason, are classified as *SBa*'s, *SBb*'s, and *SBc*'s.

It is not yet completely clear exactly what causes the existence of bars. Until the 1970's it was generally believed that barred and unbarred spiral galaxies are distinctly different sorts of objects. More recently, however, largely as a result of detailed numerical investigations, most astronomers have come to think instead that bars are probably transitory phenomena that occur in ordinary spirals. In other words, an ordinary spiral will develop a bar, the bar will then dissipate, and, perhaps, the bar will subsequently reform later on. In the early 1990's this interpretation was buttressed by a number of observations, including some from the Hubble Space Telescope, that suggest that even many allegedly unbarred spirals show indications of having a little "barlet." In other words, it may be that the distinction between barred and unbarred is one of degree rather than of kind.

Even our own Galaxy has a small bar. By 1998, it was determined that the Milky Way, which had been classified as an ordinary *Sb* spiral, actually contains a stubby bar which extends out about 10000 lightyears from the center of the galaxy (i.e., total length of about 20000 lightyears), is between 8000 and 12000 lightyears thick (in the disc), and between 6000 and 8000 lightyears high (perpendicular to the disc).

Viewed in the sky an elliptical galaxy has a two-dimensional shape that corresponds approximately to that of an ellipse. Some ellipticals look nearly circular, whereas others look distinctly flattened. This has led to a morphological classification into types *E0* through *E7*, with *E0* the most circular and *E7* most squashed. (The number after the E is related to the so-called ellipticity of the ellipse that best fits the shape. An En galaxy is characterised by the fact that the ratio of the semi-major and semi-minor axes, a and b, satisfies $n = 10[1 - (b/a)]$.) However, it is important to recognise that this classification says very little about the full three-dimensional shape. Rather, the visual appearance is dominated by so-called *projection effects* which reflect the direction from which we are looking at the galaxy. Thus, *a priori*, the true three-dimensional shape could correspond to an *oblate spheroid*, i.e., a flattened sphere (roughly like a pumpkin), a *prolate spheroid*, i.e., a stretched sphere (roughly like a football), or something more genuinely three-dimensional in the sense that it looks different in each direction.

Astronomers like to think that objects in nature have simple shapes and, for this reason, would like to believe that ellipticals are relatively simple axisymmetric objects. However, over the past fifteen years or so, detailed examinations of the distribution of light and of Döppler shifts of stars in elliptical galaxies have forced them to the conclusion that, rather than being axisymmetric, many may in fact be genuinely three-dimensional. (It is instructive to note that many objects originally thought to be nearly axisymmetric turn out to be decided three-dimensional in shape when probed more carefully. For example, recent observations suggest that, rather than being an oblate spheroid, the spheroidal component of the Milky Way is genuinely triaxial, so that, in the plane of the disc, it is some 20% thicker in one direction than in a perpendicular direction.)

In this regard, another important point needs to be stressed. Even if some elliptical galaxy is extremely flat, this is *not* because it is rotating very quickly. A blob of gas that is rotating quickly will feel a strong outward pull that tends to make it flatter in shape (which, incidentally, explains the fact that Jupiter and Saturn both deviate significantly from a completely spherical shape). For this reason, many astronomers originally believed that very flat elliptical galaxies must be rotating extremely quickly. This, however, is false. A detailed investigation of Döppler shifts exhibited by the emitted light demonstrates conclusively that, contrary to naive expectation, elliptical galaxies usually lack a strong net sense of rotation. In this sense, they are rather similar to the spheroidal component of a spiral galaxy, from which the disc has been torn away.

Irregular galaxies are defined by the fact that they do not look like either an elliptical or a spiral. They exhibit a significant degree of diversity, but are similar overall in that they look manifestly more irregular in appearance and that they never have anything approaching spiral arms. Lenticulars are galaxies that look intermediate between ellipticals and spirals. Like spirals, they have both a spheroidal component and a disc. However, the distinction between the two components is less clear cut than for a spiral and, unlike spiral galaxies, they never have spiral arms.

Overall spiral galaxies tend to be larger than elliptical and irregular galaxies. Their discs invariably contain large clouds of dust and gas, and they contain both old and new stars, typically including stars that are only now being born. Because of the clouds of dust and gas, which are easily heated to a bright glow, they tend to be brighter than elliptical galaxies of comparable size, which contain few, if any, gas clouds.

Ellipticals come in a huge range of sizes and masses. The so-called *giant ellipticals* are the largest known galaxies, whereas the *dwarf ellipticals* are the smallest, some being no larger than a globular cluster. Indeed, many astronomers would argue that, morphologically, there is no real difference between dwarf ellipticals and globular clusters, and that one should view globulars as tiny ellipticals living inside a larger galaxy.

Unlike spiral galaxies, which contain both younger Population I in the disc component and older Population II stars in the spheroidal component, ellipticals are observed to contain only Population II stars. Until the last few years, it was also believed that ellipticals typically contain little or no dust or gas, and that this fact correlates with the absence of young stars. More recently, however, it has been recognised that ellipticals do contain a goodly amount of gas. The big difference between spirals and ellipticals seems to be that, in the latter, the gas is too hot to congeal into clouds and exists instead as a low density entity filling the entire galaxy.

Irregular galaxies tend to be somewhat smaller overall. They contain lots of gas, but variable amounts of dust, as well as both old and young stars. Lenticulars too tend to be relatively small and, like ellipticals, are almost completely devoid of dust and gas clouds. This latter fact supports the theoretical expectation that spiral arms necessarily involve an interplay between gas and stars: without the presence of both stars and gas clouds, it should not be possible to generate spiral density waves.

How will the visual appearance of the stars within a galaxy change over the course of time? The answer to this question depends critically on when the stars were made. Suppose, for example, that all the gas in some galaxy was converted into stars at the same time, say 15 billion years ago, and that no new stars were ever produced thereafter. Immediately after the initial epoch of creation, the galaxy would contain a mixture of stars, presumably with a broad range of masses and hence a broad range of temperatures and colours. However, the lifetime of a star decreases very rapidly with increasing mass. This means that, over the course of time, the more massive stars will burn out and die, even though the lower mass stars have hardly begun their lives. After a million years the $40M_\odot$ stars will all have died, after 7 million years the $25M_\odot$ stars will be gone, and so on, but only after a period of 10 billion years or so will a $1M_\odot$ star have died. What this implies is that, systematically, the hotter stars are being extinguished. However these hotter stars emit shorter wavelength radiation, peaked towards the blue end of the visible spectrum. This implies in turn that, as they die, on average the visible light becomes systematically longer wavelength, and hence redder. In other words, the colour of the galaxy will change.

If the stars in a galaxy were *not* all formed at the same time, the overall appearance will change in a more complicated way. However, it is still straightforward to determine exactly what one would expect to see, using the same sort of argument. By exploiting this basic logic, astronomers have come to the conclusion that most of the stars in elliptical galaxies were produced in a single "burst," about 1 or 2 billion years in duration, some time between about 10 and 15 billion years ago. By contrast, spiral galaxies have witnessed more gradual, ongoing star formation, which has continued even to the present day. This is illustrated in Figure 27.1, which exhibits the changing stellar birth rate for typical spirals and ellipticals.

Just as stars are clumped into galaxies, so also galaxies are clumped into larger conglomerations, so-called *galaxy clusters*. These clusters can differ enormously in both appearance and composition: Some clusters contain only a few galaxies; others contain tens of thousands or more. Some clusters are very irregular in appearance; others have extremely regular and symmetric shapes. Astronomers and cosmologists have introduced a morphological classification that distinguishes between poor and rich clusters, and between regular and irregular clusters. *Poor clusters* contain relatively few galaxies, whereas *rich clusters* contain many more. *Irregular clusters* tend to be asymmetric in appearance, without a well-defined center. *Regular clusters* tend instead to be quite symmetric, usually nearly spherical with a well-defined center where the density of galaxies is particularly high. Thus, in particular, the centers of rich, regular clusters typically contain a large number of galaxies, packed very close together, usually with one or two giant ellipticals located in the very middle.

The Milky Way is located in the *Local Group*, a poor, irregular cluster about 3 million lightyears across, comprised of some two dozen or so galaxies. The Local Group is dominated by two large spiral galaxies, the Milky Way and Andromeda. The remaining galaxies, mostly irregulars and dwarf ellipticals, are all much smaller in both size and mass. In a first approximation, one can view the Local Group as a gigantic binary system, with the Milky Way and Andromeda orbiting each other. Most of the remaining smaller galaxies are *satellite galaxies*, bound to Andromeda or the Milky Way by gravity in much the same way that moons are bound to planets in the Solar System.

The nearest rich cluster is the *Virgo Cluster*, located about 50 million lightyears away. It has a diameter of about 7 million lightyears and contains about 2500 galaxies that are bright enough to be detected from the Earth. There may be many other dwarf galaxies that are simply too dim to be seen. The center of Virgo is dominated by *M87*, a giant elliptical, which, as will be discussed later, is believed by many astronomers to contain a supermassive black hole with mass of about 5 billion $M_\odot$. (The name *M87* reflects the fact that this galaxy is the 87*th* object listed in the so-called *Messier Catalogue*.) The nearest rich regular cluster is the *Coma Cluster*, roughly 350 million lightyears away, which contains at least 10,000 galaxies.

As for stars in the Milky Way, one can also use the observed motions of stars in other galaxies to estimate the masses of those galaxies. In fact, provided that one knows the distance to the galaxy, it is actually easier to do this for other galaxies, since one does not have to correct for the fact that the Earth itself is moving inside the galaxy that is being probed.

For spiral galaxies, the key tool in such mass determinations is the measurement of a *rotation curve*, which plots the average orbital speed v of stars as a function of distance r from the center of the galaxy in question. What would one actually expect to see in such a rotation curve? If all the mass in the galaxy were located at the center (which, of course, it is not!), stars would behave exactly like planets orbiting the Sun. They would therefore satisfy an analogue of Kepler's Third Law, with the orbital speed decreasing with increasing distance (in fact one would have $v \propto 1/\sqrt{r}$).

However, most of the mass in a galaxy is not concentrated in the middle, so that the observed rotation curve should exhibit a more complicated behaviour. Near the center, the typical orbital speed should be almost zero. Moving further out the speed should increase, eventually reach a maximum, and then begin to decrease again. In particular, as one approaches the very edge of the galaxy, so that most of the mass is much closer to the middle, one would expect that the rotation curve should begin to satisfy the behaviour predicted by Kepler's Third Law.

Observed rotation curves *do* start from almost zero in the center of galaxy and increase up to a maximum value. However, even when looking at objects like gas clouds well outside the bright visible portions of the galaxy, one almost never observes the rapid decrease in velocity predicted by Kepler's Third Law. Rather, v seems to either remain approximately constant or even *increase* at larger distances from the center. This is illustrated in Figure 27.2, which exhibits four typical rotation curves. The only natural interpretation of this fact is that, especially towards the edges of these galaxies, there is a large amount of matter which, for some reason or other, is essentially invisible.

Elliptical galaxies do not have a net circulation, so that a mass determination is more

complicated. However, if one knows the linear size of the galaxy and the typical speed of the stars within the galaxy, one can still estimate the amount of mass required to produce enough gravity to hold the galaxy together. Exploiting this type of argument leads to the conclusion that there is also a missing mass problem for ellipticals.
In a similar fashion, the same argument can be applied to clusters of galaxies, where the individual galaxies within the cluster are treated like the stars within an elliptical. Here too one finds that there is more mass than meets the eye. Indeed, the missing mass problem becomes progressively more acute if one looks at the Universe on larger scales, like clusters and superclusters. Most cosmologists agree that, overall, as much as 90% or more of the total mass of the Universe is in the form of various sorts of matter that have not yet been detected directly.

A number of different explanations have been proposed for this so-called *missing mass.* One of the more conservative ideas is that most galaxies are surrounded by halos of very faint stars (although no compelling explanation for the origin of such halos has been provided). Another is that, even in the neighbourhood of the Sun, there are lots of small chunks of matter which, being sufficiently small and dim, have hitherto avoided detection. A third suggestion is that there is a lot of gas floating in interstellar space, an idea that has been corroborated by the recent discovery of x-ray emission from allegedly empty intergalactic space in rich clusters. A fourth explanation is that the Universe is inundated with a very large number of dwarf ellipticals, which again are too dim to be seen.
It is likely that each of the four aforementioned candidates contributes at least somewhat to the missing mass. However, it seems difficult to explain how, either individually or collectively, they could account for *all* the missing matter. For this reason, many astronomers have come to believe that the Universe also contains a large amount of some other, more exotic, source of *dark matter*, comprised of something quite different from the protons, neutrons, and electrons which (apparently) are the dominant source of matter on the Earth and the Solar System. Recent theories of particle physics are very obliging in this regard, in that they suggest many different possibilities as to what this exotic matter might be. It should, however, be stressed that none of these very recent theories have yet been proven experimentally, and that there is as yet no direct physical evidence for the existence of such exotic stuff.
The late 1990's have seen an accumulation of experimental evidence that there *are* in fact a large number of chunks of ordinary matter in the galaxy, but that these so-called *MACHO*'s (short for MAssive Core-Halo Objects) are not sufficiently massive or abundant to account for all the missing matter. The technique used to detect these MACHO's involves an interesting application of gravitational lensing: Space is sufficiently devoid of large pieces of matter that there is usually little, except the occasional cloud of gas, between the Earth and a star in the Local Group that an astronomer might observe in the sky. Occasonally, however, a MACHO will come between the star and the Earth-bound observer. The MACHO will then serve as a gravitational lens, causing the detected image of the background star to be much brighter than before or after the lensing.
A measurement of the overall amplification and the duration of the increased brightness for a single event is not sufficient to allow one to infer the mass of the lensing MACHO. However, combining data from a number of lensing events with additional inputs such as

the expected speed with which the MACHO's move through the galaxy can yield useful estimates as to the total amount of mass that is bound up in such objects. Detecting these MACHO's is an exceedingly tedious task which would be hopeless without computers. What it involves is looking at multiple images of a star taken at different times, and determining whether the brightness varies significantly from one image to another. The painful part of the analysis is that one must literally examine tens of millions of stars in order to detect even a handful of events.
The net conclusion of this arduous task is that these MACHO's probably correspond to so-called *brown dwarfs*, chunks of matter somewhat smaller than the smallest stars, which are too small for nuclear fusion to ever occur; but that they can account for no more than a few percent of the missing mass.

What kind of galaxy is the most common? This is a difficult question to answer, because of both *environmental* and *selection effects*. A selection effect refers to an intrinsic bias that contaminates any set of observations that an astronomer might try to make. In this context the origin of the bias is extremely easy to see. Spiral galaxies are brighter overall than elliptical galaxies. However, if we look out in space what we see is limited by the fact that there is a critical threshold below which objects are simply too dim to trigger our eye or a photographic plate. It follows that, even if most of the observed galaxies are spirals, it need not be true that spirals constitute a majority of all the galaxies that are out there. There may be many more ellipticals which are simply too dim to be seen.
There is also an environmental effect in the sense that different types of environments appear to contain different proportions of spirals and ellipticals. Thus, e.g., rich regular clusters seem to be comprised almost exclusively of ellipticals, especially in their central regions, but spirals are more common in settings where galaxies are not packed so closely together.
Allowing for these effects, a reasonable estimate is that maybe 15% of all galaxies are spirals, 15% are irregulars, and the overwhelming majortiy, about 70%, are elliptical galaxies. Of the galaxies that we actually see, well over half are spirals.

Despite the pioneering arguments of the Swedish astronomer *Erik Holmberg* in 1940, until comparatively recently it was generally assumed that galaxies have only minimal interactions with one another, and that they can be treated as existing in splendid isolation. However, starting with the numerical experiments of the American brothers *Alar Toomre* and *Juri Toomre* in the early 1970's, there has evolved a growing recognition that interactions between galaxies are both frequent and important.
That they should be frequent is not hard to see. Near the Sun, the typical distance between stars is about 10 million times the size of the Sun, so that it is terribly unlikely that any two stars will ever become involved in a direct physical collision. However, in a rich cluster the average distance between galaxies is relatively small, less than 10 times their typical size. This means that, even neglecting their mutual gravitational attraction, they should run into each other from time to time: Just as billiard balls rolling randomly on a table will occasionally collide, so also galaxies orbiting in a cluster will ocasionally collide with each other.

What does such a collision entail? One thing that it will *not* involve is lots of direct physical collisions between stars, again because the distances between stars are so much smaller than the sizes of the stars themselves. However, these collisions *will* cause abrupt and significant changes in the gravitational field that any given star experiences, and this can have a substantial effect on their orbits. Thus, e.g., stars can be transferred from one galaxy to another, or even ejected into interstellar space, so that, for example, the disc component of a spiral galaxy can be stripped away. This can give rise to all sorts of interesting structures, including "bridges" between nearby galaxies.

Moreover, even if individual stars do not engage in physical collisions, much larger gas clouds *will* smack into each other, thereby compressing the clouds and generating high temperatures. This can have two different sorts of effects. (1) If a collision is sufficiently violent, the gas can be heated to the point that the atoms are moving fast enough to escape from the galaxy, so that the gas leaks out into intergalactic space. Such gas is in fact believed to be the origin of recently observed x-ray emission from intergalactic space in several rich clusters. (2) A less violent collision will not cause the gas to be expelled from the galaxy, but it *will* cause a sufficient compression to trigger accelerated star formation. This is believed to be the explanation of so-called *starburst galaxies*, spiral galaxies which are characterised by an excess number of bright young stars and a much higher rate of star formation than is observed in ordinary spiral galaxies.

When stars and gas are expelled into intergalactic space, they will carry off energy and angular momentum, so that the remaining material has less energy and less angular momentum. What this means is that it is possible for the remaining pieces of the two progenitor galaxies to stick together into a single entity, so that the two galaxies can *merge* into a single new galaxy. More massive galaxies have more gravity and, for this reason, exert a stronger attraction on other nearby galaxies. It follows that they will be involved in more frequent collisions that could cause them to capture smaller galaxies and become yet bigger. Many astronomers believe that such *galactic cannibalism* is responsible for the giant ellipticals located in the centers of rich regular clusters. Thus, e.g., the galaxy 027+352 (named for its location in the sky) is a giant elliptical galaxy with 8 distinct nuclei, at least 7 of which presumably correspond to different undigested meals.

How do galaxies evolve? At the present a clear understanding of galactic evolution does not yet exist. However, over the past twenty years a crude general picture has begun to emerge. One version of that picture is the following:

Early in the history of the Universe, billions of years ago, there were no galaxies, but there were huge clouds of rotating gas, mostly hydrogen and helium, the *protogalaxies* from which today's galaxies were born. Just as smaller clouds of dust and gas contract today to form associations and clusters of stars, so also in the past the gigantic protogalaxies presumably contracted down to smaller size in response to gravitational attraction.

Suppose in the first instance that the protogalaxy was relatively isolated, so that it could evolve on its own without being disturbed by other nearby objects. The initial gas blob, perhaps rather nearly spherical, would begin to contract, and pieces of the cloud could pinch off to form smaller substructures, including globular clusters. If there were no net rotation, the blob would remain nearly spherical. However, conservation of angular momentum implies that a spinning blob should rotate faster as it shrinks down to smaller size. But increased rotation would serve as a force tending to stretch and flatten the blob, with the net result that the material that had not yet frozen out would be converted into a disc. The blob would have been converted into a spiral galaxy, in which stars are formed only relatively slowly.

Now suppose instead that the galaxy is not isolated. Then during its formative stages the protogalaxy will probably have become involved in one or more violent collisions with neighbouring protogalaxies. Indeed, collisions in the past should have been significantly more prevalent than today, both because the protogalaxies were larger than the present day galaxies and because (as will be discussed more carefully later) early on the Universe was smaller in the sense that the typical distance between objects was smaller. At these early times, the protogalaxies were still comprised almost completely of gas, rather then stars. However, this means that the collisions would have triggered enormous bursts of star formation. Most, if not all, of the gas would be converted into stars, and much of the remainder would be ejected into intergalactic space. Moreover, such collisions would probably also serve to disrupt any nascent disc. The net result is an elliptical galaxy, with no disc component and almost no dust and gas, comprised almost completely of very old Population II stars.

This simple picture is probably not completely right. It does not, for example, explain the relatively sharp division between the spheroidal and disc components of a spiral galaxy, and it does not explain how a lenticular galaxy might come to have a disc, but no dust and gas. However, it *does* account for several observational facts in a relatively simple and plausible way. Thus, for example, it provides a simple explanation of why the types of galaxies that we see correlate with environment. Similarly, it is consistent with the fact that, in spiral galaxies today, collisions do trigger excess star formation and they can cause a galaxy to lose its disc.

This interpretation would also suggest that, even today, collisions between spiral galaxies could convert the progenitors into ellipticals. An examination of galaxies that are now, or have been recently, involved in collisions tends to support this idea in the sense that the final objects look more elliptical then spiral. However, it is not yet clear whether this picture can account correctly for the numbers and colors of stars in the post-collision objects.

Starting in the late 1980's, it was recognised that galaxy-galaxy collisions can have an important effect on the visual appearances of galaxies. today. In particular, high resolution photographs of galaxies provided by the Hubble Space Telescope led to the unambiguous conclusion, first published in 1993, that galaxies in rich clusters are much more likely to exhibit irregular features attributable to relatively recent collisions than those in lower density environments where galaxies tend to be further apart. Even galaxies in lower density environments often show signs that, at one time or another in the past, they were involved in some sort of collision. Indeed, some astronomers would go so far as to assert that almost every galaxy that has been studied carefully exhibits some sort of evidence indicating that it was involved in a collision at one time or another during its life history.

One of the most striking things about galaxies is the fact that, on the average, they are moving away from us. This effect was in fact noticed already even before galaxies were identified as objects outside the Milky Way, starting with Slipher's observations of "spiral nebulae" in 1914. With Shapley's 1918 calibration of Leavitt's period-luminosity relation for Cepheid variables, one also had a way in which to estimate distances; and, by the early 1920's, it was possible to search for correlations between the recession velocity of some object and its distance away from the Earth. Indeed, in 1924, the same year in which Hubble announced his proof that other galaxies exist, the German *Carl Wirtz* recognised explicitly that, overall, more distant galaxies are moving away from us more quickly. Four years later, motivated by the earliest ideas about the Big Bang Model of an expanding Universe, the British cosmologist *H. P. Robertson* conjectured that there should be a linear relation between the average recession velocity v and distance r. The following year, in 1929, Hubble presented the first careful analysis of observations, indicating the existence of such a linear relation,

$$v = Hr,$$

with H a constant. This relation has since come to be known as the Hubble Law.
What this law says is that, on the average, a galaxy separated from us by a distance r will be moving away with a velocity v that is linearly proportional to r. Individual galaxies will show some deviations from this law but, overall, there is a linear trend. In the intervening sixty-plus years, this relation has been subjected to increasing scrutiny, but the conclusion has remained firm: the Hubble Law does in fact provide an extremely good characterisation of what one actually sees. Typical data for some relatively nearby galaxies are exhibited in Figure 27.3.

How accurately is this relation actually known to be true? Recession velocities can be measured extremely accurately using Döppler shifts, so that there is no problem there. However, there *is* a fair amount of uncertainty associated with measuring absolute distances. Thus, e.g., there was the problem described already regarding Shapley's normalisation of the period-luminosity relation. However, even though absolute distances are hard to determine with great certainty, *relative distances* are comparatively easy to ascertain. For example, even if one does not know the distance to either of two objects, it is relatively simple oftentimes to determine that one is (say) about 3.5 times as far away as the other. The net result is that, even though the linear relation between v and r implied by the Hubble Law has been confirmed very accurately, the value of the Hubble constant H is considerably less well determined. Indeed, as late as the mid 1990's, different groups of cosmologists, trying to estimate distances in different ways, claimed values of H ranging from 50 to 100 km/sec/Mpc, where *Mpc* denotes 1 million parsecs, a distance of about 3.26 million lightyears. (A *parsec* is defined as the distance to an object which, over the course of a year exhibits a parallax angle of one second of arc. It corresponds to about 3.26 lightyears.) Towards the late 1990's, there has emerged a growing consensus that the correct value is about 70 km/sec/Mpc, with an uncertainty of 10 km/sec/Mpc or so, although agreement is not yet universal. The range 50 to 100 km/sec/Mpc translates roughly into a value of between 16 and 32 kilometers per second per million lightyears.
Assuming, for example, that the lower value is correct, the Hubble Law says that a galaxy separated from us by ten million lightyears will, on the average, be moving away at a speed of 160 km/sec, and that a galaxy separated by 1 billion lightyears will be moving away at 16,000 km/sec.

Lectures 29 and 30

QUASARS AND ACTIVE GALAXIES

In order to characterise the large scale structure of the Universe as a whole and, especially, to determine the correct value of the constant H entering into the Hubble Law $v = Hr$, one must estimate as exactly as possible the distances to various objects outside the Milky Way. The basic trick used in such distance estimates is the identification of so-called *standard candles*, easily identified objects which (hopefully) have a known intrinsic brightness. The important point is that if one knows both (1) the intrinsic brightness of some object (because it is a standard candle) and (2) the apparent brightness (which is readily measured) one can determine immediately (neglecting the effects of any intervening dust and gas) how far away the object must be to appear as bright or dim as it actually does.

The simplest, and best known, example of standard candles is provided by Cepheid variables, for which luminosity or intrinsic brightness can be inferred using the period-luminosity relation. Because they are extremely bright stars, individual Cpeheids can be detected from the Earth out to distances of about 20 million lightyears or more, which means in particular that they can be used to map out the distances of various objects within the Local Group. Unfortunately, the next cluster out is the Virgo Cluster, about 50 million lightyears away, too far away to permit the detection of Cepheids using Earth-based observations. Recently, however, using the Hubble Space Telescope in space it has proven possible to detect Cepheids in Virgo, which means that, there again, they can be used as standard candles to effect accurate distance estimates.

At yet larger distances, Cepheids are simply too dim to see. However, one *can* detect the very brightest stars in galaxies out to a somewhat larger distance, about 80 million lightyears. The key point then is that if one *assumes* that the brightest stars in every galaxy are comparable in luminosity, they can be used as standard candles. By measuring the apparent brightness of these stars and then assuming a value for their intrinsic brightness, it is again possible to estimate a distance to the galaxies in which they are contained.

A similar argument can be used for other sorts of objects, including individual *hydrogen clouds*, which can be detected out to about 300 million lightyears. Indeed, if one is lucky enough to see a rare supernova, one can estimate a distance as large as 8 billion lightyears or more. However, at very large distances, it eventually becomes impossible to find any standard candles to use: even entire galaxies are too dim to be detected. What this means is that, for the most distant objects we see, a direct distance determination is impossible. Rather, as described more carefully below, one proceeds by measuring a Döppler shift, inferring a recession velocity, and then using the Hubble Law to convert that recession velocity v into a distance r.

Although a powerful tool, the use of standard candles *does* have problems. When looking at relatively nearby objects, there will be a number of different standard candles that one can use and, if they all give comparable distances, one can be reasonably confident that the inferred distances are at least approximately correct. However, at larger distances fewer standard candles are available, so that there are fewer cross checks and the likelihood of error increases.

Another problem historically has been the misidentification of various standard candles. One example, mentioned already, involved Shapley's failure to distinguish between Type I and Type II Cepheids, which led to an error in distances by a factor of about 2.6. Another classic example was when Hubble confused several very bright stars in Virgo with hydrogen clouds, which are about 5 times more luminous, and hence seem only $1/\sqrt{5} \approx 1/2.2$ times as far away. The net result was that the earliest estimated distance to Virgo, and hence all more distant objects, was about $2.6 \times 2.2 \approx 5.7$ larger than the value recognised today. This, in turn, was in large part responsible for the fact that Hubble's earliest estimates of the Hubble constant ranged between 465 and 513 km/sec/Mpc, much larger than the values between 50 and 100 km/sec/Mpc typically quoted at the present time.

A third source of error is associated with the fact that, just like interstellar space within a galaxy, intergalactic space contains gas which can make a distant object appear dimmer than it would otherwise. The key point is that a bright object can appear dim either because it is far away or because it is being obscured by some intervening "fog." Only by correcting for the effects of this *intergalactic extinction* of light can one infer correct distances.

A final point is that, as one is looking out towards more distant objects, one is also looking into the past. Thus, for example, light from a galaxy one billion lightyears away left that object about a billion years ago and has been travelling through space since that time. However, there is no guarantee that galaxies a billion years ago looked like the galaxies of today, so that it may be dangerous to use them, or any objects in them, as standard candles. Indeed, as discussed in an earlier lecture, one knows that elliptical galaxies must have looked substantially bluer and brighter in the past.

As stated already, direct measurements of distance are impossible for objects that are very far away, so that one is forced to use the Hubble Law itself to obtain a distance estimate. This is done by recognising that, if a galaxy or some other object emitting electromagnetic radiation is moving away from us, the spectrum of radiation which is observed will be systematically dilated towards longer wavelengths by an amount that depends on the speed with which it is moving away. All that one need do, therefore, is identify one or more spectral lines which were emitted with wavelengths λ_{emit} and observed at wavelengths λ_{obs}, and, for any given line, compute the *redshift*

$$z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}}.$$

Different wavelengths are all dilated by the same fractional amount, so that the redshift is independent of which line one examines. Provided that the object is moving away, $\lambda_{obs} > \lambda_{emit}$, so that $z > 0$.

To the extent that the recession speed v is much smaller than the speed of light, the effects of special and general relativity can both be ignored, and Newtonian mechanics is approximately correct. In this limit, the redshift z is related to the recession velocity v by the simple relation

$$z = \frac{v}{c}.$$

If, however, one *assumes* that the entire recession velocity is due simply to the Hubble Law, one can set $v = Hr$, so that

$$z = \frac{Hr}{c} \qquad \text{or} \qquad r = \frac{cz}{H}.$$

In other words, by measuring z and assuming a value for H, one can infer a distance r. This connection between z and r is based on the assumption that the *entire* relative velocity between the object and the observer can be attributed to the overall expansion of the Universe, i.e., the systematic sense in which distant objects are moving apart. However, one knows from an examination of nearby objects that this is not exactly true. The Hubble Law is only an average relation, which neglects so-called *peculiar velocities* associated with gravitational interactions amongst relatively nearby objects. However, the Hubble Law implies that, at larger distances, the average recession velocities are extremely large, whereas the peculiar velocities should be comparable to what is observed relatively nearby (assuming that our place in the Universe is not special!). It follows, therefore, that at large distances the neglect of peculiar velocities should not be a major source of error, so that it *is* reasonable to use the Hubble Law itself as a distance indicator.
One final point: The simple relation between z and r quoted above only holds for relatively nearby objects with v much smaller than c. When v becomes large, relativistic effects become important and this relation is modified. It follows that, in fact, the observed z can be much larger than 1. Indeed, as will be discussed later one actually observes redshifts as large as $z \approx 5$, these corresponding to objects moving away at speeds well in excess of 90% the speed of light.

But what does this have to do with quasars and active galaxies? The radio telescope was first invented in 1937 by *Grote Reber*, an engineer and amateur astronomer from Illinois who built a small one in his backyard. By 1944, he had found three different sources outside the Solar System, known now as Sagittarius A (the first radio emitter discovered in the constellation Sagittarius A), Cassiopeia A (the first in Cassiopeia), and Cygnus A (the first in Cygnus). Sagittarius A corresponds to the nucleus of the Milky Way which, as described already, is a strong radio emitter. Cassiopeia A is associated with a supernova remnant inside the Milky Way. But what about Cygnus A?
At that time, the identity and precise location of Cygnus A were completely unknown. However, several years later, in 1951, an optical counterpart was discovered, this corresponding to a very dim galaxy. The measured redshift of the galaxy was $z \approx 0.057$, which translates into a recession velocity of about 17,000 km/sec. Assuming a Hubble constant H=50 km/sec/Mpc, this implies a very large distance of about 1 billion lightyears, which explains why the visible galaxy looks so dim. However, despite this enormous distance Cygnus A is one of the brightest radio sources in the sky, which means that its intrinsic brightness must be absolutely enormous. Indeed, to seem as strong a radio emitter as Cygnus A, an object 1 billion lightyears away must have a total radio luminosity about 10 million times the radio luminosity of the Andromeda galaxy. It must give off 10 million times as much energy in radio waves as does an entire ordinary galaxy!
Over the next several years, a number of other, quite similar objects were found, each corresponding to a strong radio source associated with a tiny "starlike" object that is emitting

bizarre spectral lines. What these objects were was unclear, and the unusual character of the spectral lines went unexplained. However, in 1963 the American astronomer *Maarten Schmidt* finally recognised why it was the lines looked so weird: one was seeing the effects of an enormous redshift, for which the fractional changes in wavelength for familiar spectral lines was very large. Thus, in particular, Schmidt analysed the spectrum of one of these objects to show that it has a redshift $z \approx 0.16$, which, assuming $H = 50$ km/sec/Mpc, implies a distance of about 3 billion lightyears.
By the present time, literally thousands of these "quasi-stellar sources," or *quasars*, have been discovered.

If the observed redshift z does indeed correlate with distance, as was assumed in the preceding, these quasars must be extremely far away. For example, by the mid 1990's astronomers had measured redshifts z in excess of 5, which implies a recession velocity well in excess of 90% the speed of light. This in turn implies that the object must be extremely far away, well over 10 billion lightyears. But this means that the light which we see today was emitted many billions of years ago, when the Universe was very young, presumably long before galaxies as we currently know them even existed.
However, if these objects are really that distant, the fact that we can still see them at all means that they must be incredibly bright. The intrinsically brightest ordinary galaxies are about ten times as luminous as the Milky Way, and these can only be seen from the Earth out to a redshift of $z \approx 1$. By contrast, the typical total luminosity of a quasar is 100 times the total luminosity of the Milky Way. Not all this energy is emitted in the form of radio waves. Rather, telescopes above the Earth's atmopshere indicate that quasars are also strong emitters in infrared and x-rays.

Although quasars are incredibly luminous, the size of the emitting region in a quasar must be extremely small. Quasars can exhibit substantially changes in luminosity, 20% or more, over the course of one or two hours, and their overall brightness can change by as much as a factor of 2 or 3 within a couple of years. The net inference is that the size of the emitter cannot be more than a few *lighthours* in size, i.e., the distance light can travel in a few hours. The basic idea, illustrated schematically in Figure 29.1, is easily understood: Even if the object in this Figure, shown to be 1 lightday across, were to emit a sudden flash of light, photons from Point A would arrive at the Earth a full day before photons from Point C. It follows that, as seen from the Earth, the flash would entail a gradual change in brightness lasting an entire day.
Given that light travels at about 300,000 km/sec, and that 1 AU corresponds to a distance of about 150 million km, one lighthour corresponds to a distance of about 7.2 AU, which is substantially less than the distance from the Sun to Saturn. In other words, one is led to the seemingly incredible conclusion that, in a quasar, an amount of energy 100 times as large as that emitted by an entire galaxy is being emitted from a region smaller in size than the Solar System!
The only way to avoid this conclusion is to argue that, in fact, quasars are not that far away. In this case, the total luminosity required to explain the observations would go down, even though the size of the emitting region would remain small. However, if quasars are really not all that distant, their large redshifts cannot be attributed to the Hubble Law

and, as such, one must come up with an alternative explanation of the large values of z that are observed.

Back in the late 1960's and early 1970's, some astrophysicists thought that one might be seeing a gravitational redshift, rather than a Döppler shift, the idea being that quasars are somehow regions in space where gravity is extremely strong, so that light escaping out of the quasar loses a lot of energy. However, detailed numerical investigations performed since then have made this alternative explanation seem implausible. Basically, what has been found is that any collection of objects sufficiently dense to cause a gravitational field capable of producing such enormous redshifts will likely end up collapsing to form a gigantic black hole, so that the high redshift light could never escape. As described below, one *does* believe today that quasars contain a supermassive black hole, but the hole itself is probably *not* responsible for the observed redshifts.
The one definitive way to prove that the redshifts observed for quasars are not a result of the fact that they are very distant would be to find a high z quasar attached somehow to another nearby low z object. Over the past thirty-plus years, systematic investigations have been made with precisely this aim. However, despite an enormous amount of effort, no such connections have ever been found. Rather, these investigations have only reinforced the expectation that quasars are extremely distant objects. In particular, a fair number of the quasars with the smallest z's appear to be located at or near the centers of very faint galaxies with almost identical redshifts.

Another important fact which would suggest that quasars really are what they seem to be is that one sees similar, "toned-down" versions in the form of so-called *active galaxies*.
One type of active galaxy is a *Seyfert galaxy*, the first example of which was discovered in 1943. In terms of visual appearance, a Seyfert galaxy corresponds to a bright "starlike" nucleus superimposed in the middle of what seems, otherwise, to be a more or less ordinary spiral galaxy. As might be expected, Seyferts contain lots of hot gas, this material having been heated by the central energy source. Seyferts are known to exhibit significant changes in luminosity, the apparent brightness changing by as much as a factor of 2 or 3 within the course of a few months. They are in fact relatively common objects. About 2% of all spirals appear to be Seyferts; they constitute about 10% of all the brightest galaxies. The fact that they look so much like ordinary spirals has led to the suggestion that Seyferts are not a special kind of galaxy. Rather, it may be that the large energy emission represents a transitory event in the lifetime of an otherwise ordinary galaxy.
Another type of active galaxy is an *N-galaxy*. In visual appearance, an N-galaxy corresponds to a bright "starlike" nucleus superimposed in the middle of an otherwise normal elliptical galaxy. These differ from Seyferts in that, being ellipticals, they show no signs of concentrated accumulations of gas, hot or otherwise. Some N-galaxies evidence particularly large changes in luminosity over the course of a few years or less. The classic example thereof is the galaxy *BL Lacerta*, located in the constellation Lacerta, which can change in brightness by a factor of 15 in less than half a year. These N-galaxies are known as *BL Lacertae objects*, or *blazers* for short.

One particularly dramatic example of active galaxies is provided by so-called *double radio sources*. These double radio sources have a bright "starlike" nucleus like a Seyfert or N-galaxy. However, they also have two huge radio lobes, displaced far outside the galaxy, from which radio waves are emitted into intergalactic space. Each of these lobes can be as long as 200,000 lightyears, twice the size of the Milky Way, and the distance between the lobes can be as much as 20 million lightyears, more than 5 times the diameter of the Local Group.

A good example of a double radio source is provided by the galaxy *NGC 5128* (named as the 5,128*th* object listed in the so-called *New General Catalogue*), which is the visible galaxy containing the radio source Centaurus A. *NGC 5128* is a giant elliptical galaxy, surrounded by a rotating ring of dust and gas that shows evidence of active star formation. The galaxy contains two large lobes, approximately 2 million lightyears apart, which are directed perpendicular to the ring. It also contains a so-called x-ray *jet*, this corresponding to a relatively sharply collimated stream of x-rays ejected into intergalactic space towards one of the radio lobes. The presence of the ring, and the observed star formation therein, suggest that this galaxy is currently recovering from a major galaxy-galaxy collision, probably resulting in a merger, about 1 billion years ago.

When quasars and active galaxies were first discovered, it was not at all clear what they are or how they work. In particular, it was difficult to imagine a mechanism that could generate such enormous amounts of energy within a relatively small region, comparable in size to the Solar System. However, over the past twenty years or so, most competent astrophysicists have come to believe that the observed behaviour of both active galaxies and quasars can and should be attributed to the presence of a supermassive black hole located in the center of an otherwise ordinary galaxy.

As a concrete example, one can consider the accumulating evidence that there exists such a supermassive black hole in *M87*, a giant elliptical galaxy about 40 times as massive as the Milky Way, located at the center of the Virgo Cluster. This galaxy is a typical active galaxy, with a bright starlike nucleus. It also has a jet that extends outwards for about 6000 lightyears, emitting in both visible and radio wavelengths, which was first discovered at visible wavelengths in 1918. Because this galaxy is relatively nearby, only about 50 million lightyears away, it is comparatively easy to see; and, because it does not seem to have been disturbed by any significant collisions in the relatively recent past, astronomers would expect that it should be comparatively easy to characterise.

Starting in the middle of the 1970's, astronomers studying the very center of M87 discovered that its spectral lines were broadened, i.e., that what should have been single well-defined frequencies were "fuzzed" out to include fairly substantial ranges of frequencies. The obvious implication was that the observed starlight was being emitted by a number of different stars moving at very high speeds. However, if the stars are moving very fast, they will rapidly escape from the core out into the remainder of the galaxy unless there is some significant source of gravity to keep them trapped in the center. In particular, a careful examination of the lines suggested that there must be at least 5 billion $M_\odot$ of material confined within a very small volume.

Towards the end of the 1970's, astronomers also discovered that the stars seemed to be "distributed wrong." Starting in the 1960's, astronomers had developed the so-called *Mitchie-King models*, which seemed very successful in characterising the distribution of stars inside most elliptical galaxies. However, these models simply could not fit the observations of the core of M87. There are far more stars in the center than expected. Exactly how many was not completely clear at the time, since it was hard to resolve the core adequately to see exactly what was going on in the middle. However, the best estimate corresponded to an excess of several billion $M_{\odot}$. This number has been corroborated by recent high resolution observations made using the Hubble Space Telescope which indicate that the density of stars at the very center of M87 is at least three hundred times higher than what would be predicted by the Mitchie-King models.

Back in 1980, this sort of high density concentration seemed exceptional. More recently, however, high resolution images obtained from the Hubble Space Telescope, as well as several ground-based telescopes, have provided indications that such high density central concentrations may actually be quite common. In the early 1980's, it was difficult to resolve features in an observed object on scales much below half a second of arc. However, with the advent of space-based telescopes which avoid distortions associated with the Earth's atmosphere and, particularly, with the advent of improved signal analysis techniques facilitated by the development of faster computers, astronomers can now do significantly better. Specifically, under optimal conditions it is possible to resolve structures down to an angular scale of about 0.05 seconds of arc which, for relatively nearby galaxies, corresponds to a physical distance of only a few lightyears. In other words, it is possible to detect structures even in nearby galaxies as small as ten lightyears or so! The important point, then, first appreciated fully in 1995, is that many nearby galaxies appear to have a bright central cusp, corresponding to a high density collection of stars, that is much brighter than what would have been inferred from Mitchie-King models or any similar alternative.

(In the late 1990's, understanding the implications of these cusps was one of the single most important conceptual issues confronting galactic astronomers. Combining ideas from theories of chaos with numerical investigations of model galaxies suggests that it is impossible to construct an equilibrium model of a galaxy that simultaneously (1) is nonaxisymmetric and (2) contains a very steep cusp. However, this had led galactic astronomers to rethink one of their most cherished beliefs, namely that galaxies can be viewed as being in, or very close to, equilibrium, and that, at least in the absence of galaxy-galaxy collisions their overall structure will remain essentially unchanged over the course of billions of years.)

Both high velocities and an excess number of stars in the middle of a galaxy can be explained very simply by assuming that there is a supermassive black hole at the very center. The presence of a supermassive black hole would tend to attract stars further towards the core, thereby increasing the density. However, the stars in the core would then feel excess gravity, both from the hole and from their mutual attraction, which would lead to higher velocities. In fact modifying the Mitchie-King models to allow for the effects of a 5 billion $M_{\odot}$ black hole yields a new model which seems to fit the observations of M87 extremely well.

The supermassive black hole in M87, assuming that it exists, is relatively large. However, it is almost certainly not the only supermassive black hole located in our neighbourhood of the Universe. The Hubble Space Telescope has found compelling evidence for supermassive black holes in a number of different galaxies, and it may even be true that nearby Andromeda contains a central black hole with a mass of about $5 - 10$ million $M_{\odot}$. It may therefore be that supermassive black holes are quite common, residing in the centers of many, if not all, galaxies. What does, however, seem to be true is that the black hole in M87 is substantially larger than any present in other relatively nearby galaxies.

But how could such a supermassive black hole explain the observed luminosity associated with a quasar or active galaxy? Just as for the case of activity at the center of the Milky Way, this may be associated with the accretion of gas onto the supermassive hole. Suppose that a 5 billion $M_\odot$ hole were inserted into the center of a galaxy, where it was surrounded by dust, gas, and stars. What would happen? The gravitational field associated with the hole would attract this nearby material towards the center and, if no angular momentum were involved, much of the material would fall directly onto the hole. However, if either the hole or the galaxy rotates, there will be a net angular momentum and, as a result thereof, the matter will tend instead to settle down into an accretion disc, with gas orbiting the hole along near-circular orbits. The relativistic analogue of Kepler's Third Law implies that the orbital period of the gas increases as one moves further away from the hole, so that bits of gas at slightly different radii revolve at different rates. It follows that there will be friction between successive layers of gas. However, this friction implies that the gas will radiate energy and angular momentum into space, eventually spiraling down towards the hole.

This much of the picture is virtually identical to the situation believed to exist at the center of the Milky Way. However, one has not yet explained the origin of lobes and/or jets. These are explained through the following additional refinement of the model: As the gas spirals in towads the hole it will be heated to very high temperatures, high enough that it emits x-rays and/or gamma rays. Moreover, if a substantial amount of material is accreting towards the black hole, the ambient density of the gas can become quite high. However, the combination of high temperature and relatively high density leads to high pressure; and, if the pressure is sufficiently high, it can prevent the black hole from accreting the infalling material, which must instead be expelled somehow. Basically the black hole cannot "swallow" the infalling matter fast enough.

As the gas is pushed away from the hole, it will try to follow the path of least resistance. This means that, rather than escaping through the relatively dense accretion disc, it will instead tend preferentially to be focused perpendicular to the disc, being ejected in two more or less collimated beams.

But why are the jets so sharply collimated? Here it is probable that several different effects play a role. Part of the answer is related to the fact that, when a high speed beam of material is squirted into another medium, rather than through empty space, it tends to remain collimated. This is a physical phenomenon which is easy enough to test in your own bath tub: If you squirt water out of the nozzle of a hose into the air at a high speed, it will typically spread out over a wide angle. If, however, you instead spray the water into your tub, the water in the tub will reduce the spreading, and the ejected stream of water will remain far more focused.

Gas being ejected from near a black hole initially has to pass through parts of the accretion disc, which will help to collimate the jets initially. Once they have escaped the disc, this influence is gone. However, the presence of even very low density gas in interstellar and intergalactic space also helps keep them focused. Indeed, the fact that the jets *are* collimated as much as they are is one concrete piece of evidence suggesting that there is a substantial amount of gas in so-called empty space.

There is also another, even more compelling, piece of evidence indicating the presence of this intergalactic gas, namely the fact that the lobes in a double radio source need not be aligned. Rather, one observes some so-called *head-tail sources*, which appear to correspond to objects moving at high speeds through the intergalactic medium. As illustrated schematically in Figure 29.2, the central object leads, and the lobes trail behind like the smoke emitted from a rapidly moving steam locomotive.

Another important point is that, even though the black hole itself cannot support a magnetic field, the material around it will in fact generate a huge magnetic field. This field can also contribute to collimating the jet. Even more significantly, however, this field is responsible for the intense synchrotron radiation emitted from the lobes. Synchrotron radiation typically results when high speed electrons moving with speeds close to the speed of light are subjected to huge accelerations as a result of a strong magnetic field. The lobes themselves are presumably generated when the high speed material originally ejected from the hole crashes into one or more higher density gas clouds, at which point they are decelerated dramatically.

As discussed already, accretion onto a black hole is a very efficient source of energy. For the case of a spherical hole, each little chunk of mass m will yield an amount of energy just under $0.06mc^2$, and, for rotating black holes, the efficiency can be even higher. It follows that one could power even one of the most luminous quasars with a relatively modest accretion rate of less than $10M_\odot$ per year.

Note that this explanation of quasars and active galaxies implies that whether or not jets and lobes are present depends on the overall accretion rate. If this rate is sufficiently low, the hole can "swallow" everything aimed at it, so that nothing is left over. If, however, the accretion rate is too high, the black hole is unable to "swallow" all the infalling material, and the remainder is ejected into intergalactic space.

One other interesting point is that although many active galaxies are seen relatively nearby, there are few bright quasars observed nearby with small redshifts. In the context of the preceding model, this is explained as follows: Billions of years ago, when galaxies (or protogalaxies) were very young, free gas (not tied up in stars) was more abundant than it is today. It follows that there was more material for supermassive black holes to accrete and, as such, more luminosity was generated. However, as galaxies become older, the amount of free gas decreases. This means that there is less material for the hole to capture and, consequently, less energy gets emitted. The central engine has not really changed. The only difference is that there is less fuel to stoke the flames.

One other point is worth making: If a quasar were a completely isolated entity, it would rapidly swallow all the ambient gas and then turn off relatively abruptly. The net result is that, even if supermassive black holes are relatively common, only very few quasars would be seen. It is only because of collisions between (proto-)galaxies that jiggle the gas inside, pushing some of it in towards the hole, that a significant accretion rate could be maintained for times as long as a billion years.

Lectures 31, 32, 33, and 34

THE STANDARD HOT BIG BANG MODEL

Normally astronomy, like other branches of science, proceeds by means of extrapolation: One performs some set of experiments and/or observations and, based on the results of these observations, formulates a general theory that is used as a basis for making additional predictions, which can then be tested experimentally. However, for various reasons twentieth century cosmology has not always adopted this tact. Instead, the standard Hot Big Bang Model, which constitutes the conventional paradigm of modern cosmology, was formulated originally by postulating a collection of simple, not implausible assumptions, at the time not yet particularly well motivated observationally, and then exploring systematically the implications of these assumptions.
One major reason for this was the fact that, back in the 1920's, 1930's, and 1940's, when the Big Bang Model was originally being developed, adequate data did not exist to permit a direct extrapolation from observations to a single all-encompasing picture. Indeed, the earliest version of the Big Bang Model was formulated by the Russian *Alexander A. Friedmann* in 1922, several years before Hubble's definitive demonstration that other galaxies actually exist. Even as recently as the early 1960's it was relatively straightforward to concoct a number of different theories, none patently absurd, each of which was capable of "explaining" much of what one actually sees. Only during the past three decades or so have detailed observations and experiments provided concrete evidence which excludes more or less definitively different alternatives to the Hot Big Bang Model that have been proposed at various times during the twentieth century.
A second important reason for the unusual approach adopted by cosmologists has been the fact that cosmology aims to be an all-embracing discipline: it is, more or less by definition, the science of everything. However, given the broad aim of explaining everything, it is hard to really know where to begin. To secure a concrete starting point, cosmologists embracing the Big Bang Model have chosen typically to start from a collection of three, mathematically precise, assumptions.

The first standard assumption entering into the Hot Big Bang Model is the assumption of *homogeneity* on large, i.e., the idea that, on the average, the Universe is the same everywhere. This assumption implies, in particular, that, at any given instant, the average mass density of the Universe is the same everywhere in space, although the density need *not* be the same everywhere in time. For example, an observer sitting on a typical planet in a distant galaxy 10 billion lightyears away from the Earth should, when probing the Universe on sufficiently large scales, see the same kind and amount of stuff as does an observer sitting on the Earth. However, this assumption does *not* imply that an observer looking out at the Universe five billion years ago or ten billion years hence will also see the same average density as today.
The second assumption is an assumption of *isotropy* on the large, i.e., the idea that, overall, things look the same in all directions. Thus, e.g., if an observer looks out to large distances in different directions he or she should, on the average, see the same numbers and types of things. It is an interesting exercise in logic to show that if every observer in the Universe sees an isotropic distribution of stuff, then the Universe as a whole must in

fact be homogeneous.

A third, perhaps more insidious, assumption is an assumption of *universality*, the idea that the laws of science are the same everywhere in the Universe, now and for all times. This assumption, dating back to Newton's *Rules of Reasoning in Philosophy*, is implicit already in everything described hitherto in this course. Unless one makes *some* assumption about the applicability of the known laws of science, one has no idea how to begin in trying to understand what one actually sees.

It should perhaps be emphasised that the assumption of *universality* is *not* forced upon one purely on the basis of observations, and that other alternatives are possible, at least in principle. Thus, e.g., in the 1940's Dirac tried to develop a theory of physics and cosmology which assumed that the overall strength of gravity, as measured by the gravitational constant G, actually varies over the course of time. Recent high precision measurements of the binary pulsar serve to constrain possible variations in G, but one cannot exclude entirely the possibility that G changes, although there is no reason to assume that it does. A more radical speculation, due to the Russian physicist *Andrei Sakharov* in the late 1980's, was that things like the dimensionality of spacetime could be different at different places and/or times in the Universe.

In this context, one is *not* assuming that cosmologists know all the laws of physics and chemistry that govern the Universe as a whole. Indeed, it is unquestionably true that our knowledge of physics and chemistry fails when densities and temperatures become too high. However, one *is* assuming that the laws of physics and chemistry known to be true in the laboratory will also be true in outer space when the same physical conditions are realised. Thus, e.g., one supposes that hydrogen at a temperature of 1 million K and a density of 1 gm/cm^3 should behave the same, whether on the Earth, in the Sun, or in intergalactic space, and whether today or ten billion years ago. Moreover, one is also assuming that the broad world view suggested by Einstein's theory of general relativity is correct and that the broad general features of various modern theories of particle physics and the like are at least approximately true, except in very exceptional circumstances where the Universe is very different from how it is in our environs today.

The assumptions of homogeneity and isotropy are often combined to yield what is sometimes termed an anti-anthropogenic Cosmological Principle, which articulates the idea that we on the Earth do not occupy a special place in the Universe. Thus, in particular, two observers at different places in the Universe should, statistically, see completely identical surroundings. The cosmologist *Alexander Vilenkin* has (somewhat jokingly) referred to this as a principle of cosmic mediocrity. Historically, this Cosmological Principle arose as an outgrowth of the realisation in the 1910's and 1920's that the Sun is not located at or near the center of our galaxy, and that our galaxy does not seem special in any obvious way.

It is of course clear that the Universe as a whole is not *exactly* homogeneous and isotropic. Indeed, the presence of objects like people is a (potentially important) counter-example. However, what the cosmologist assumes is that, if one views the Universe on sufficiently large scales, such local irregularities "wash out," so that, in a first approximation, the Universe as a whole can be viewed as a smooth, homogeneous and isotropic, entity. The next approximation, to be addressed in later lectures, then entails going beyond the assumptions of large scale homogeneity and isotropy to characterise and understand shorter scale irregularities. *John Wheeler*, the physicist who first coined the term "black hole," likes to use the following analogy: Viewed from an airplane $30,000$ feet or more above sea level, an ocean looks like a smooth, structureless entity even though, when viewed up close, the water exhibits a complex pattern of foam, waves, and turbulence.

It should be noted explicitly that the Hubble Law $v = Hr$ is completely consistent with the assumptions of homogeneity and isotropy. Indeed, it is a straightforward exercise in analytic geometry or vector analysis to show that if an observer in galaxy 1 sees galaxies 2 and 3 each move away from her in a fashion described by the Hubble Law, it follows necessarily that another observer on galaxy 2 will see both galaxies 1 and 3 moving away from him in a fashion described by the Hubble Law.

It should also be recognised that if everything is moving apart uniformly in a fashion described by the Hubble Law, it must be true that, at earlier times, things were closer together. Indeed, the assumption that the same Hubble Law was true at earlier times leads to the retrodiction that, at a finite time in the past, everything that we see around us was piled up very close to us in an essentially infinite density state. This in turn leads to the key tenet of the Hot Big Bang Model, the idea that the Universe started in a very hot, high density state and subsequently expanded into what we see today.

As will be discussed below, there is substantial observational evidence that, overall, the Universe is in fact homogeneous and isotropic. However, it is important to recognise that, according to our current understanding, this overall homogeneity and isotropy of the Universe is *not* something that had to be true in the sense that it was forced upon the Universe by the laws of physics. Rather, within the context of the standard Hot Big Bang Model, the observed homogeneity and isotropy are unexplained mysteries. Other variants of this standard model, including so-called *Inflationary Cosmologies*, aim to provide a physical explanation of why it is that the observed Universe does in fact seem to be rather nearly homogeneous and isotropic.

The three assumptions denumerated above – homogeneity, isotropy, and universality – seem reasonably innocuous, so that one might be willing to accept them without serious question. However, it is not at all obvious that such plausible assumptions will always lead to simple, physically sensible conclusions. As an instructive salutary counter-example, which illustrates the kinds of difficulties that can arise if one introduces even seemingly innocuous assumptions, consider what has come to be known as Olbers' Paradox, first discussed in 1826 by the German astronomer *Heinrich W. M. Olbers* (1758 - 1840). What Olbers demonstrated is that if one assumes that the Universe is homogeneous and isotropic and, in addition, static and infinite in extent, he or she is led to the problematic conclusion that the night sky must be essentially infinitely bright!

Olbers' conclusion follows simply from adding up the amount of light that we can see if, consistent with his assumptions, the Universe is homogeneous, isotropic, and unchanging, and we can see infinitely far away. Let $B(r)$ denote the total amount of light reaching an observer on the Earth from objects a distance r away. Granted Olbers' assumptions, how should this compare with $B(2r)$, the amount of light received from objects a distance $2r$

away? An observer on the Earth only sees $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ as much light from each object at $2r$ as from an identical object at r, since the light which the objects are emitting is spreading out over a sphere of radius $4\pi r^2$. However, there are $2 \times 2 = 4$ times as many objects separated by a distance $2r$ as are separated by a distance r, since the objects we see at a distance r are distributed uniformly over a sphere, again of radius $4\pi r^2$. The net result is that the observer sees just as much light from objects separated by a distance $2r$ as from objects separated by r. Each object at $2r$ gives off only one fourth as much light, but there are four times as many objects.

The obvious point then is that the number 2 was in no way special: The same logic dictates that, on the average, every distance contributes exactly the same to the total amount of light observed on the Earth. According to Olbers' assumptions, we should see just as much light coming from stars one billion, or even one trillion, lightyears away as from one hundred lightyears away! The more distant stars will appear much dimmer individually, but there are many more of them, so that their contribution to the total observed light should be the same. Everything that we see should therefore be dominated by an infinite number of infinitely distant stars.

Around 1900, a number of different thinkers suggested an ingenious resolution to this dilemma, namely that most of the light from very distant stars is absorbed by intervening dust and gas, so that these stars contribute less to the total observed brightness than they would otherwise. However, further thought indicates that this explanation is a swindle! Light carries energy, and if some of this energy is dumped into intervening gas clouds, these clouds will eventually be heated to the point that they themselves will be just as hot and bright as the stars that originally emitted the light.

The standard Hot Big Bang Model provides two resolutions to Olbers' Paradox. The first involves the idea that, contrary to Olbers' assumptions, the Universe is not infinitely old. One might thus expect to see light only from a finite region even if the Universe as a whole is infinitely large. For example, if the Universe is only about 20 billion lightyears old, one might expect the finite speed of light to imply that one cannot see anything more than about 20 billion lightyears away. (Actually, as discussed in a later lecture, the fact that the Universe is expanding while the light is travelling makes things a bit more subtle!) Even if there is an infinite amount of light being emitted by more distant stars, this light will not be observable. The other resolution is that, according to the Hubble Law, very distant objects are moving away from us at high speeds so that the light which we see emitted from them is subjected to enormous redshifts. This redshifting implies that the light detected from very distant objects has very low energy so that, even if it does reach us, it will contribute less to the total observed brightness than it would if it were not moving away.

Despite these simple resolutions, Olbers' Paradox does emphasise an important point, namely that, although distant objects appear individually to be very dim, there are so many of them that they can make an appreciable contribution to the totality of what we see. Indeed, in recent years an examination of the light emitted by very distant objects has played an important role in confirming the validity of the assumption of large scale isotropy. Back in the 1950's the principal evidence for large scale isotropy was the fact that, on the average, galaxies are moving apart in a fashion described by the Hubble Law, $v = Hr$, and, in particular, that the value of the Hubble constant H is independent of direction in the sky. However, since then three other, significantly stronger, tests of isotropy have emerged.

If one probes the heavens at x-ray wavelengths, one typically observes two sorts of signals, namely (1) concentrated emissions which can be associated with specific, often relatively nearby, objects (like x-ray binaries and/or quasars) and (2) a diffuse background associated with objects that are so far away that they cannot be resolved individually. A detailed analysis of the diffuse background demonstrates that, consistent with the assumption of isotropy, it is statistically identical in different directions. Thus, e.g., different cells in the sky of size 3 degrees by 3 degrees are statistically identical to better than 3%, and, if one looks at cells of size 90 degrees by 90 degrees they are identical to better than 1 part in 1000.

Another test of isotropy can be performed by mapping out the distribution of distant high redshift radio galaxies. In general very distant galaxies with $z > 1$ or so are too dim to be easily detected from the Earth. However, as discussed in an earlier lecture, some of these galaxies are strong emitters of radio emission, and this radio emission can in fact be detected out to large distances, corresponding to redshifts z significantly larger than 1. An analysis of the distribution of these bright radio galaxies again confirms the overall isotropy of the observable Universe, as well as the large scale homogeneity.

A third test of isotropy, which will be discussed more carefully below, is based on the observation that so-called empty space is filled with cold blackbody radiation at a temperature of about 2.7 K, this corresponding primarily to microwave wavelengths. This radiation is an important relict of an earlier, much hotter, Universe and its detailed properties can in principle provide important information about how galaxies originally formed. However, in the present context the key question is the degree to which the blackbody radiation is completely isotropic in the sense that the observed temperature is the same in different directions in the sky. The short answer to this is that the radiation is *not* completely isotropic. However, it *is* isotropic to about one part in a thousand, and most of the observed anisotropy can be attributed to the fact that the Milky Way galaxy is moving through space. Indeed, if one corrects for the motion of the Galaxy, one infers that, overall, the black body radiation is isotropic to a few parts per hundred thousand.

The basic idea underlying the Big Bang Model is that the Universe as we know it began some 10 to 20 billion years ago as an extremely hot and dense entity, the Big Bang or primordial fireball, and that it has been expanding and cooling ever since. Because the Universe was extremely hot at early times, the matter content was very different from what we see today. Thus, in particular, there were no atoms or nuclei, although there *were* large numbers of various sorts of subatomic particles which, today, are only observed in high energy accelerators. However, this high temperature, high density blob of material expanded and, like any expanding gas cloud, cooled as a result of the expansion.

As the temperature dropped, the exotic particles disappeared, e.g., by decaying into more standard particles or by matter-antimatter annihilation, until presumably the Universe came to be dominated by the protons, neutrons, electrons, photons, neutrinos, and antineutrinos which are believed to be the dominant sources of matter today (modulo of course the elusive dark matter!). The protons and neutrons then combined to form the nuclei of light elements, including deuterium, lithium, beryllium, and, particularly, helium. This led to a Universe comprised, as it is now, primarily of hydrogen and helium.

As the temperature dropped still more, the nuclei combined with electrons to make neutral atoms and, once these neutral atoms formed, it was possible for the matter to begin to evolve the sorts of structures that are currently observed, including stars, galaxies, and clusters of galaxies. The blackbody radiation observed today in so-called empty space is the relict of the primordial fireball.

At this stage, it is important to stress a serious potential misconception, namely the idea that the Universe is expanding *in something.* According to the standard Big Bang Model, the Universe is (by definition) everything and, as such, it is meaningless to think of it as evolving in something else. The incorrect idea that the Universe is expanding *in something* is predicated upon our ordinary Newtonian intuition, which presupposes fixed unchanging notions of space and time. However, a correct description of the large scale structure of the Universe necessarily involves general relativity. The crucial point then is that general relativity provides a self-contained unified description of the Universe, whereby the evolving spacetime is everything. It thus makes no sense to ask in what the Universe is expanding, any more than, in the context of general relativity, it makes sense to ask what precedes the initial high density Big Bang. Cosmologists can and have addressed the question of what preceded the Big Bang. However, this question must necessarily be asked outside the context of ordinary general relativity, since a consideration of what preceded the Big Bang entails a consideration of what preceded space and time as they are currently understood.

(Earlier in the twentieth century, many astronomers, including such greats as Eddington, tried to visualise the curved three-dimensional space that constitutes the Universe as expanding in some flat four-dimensional space, insisting that this facilitated their intuition. Maybe. However, they were forced to recognise that this surrounding four-dimensional space has no physical reality, and that this picture is no more than a mental crutch. Indeed, mathematicians know that there exist curved three-dimensional spaces which cannot be embedded in a surrounding flat four-dimensional space.)

As a useful analogy, one can view the three-dimensional spatial Universe as being like the two-dimensional surface of a spherical balloon which is being inflated (neglecting the hole through which the air is injected). Mathematically, one can think of this expanding sphere as an entity in its own right, independent of the space around it, just as one can study Euclidean geometry on the plane without invoking anything about a third dimension. More physically, a dumb ant stuck to the two-dimensional surface of the balloon might notice that the balloon is getting larger, e.g., in the sense that some nearby food residing on the surface is moving away. However, a recognition of this "Hubble Law" associated with the expansion of the ant's Universe would not necessarily entail any recognition of a third dimension.

Another important point is that if one views the surface of the balloon as everything, there is no meaningful notion of a "center" from which the balloon expanded. (Neglecting the hole into which air is blown to inflate the balloon) every spot on the surface of the balloon is completely identical. In the balloon analogy, the Big Bang corresponds to a point when the radius of the balloon is essentially zero in size, before it began to expand.

Despite these caveats, it *is* possible to construct a (not quite self-consistent) picture of the expanding Universe in more conventional Newtonian language. What this entails is the idea that the Big Bang involved a gigantic explosion, whereby material initially packed very close together expands and cools. The important point to emerge from this picture is that, as the material expands, it is fighting against the gravitational attraction which is trying to hold all of the material together. The net effect of this gravitational attraction is that the expansion slows down. The ultimate fate of the Universe therefore depends on the strength of the explosion: If the explosion is sufficiently strong, the gravitational attraction of the material will never succeed in halting the expansion and the Universe will continue to expand forever, becoming progressively larger. If, however, the explosion is relatively weak, the gravitational attraction will eventually halt the expansion and the material will recontract to a high density state. In this case, the Universe will only last for a finite period of time, eventually terminating in a Big Crunch. As discussed more carefully below, whether or not the Universe will expand forever is in principle a question that can be answered one way or the other on the basis of astronomical observations.

In trying to assess the extent to which the Big Bang Model has been confirmed observationally, it is important to distinguish between the basic paradigm and specific implementations. Given the accumulated evidence of the past seventy years, the basic Hot Big Bang paradigm – the idea that the Universe is expanding and cooling from a hot dense state, leaving behind the blackbody radiation as a relict of the primordial fireball – seems reasonably secure. However, various concrete models which have been proposed to implement this paradigm – including "spatially flat" Universes, inflationary cosmologies, hot and cold dark matter models, and the like – are significantly less secure. It seems likely that our detailed models of the Universe will change significantly in the next several decades. Indeed, recent observations facilitated by the Hubble Space Telescope and various other probes suggest that many of our models are too simplistic. However, most cosmologists would be quite surprised if the basic paradigm were shown to be grossly wrong. As recently as the mid 1960's, it was relatively easy to construct alternatives to the Big Bang Model that could account for most, if not all, the available observational data as well as the Big Bang Model. However, as more and more data has accumulated, these different alternatives have been ruled out observationally. At the present time, the Big Bang Model is the only known paradigm that has been developed into a successful theory which seems capable of explaining most, if not all, of what one sees.

As stated already, the observed microwave background, i.e., the ambient blackbody radiation at a temperature of about 2.7 K, is interpreted as a strong piece of evidence corroborating the Hot Big Bang Model. The reason for this is simple. The fact that we can see radio and microwave radiation from distant galaxies and quasars is concrete evidence that the Universe as a whole is not opaque to radio and microwaves. However, if the Universe is not opaque how is it that the ambient radiation assumes a near-perfect blackbody form? As discussed in an earlier lecture, black body radiation is presumably generated only in environments which are essentially opaque.
The Hot Big Bang Model provides a simple answer to this question. This model implies

that, when the Universe was younger, it was also much denser and much hotter; and that, sufficiently early on, it was hot enough that all the atoms were ionized. However, when atoms become ionized into charged particles, their electromagnetic interactions with photons are much stronger than when they are neutral, so much so that, as a result of collisions, one would expect the photons to evolve towards a blackbody spectrum. In other words, one expects that, when it was young and hot, the Universe was opaque, and that interactions between photons and charged particles led to a blackbody spectrum. This spectrum would of course correspond to a very high temperature. However, as the Universe expands and cools, the high temperature blackbody spectrum, characterised overall by high frequency photons, will be systematically redshifted down to a blackbody spectrum with lower frequency photons and, consequently, a lower temperature. In this sense, the microwave blackbody spectrum observed today is the relict of a much hotter blackbody spectrum generated when the Universe was young.

It is essentially impossible to explain the presence of the observed microwave background without supposing (a) that, for reasons unknown, the Universe was created with a near-perfect blackbody spectrum or (b) that, at some time in its early history, the Universe was much hotter than it is today. The first of these alternatives seems highly contrived and, as such, most cosmologists have embraced the latter alternative. The microwave background was first discovered independently in 1964 by two different groups: *Arno Penzias* and *Robert Wilson*, two scientists working on microwave detectors at Bell Laboratories, and *Robert Dicke, James Peebles, P. G. Roll*, and *David Wilkinson*, a professor and three students at Princeton University, who were the first to implement a systematic search for this radiation.

(The sociology of this discovery is interesting. When Penzias and Wilson, working at Bell Labs in New Jersey, first discovered their microwave signal, they had absolutely no idea what it meant; and it was only after conversations with Dicke and several other physicists down the road in Princeton that they recognised the importance of what they had found. The discovery was reported in two consecutive letters that appeared in *The Astrophysical Journal*, the first by Penzias and Wilson that reported their original discovery and the second by Dicke *et al*, which confirmed the discovery and explained what it implied. Penzias and Wilson subsequently received the Nobel Prize for the discovery of the cosmic microwave background.)

Before the discovery of the microwave background, the Hot Big Bang Model was one of several competing theories that were being advocated by different cosmologists. However, unlike its competitors, the Big Bang Model actually predicts that a microwave background should exist; and, for that reason, the discovery of the microwave background caused most cosmologists to reject those alternatives as unviable. Since that time, the general consensus in the cosmological community is that at least the general paradigm of a Hot Big Bang must be correct.

But how perfectly blackbody is the observed microwave background? To answer this question, it is useful to characterise what one would expect to see observationally if she were immersed in the midst of a region of space characterised by a completely perfect blackbody spectrum. If the observer in question were at rest relative to the radiation, she would measure the temperature of the radiation to be exactly the same in all directions of space. Thus, e.g., if she were to look at opposite directions in the sky and analyse the ambient radiation, she would see absolutely identical spectra. However, this would no longer be true if the observer were moving relative to the radiation. In that case, the Döppler effect would imply that the temperature in different directions should be slightly different. If the observer looks at the radiation in the direction in which she is moving, she will find the radiation systematically blueshifted to shorter wavelengths, so that the measured temperature will be slightly higher. Alternatively, if she looks in the opposite direction, she will find the radiation redshifted to longer wavelengths, this corresponding to a slightly lower temperature. In general, for an observer moving with a speed $v \ll c$, the change in temperature ΔT away from the value T observed by an observer at rest would satisfy

$$\frac{\Delta T}{T} = \frac{v}{c} \cos\theta,$$

where θ represents the angle between the direction of motion and the direction in which the temperature is being probed. This observer-dependent effect is known as a *dipole anisotropy.*

Careful analysis of the observed microwave background by a number of different groups, starting in 1969, indicates (a) that the observed spectrum in different directions is always almost exactly blackbody, but (b) that there *is* a measurable dipole anisotropy, with a fractional amplitude of about one part in a thousand. Assuming that this anisotropy reflects the motion of observers on the Earth relative to the radiation, one infers that the Solar System as a whole must be moving with a speed of about 370 km/sec, roughly in the direction of the constellation Leo. Correcting for the known motion of the Sun in the Milky Way and the motion of the Milky Way in the Local Group, this implies that the Local Group as a whole is moving with a speed of approximately 600 km/sec, roughly in the direction of Centaurus.

These observations are interesting at a fundamental level in that they suggest the possibility of identifying a preferred frame of reference for the Universe as a whole. General relativity, like Newtonian mechanics, implies that there is no way in which to determine fundamentally whether or not an object is at rest or simply moving with a constant velocity. In this sense, there is no preferred rest frame for the Universe. However, the microwave background as a whole does provide a preferred frame, the closest perhaps to a fundamental rest frame that can ever be identified. Aside from this philosophical implication, the observed velocity of 600 km/sec is interesting in that it is a relatively large value. The implication is that a fairly substantial chunk of the Universe, millions of lightyears across, is moving coherently with a relatively high speed with respect to the microwave background.

It is also important to determine whether, if one corrects for the dipole anisotropy, the observed distribution of radiation is a perfect blackbody. Dating back to the pioneering work of the Russian *Evgenii M. Lifshitz* in the 1940's, proponents of the Big Bang Model have recognised that the answer to this question should be: no! Cosmologists believe that the observed blackbody radiation must be slightly lumpy in order to explain the origins of structures like galaxies, clusters, and the like; and, by the late 1980's, they had predicted (for reasons to be described in a later lecture) that there must be additional irregularities in the microwave background at the level of one part in ten thousand or a bit less. These so-called *quadrupole anisotropies* were first detected in 1991 by the *COBE* (*COsmic Background Explorer*) satellite. At present, the best estimate for the temperature of the microwave background is $T = 2.731$ K, with quadrupole irregularities corresponding to variations in temperature of about 0.0003 K. The discovery of these irregularities is another important piece of evidence supporting the standard Big Bang Model. If these irregularities had not been observed, it would have been difficult to explain the origins of the observed large scale structure of the Universe. The particular form of the irregularities, an issue still under investigation, provides important clues as to the mechanism whereby galaxies formed in the early Universe. More on that in a later lecture.

As stated already, viewed Newtonianly the expansion of the Universe involves pieces of matter moving apart from each other and, as such, overcoming their mutual gravitational attraction. It follows that, to the extent that it is populated by ordinary sources of material, as the Universe expands and more gravity has been overcome, the overall expansion rate should decrease. In other words, objects should have moved apart more quickly in the past. The assumptions of homogeneity and isotropy imply that, at any instant in the history of the Universe, there must be a Hubble Law $v = Hr$ connecting distance r and average recessional velocity v. The fact that the rate of expansion is decreasing implies, however, that the value of H is changing. When the Universe was smaller, objects moved apart more quickly, so that the Hubble constant H was larger in value than it is today: although constant in space, H actually changes in time.
If an object is travelling at a constant speed v, it will, in a time interval T, travel a distance $r = vT$. It follows that a galaxy (or any other object) currently separated from us by a distance r and moving away with a constant velocity v must have been separated by zero distance at a time $T = r/v$ in the past. However, if one were to assume that v is given by the Hubble Law $v = Hr$, with H the current value of the Hubble constant, this would imply a zero distance at time $T = r/(Hr) = 1/H$. Independent of the separation today, the object would have been separated by zero distance at a time $T = 1/H$ in the past.
This assumes a constant speed v. If, however, the rate of expansion is decreasing, the object was moving away faster in the past. A little bit of mathematics then allows one to conclude that the object would have been separated by zero distance at a somewhat different time $T < 1/H$, the precise value of which is again independent of the current value of r. Presuming that T corresponds to the time elapsed from the Big Bang, when everything was piled up on everything else, this means that a knowledge of the current value of the Hubble constant enables one to infer an upper limit on how old the Universe can be!

Recall from a previous lecture that observations indicate a value of H between 50 and 100 km/sec/Mpc. It is therefore useful to write $H = 100h$ km/sec/Mpc, where h is a dimensionless number, presumably between 0.5 and 1.0. However, if one translates the quantity $1/H$ into years, one then infers an age for the Universe

$$T < 9.8h^{-1}\text{billion years} \approx 10h^{-1} \text{ billion years.}$$

If the true value of h is about 0.5, corresponding to a Hubble constant of about 50 km/sec/Mpc, this implies an age less than about 20 billion years. This is a number

consistent with theories of stellar structure, which indicate that globular clusters are between 10 and 16 billion years old, with (as of 1999) 10 to 12 billion years being the most probable range. However, if h is as large as 1.0, this corresponding to a Hubble constant of about 100 km/sec/Mpc, it follows that the Universe can be at most about 10 billion years old. This result would, if true, seemingly imply that globular clusters are older than the Universe, which of course is nonsensical. It follows that a value of h as large as 1.0 would pose a major problem for astronomers. However, a smaller value of $0.5 - 0.7$ or so *is* completely consistent with other estimates of the age of the Universe
The conclusion that $T < 1/H$ was particularly problematic back in the 1930's and 1940's when, based on the erroneous estimates by Hubble, H was believed to be as large as 500 km/sec/Mpc, this corresponding to an age $T < 2$ billion years. Indeed, this was one of the major motivations for the so-called Steady State Universe, an alternative to the Big Bang Model originally proposed in 1949 and accepted by many cosmologists until the discovery of the microwave background in 1964. More on that in a later lecture.

So will the Universe expand forever or will it instead recontract to a Big Crunch? Formulated in Newtonian language, the answer to this question involves a comparison of the typical speed with which objects are moving apart and the overall strength of the gravity which is trying to draw them back together. To answer this question observationally, one therefore needs to determine two pieces of data, namely (1) the value of the Hubble constant H, which sets the overall scale as to how fast things blow apart, and (2) the average density of the Universe, ρ, which determines how much gravity is operative. In terms of H, the current value of the Hubble constant, one can compute a critical density

$$\rho_{crit} = \frac{3H^2}{8\pi G},$$

where G represents the gravitational constant. The important question then is whether ρ, the average density of the Universe, is smaller or larger than ρ_{crit}, i.e., whether the ratio $\Omega = \rho/\rho_{crit}$ is bigger or smaller than 1. If $\rho > \rho_{crit}$ and $\Omega > 1$, gravity will eventually win and the Universe will recontract to a Big Crunch. If, however, $\rho < \rho_{crit}$ and $\Omega < 1$, the Big Bang Model predicts instead that the Universe will expand forever. A useful number to keep in mind is that, if $\Omega = 1$, the predicted age of the Universe $T = 2/(3H)$. This implies that a Hubble constant $H = 50$ km/sec/Mpc translates into an age of between 13 and 14 billion years. As the value of Ω increases, the predicted age of the Universe decreases. Big Ω means lots of mass, which means lots of gravity, which means the expansion is slowing down relatively quickly, which means that the time since the Big Bang is shorter.

So is the average density ρ bigger or smaller than ρ_{crit}? Unfortunately, this is very hard to say! The factor of 2 uncertainty in H implies a factor of 4 uncertainty in H^2 and there are even more important uncertainties in observational estimates of the average mass density of the Universe. An earlier lecture addressed the fact that, already in individual galaxies, there appears to be much more mass present than can be attributed to stars and other luminous sources of matter. This problem becomes even more acute if one probes the Universe on larger scales, corresponding to clusters of galaxies or superclusters. However, one thing is definitely true: If one only counts the luminous sources of material, the inferred mass density ρ_{lum} is much smaller than ρ_{crit}.

By inserting the proper value of the gravitational constant G, one concludes that the critical density

$$\rho_{crit} \approx 1.88 \times 10^{-29} h^2 \text{ gm/cm}^3,$$

where, presumably, h is between 0.5 and 1.0. This would seem an incredibly low density, corresponding to about $0.05h^2$ protons per cubic meter. However, the best estimates of the total amount of luminous matter yield a still lower density,

$$\rho_{lum} \approx 8 \times 10^{-32} h^2 \text{ gm/cm}^3.$$

Fortunately, the estimated value of ρ_{lum} also depends on h^2 and, hence, the unknown value of the Hubble constant. Why this is so is easy to understand. To determine the amount of luminous matter, one must estimate the brightness of distant galaxies containing the mass. However, the estimated brightness depends on the assumed distance to the galaxies, which in turn is typically estimated using the Hubble Law. The net conclusion is that, independent of the assumed value of H, $\rho_{lum}/\rho_{crit} \approx 0.004$.

If one estimates the total density of the Universe using tools like rotation curves which probe the dynamical effects of gravitational attractions, one invariably finds much larger values for the density. In this sense, there appears to be much more matter out there than actually meets the eye. Moreover, even these much larger values may be an underestimate of what is actually out there. These dynamical mass estimates aim to probe lumps of material on a given scale, be these galaxies or clusters thereof. However, if the matter in the Universe is so distributed as to be roughly homogeneous on that scale, it will not be probed by these dynamical tests.

This may perhaps account for the striking fact that the estimated value of ρ, and hence the inferred value of Ω, increases as one probes the distribution of matter on progressively larger scales. On the scales of individual galaxies, both spirals and ellipticals, the estimated density typically corresponds to a value $\Omega \sim 0.01$. On the scale of clusters, this number increases to about $\Omega \sim 0.1$, and, on the largest scales probed hitherto, corresponding to superclusters, $\Omega \approx 0.2$, with an uncertainty of about 50%.

At present, the best observational evidence suggests that $\Omega < 1$, so that the Universe will expand forever. However, mass estimates on larger scales yield values of Ω progressively closer to the special value $\Omega = 1$, and there are recent unconfirmed claims that, on the very largest scales, Ω is in fact very close to 1. In this sense, the issue of whether the Universe will expand forever must be considered unsettled at the present time. In any event, what *is* curious is the fact that the value of Ω is relatively close to 1. According to the standard Big Bang Model, there is no reason *a priori* that this should be the case. The average density of the Universe could equally well be much larger or much smaller than ρ_{crit}. The fact that Ω is as close to 1 as it is is one of the big puzzles of the standard model.

Turn now to a short thermal history of the early Universe. Here the first point to stress is that, sufficiently close to the initial Big Bang, our cosmological models, based on general relativity, must inevitably fail. Indeed, as formulated in the context of general relativity, the quantum mechanical Uncertainty Principle implies that, when the Universe was sufficientlly dense and hot, characterised by a temperature $T \sim 10^{32}$ K, the very notions of space and time break down. Just as quantum physics "fuzzes out" the ordinary notion of a particle in the sense that one cannot determine precisely where an object is or how fast it is moving, so also, when applied to general relativity, quantum physics must "fuzz out" the ordinary notion of spacetime.

Actually, our understanding of the Universe must fail already at much lower temperatures. Different modern theories of subatomic physics have made many competing predictions as to what one might expect to see at very high temperatures. However, none of these theories has yet been confirmed in detail by laboratory experiments, so that there is no way to know which, if any, is correct. The only thing about which cosmologists and physicists all agree is that, at very high temperatures, the energy density of the Universe was dominated by various sorts of particles that are "nearly massless" in the sense that their rest mass energy mc^2 is much smaller than the energy associated with their very high speed motions.

The highest temperature where we have concrete experimental confirmation of modern physical theories is roughly $T \sim 10^{11}$ K. According to the Big Bang Model, this temperature was reached when the Universe was approximately a second old, i.e., about a second after the Big Bang. At this stage, the Universe was probably composed primarily of photons, neutrinos, antineutrinos, protons, neutrons, electrons, and, in addition, electron-positron pairs. The first six of these are particles that are still abundant today. Although no longer present in appreciable quantities at the present, electron-positron pairs were abundant when the temperature was as high as 10^{11} K because there were enough high energy photons with energy in excess of the electron rest mass energy to permit appreciable pair production: simultaneously photons were being converted into pairs and pairs were reannihilating back into photons.

At this time, the density of the Universe was much larger than it is today, so that particles were packed much more closely together. This, together with the high velocities of the particles, meant that there were large numbers of collisions amongst the different particles (things close together that run around quickly tend to bang into each other a lot!), which facilitated a well-shuffled, near random distribution, i.e., a so-called *equilibrium distribution*. The net result was that the photons were characterised by a near-perfect blackbody spectrum, and that the neutrinos, antineutrinos, and other particles were characterised by the analogue of a blackbody spectrum appropriate for fermions. In each case, the form of the distribution was determined completely by the temperature.

As the Universe continued to expand and cool, the particles tended overall to move apart and those with a nonzero mass tended to slow down (recall that, for massive particles, high temperature implies large random motion!). This meant, however, that interactions between them occurred less frequently. This had an important consequence, namely that, at a temperature slightly below 10^{11} K, the weak interaction, the only non-gravitational interaction experienced by neutrinos and antineutrinos, became too weak to maintain an equilibrium between the neutrinos/antineutrinos and the remaining fermions. The net result was that the neutrinos and antineutrinos were effectively decoupled from the remaining material in the Universe. Since that time, they have simply expanded and cooled with their own fermionic version of a blackbody spectrum. According to the Hot Big Bang Model, there should be a microwave background of neutrinos and antineutrinos analogous to the microwave background of photons. However, because these particles interact so weakly with other types of matter, there is little hope of detecting them in the near future. The decoupling of the neutrinos from the other fermions was of crucial importance in yielding a Universe anything like what we actually see today. When the weak interaction is important, it facilitates such processes as

$$n \leftrightarrow p + e + \overline{\nu}_e,$$

where neutrons (n) can be created from, or converted into, protons (p), electrons (e), and electron antineutrinos ($\overline{\nu}_e$). The important point then is that as the temperature drops, the relative abundance of neutrons decreases, as more and more of them are converted into protons. Indeed, if the weak interaction remained important until the present day, there would be almost no neutrons in the Universe. This means that no atoms heavier than hydrogen could exist, so that life as we know it would be utterly impossible.

When the weak interaction "turned off" and the neutrinos became decoupled, the relative abundance of neutrons to protons was roughly 1 to 5, i.e., there was about 1 neutron for every 5 protons. However, this ratio slowly changed. A free neutron not stuck in a nucleus will decay into a proton, an electron, and an anti-neutrino with a *half-life* of about 1000 seconds, i.e., on the average, given any sample of free neutrons half will decay within about 1000 seconds. However, if a neutron gets stuck into a nucleus before this happens, it will remain safe from subsequent decay, remaining a neutron even to the present day.

What was the dominant source of energy in the Universe at a temperature of about 10^{11} K? In answering this question, it is important to recognise that the Universe contained many more photons, neutrinos, and antineutrinos than protons, neutrons, and electrons. Indeed, even today there are roughly a billion photons for every proton or neutron in the observable Universe, and six times as many neutrinos and antineutrinos (the number 6 comes from the fact that there are 3 generations of neutrinos, each with a corresponding antineutrino, and that all 6 of these different types of particle should be about as abundant as photons). Because photons, neutrinos, and antineutrinos were so abundant, they contribute significantly to the total energy density and, because there were large numbers of photons being converted into electron-positron pairs, there was also a large energy density associated with these pairs. The net result is that, at a temperature of 10^{11} K, there were three principle sources of energy, namely photons, neutrinos and antineutrinos, and electron-positron pairs. Protons and neutrons also contributed to the total energy content of the Universe, but their relative contribution was much smaller. Because they are much less abundant, their rest mass energy turned out to be relatively unimportant and, because they are much more massive than electrons and positrons, they moved relatively slowly, so that there was not very much energy associated with their motion.

If, at that time, one were to have examined a little cube in the Universe of side R, with volume $V = R^3$, one would have found that, on the average, each of the three principal sources of energy – photons, neutrinos and antineutrinos, and electron-positron pairs – would contribute a total amount of energy proportional to VT^4. Here the dependence on V just says that if one looks at a volume twice as large one should see twice as much stuff and hence twice as much energy. The dependence on T^4 reflects the Boltzmann Law, which states that, for a blackbody, the total amount of energy emitted from a given region in space is proportional to the fourth power of the temperature.

The important point now is that, according to the Big Bang Model, the temperature T of the Universe varies inversely as the size of the Universe. Thus, as the volume $V = R^3$ containing a given amount of stuff grows with the expanding Universe, its temperature $T \propto 1/R$. The net result is that the total energy within that *comoving volume* associated with photons, neutrinos, and pairs scales as $VT^4 = R^3T^4 \propto T^{-3}T^4 = T$. In other words, the energy within the given comoving volume is proportional to the temperature T and decreases as the Universe expands and cools.
As the temperature decreases, the photons will of course become less energetic and, once the temperature approaches 10^{10} K, they are sufficiently low energy that they can no longer trigger abundant electron-positron pair formation. The net result is that the existing electron-positron pairs decay into photons without being replaced by new electron-positron pairs. After this has happened, the photons and neutrinos still contribute significantly to the total energy density, but pairs no longer play an appreciable role.

Even at very high temperatures nuclear fusion can happen in the sense that nucleons, i.e., protons and neutrons, can stick together temporarily. But at these high temperatures there are so many high energy photons that larger nuclei, once assembled, are almost immediately broken to pieces by interactions with one or more photons. However, once the temperature dropped to about 1 billion K or so, the number of high energy photons decreased to the point that nuclei could remain stuck together, so that protons and neutrons could be combined into larger entities. This *nucleosynthesis* involves a complicated collection of nuclear reactions whereby hydrogen nuclei, i.e., individual protons, are converted first into deuterium (one proton plus one neutron), and then into various other elements including tritium (one proton and two neutrons), helium-3 (two protons and one neutron), ordinary helium-4 (two protons and two neutrons), lithium-7 (three protons and four neutrons) and beryllium-9 (four protons and five neutrons).
In point of fact, what happened is that almost all of the remaining neutrons were converted into helium-4, and that only small traces of the remaining heavier elements were made. The reason for this involves nuclear chemistry. It turns out that ordinary helium-4 is an extremely stable nucleus which, once made, tends to remain helium unless it is converted into carbon (as is the case for helium burning in high mass stars). The net result is that lighter stuff gets processed relatively efficiently into helium-4. However, once the material has been converted into helium-4, it is hard to get any bigger. There are no stable nuclei containing five nucleons, so that there is no way to stick a single proton or neutron into a helium-4 to make something else. Similarly, there are no stable nuclei containing eight nucleons, so that it is impossible to stick two helium-4's together to make something else. One *can* make traces of other nuclei by somehow bypassing these roadblocks, e.g., by combining two nuclei each containing three particles, but overall almost all the neutrons got channeled into helium-4. In particular, essentially no carbon, or any heavier nuclei, were made.
If the relative abundance of neutrons and protons had stayed constant since the time when the neutrinos decoupled from the other fermions, the net result of this epoch of nucleosynthesis would be a Universe which, by mass, was about 2/3 hydrogen and 1/3 helium. Detailed numerical calculations show that, in point of fact, the proportion of helium was somewhat lower, between 23 and 26% by mass, this because an appreciable

number of the neutrons had decayed before they were trapped in nuclei.

This nucleosynthesis, which began at a temperature of about 1 billion K, terminated relatively abruptly at a temperature of about 400 million K, when the Universe as whole was a few minutes old. The principal reason for this was that, as for the case of stars, when the temperature drops too low the electric repulsion between small positively charged nuclei prevents their sticking together to form yet larger nuclei. Only at high temperatures will lighter nuclei move with enough energy to over come the electric repulsion, allowing them to exploit the attractive character of the shorter range strong interaction.

The predicted relative abundances of the various light elements depends critically on inputs from modern theories of particle physics. For example, the apparent fact that there are three kinds of neutrinos, rather than two or four, is important in that it determines how much gravitating material the Universe contains, which is turn determines the rate at which the Universe expands and cools. As the Universe expands, particles move apart, so that they interact less frequently, which, together with decreasing temperatures, serves to decreases the overall efficiency of nuclear fusion.

The predicted relative abundances of the lighter elements can actually be tested by present day observations. What this entails is determining the current elemental abundances and, by correcting for what one anticipates would have happened as a result of nuclear fusion in stars, estimating what must have been present before nuclear fusion began in stars. Such a detailed analysis yields an excellent agreement between observations and theory, a fact which is decidedly *nontrivial*. If, e.g., there were only two generations of neutrinos, rather than three, the agreement between observation and theory would at best be marginal; and, if there were four generations observations and theory would definitely disagree. Thus, e.g., if particle physicists were to discover another generation of neutrinos, or some other massless, or near massless, particle which should again be characterised by a fermionic analogue of a blackbody distribution, this would pose a major problem for the Hot Big Bang Model.

After nucleosynthesis, the Universe continues to expand and the temperature continues to drop, but for a long time nothing very significant happens. However, once the temperature drops to around 10,000 K, an important transition occurs: At this point, the total energy associated with photons and neutrinos, which is proportional to the temperature T, become smaller than the rest mass energy associated with protons and neutrons. The net result is what is described as a transition from a *radiation-dominated Universe,* where the dominant source of energy is massless photons and massless, or nearly massless, neutrinos, to a *matter-dominated Universe*, where, at least for the stuff that is actually seen, the dominant source of energy is the rest masses of particles of ordinary matter – although there might still be other sorts of exotic material contributing to the so-called missing mass! Today, the photons have dropped to a temperature of about 2.7 K, which means that their energy, like that of the neutrinos and antineutrinos, is completely negligible compared with the rest mass energy of protons and neutrons. (These conclusions were derived originally at a time when physicists believed that neutrinos and antineutrinos are strictly massless. However, as noted in an earlier lecture, there is growing evidence that these particles *do* have a small albeit nonzero mass; and one might worry that such a mass would alter the basic conclusions. Strictly speaking, endowing neutrinos and antineutrinos with a nonzero mass *will* change these sorts of predictions, but these changes will be extremely small if the neutrino masses are as small as seems most likely in 1999. Unless the masses are appreciably larger than what is currently suspected, for most practical purposes neutrinos and antineutrinos behave as if they are essentially massless.)

At a slightly lower temperature, about 4000 K, it becomes possible for nuclei and electrons to combine to form neutral atoms. As discussed already, this so-called *recombination* means that photons and nuclei no longer interact very strongly, so that the Universe, originally opaque to electromagnetic radiation, become essentially transparent. At this stage, when the Universe is a few hundred thousand years of age, the matter and photons are largely decoupled, and it is possible for the ordinary neutral matter to begin evolving into the structures which we see today, including stars, galaxies, and clusters of galaxies.

Mathematical Appendix: How Fast Is the Universe Expanding?

According to the Big Bang Model, the rate at which the Universe has expanded over the course of time depends on both how much mass/energy is out there to gravitate and of what that mass/energy is comprised. Assuming, for simplicity, that the amount of material is such that $\Omega = 1$, the predictions are easily described. To the extent that the energy content of the Universe is dominated by the rest masses of ordinary particles like protons and neutrons, one infers that the size $R(t)$ associated with a comoving volume $V = R^3$ grows as

$$R(t) = R_0(t/t_0)^{2/3},$$

where R_0 is the size of that volume at the present time, t_0. Moreover, the current age of the Universe

$$t_0 = \frac{2}{3H},$$

where H is the Hubble constant, which determines the rate at the which the Universe is expanding today. If $\rho < \rho_{crit}$ and $\Omega < 1$ the rate at which the Universe expands decreases more slowly since there is less gravity to fight the expansion. It follows, therefore, that the present age of the Universe, i.e., the time since the Big Bang, $t_0 > 2/(3H)$. If $\Omega > 1$ the expansion rate decrease more rapidly and $t_0 < 2/(3H)$. Irrespective of the value of Ω, R grows more slowly than linearly in time, i.e., the rate of expansion is slowing down, so that $t_0 < 1/H$.

Other assumptions about the energy content yield different results. If, e.g., the energy content of the Universe is associated primarily with massless or near-massless particles like photons and neutrinos, as was presumably the case early on, for $\Omega = 1$ one infers a different expansion rate, $R \propto t^{1/2}$. $\Omega < 1$ again means more rapid expansion, and $\Omega > 1$ again slows things down. As before, irrespective of the value of Ω, R grows more slowly than linearly in time.

That the expansion of the Universe is slowing down, so that R increases more slowly than linearly in time, is guaranteed for any "normal" matter with the properties that the mass density ρ and the pressure P are nonnegative. According to general relativity, it is the combination $\rho c^2 + 3P$ that drives changes in the rate at which the Universe expands, and this rate must always decrease provided that $\rho c^2 + 3P > 0$.

These conditions seem so reasonable that one might perhaps expect that the expansion of the Universe has always been slowing down. However, various theories from particle physics suggest that this need not be so. Quantum mechanics makes the seemingly bizarre prediction – confirmed experimentally – that, under certain circumstances, the combination $\rho c^2 + 3P$ can be negative. This is, e.g., the fact underlying conventional theories of an Inflationary Universe, which assume that, at some early epoch in the history of the Universe, $P = -\rho c^2$, so that $\rho c^2 + 3P = -2\rho c^2 < 0$. This particular combination implies that, rather than growing more slowly than linearly in time, R will grow exponentially in time.

Modifying the laws of general relativity by introducing a so-called *cosmological constant*, as described in a later lecture, is equivalent physically to assuming that the vacuum, i.e., empty space, is really populated by a completely homogeneous distribution of some exotic matter for which $P = -\rho c^2$.

Lectures 35 and 36

THE LARGE SCALE STRUCTURE OF THE UNIVERSE

How truly homogeneous and isotropic is the Universe? As discussed in a preceding lecture, there is substantial evidence for homogeneity and isotropy on very large scales, deriving from the distributions of diffuse x-ray sources and radio galaxies and from the overall form of the microwave background. However, on smaller scales one observes substantial deviations from bulk homogeneity and isotropy. The existence of individual stars, galaxies, and galaxy clusters are all examples of structures inconsistent with the assumptions of homogeneity and isotropy, which indicate that, at least up to the scale of galaxy clusters, the current Universe exhibits significant irregularities.

That much has been well known for over fifty years. More recently, however, it has also been discovered that the Universe exhibits substantial structure even on somewhat larger scales. Specifically, concrete evidence has emerged indicating that, even on scales as large as 300 million lightyears or so, i.e., roughly 100 Mpc, there are large scale irregularities in the observed distribution of galaxies. Moreover, there also exists considerable evidence indicating that portions of the Universe of comparable size, 300 million lightyears or so, are moving coherently relative to the systematic expansion associated with the Hubble Law. Only on scales larger than 300 million lightyears does the Universe appear to be nearly homogeneous and isotropic.

In order to characterise the large scale structure of the Universe, as probed by galaxies, one needs to map out the locations of galaxies in three-dimensional space. It is relatively simple to determine the angular positions of objects in the sky, which yields information about two of the three dimensions. However, probing the third dimension requires an accurate determination of the distances to the various objects in question which, if they are far away, is difficult to obtain.

Because a direct determination of distances is difficult, most cosmologists trying to probe the large scale structure of the Universe have used the Hubble Law itself to obtain estimated distances. As discussed in an earlier lecture, this involves (a) measuring a Döppler shift for the spectral lines of some object of interest, (b) translating the observed redshift z into a velocity $v = cz$ (or, for larger z, using the appropriate relativistic generalisation of this formula), and then (c) assuming that the entire velocity can be attributed to the overall expansion of the Universe, so that $v = Hr$. Given this assumption, two angular positions and a relative velocity can be converted into a position in three-dimensional space.

There is an obvious potential danger in proceeding in this simplistic way. Specifically, if the Universe exhibits large scale irregularities in the distribution of galaxies, i.e., significant deviations from homogeneity, it could equally well exhibit large scale deviations from the Hubble Law. In other words, there could be large scale *peculiar velocities*. This has in fact led to slight problems for a number of different analyses of large scale structure. If for example, one uses this technique to estimate the distances to a number of different galaxies in the same cluster, the fact that different galaxies have substantially different peculiar velocities can make them seem to be significantly further apart than they actually are. However, it would appear that, in most cases, this naive translation from redshift z to distance r is not that serious a source of error.

The first significant concrete evidence indicating that the Universe is irregular even on very large scales was provided in 1981 by *Robert Kirschner, Augustus Oemler, Paul Schechter*, and *Stephen Schechtman.* What these workers found was the apparent existence of a so-called *void*, a large region in space which is significantly underpopulated in the sense that there are many fewer galaxies than one might expect in a region of that size. The original claim was for an enormous void located in the direction of the constellation Boötes, corresponding to a volume of about 1 million cubic Mpc, i.e., a region roughly 100 Mpc or 300 million lightyears on a side, which contains almost nothing. Later, more systematic, investigations have revealed that the void is not quite so large or empty as had been claimed initially. However, what *is* true is that the void corresponds to a large region of space, well over 100 million lightyears on a side, which contains very few large bright galaxies, although it does contain a substantial number of smaller, relatively faint galaxies.

An improved, systematic attempt to probe the large scale structure of the Universe was subsequently undertaken by three cosmologists at the Center for Astrophysics at Harvard University, *Valerie de Lapparent*, *Margaret Geller*, and *Jon Huchra.* Their basic aim was to effect a systematic investigation of a fairly large portion of the sky, performing a so-called *magnitude-limited survey.* The idea underlying such a survey is to focus on a given region in the sky and then measure the angular position and redshift of *every* galaxy that looks brighter than some critical minimum brightness. The advantage of such a survey is that, if the Universe were truly homogeneous and isotropic, one would know exactly what to expect.

Assuming a homogeneous distribution, as one looks further out to larger redshifts z and distances r, the observed number of galaxies at a given distance brighter than the critical minimum brightness will initially increase because (as in the discussion of Olber's Paradox) the total number of galaxies grows with distance as $4\pi r^2$. However, when one looks far enough away, an appreciable portion of the galaxies are dimmer than the minimum brightness for the survey, so that the observed number begins to decrease. The net result is that a plot of the expected number of galaxies as a function of distance takes a simple, easily calculated form. Comparing the actual observations with the theoretical expectation for a homogeneous distribution of galaxies provides a concrete way of quantifying the extent to which the distribution of galaxies is, or is not, nearly homogeneous.

Looking out into the sky, one can visualise a large *celestial sphere* centered around the Earth and extending out into space. By analogy with cartography on the Earth, one can then endow this celestial sphere with analogues of longitude and lattitude, which the astronomer quaintly terms *right ascension* and *declination.* Selecting a region in the sky to investigate corresponds to choosing a given range of right ascension and declination. The Center for Astrophysics (CfA) survey entailed focusing on a range in right ascension/longitude corresponding to 135 degrees, and then considering a number of different narrow wedges each corresponding to an interval in declination/lattitude of 6 degrees. The one additional complication here is one of terminology: For historical reasons, right ascension is typically measured in *hours*, rather than degrees, the 360 degrees of a circle being divided into 24 hours, each containing 15 degrees.

The data from the first wedge, published in 1986, are illustrated in Figure 35.1a, which includes 997 different galaxies with right ascension and declination in the specified 6 degree interval. In this Figure, the Milky Way galaxy is located at the sharp corner at the bottom. Right ascension sweeps from right to left, covering an interval of 9 hours or 135 degrees. The recession velocity v (which, from an earlier lecture, equals cz, with c the velocity of light and z the redshift) is plotted moving away from the Milky Way. Assuming a Hubble constant $H = 50$ km/sec/Mpc, a velocity $cz = 5000$ km/sec corresponds to a distance of 100 Mpc or about 326 million lightyears.

It is clear that this plot exhibits a significant amount of structure. In certain regions large numbers of galaxies are concentrated, whereas others contain few if any galaxies. The strong concentration at 13 hours right ascension, stretching out almost vertically, corresponds to the Coma Cluster, located roughly 350 million lightyears away. The fact that the galaxies in the Coma Cluster exhibit a large spread in redshifts is a reflection of the fact that the different galaxies in the cluster have a broad range of peculiar velocities. The resulting pattern of galaxies aligned so as to point towards us has been termed the *Finger of God* in the popular press.

More significant than the Finger of God is another elongated pattern, known as the *Great Wall*, which extends from a right ascension of 17 hours and a velocity of 10,000 km/sec to a right ascension of 8 hours and a velocity of 7,000 km/sec. The Great Wall looks to be a single coherent structure, although its three-dimensional character is unclear: On the basis of this single slice of the Universe, it is not possible to characterise completely the form of the three-dimensional clustering. In particular, it is not clear whether the observed concentrations of galaxies correspond to pieces of filamentary, nearly one-dimensional, structures or whether instead one is observing slices through something that is more sheetlike and nearly two-dimensional.

De Lapparent and collaborators originally argued on statistical grounds that one was most likely seeing cross sections of two-dimensional structures. However, this was only demonstrated unambiguously in 1988 and 1989 when Geller and Huchra had finished analysing two additional wedges. Combined with the first, these additional wedges, exhibited in Figures 35.1*b* and *c*, establish conclusively that, on the scale of hundreds of million of lightyears, galaxies tend to be distributed along a pattern of intersecting sheets, reminiscent of a honeycomb. Along the walls of the honeycomb, the density of galaxies is disproportionately high. Away from the walls, the density is much lower although some galaxies are still to be found.

Given the existence of these large scale irregularities in the observed distribution of galaxies, manifesting deviations from complete homogeneity, it is also natural to ask whether there are bulk streaming motions representing peculiar velocities of galaxies superimposed on the overall expansion of the Universe associated with the Hubble Law. Answering this question is much more difficult observationally. Since the whole point here is to look for local deviations from the Hubble flow, one cannot use the Hubble Law itself to determine the distances to various objects. Rather, one needs to come up with alternative techniques in terms of which to estimate distances.

Until recently this would have seemed impossible. However, starting in the mid 1970's it was discovered that, for both spiral and elliptical galaxies, there is a relatively tight correlation between the typical speed with which stars move in a galaxy and the intrinsic brightness of the galaxy: in galaxies that are brighter overall stars tend to move more quickly. This has led to the so-called *Tully-Fisher relation* for spiral galaxies and a corresponding *Faber-Jackson relation* for ellipticals.

These correlations are *not* as strong, e.g., as the correlation between period and luminosity for Cepheid variables but, nevertheless, they are strong enough to permit a reasonable estimate of distances to various galaxies. In principle, the idea is quite simple. Given a determination of the typical speed with which stars move, derived from measured Döppler shifts, one uses the Tully-Fisher or Faber-Jackson relation to estimate the intrinsic brightness. However, by comparing the intrinsic brightness with the measured apparent brightness, one can infer a distance to the galaxy.

Using the Tully-Fisher and Faber-Jackson relations, a group of seven cosmologists, the self-proclaimed *Seven Samurai* (also known as Alan Dressler, Sandra Faber, David Burstein, Roger Davies, Donald Lynden-Bell, R. J. Terlevich, and Gary Wegner) set out to measure the bulk motions of a large number of nearby galaxies relative to the overall expansion implied by the Hubble Law. Their basic conclusion, first announced in 1987, was that almost all the observed galaxies within a very large region of space, with a size of 300 million lightyears or more, seem to be moving relative to the Hubble flow in a more or less coherent fashion, with a velocity of about 600 km/sec. In particular, they observed that the data are consistent with the interpretation that all the galaxies in the sample are being attracted gravitationally towards some single enormous mass concentration, about 300 - 500 million lightyears away from the Milky Way, in the general direction of the constellation Centaurus (although other interpretations were equally valid). This they dubbed the *Great Attractor*, a name which, not surprisingly, played very well in the popular press.

When this result was first discovered, it was considered extremely surprising and, since the measurements required to obtain it were very difficult, it was greeted with a fair amount of scepticism. Particularly for this reason, over the past several years a number of other groups have also tried to probe these bulk motions, looking at different galaxies and using alternative techniques. These later investigations have served to corroborate the basic conclusion of the Seven Samurai, namely the existence of the large scale flow, although the details do vary somewhat from survey to survey. For this reason, there is an emerging consensus amongst cosmologists that large scale motions do indeed exist on the scale of a few hundred million lightyears, the same scale as that associated with large scale inhomogeneities. Whether or not these streaming motions are associated with a single mass concentration is unclear at the present time: enlightened betting seems to be against a single Great Attractor.

But how did these different large scale structures evolve? The basic theory of structure formation, accepted by almost all cosmologists, dates back to the ideas of the Russian *Evgenii M. Lifshitz* in 1946, which entailed an application to the Big Bang Model of earlier ideas of the Englishman *James H. Jeans*, originally formulated in 1900. This so-called *gravitational instability theory* seeks to interpret the formation of structure in the Universe as reflecting the effects of a very slight initial imbalance between pressure and gravity in the early Universe. The physical principles are the same as those involved in understanding the fact that, even today, in galaxies like the Milky Way clouds of gas are contracting and fragmenting into collections of stars.

Suppose that, early on, the Universe was nearly but not exactly homogeneous, so that it contained various irregularities. The basic question then is whether these irregularities would tend to smooth themselves out or whether instead they would evolve in such a fashion as to become more pronounced. Just as for the case of gas clouds in galaxies, this involves a competition between pressure, which would tend to blow apart higher density regions, and gravity, which would tend to make these higher density regions contract.

If no gravity were operative, small irregularities would simply execute low amplitude oscillations, qualitatively similar to the sound waves that propagate through a room when, using your vocal chords, you compress some air in your mouth. Allowing for a small amount of gravity alters the details of these acoustic waves, changing the frequency with which they oscillate but, overall, the qualitative behaviour remains the same. However, if gravity is strong enough, it will instead force slightly higher density regions to contract and become yet higher density and, similarly, make lower density regions become less dense. But the fact that the region has become higher density means that gravitational effects will be even stronger, so that the contraction will continue and actually accelerate. The net result, presumably, will be the formation of various structures like galaxies and clusters of galaxies which are much denser overall than the average density of the Universe.

To determine whether pressure or gravity wins, i.e., whether an initial higher density irregularity simply oscillates or instead starts to contract, one needs to assess the total amounts of gravity and pressure that are operative. The amount of gravity depends on the total mass of the higher density region, since this determines the amount of material that is available to gravitate. The total amount of mass within the given region depends on two inputs, namely the mass density and the size of the region. The amount of pressure depends on the temperature of the material in that region, which determines the typical speed with which the particles therein are moving, since, as described in an earlier lecture, higher temperatures and larger random velocities correspond to higher pressures.

What this implies is that, for a given temperature T and a given density ρ, there is a minimum size for a high density irregularity, both in length and total mass, in which gravity is strong enough to overcome pressure. In other words, the gravitational instability expected to lead to the formation of large scale structure can only proceed on scales larger than a certain critical scale. For given T and ρ, one identifies a minimum mass M_J, the so-called *Jeans mass*, and a minimum length R_J, the *Jeans length*.

But when can this process begin? Recall that, overall, there are many more photons than nucleons, roughly a billion times as many, and that, before the time of recombination, when atoms were ionized into charged particles, the photons and nucleons were strongly coupled to one another by the electromagnetic interaction. This strong coupling between nucleons and photons means that, if they clumped at all, the photons and nucleons had to clump more or less identically. However, the fact that there were, and are, many more photons than nucleons meant, effectively, that it was the photons, rather than the nucleons, that would decide whether they would clump. Before the time of recombination, the nucleons in the Universe were essentially slaves to the much more abundant photons. However, being massless particles, photons travel at the speed of light, which implies that they are very effective in resisting the attractive gravity. The net result is that, before the time of recombination, gravity cannot force a slightly overdense region to contract unless the region in question is extremely large. Specifically, at a time t after the Big Bang the Jeans length R_J, the size of the smallest region that will be susceptible to gravitational instability, is roughly $R_J \approx ct$, comparable to the total distance that a photon could travel within the entire history of the Universe up to that point.

However, after recombination the photons and the now neutral matter are essentially decoupled, so that the nucleons *can* evolve independently. Even though overdense regions of photons still cannot contract gravitationally unless the region is comparable in size to ct, the nucleons most definitely *can*. For these nucleons, the pressure with which they can resist gravity is now fixed by the relatively low temperature of the matter, $T \approx 4000$ K, and this leads to a much smaller Jeans mass and Jeans length. Indeed, assuming a temperature $T = 4000$ K and plausible estimates as to the average density of the Universe at that time, one infers that the Jeans mass, corresponding to the smallest scale on which structures can begin to form, is roughly $M_J \approx 10^6 M_\odot$. It is interesting, and probably not accidental, that this Jeans mass is comparable to the mass of typical globular clusters, which are known to be the oldest objects in the Milky Way and other galaxies.

In all of this, one still has to recognise that, if the Universe were completely homogeneous and isotropic, there would be no irregularities, so that gravitational instability could never lead to the formation of structure. Only by hypothesizing the existence of some small initial irregularities, present at the time of recombination, can one provide the "seeds" that subsequently evolve into the structures that exist today.

Precisely why such irregularities should have existed is, within the context of the standard Hot Big Bang Model, an unexplained mystery. Typically, the cosmologist tries to invoke some sort of unknown physical processes which were operative in the very early Universe, at a time when the Universe was so hot that we do not really understand what was going on. Indeed, this problem is more acute than one might initially recognise in the sense that, to account for what is seen, the amplitudes of the small irregularities that existed must be rather finely tuned! If, at the time of recombination, the amplitude of these irregularities were too small, gravitational instability would not have succeeded in evolving as much structure as one observes to exist today. Alternatively, if the amplitude of the irregularities were too large, the Universe would have evolved to become much more lumpy than what is observed today.

At the present time, one does not yet know exactly what the form of these initial irregularities must have been. However, detailed numerical modeling has demonstrated conclusively that, in order to evolve structures even remotely similar to what is observed today, the amplitude of the "seeds" at the time of recombination must have been about one part in ten thousand, or perhaps a bit less. If the amplitude were significantly larger or smaller the Universe would have evolved into something looking very different from what we actually see.

Within the context of the standard Hot Big Bang Model, there is no concrete reason why irregularities of precisely such an amplitude must have existed; and, as such, one might worry that the whole idea of large scale structure evolving via gravitational instability from a near-homogeneous Universe could be incorrect. Is there any real evidence indicating that, at the time of recombination, such finely tuned irregularities actually existed? Until recently the answer to this was: no! However, the quadrupole anisotropy in the microwave background detected by *COBE* has provided significant concrete evidence that these irregularities almost certainly did exist.

The idea here is simple. Before the time of recombination, nucleons and photons were strongly coupled, so that if the nucleons in the Universe were to exhibit irregularities of given form and amplitude, the photons should exhibit comparable irregularities with roughly the same form and amplitude. After recombination, the nucleons and photons were decoupled, and the nucleons began to evolve structures. However, the photons still could not evolve structures, any irregularities in their density simply propagating as waves that neither increase nor decrease in amplitude. What this means is (1) that the irregularities in the microwave background observed today should, modulo the overall cooling associated with the expansion of the Universe, be identical to the photon irregularities that were present at the time of recombination; and (2) that those photon irregularities at the time of recombination were very similar to the irregularities in the nucleons at the time of recombination. In other words, the irregularities in the microwave background detected by *COBE* are fossil remnants of the irregularities in the distribution of ordinary matter which must have existed at the time when galaxies first began to form.

The fact that the microwave background exhibits irregularities with a typical amplitude of a few parts in one hundred thousand thus implies that, at the time of recombination, the distribution of nucleons must have exhibited irregularities of comparable magnitude. This number is completely consistent with the previous assertion that, to form the kinds of structure that are observed today, the nucleons must have exhibited irregularities of amplitude one part in ten thousand or a bit less. The detection of the quadrupole anisotropy by *COBE* thus provides an important observational confirmation of the idea that large scale structures formed via gravitational instability starting at the time of recombination. If *COBE* had *not* detected these irregularities, cosmologists would have been forced to the conclusion that there is some serious flaw in the standard picture of structure formation.

Another important point is the fact that, by probing in detail the form of the present irregularities in the microwave background, one is essentially probing the detailed form of the irregularities in the distribution of ordinary matter at the time when structures first began to evolve. Indeed, a complete analysis of the irregularities of the microwave background, which has not yet been effected, would yield extremely strong constraints on the overall form of the initial irregularities which must have evolved into what we see today.

Proceeding within the context of the Hot Big Bang Model, and assuming that large scale structure evolved as a result of gravitational instability, understanding why we see what we see entails answering two basic questions: What is the physical origin of the irregularities that were present at the time of recombination, and how precisely did these irregularities evolve into what is observed today?

As stated already, most cosmologists have typically avoided addressing directly the first of these questions, supposing simply that the primordial irregularities were generated somehow or other, via some unknown processes, in the very early Universe. Until recently,

there seemed little alternative. However, over the past fifteen years a number of different theories of subatomic particle physics (none yet confirmed in the laboratory) have been elaborated to the point that they can, at least in principle, make concrete predictions as to the form of the primordial irregularities. One concrete setting in which this has been done involves the so-called inflationary Universe, which will be considered in a later lecture.
Mostly, however, the focus has been on how precisely these initial irregularities evolved into what one sees. In addressing this question, the cosmologist typically proceeds by making some specific assumptions about what the Universe was like at the time of recombination and then asking whether a Universe that had those general properties could evolve by the present time into a Universe resembling what is observed today. This work has involved detailed numerical calculations, using high speed computers that have been developed over the past forty years.

Broadly speaking, the kinds of structures that will evolve depend on three types of inputs: (1) certain bulk properties of the Universe, such as the average mass density and the value of the Hubble constant; (2) the detailed form and amplitude of the small initial irregularities which were present at the time of recombination; and (3) the matter content of the Universe, i.e., of what the Universe is made.

A knowledge of the average mass density and the Hubble constant is important because it permits an explicit prediction as to what the Universe would be like in the absence of any deviations from complete homogeneity and isotropy. Indeed, the whole idea of modeling structure formation in terms of gravitational instability is to understand precisely how small deviations from such an idealised homogeneous and isotropic Universe can evolve into what is observed today.
As a concrete example of why the details of the idealised homogeneous evolution matter, consider the importance of the rate at which the Universe expands: If the Universe were not expanding, or if it were expanding only very slowly, overdense regions would collapse very quickly and significant structures would form much earlier on. Indeed, in this case initial irregularities of amplitude a few parts in one hundred thousand would yield a Universe much lumpier than what is actually seen. However, it is an observational fact that the Universe is expanding quite rapidly, and this expansion serves to resist the tendency of overdense regions to collapse. One way of visualising this involves recognising that, while self-gravity is trying to pull bits of matter in an overdense region closer together, the overall expansion of the Universe is trying to stretch them apart. The net result is that these regions contract more slowly and structures only form significantly later on.
The critical input here is $\Omega = \rho/\rho_{crit}$, the average density of the Universe relative to the critical density (the value of which depends on how fast the Universe is expanding!). As discussed already, the best observational evidence would suggest that $\Omega \approx 0.2$. However, the inflationary model suggests instead that Ω should be almost exactly equal to 1. For this reason, cosmologists have explored systematically a range of possible values, extending from $\Omega \approx 0.1$ to $\Omega \approx 1$.

Specifying the overall amplitude of the initial irregularities is not enough. One also needs to determine the *length scale* of these irregularities. Are there only large scale irregularities, with very large characteristic lengths, or are there only shorter scale irregularities, or is there a more complicated pattern including irregularities over a variety of different length scales? The theory of gravitational instability implies that any irregularities larger than the Jeans length can evolve to form significant structures, but it does not say anything about whether such irregularities are present. If, e.g., one were to assume that only very large scale irregularities are present initially, he or she would guarantee that very large scale structures form first, i.e., that clusters of galaxies would form before individual galaxies or globular clusters.

Today most cosmologists proceed by assuming that, at the time of recombination, there were irregularities of comparable power over a broad range of length scales, extending from very small to very large scales. There are at least three different motivations for this, two older and one much more recent. The oldest motivation is associated with various speculations that these irregularities could have been formed via some sort of quantum mechanical processes in the very early Universe. This idea dates back to the 1960's, and has assumed increased importance in view of concrete predictions based on inflationary models. A second motivation is related to the observation that the Universe shows structure over a broad range of scales and that, in a certain mathematical sense, the amount of structure is the same from the scale of individual galaxies up to the scale of superclusters (the buzz word is *fractal*). It is natural to conjecture that the existence today of structures over a broad range of scales is a reflection of the fact that the initial "seeds" for these structures also came with a broad range of scales. More recently, this idea of a *scale-invariant spectrum of initial perturbations*, originally proposed by the American *Edward Harrison* and the Russian *Yakov B. Zel'dovich*, has also received corroboration from investigations of the microwave background. Specifically, the irregularities observed by *COBE* seem largely consistent with such a scale-invariant form, although the details of the irregularities in the microwave background have not yet been completely disentangled.

Determining what the Universe is made of is, at the present time, perhaps the most difficult problem for the cosmologist. As discussed in several earlier lectures, there is considerable dynamical evidence for some sort of dark matter but, since this mass has not yet been detected directly, one has no idea as to what it really is. Some of this dark matter may be ordinary stuff like protons and neutrons which, for some reason, have not been converted into luminous clouds or stars. However, many cosmologists believe that this matter may be something more exotic. If, e.g., one believes that Ω is larger than $0.1-0.2$, he or she seems forced to the conclusion that much of the dark matter is not comprised of ordinary protons and neutrons since, otherwise, the relative abundances of the light elements produced at the time of nucleosynthesis would be significantly different from what is actually seen.

One possibility, that was particularly popular during the 1980's, is the idea that neutrinos are *not* exactly massless. Indeed, in the late 1970's a group of physicists in the Soviet Union claimed to have demonstrated that electron neutrinos have a nonzero mass, being roughly one thousandth as heavy as electrons. The important point then is that, since there are so many more neutrinos than protons, neutrons, and electrons – roughly a billion of each type of neutrino and antineutrino for each proton! –, the total rest mass of all the neutrinos could in fact be appreciable, significantly larger than the total rest mass of these heavier particles, and thus account for most of the missing mass.

This initial claim has not been corroborated by experiments performed by other groups, so that a neutrino mass that large seems unlikely. However one cannot rule out the possibility that neutrinos *do* have a small but nonzero mass. Indeed, several different experiments performed in the late 1990's suggest that all three types of neutrinos have nonzero – and unequal – masses, albeit substantially smaller than what was believed in 1980. These experiments seem to corroborate an interesting speculation from particle physics, namely that neutrinos can actually change from one type to another, e.g., from muon neutrino to electron neutrino and visa versa, and that by observing these changes one can infer differences between the masses of different neutrinos like $\Delta m = m_{\nu_\mu} - m_{\nu_e}$.

But why does it matter what the missing matter really is? When structures first start to evolve, the details of what they are made of are largely irrelevant. All that really matters is the total amount of mass, which determines the amount of gravity that is operative, and the temperature, or typical speed of the individual particles, which determines the total amount of pressure. However, as gravitational instability proceeds and the density of the lumps increases, the detailed form of the matter eventually becomes quite important.

Ordinary atoms, which contain protons and electrons, interact via the electromagnetic interaction and, even though this interaction is relatively unimportant for low density material, it *does* play an important role when the density is sufficiently high. In particular, the electromagnetic interaction can play an important role in allowing a contracting blob of ordinary atoms to settle down into a relatively compact equilibrium configuration. By contrast, particles like neutrinos do not feel the electromagnetic interaction, which implies that, at high densities, they will behave rather differently.

If much of the missing mass were in the form of *massive neutrinos*, another important effect would also come into play. Neutrinos, like protons, neutrons, and electrons, are fermions, so that they satisfy the Exclusion Principle. However, as discussed earlier in the context of white dwarfs and neutron stars, this implies that, if they are packed too close together, they will experience a degeneracy pressure which resists any further compression. For protons, neutrons, and electrons, this degeneracy pressure is essentially irrelevant on scales as large as galaxies since this pressure only becomes substantial for extremely high densities and very small sizes.

However, if neutrinos do have a nonzero mass, their masses must be very small, much smaller than that of electrons; and, as noted already, they are extremely abundant, a billion or so of each type for every proton in the Universe. The crucial point then is that these facts imply that the conditions required for significant neutrino degeneracy pressure would be very different. Suppose, for example, that neutrinos have a mass that is about two thousandths the mass of an electron, i.e., $0.0005m_e$, and that there are enough neutrinos in the Universe to yield a neutrino density corresponding to $\Omega \approx 0.2$. One then infers that, because of degeneracy pressure, neutrinos cannot clump on scales smaller than several hundred thousand lightyears, which is bigger than the visible part of the Milky Way galaxy.

This latter fact has a striking implication: Suppose that some initial irregularity is comprised of both ordinary nucleons and massive neutrinos. In response to gravitational instability, both types of particles would initially contract in an essentially identical fashion. However, although the nucleons would continue to contract down to small sizes and high densities, forming galaxies, globular clusters, and the like, the neutrinos would stop contracting at a larger scale, somewhat larger than the luminous component of an ordinary galaxy. The net result would be a galaxy comprised of "normal" nucleonic matter, surrounded by an extended *halo* comprised of more "exotic" massive neutrinos.

This particular scenario is only one of many that have been proposed to explain the structure and origin of galaxies and, since the form of the missing mass has not yet been determined, one has no idea whether it is correct. Indeed, recent numerical work suggests strongly that, by themselves, massive neutrinos could not explain the types of structure that have been detected by the CfA survey and other similar investigations (even assuming that their masses were as large as believed in 1980!). This underscores what is perhaps the biggest dilemma of modern cosmology: Until one knows the identity of the dark mass which, apparently, dominates the overall mass density of the Universe, it will not be impossible to understand completely the origins of the structures that we actually see.
In principle, the average mass density of the Universe and the value of the Hubble constant can be obtained from the kinds of observations that are currently being performed. Moreover, the form of the primordial irregularities at the time of recombination can in principle be extracted from an analysis of irregularities in the microwave background by exploiting the results from *COBE* and other probes planned for early in the next millenium. However, until the identity of the missing mass has been established, a complete understanding of the origins of the observed large scale structure will not be possible.

Lectures 37 and 38

RECENT IDEAS ABOUT THE VERY EARLY UNIVERSE

Although the Hot Big Bang Model has become the generally accepted paradigm of cosmology, a small number of cosmologists still entertain other alternatives. Moreover, even after Einstein's development of general relativity in 1916 and 1917 and Hubble's subsequent discovery that the Universe is "expanding," the Big Bang Model has not always been the most popular cosmological model. Thus, in particular, during the 1950's a more popular alternative was the so-called *Steady State Model*, proposed originally in 1949 by three Englishmen, *Herman Bondi*, *Thomas Gold,* and *Fred Hoyle*.

The Steady State Model is based upon what its proponents called the *Perfect Cosmological Principle.* This Principle asserts that, on the largest scales, the Universe is homogeneous and isotropic in space, so that every point in space is statistically identical. In this sense, it is consistent both with observations and with the Big Bang Model. However, it also asserts that the Universe is homogeneous and unchanging in time, so that the statistical properties of the Universe are the same today as they were ten billion years ago and as they will be ten billion years in the future.

The fact that the Universe is expanding implies that, according to the usual laws of physics, its average density must decrease, so that its statistical properties will change in time. In order to reconcile the observed expansion of the Universe with statistical homogeneity, the Steady State Model is therefore forced to alter the basic laws of physics. This it does by hypothesizing the *continuous creation of matter*, i.e., assuming that, as the Universe expands, matter is continually being created from empty space in such a fashion as to ensure that the average density of the Universe remains unchanged.

This idea may perhaps sound completely crazy. However, to many scientists in the 1950's it sounded no crazier than the idea that the Universe had a beginning at some finite time in the past. Moreover, because most of the Universe is empty space, with a density *much* lower than the density of our environs, the amount of matter that needs to be created to maintain a constant average density would be relatively small, much smaller than anything that one might easily detect experimentally. In this sense, the hypothesis of continuous creation is not easy to either verify or disprove.

Why was the Steady State Model actually proposed? Here one can give at least two answers, one largely philosophical and the other extremely pragmatic. The philosophical motivation for the Steady State Model was, as hinted already, a strong repugnance on the part of many astronomers and physicists to accept the idea that the Universe has a beginning. Indeed, already back in the 1920's many of the leading cosmologists were involved in decidedly nonscientific debates over whether it were appropriate to assume that the Universe really is finite in age. Thus, e.g., in an exchange published in the journal *Nature* in 1931, Eddington argued that "philosophically, the notion of a beginning to the present order of Nature is repugnant," whereas his former student, *George Lemaitre*, since 1923 a priest in the Catholic Church, found this possibility "intriguing" in view of the new physics which such a possibility could teach us. Interesting, the name "Big Bang" was originally coined by Hoyle and other proponents of the Steady State Model as a mocking epitaph for the alternative theory.

More pragmatically, there was a major problem with the estimated age of the Universe. Until the 1950's, it was generally accepted that the correct value of the Hubble constant H is about 500 km/sec/Mpc. However, it was well known that, according to the Big Bang Model, the age of the Universe $T < 1/H$, so that, assuming this value for the Hubble constant, one concludes that the Universe is at most about 2 billion years old. This was a problem because, by 1930, physicists like *Ernest Rutherford* had used arguments involving radioactive decay to conclude quite definitively that the Solar System must be at least about 3 billion years old. By making the Universe unchanging in both space and time, and hence infinitely old, the Steady State Model avoided the potential embarrassment of a Solar System that is older than the Universe.

So how can one prove that the Steady State Model is in fact incorrect? Conceptually, the easiest way to do this is to demonstrate that, rather than remaining unchanging, the Universe has in fact changed systematically over the course of time. Until the 1960's there was little observational evidence that the Universe really is changing. However, since that time several different indications have emerged.

The first concrete evidence was provided by the fact that, when corrected for redshifts associated with their motion away from the Earth, distant galaxies look substantially brighter and bluer than do more nearby galaxies. When we observe very distant galaxies, we are of course seeing them the way that they were in the distant past, so that the fact that they look bluer implies that, at early times, galaxies were much brighter and bluer than they are today. Indeed, these observations are consistent with the hypothesis that galaxies were formed roughly 10 to 15 billion years ago (the exact time depends on what is assumed about the Hubble constant!), star formation happening rapidly in ellipticals but more gradually in spirals.

Another related piece of evidence reflects the distribution of quasars. One sees a large number of quasars at redshifts z between about 1 and 4 or so, but very few at much larger or smaller redshifts. The paucity of low redshift quasars is interpreted as implying that quasars have "turned off" at a relatively recent time. The absence of quasars at much larger redshifts is interpreted as reflecting the fact that, very early on, quasars had not yet formed.

A proof that the Universe is changing is, logically speaking, the most direct demonstration that the Steady State Model cannot be correct. Historically, however, a more compelling piece of evidence was provided by the discovery of the microwave background. If, as is assumed in the Steady State Model, the Universe was always the way that it is today, there is no way to explain the ambient low temperature blackbody radiation with which the Universe as a whole is innundated. However, according to the Big Bang Model the Universe was much hotter early on, hot enough that atoms were ionized. This implies that, early on, the Universe was opaque, thereby providing a mechanism to account for the evolution of a near-perfect blackbody spectrum.

That the Universe should be innundated with blackbody radiation at a very low temperature was first recognised in the mid 1940's by a number of different individuals, including *George Gamov, Robert Alpher*, and *Robert Herman,* who also speculated that many of the heavier elements had first formed very early on when the Universe had a temperature of about a 1 billion K. At that time, there was no evidence that such a microwave background actually exists, and the best technology was inadequate to facilitate a reasonable search. However, with the advent of improved microwave technology, such searches became possible; and, in the early 1960's, *Robert Dicke* and his graduate students began a systematic search for this radiation. They first found a signal in 1964, but were beaten out by a couple of months by *Penzias* and *Wilson*, who, while analysing the properties of a new microwave antenna, had discovered what they originally thought to be some spurious "machine noise."

The discovery of this microwave background convinced most cosmologists that the Steady State Model is wrong. However, a number of die-hards have still sought to defend this model, even to the present day. In order to do this satisfactorily, one needs to provide an alternative explanation for the origin of the observed blackbody radiation. To do this, a number of different ideas have been proposed, e.g., having long wavelength radio and microwave radiation interact with grains of dust in interstellar or intergalactic space. However, there appears to be no way in which to have a Universe sufficiently opaque to explain the observed, near-perfect blackbody form of the spectrum without making the Universe too opaque to allow us to see radio and microwave emission from distant radio galaxies.

Another alternative to the Hot Big Bang Model was proposed by Albert Einstein in 1922, and is sometimes known as the *Einstein world model*. As discussed in an earlier lecture, Einstein's original formulation of general relativity was effected in 1916. Later that year Karl Schwarzschild presented the first explicit example of a spacetime satisfying the equations of general relativity, namely the spacetime associated with a spherical black hole. Six years later, in 1922, Alexander A. Friedmann provided another explicit solution, this time corresponding to a spacetime associated with a Universe that is homogeneous and isotropic. The most striking feature about this solution was that it corresponds to a world model that necessarily expands or contracts. In particular, it was evident that general relativity implies that a homogeneous and isotropic Universe cannot persist for any finite time in a static state, without either expanding or contracting. That a homogeneous and isotropic Universe cannot be static had in fact been recognised in a more restricted setting by Einstein himself in 1917.

Today Friedmann's original solution is understood as providing the framework for one simple implementation of the Hot Big Bang Model, and it is assumed that the expansion calculated by Friedmann corresponds to the overall expansion associated with the Hubble Law. However, when it was first recognised that general relativity does not allow for a Universe that neither expands nor contracts, Hubble had not yet demonstrated that most galaxies are moving away from us. The theoretical prediction of an expanding Universe therefore seemed awkward and unnatural, and Einstein viewed the impossibility of a static cosmology as a potential embarassment for his theory.

For this reason, Einstein was led to propose a modified version of his theory of general relativity which *does* allow for the possibility of a Universe that, albeit homogeneous and isotropic, is also static. The way he did this was to modify the equations of general relativity by incorporating what he called a *cosmological constant*, usually denoted Λ.

Formally, Einstein's modification is simply a convenient mathematical trick. Physically, it corresponds effectively to the assumption that so-called empty space is endowed with a nonzero pressure and energy density. The important point then is that, if the relative sizes of this cosmological constant and the ordinary energy density of the Universe are selected judiciously, it is in fact possible to construct a solution to the equations of general relativity corresponding to a homogeneous and isotropic Universe that is also static.

Several years later, Hubble announced his first concrete evidence indicating that the Universe overall is in fact expanding. At that stage, it became apparent that the Einstein world model is wrong, and that Friedmann's spacetime model, computed using the original theory of general relativity, could perhaps serve as the basis for a more reasonable description of what is actually observed. Einstein's reaction was to refer to his proposed cosmological constant as the "biggest blunder" of his life. He had always viewed its introduction as an unnatural modification of the original, extremely elegant, theory of general relativity; and he was particularly chagrined by the fact that, had he stuck to his original theory, he would have been led to *predict* that the Universe is expanding before that fact was discovered observationally.

Even more embarrassing, perhaps, was the fact that, as was recognised within a couple of years, the Einstein world model is unstable. Precisely what does this mean? Neglecting the effects of quantum mechanics, it is possible at least in principle to balance a pencil on its point if the pencil is oriented in *exactly* the right way. However, as a practical matter balancing a pencil on its point is impossible since even the tiniest perturbation, associated with an incredibly small air current or anything else, will cause the pencil to fall over; and, for this reason, one does not expect to ever see pencils balanced on their points. A pencil balanced on its point is an example of an *unstable equilibrium*, and one does not anticipate that such unstable equilibria will ever arise in the real world.

In the modified theory of general relativity, allowing for a cosmological constant, the possibility of a static Universe arises for essentially the same reason that, according to either Newtonian gravity or ordinary general relativity, static stars can exist: An appropriate balance, or equilibrium, has been achieved between gravity and pressure. The problem, however, is that for the Einstein world model this balance is an unstable equilibrium. Even the tiniest mismatch between the values of the cosmological constant and the average mass density would alter the equilibrium, forcing the Universe to either expand or contract. Moreover, even assuming that the average mass density managed somehow to be chosen just right, irregularities reflecting the existence of stars and galaxies would quickly evolve into a Universe that is *much* lumpier than what is observed today,

With the discovery that the Universe is expanding in a fashion described by the Hubble Law, the original motivation for a nonzero cosmological constant evaporated. However, much to Einstein's annoyance, the idea of a nonzero Λ, once introduced, never quite disappeared. The standard implementations of the Hot Big Bang Model all assume that the cosmological constant $\Lambda = 0$, but it *is* possible to construct variants thereof that incorporate a nonzero Λ. These modified models do not yield a static Universe, but they do entail detailed predictions somewhat different from those constructed assuming $\Lambda = 0$.

The Hot Big Bang Model is a very successful theory in that it provides a relatively simple framework in terms of which to understand much of what is observed in the Universe as a whole: It provides a natural setting in terms of which to interpret the observed fact that, overall, galaxies are moving away from us in a fashion described by the Hubble Law. In addition, by virtue of its retrodiction that the Universe was hotter at earlier times, it provides a sensible explanation of the origin of the microwave background. Combined with inputs from accepted, and experimentally corroborated, theories of particle physics, it also yields concrete predictions as to the relative abundances of the light elements produced during a primordial epoch of nucleosynthesis, predictions which are consistent with observations.

Moreover, the Hot Big Bang Model facilitates concrete estimates for, and bounds on, the age of the Universe which, at least for relatively small values of the Hubble constant, are consistent with the ages of old objects in the Universe, such as globular clusters. By postulating an initial Big Bang some 15 or so billion years ago, it also provides a framework in terms of which to understand the observed fact that, over the course of several billions of years, the contents of the Universe, as characterised, e.g., by galaxies and quasars, are changing systematically.

Less compelling, perhaps, but also important is the fact that, given primoridal "seeds" of the right type at the time of recombination, it would appear possible to explain the large scale structure of the Universe, as inferred from the CfA survey and other similar analyses.

However, despite these successes there are several lacunae in the Hot Big Bang Model, not things that it gets wrong but, rather, things that it simply cannot explain.

The first of these is the so called *flatness problem*: Why is it that the average density of the Universe is close to the critical density, so that Ω is relatively close to 1? We do not know exactly what the value of Ω actually is, estimates ranging from about 0.1 to 1.0, but we *do* know that Ω is not very much different from 1. In particular, astronomical observations seemingly demonstrate conclusively that the value of Ω could not be as small as 0.001 or as large as 10. This, however, is curious because, *a priori*, there is no reason that Ω had to be anywhere close to 1. A value $\Omega = 10^{-12}$ or $\Omega = 10^{6}$ is equally consistent with Einstein's general relativity.

In fact, the problem is even more acute than one might first expect since, according to the standard Big Bang Model, the value $\Omega = 1$ is actually unstable. If Ω is exactly equal to 1 at some time in the history of the Universe, it will remain equal to 1 for all later times. However, if at early times Ω deviated even slightly from 1, the magnitude of the deviation $|\Omega - 1|$ will rapidly increase. The fact that today Ω is relatively close to 1 therefore implies that, early on, Ω was *extremely* close to 1. Thus, e.g., if $\Omega = 0.2$ today, so that $|\Omega - 1| = 0.8$, the value of $|\Omega - 1|$ must have been about 1/2000 at the time of recombination and as small as about 2×10^{-17} at the end of the epoch of primordial nucleosynthesis.

To see why this odd result is known as the flatness problem, one needs to understand that, according to general relativity, the value of Ω at any given instant is actually related to the curvature of space at that time. If $\Omega = 1$, space is flat in the sense described in an earlier lecture: initially parallel lines remain parallel, the shortest distance between two points is a straight line, and every triangle contains three angles adding up to 180 degrees. If instead $\Omega > 1$, the space is curved in on itself like the surface of a sphere, with a positive curvature. Indeed, the three-dimensional space predicted by general relativity is precisely the analogue of the two-dimensional surface of an ordinary sphere, only in one higher dimension. In this case, initially parallel lines can converge and the sum of the angles of a triangle is always larger than 180 degrees. Alternatively, if $\Omega < 1$, the space is curved out on itself like a saddle or hyperboloid, with a negative curvature, so that initially parallel lines diverge and the sum of the angles in a triangle is always less than 180 degrees.

There is a fundamental distinction to be made between the three-sphere, with positive curvature, and the other two cases. Flat space, with zero curvature, and a hyperbolic space, with negative curvature, both extend forever in all directions, thus corresponding to Universes that are infinite in volume. By contrast, the three-sphere, like the ordinary surface of a two-dimensional sphere, closes in on itself, so that it corresponds to a Universe that has a finite volume! Recall further that the positive curvature three-sphere obtains for a Universe with an average density above the critical density, and hence $\Omega > 1$. This implies that in this case the Universe only has a finite duration, eventually recontracting to a Big Crunch. Alternatively, the negative and zero curvature cases correspond to a Universe with $\Omega \leq 1$, this leading to world models that expand indefinitely. In this sense one can say that $\Omega > 1$ yields a *closed Universe*, of finite extent in both space and time, whereas $\Omega \leq 1$ yields an *open Universe*, of infinite duration in both space and time.

The fact that $\Omega > 1$ yields a closed Universe suggests the intriguing possibility that if one were to move far enough in some given direction, he or she would eventually travel all the way around the Universe and return to the point in space from which he or she began. The only problem with that idea is that, since the Universe is finite in duration as well as extent, it could be that there would be no time to complete this hypothetical trip. In fact, a careful calculation shows that this is the case: if, in a closed Universe, you were to start on a trip to circumnavigate the Universe, the Big Crunch would occur before you returned from whence you began. The only thing that can manage to circumnavigate the entire Universe is a photon that does not "see" any matter which, if it starts on its trip right at the Big Bang, will return to its starting point precisely at the time of the Big Crunch.

In the past, cosmologists have adopted various tacts to accomodate the fact that, apparently, Ω is so close to 1. The oldest, and perhaps least satisfactory, resolution is simply to say: That's the way it is! The Universe just happened to be created in such a fashion as to ensure that it is very nearly flat. This approach buries the problem by attributing the near-flatness to whatever or whoever decided what happened at the Big Bang.

An alternative tact, which was especially popular during the 1970's, was to invoke what is sometimes termed the *Anthropic* or *Anthropogenic Principle.* The idea here is to argue that the Universe that we see had to be relatively similar to what is actually seen since, otherwise, we would not be here to see it. If, e.g., the Universe were much denser, with $\Omega \gg 1$, its overall duration would be much shorter, too short perhaps for complex life like our own to evolve. Alternatively, if the Universe were much less dense, with $\Omega \ll 1$, structures like stars, galaxies, and clusters would have formed much more slowly so that, perhaps, life as we know it would be impossible at the present time. This particular viewpoint does not necessarily require the assumption that we constitute the only form of life possible in the Universe. What it does, however, assume is that, since we apparently exist, conditions in the Universe were such as to permit the development of life like our own.

Most cosmologists find the preceding type of argumentation too vague and philosophical for comfort, but it was begrudgingly accepted by many until around 1980 since the "that's the way it is" answer seemed the only available alternative. However, starting around 1980 a number of cosmologists have tried to concoct different variants of the standard Big Bang Model which could provide a dynamical explanation of the fact that Ω is close to 1. The basic tact here is to invoke recent ideas from the theory of microscopic particle physics to alter the assumed dynamics of the Universe at very early times, long before the time of nucleosynthesis, so that $\Omega = 1$ was a stable, rather than unstable, value. The key point here is that if, early on, the Universe was forced to evolve towards a value *extremely* close to $\Omega = 1$, it would today still remain very close to the critical value even though it has been evolving away from $\Omega = 1$ for a long time. One concrete implementation of this idea is provided by Inflationary Universes, which will be described below.

A second limitation of the standard Hot Big Bang Model is the so-called *horizon problem.* The key realisation here is that, because of the finite speed of light, if the Universe is of finite age we should only be able to see a finite piece of it, even if the Universe as a whole is infinitely large. (Strictly speaking, as described below, this conclusion can be violated if one allows for the existence of a cosmological constant or assumes a Universe where the expansion is driven by bizarre matter where the mass density or the pressure is negative.) Even Newton would agree that, if the Universe is only 10 billion years old, we should not be able to see anything further than 10 billion lightyears away. Similarly, when the Universe was much younger, the size of the region that could have been observed by some hypothetical observer, the so-called *cosmic particle horizon*, should have been much smaller. (See Figure 37.1.)

This leads, however, to an awkward point: Why should the Universe overall be very nearly homogeneous and isotropic? Normally in physics and chemistry, if one sees an object that is in a well-mixed and near-uniform state, what one sees can be attributed to interactions between the atoms and molecules of which the object is made. For example, the air in the room in which you are sitting is distributed in a roughly homogeneous fashion as a result of near-random collisions amongst the nitrogen and oxygen molecules and between these molecules and the walls of the room, which lead to the gas being in a well-shuffled state.

The problem, however, is that, within the context of the standard Hot Big Bang Model, these sorts of effects cannot be invoked to account for the observed homogeneity and isotropy on very large scales. If the homogeneous and isotropic region of interest at some given time is larger than the distance that light could have travelled within the entire history of the Universe up to that point, there is no way in which anything happening on one side of the region can have any effect on what happens on the other: since nothing travels faster than light, there is no mechanism to facilitate any *causal contact.* But if this be true, there is no physical mechanism which could cause the contents of that region to get completed shuffled up. So why do the two sides look the same? Once again, the only apparent explanation is that, for reasons unknown: That's the way it is! The Universe was created initially in a nearly homogeneous and isotropic state.

At this stage, one might perhaps ask: How really strong is the evidence that the Universe is and was completely homogeneous and isotropic over distances larger than the cosmic horizon? In fact, there is strong indirect evidence that, even as early as the time of nucleosynthesis, when the temperature of the Universe was as high as 1 billion degrees K, the Universe was homogeneous and isotropic over scales much larger than the cosmic horizon.

At the time of nucleosynthesis, the Universe was only a couple of minutes old, so that, according to the standard picture, light could only have travelled a few lightminutes since the Big Bang. One might therefore suppose that the Universe should be roughly homogeneous and isotropic on the scale of a lightminute or so, but more irregular on larger scales. However, given such irregularities the details of primordial nucleosynthesis would be different than for a completely regular Universe. Different parts of the Universe would have produced different relative abundances of the various light elements and, even more importantly, the "average" relative abundances produced by all these different regions would be different than what would result in a nearly homogeneous and isotropic Universe: If, e.g., half the Universe were much denser than average, and half much less dense than average, the net result of nucleosynthesis would be relative amounts of helium, deuterium, lithium, and so forth much different than what would arise if the entire Universe had the same density.

After the period of nucleosynthesis ended, as the Universe grew older different regions which early on were not in causal contact would eventually communicate and, as such, their contents could perhaps be shuffled up to yield a roughly homogeneous distribution. However, because the average effects of the nucleosynthesis were different than for an initially homogeneous Universe, the observed relative abundances would still be quite different from the standard predictions which, recall, are consistent with astronomical observations. If, at the time of nucleosynthesis, the Universe were not nearly homogeneous and isotropic over scales much larger than the expected cosmic horizon, the resulting abundances of the elements would be substantially different from what we see today.

Another indication of primordial near-perfect regularity is provided by the microwave background which, recall, is a fossil remnant of the Universe at the time of recombination, when neutral atoms first formed. The important points then are (1) that, especially correcting for the dipole anisotropy, the background looks almost identical in widely separated directions in the sky and (2) that, according to the standard picture, widely separated regions of space had not been in causal contact by the time of recombination. Once again, if, at the time of recombination, the Universe were not nearly homogeneous and isotropic over scales much larger than the expected cosmic horizon, we should see something much different from what we see today.

Finally, there is more direct evidence for near-regularity beyond the cosmic horizon even today through an examination of the distributions of quasars and radio galaxies at relatively high redshifts. The important point here is that, according to the standard picture, the distribution of high redshift objects in opposite directions of the sky probe regions which, even today, have never had any causal contact. Light has had a chance to travel to us from distant objects on one side of the sky, but it has not had time to reach distant objects on the other side.

Another problem, hitherto avoided in these lectures, is the origin of the finely tuned fluctuations, with an amplitude of a few parts in a hundred thousand, which apparently had to exist at the time of recombination to lead to the types of large structure that are actually observed. From whence do they ultimately derive? The cosmologist might envision some law of nature which implies that the Big Bang created a Universe that was completely homogeneous and isotropic. That would be deemed somewhat awkward, but something which one might be able to accept. However, assuming that the Universe was almost, but not quite, homogeneous and isotropic, and that the primordial irregularities were of just the right amplitude to give rise to what we actually see, strikes most cosmologists as exceptionally contrived. Much less awkward would be the idea of a concrete theory which would predict the existence of the irregularities which we know were required to yield the observed large scale structures and which, from recent observations of the microwave background, we know were there at the time of recombination.

The so-called *Inflationary Universe*, first proposed by a number of different workers around 1980, is at present the best developed theory that has been constructed that can (perhaps) explain these three otherwise unsolved puzzles. The earliest concrete implementation of an Inflationary Universe was probably that provided in 1978 by the Russian *Alexei A. Starobinsky*, who was led to this new picture largely in his attempts to understand what a quantum theory of gravity might entail and was not especially concerned about the fact that such a theory could solve these mysteries. A couple of years later, the idea was proposed independently by the American *Alan Guth*, who was apparently the first to fully appreciate that inflation could explain the flatness and horizon problems.

The basic idea of an Inflationary Cosmology is very simple. Stated suscinctly, one assumes that, at some time very early in the history of the Universe, long before the time of nucleosynthesis when the Universe was only a fraction of a second in age, there was a period in which the overall dynamics of the Universe was dominated by a source of matter and energy very different from the sources that are readily observed in nature today. In particular, one assumes that the dynamics was dominated by a so-called *vacuum energy density* which, as a practical matter, acts exactly like the cosmological constant Λ introduced, and then abandoned, by Einstein in the 1920's. The presumed physical origins of this vacuum density are difficult to explain without a fair background in particle physics. However, a rough explanation is provided by the statement that, under certain circumstances, the so-called vacuum, which is comprised of an infinite number of virtual particle-antiparticle pairs, can actually serve as a source of gravity, like other sources of matter and energy.

Einstein assumed the existence of a cosmological constant chosen to precisely balance the effects of ordinary matter, so as to facilitate the possibility of a static world model. This inflation does *not* do. Rather, what inflation assumes is that the effective cosmological constant associated with the vacuum energy density completely dominates the overall dynamics of the Universe. Expressed in Newtonian language, the vacuum density behaves like a source of negative mass whose gravitational effect is repulsive, rather than attractive. The net result is that the Universe blows apart extremely fast, in fact exponentially fast, so that even a very small spatial region in the Universe rapidly inflates to an absolutely enormous size. In particular, the dynamics associated with this effective cosmological constant has the strange property that, during inflation, the Universe expands faster than the cosmic horizon. This implies that, over the course of time, as inflation proceeds progressively smaller fractions of the Universe are in principle observable.

In order to understand how inflation solves the horizon problem, it is necessary to consider a bit more carefully what it is that we actually see when we look out into the Universe. If the Universe were not expanding, so that the distances between us and distant galaxies are essentially constant, the answer is more or less obvious. In this case, Newtonian physics and special relativity would both agree that, for example, the light that we see from a galaxy one billion lightyears away left the galaxy a billion lightyears ago and took a billion years to get here. However, it is observational fact that distant galaxies are moving away from us in a fashion described by the Hubble Law, a phenomenon interpreted by general relativity as implying that the very fabric of spacetime is expanding. What this means is that the object which we infer today as being one billion lightyears away was closer in the past, so that the light from it that we see had to travel a shorter distance. The net result is that, according to the Big Bang Model, light from an object currently one billion lightyears away took somewhat less than one billion lightyears to reach us, the details depending on precisely how fast the Universe expanded while the light was travelling through space. Assuming, e.g., that $\Omega = 1$ and (not correctly!) that the expansion has always been driven entirely by the rest mass of ordinary particles like protons and neutrons, we should today be able to see out to a distance $r = 2c/H$, where H is the Hubble constant.

One important point in all of this is that how far out we can see involves an interplay between two effects, namely (1) that light travels towards us with a velocity which must equal c and (2) that, because the Universe is expanding, objects emitting the light were closer to us in the past than they are today. The other important point is that, although general relativity implies that nothing, including light, can travel through space with a speed in excess of c, there is nothing in the laws of general relativity that preclude the possibility that *de facto* the very fabric of spacetime itself expands faster than the speed of light. If, however, spacetime expands that fast it follows that, over the course of time, one actually sees a progressively smaller, rather than larger, region of comoving volume: pieces of the Universe that were visible at some initial time blew away so fast that, later on, they were no longer visible.

Consider now some tiny piece of the Universe, much smaller than the size of the cosmic horizon at the outset of inflation. Because this region is much smaller than the horizon size, it is easy enough to invoke ordinary physical processes like collisions between particles to explain why it should correspond to a nearly homogeneous and isotropic entity. Now allow for a period of inflation during which this tiny region expands exponentially into an incredibly large volume. If the inflation proceeds sufficiently rapidly and for a sufficiently long time, the net result is that the originally tiny region is inflated to a size much bigger than the cosmic horizon at that time or even the size of the cosmic horizon today.

This basic idea provides a simple and elegant solution to the horizon problem: Inflationary models assume that the entire observable Universe of today was actually in causal contact at very early times, thereby permitting it to evolve to a near-uniform homogeneous and isotropic state. However, as a result of the extremely rapid expansion during an epoch of inflation, most of it fell out of causal contact and is only now coming back within the cosmic horizon. This inflationary scenario implies that the Universe should *not* be completely homogeneous and isotropic on the very largest scales. However, the very large scales on which the Universe is lumpy are much larger than the cosmic horizon today, so that they are inaccessible observationally. In this picture, the Universe is lumpy on short scales, because of gravitational instability, and on the very largest scales, because of an absence of causal contact. However, on intermediate scales, corresponding to the largest scales that are accessible observationally, the Universe should be very nearly homogeneous and isotropic.

Inflation also provides a simple resolution for the flatness problem. Suppose, e.g., that the Universe is closed, so that, at any instant of time, space corresponds to a three-sphere. If the radius of this sphere is very small, the fact that space is curved will be readily apparent and, according to general relativity, the density parameter $\Omega \gg 1$. However, if that same sphere is much larger, the curvature will be much less apparent and the density parameter Ω will be much closer to 1. This is completely analogous to what one would infer for the two-dimensional surface of an ordinary sphere: The Earth is of course nearly spherical but, since its radius is relatively large, in everyday life we are seldom aware of the fact that the Earth is not flat. If, however, we lived on the moon Deimos of Mars, which is only a few kilometers in radius, the curvature would be apparent and we would perceive a world that is far from flat. The same argument also works for $\Omega < 1$. When any smoothly curved space is dilated to a larger volume, it will, when viewed on the same scale, appear more nearly flat than it did originally.

Although this was not recognised by Guth and other early proponents of the idea, inflation also holds forth the prospect of explaining the origins of the irregularities that must have been present at the time of recombination. The same theories of particle physics that predict that an epoch of inflation should happen also imply that, as that epoch ends, the nearly homogeneous and isotropic Universe generated by the rapid expansion should develop small but finite irregularities as a result of quantum mechanical effects.
The precise details are difficult to calculate, and the results of the calculations depend to a certain extent on the specific way in which inflation is presumed to work. Two things, however, appear to be clear: (1) The irregularities generated at the end of inflation should be characterised by a nearly scale-invariant spectrum, consistent with earlier thinking, and, more pragmatically, with observations of the microwave background. (2) It *is* possible that the amplitude of these perturbations will be consistent with what is required for galaxy formation and what is inferred from the microwave background, although such an agreement is not necessarily guaranteed.
(Making the period of exponential expansion end in a reasonable fashion, known as the "graceful exit problem," is one of the hardest tasks confronting any theorist interested in constructing an inflationary model of the Universe. As the Universe expands exponentially, the mass density is diluted exponentially, so that the Universe rapidly approaches a cold, near-vacuum state. The end of inflation must somehow transform the vacuum energy that powered the exponential expansion into particles which can replentish and reheat the Universe.)

It should, perhaps, be stated explicitly that the idea of an exponentially expanding Universe is not new to this inflationary scenario. Indeed, shortly after Einstein proposed his cosmological constant, it was recognised explicitly by the Dutch astronomer and mathematician *Willem de Sitter* (1872 - 1934) that, neglecting all ordinary sources of matter, Einstein's modified theory of relativity leads to a Universe that expands exponentially in time. What makes the Inflationary Universe new is (1) that the effective cosmological constant is associated with properties of matter assumed to exist in the Universe, rather than with some universal property of spacetime, and (2) the recognition that a phase of exponential expansion can provide a natural explanation of the horizon and flatness problems.

All of this sounds extremely promising. However, it should be stressed that, at the present time, there is no concrete evidence from experiments in accelerators and the like that the theories of particle physics that predict an Inflationary Universe are in fact correct. Nevertheless, it *is* significant that most generalisations of the Standard Model of elementary particles to have been proposed in the last twentyfive years or so do lead to the concrete prediction that, under certain conditions, the vacuum will behave like a cosmological constant. As stressed, e.g., by Yakov B. Zel'dovich, this would suggest that, if the current speculations of particle physicists are even roughly correct, something like inflation should have existed at some time or other in the early Universe. In this sense, inflation should perhaps be viewed as a general paradigm, which would be expected to arise in one form or another, rather than a concrete model, the details of which one has reason to trust.

Lecture 39

A RECAPITULATION OF THE COURSE

The earliest cosmological models were based typically on religious and/or mythological considerations, rather than on anything that would resemble what is currently understood as science. They were almost always geocentric, postulating a relatively small Earth- and people-centered Universe. The Sun, the Moon, and the other planets were recognised as distinct individual entities moving through the heavens. However, the stars were usually viewed as "frozen," e.g., attached to a sphere in the sky. Viewed on the large, most advances in cosmology since that time can be interpreted as involving the development of a more scientific approach, the use of various advances in science to better understand what it is that the Universe contains, and as a result thereof, the development of a progressively less anthropic cosmology.

The Pythagorean philosophers in ancient Greece, and their various followers, were apparently responsible for the first effort (at least in the West) to replace these earlier mythological, anthropically-oriented cosmologies by something more fundamentally scientific in nature. This culminated with the world model introduced by Aristarchus, which postulated correctly the fact that it is the Sun, rather than the Earth, that is at the center of the Solar System. However, later Greek thinking was dominated by more conservative and anti-empirical philosphers, especially Plato and Aristotle. This intellectural transformation led to the rejection of Aristarchus' heliocentric Universe, although Aristarchus's ideas were not completely forgotten. At the end of the ancient world, before the fall of the Western Roman Empire, the accepted cosmological world model was the geocentric model perfected by Ptolomy, which pictured a complicated, mechanistic Universe, based on an intricate pattern of epicycles and deferents.

After the fall of Rome, much of ancient thinking was lost and forgotten, at least temporarily, and the West was dominated by a distinctly anti-pagan bias. However, the work of many of the ancient Greek and Roman thinkers was eventually recovered and, by the late Middle Ages, the Ptolomaic world model and the physics of Aristotle were firmly re-entrenched as the accepted orthodoxy. However, Ptolomy and Aristotle were *not* accepted universally by all thinkers of the time, and the alternatives proposed by Aristarchus and others were not completely unknown.

The transformation to a more modern cosmology, based on a heliocentric Solar System, began with Copernicus, who showed in particular that it is possible to construct a heliocentric version of the Ptolomaic world model, again based on an elaborate set of epicycles and deferents. The Copernican world model displaced the Earth from its preferred location at the center of the Universe. However, it was still archaic, and arcane, from a modern perspective: it was still motivated by the physics of Aristotle, attaching the planets and stars to heavenly spheres, and it still failed to appreciate fully the overall size of the Universe or how stars fit into the Big Picture.

Transforming the Copernican idea of a heliocentric Universe into a realistic description of (at least) our Solar System required both improved astronomical observations and new physical ideas. The former were provided by Tycho Brahe who, over a twenty year period, assembled a careful data base which was then exploited by Johannes Kepler to derive

his three empirical laws that characterise how the planets move around the Sun. With his pioneering use of the telescope, Galileo also played a significant observational role by discovering concrete evidence that certain basic tenets of the Ptolomaic world model could not be correct. However, even more importantly, his simple but ingenious experiments in mechanics led him to the principle of Galilean relativity and, thereby, set the stage for the formulation of a systematic, predictive theory of physics.
This systematic formulation of physics was provided by Newton, who postulated a specific kinematical description of the Universe, based on absolute notions of space and time, and then used this kinematics to formulate a theory of dynamics which was able to explain *why* it is that objects on the Earth, and in the heavens, do what they appear to do. In particular, Newton showed that his theory of physics predicts that, to a high degree of approximation, the orbits of the planets are ellipses; that it explains changes in velocity as a planet moves closer to or further away from the Sun; and that it accounts for the relative periods with which different planets orbit the Sun.

This Newtonian formulation of physics facilitated a concrete view of the Solar System, relatively similar to what is accepted today. However, although it made the notion of stars attached to a celestial sphere appear an absurdity, it did not yet provide any alternative interpretation as to how stars fit into the Big Picture. Without an understanding of the distances to the stars, one could have no idea as to what they are or how they are distributed throughout the Universe.
Direct determinations of the distances to nearby stars rely on the measurement of very small parallax angles and, as such, were impossible until the development of improved optics in the early part of the nineteenth century. However, by 1840 parallaxes had been measured for a number of relatively nearby stars, thus providing a first concrete indication as to the enormous size of the Universe. The second half of the nineteenth century witnessed the first serious scientific speculations on the problem of large scale structure, e.g., the recognition that stars could be grouped in clusters, the notion that we are located in something like a galaxy, and the speculation that other galaxies might also exist.

By 1900 there had evolved a cosmology which, although very different from that of Copernicus or even Newton, was still wrong in several important ways. The cosmology of 1900 missed the correct overall shape of the Milky Way, and assumed incorrectly that the Sun is at or near the center of our galaxy. It provided an estimated size of 30,000 lightyears for the galaxy, substantially smaller than what is currently accepted. It also had nothing concrete to say about whether other galaxies exist. Moreover, there was as yet no clear sense as to whether the Universe is infinitely old, or whether instead it began at a finite time in the past. All that one really knew there was that, if evolving theories of geology were correct, the Earth should be a billion years of age or more.
To correct these lacunae and misconceptions required developing new physical ideas and then applying them to the Universe as a whole.

Three basic revolutions in physics played an important role. The first was the development of special relativity, as formulated for the first time in 1905. This theory involved a fundamental change in the Newtonian world model in the sense that Newton's notions of absolute time and space were replaced by a combined entity, spacetime, this reflecting

the fact that different observers will actually perceive time and space in different ways. Newtonian physics was built from two principal ideas, the notion of absolute space and time and the notion that only relative motion is fundamental. Accepting both of these implies that different observers will measure light travelling at different speeds. However, it is an empirical fact that different observers will not measure the speed of light to be different, so that one of Newton's two ideas had to be surrendered. It is still believed that the principle of relative motion is fundamental, but the notion of absolute space and time is no longer accepted.
Special relativity leads to several important predictions which have direct astronomical implications. Most obvious, perhaps, is the fact that nothing can travel faster than the speed of light, which implies an absolute limit as to what parts of the Universe we can ever hope to probe. Special relativity also leads to the prediction that matter is a form of energy which can be transformed into other forms. This leads in turn to an explanation of the phenomena of fission and fusion, the latter playing a fundamental role in the structure and evolution of stars like the Sun.

The second new ingredient was general relativity, as formulated in 1916. This theory subverts the Newtonian world model in an even more profound sense in that it rejects the idea that there is a fixed arena of space and time, in which the physics of the Universe unfolds. Rather, it leads to a picture whereby the very fabric of spacetime changes in response to the presence of matter and/or energy. This theory was motivated by the need to incorporate the Equivalence Principle, first appreciated fully by Galileo and tested with increasing precision over the next several centuries, into special relativity, which had no place for gravity. It leads to the prediction of black holes and other exotica which apparently exist in nature. It also provides a concrete mathematical framework in terms of which to formulate the theory of a dynamical Universe.

Despite their radical innovations, both special and general relativity are still classical theories in the sense that they embrace the ordinary notion of particles at fixed locations moving with specific velocities, although these particles evolve in a different spacetime continuum. However, even that piece of Newtonian thinking fails with the introduction of quantum mechanics, which was first formulated systematically in 1927. Quantum mechanics is absolutely crucial to our understanding of the microscopic world of atoms and molecules, providing a setting in which to understand the physical significance of temperature as random motion; the fundamental nature of light and how it interacts with other forms of matter and energy; and, thereby, the origins of the blackbody radiation emitted by most objects which we see, including both planets and stars, as well as so-called empty space.

Using these new ingredients – special relativity, general relativity, and quantum mechanics – twentieth century physicists, astronomers, and cosmologists have been wildly successful in piecing together a unified theory of the large scale structure of the Universe that seems largely in accord with what one actually sees.
The basic observable properties of stars had been reasonably well established by about 1915, with the development of the Hertzsprung-Russell diagram. Moreover, it was recognised that, at one level or another, whatever is happening in a star must reflect a com-

petition between gravity and pressure. However, only in the 1940's and 1950's, with the development of comprehensive theories of nuclear fusion, did a clear understanding of stellar evolution begin to emerge. It is now recognised that the entire life history of a star can be understood as an ongoing fight between gravity, which is always present, and various sorts of pressure, provided by nuclear fusion and/or degeneracy. In particular, it is recognised that a single star will, over the course of its life history, start as a main sequence star, then evolve into a red giant, next transform itself into a supergiant, and, finally, end its existence as a white dwarf, a neutron star, or a black hole. Moreover, one knows that the dominant factor that sets the overall rate at which a star evolves, and determines the final corpse into which it will evolve, is the mass of the star, which fixes the amount of gravity that some source of pressure must resist.

By using observed and inferred properties of stars, it has also been possible to map out the large scale structure of the Universe. For example, the period-luminosity relation for Cepheid variables, developed by Leavitt, played a critical role in Shapley's pioneering work in 1918 on determining the size and shape of the Milky Way and on Hubble's decisive proof in 1924 that Andromeda is a distinct galaxy, outside the Milky Way.

The fact that the Universe is "expanding," i.e., that distant galaxies are systematically moving away from us in a fashion described by the Hubble Law, implies that, unless the currently accepted laws of physics are grossly in error, the Universe was very different in the past. At earlier times, the Universe must have been both hotter and denser, with the temperature and density both becoming infinitely high at a finite time in the past. The time since this original infinite density state is bounded by the current value of the Hubble constant. Assuming a Hubble constant of about 50 km/sec/Mpc, as suggested by many recent measurements, it follows that the Universe must be at most 20 billion years of age. This number is consistent with work in stellar evolution that indicates that the oldest stars that we see are at most 14 to 16 billion years of age, and probably less.

The founders of the Big Bang Model had relatively little direct evidence to justify their theory. However, motivated largely by philosophical biases, notably the ideas that the Universe as a whole should be homogeneous and isotropic and that the laws of physics are unchanging, and a belief in the mathematical elegance – and rightness – of general relativity, they constructed a unified framework which has, since that time, led to numerous physical predictions that were subsequently confirmed.

Combined with various assumptions about particle physics, the Big Bang Model makes specific predictions as to the relative abundances of the light elements that were present when the Universe was young, predictions that are confirmed by observations. By the mid 1940's, it was also recognised that the Big Bang Model predicts that so-called empty space should be innundated with blackbody radiation at microwave wavelengths, a fact that was only proven to be true in 1964, nearly twenty years later.

Granted the assumption that the Universe was slightly irregular at the time of recombination, the Hot Big Bang Model also constitutes a framework in which, by invoking gravitational instability, one can develop a theory for the origins of the large scale structures that are actually seen. Detailed numerical experiments over the past twenty years suggest that the gravitational instability picture can probably explain the large scale structure that is actually observed. However, the details are not completely understood and, most probably, will not be understood until the identity of the missing mass has been determined.

The most recent trend in cosmology, implemented over the past twenty years or so, has been to exploit various ideas proposed in particle physics to speculate on what the Universe was like at very early times. This has led to (hopefully) realistic attempts to explain some of the more puzzling facts about the Universe. Why is it that the Universe is rather nearly flat, even though, according to general relativity, it could have a very large curvature, either positive or negative? Why is it that different regions of the Universe which were never in causal contact are, and apparently were, characterised by nearly identical physical conditions? And why is the Universe precisely as lumpy as it is? If the seeds of irregularity at the time of recombination were appreciably different the Universe would look exceedingly unlike what we actually see.
Many different potential answers to these sorts of questions are being discussed, including various versions of inflationary cosmologies. At the present time, it is probably fair to say that none of these has been developed into a comprehensive model that can make careful, detailed predictions. However, it is quite possible that, in the near future, cosmologists will succeed in developing realistic models which answer these questions, as well as predicting what the Universe must have been like at even very early times.

The current aim of much of work at the interface between cosmology and microscopic particle physics is to understand what the Universe was like at increasingly early times, pushing our horizon of ignorance progressively closer to the original Big Bang.
This leads to a consideration of even more bodacious questions such as: What is the Universe like on the very smallest scales? Is it reasonable to believe that our classical notion of a smooth spacetime continuum is really correct or, as suggested originally by *Andrei Sakharov* and *John A. Wheeler* in the 1950's, is the Universe more chaotic and frothy on very small scales? It may be that, when probing the Universe on sufficiently short scales, space and time are not the right concepts to use; and that, to the extent that we still try to use them, one will be led to visualise a spacetime that is exceedingly irregular, what is sometimes called *spacetime foam*.
Or what preceeded the Big Bang? And from whence did the observed matter in the Universe actually come? If the Universe eventually recontracts to a Big Crunch, will a new Universe arise, phoenixlike, from the ashes of the old? For example, there is substantial evidence from recent attempts to reconcile general relativity with quantum mechanics that much, if not all, of the matter that we see in the Universe today was actually created from empty space as a result of the expansion of the Universe. Indeed, this idea is simply a scaled up version of relatively more modest ideas which relativists are trying to explore in the context of black holes.
Attempts to incorporate quantum mechanical effects into the context of black hole physics lead to the concrete, albeit bizarre, prediction that black holes will behave like blackbodies, radiating photons and other particles into space at a temperature T that varies inversely as their mass M. As this energy is radiated, the total mass and energy of the black hole will decrease. This implies, however, that the black hole will shrink in size and that, eventually,

it may evaporate into nothing, leaving behind a singularity in the fabric of spacetime.

What *can* be said is that, for the first time ever, scientists have developed a concrete cosmological model which provides a general framework in terms of which to explain many of the properties of the Universe which we observe on large scales, and which makes definite, testable, predictions, as to what the Universe was, is, and will be like, both at the present and at other times.

STUDY GUIDE FOR LECTURES 27 - 30

Words and names that you should recognise and understand:

Harlow Shapley	Edwin Hubble	Vesto M. Slipher
spiral nebulae	E. Öpik	Andromeda galaxy
spiral galaxies	barred spiral galaxies	elliptical galaxies
irregular galaxies	lenticular (or S0) galaxies	Sa's, Sb's, Sc's
galactic bars	SBa's, SBb's, SBc's	E0, E1, E2, E3, E4, E5, E6, E7
projection effects	oblate spheroid	prolate spheroid
rotational flattening	giant ellipticals	dwarf ellipticals
galaxy cluster	poor cluster	rich cluster
regular cluster	irregular cluster	Local Group
Virgo Cluster	Coma Cluster	rotation curve
missing mass	environmental effects	selection effects
Erik Holmberg	Alar and Juri Toomre	galaxy collisions
starburst galaxies	galaxy mergers	galactic cannibalism
protogalaxies	Carl Wirtz	H. P. Robertson
Hubble Law $v = Hr$	Hubble constant	standard candles
intergalactic extinction	redshift z	$r = cz/H$
peculiar velocities	Grote Reber	quasi-stellar sources or quasars
Maarten Schmidt	active galaxies	Seyfert galaxies
N-galaxies	BL-Lacertae objects/blazers	double radio sources
jets	M-87	Mitchie-King models
head-tail sources		

Recommended Questions:

1. What was the chain of reasoning that suggested that the "spiral nebulae" observed by V. M. Slipher are actually other galaxies outside the Milky Way?
2. What was the basis of E. Öpik's argument that Andromeda must be a distinct galaxy like the Milky Way?
3. What observations by Edwin Hubble in 1923 were interpreted as providing conclusive evidence for the existence of other galaxies?
4. Denumerate the principal components of a spiral galaxy.
5. What observed properties of spiral galaxies are used as the basis for the morphological distinction between Sa, Sb, and Sc (as well as SBa, SBb, and SBc for barred spirals)?
6. How do barred spirals differ from ordinary spiral galaxies?
7. What is believed to be the origin of bars in barred spirals?
8. What observed property of elliptical galaxies is used as the basis for the morphological classification into $E0$ - $E7$? What does this classification imply about the true three-dimensional shape of the galaxy?
9. What do astronomers currently believe to be the true three-dimensional shapes of elliptical galaxies?
10. How does an irregular galaxy differ in appearance from both spirals and ellipticals?
11. How does a lenticular galaxy differ in appearance from both spirals and ellipticals?

12. Distinguish between spirals, ellipticals, irregulars, and lenticulars in terms of (a) the ages of the stars that they contain, (b) the relative abundances of dust and gas, (c) their typical size, and (d) their typical brightness.
13. Which galaxies are the smallest? Which are the largest?
14. What is the physical reasoning used to infer that, in elliptical galaxies, almost all the stars were made in a single "burst" about 15 billion or so years ago, but that star formation in spirals has been more gradual?
15. Describe the morphological distinctions between rich and poor galaxy clusters. between regular and irregular clusters.
16. Describe how an analysis of rotation curves leads to the conclusion that, especially in the outer protions of spirals, there is a large amount of missing mass.
17. What is the evidence for missing mass in elliptical galaxies? in galaxy clusters?
18. Denumerate five different sources that have been suggested to account for the missing mass.
19. Describe what are meant by environmental and selection effects.
20. In terms of what we see, what is the most common type of galaxy?
21. Correcting for selection effects, what is believed to be the relative abundances of the different types of galaxies?
22. Why are galaxy-galaxy collisions a relatively frequent occurance?
23. Denumerate some of the possible results of a galaxy-galaxy collision.
24. What is the physical origin of starburst galaxies?
25. According to current thinking, why is that some protogalaxies evolved into spirals whereas others evolved in ellipticals?
26. How does this explanation account for the observed environmental effects associated with the observed distributions of galaxies?
27. What precisely does the Hubble Law say?
28. Why is it that the value of the Hubble constant H is relatively poorly known?
29. Describe how astronomers and cosmologists use various standard candles to estimate the distances to other galaxies.
30. Denumerate some of the difficulties associated with distance estimates using standard candles.
31. Describe how astronomers and cosmologists use the Hubble Law to estimate the distances to far away galaxies. What is the basic assumption underlying this technique? Why is it believed that this assumption is reasonably well justified for very distant galaxies?
32. What is the evidence suggesting that most quasars are extremely distant?
33. What is the evidence suggesting that quasars are extremely luminous?
34. What is the evidence indicating that quasars are extremely small?
35. How, physically, might one try to show that quasars are actually not so distant? Have such attempts met with any success?
36. What are Seyfert galaxies? How abundant are they?
37. What are N-galaxies? What are BL-Lacertae objects (or blazers)?
38. What is a double radio source?

39. Describe the observed features of the galaxy M-87 that are interpreted as indications that M-87 contains a supermassive black hole. How would such a black hole account for what is seen?
40. Describe in detail how accretion onto a black hole can explain the luminosity emitted from a quasar at the center of an active galaxy.
41. Why is it that some, but not all, quasars and active galaxies have jets?
42. What are the physical mechanisms that serve to keep jets as sharply collimated as they appear to be?
43. What are head-tail sources, and how are they interpreted as evidence for the existence of intergalactic gas?
44. About how much matter would a supermassive black hole have to accrete to explain the observed luminosities of quasars?
45. Why are galaxy-galaxy collisions believed by some astronomers and cosmologists to play an important role in allowing a quasar to emit radiation for long periods of time?

STUDY GUIDE FOR LECTURES 31 - 34

Words and names that you should recognise and understand:

homogeneity	isotropy	universality
Alexander A. Friedmann	P. A. M. Dirac	Andrei Sakharov
Cosmological Principle	Heinrich Olbers	Olber's Paradox
diffuse x-ray background	high redshift galaxies	microwave background
Big Bang Model	the Big Bang	primordial fireball
Arno Penzias	Robert Wilson	Robert Dicke *et al*
dipole anisotropy	Evgenii Lifshitz	quadrupolse anisotropy
Cosmic Background Explorer	critical density	density parameter Ω
Big Crunch	luminous matter density	total mass density
comoving volume	nucleosynthesis	radiation-dominated Universe
matter-dominated Universe	recombination	

Recommended Questions:

1. Describe carefully what is meant by the three underlying assumptions – homogeneity, isotropy, and universality – entering into the standard Hot Big Bang Model.
2. What alternatives to universality have been considered in the past by different scientists?
3. Describe what is meant by the Csomological Principle and discuss the historical context in which it evolved.
4. Describe the sense in which the Hubble Law is consistent with the assumptions of homogeneity and isotropy.
5. Describe carefully Olber's Paradox, indicating (a) the specific assumptions that are made and (b) precisely how these assumptions lead to the prediction of an infinitely bright night sky.
6. Describe the two resolutions to Olber's Paradox provided by the standard Hot Big Bang Model.
7. Denumerate three pieces of evidence indicating that, on very large scales, the Universe is indeed very nearly homogeneous and isotropic.
8. Within the context of the Hot Big Bang Model, why is it wrong to visualise the Universe as expanding "in something"? to visualise the Universe as expanding from a specific point? Why is it impossible to identify a "center" of the expanding Universe?
9. Why is it difficult to explain the ambient blackbody radiation at a temperature of about 2.7 K without assuming that the Universe was fundamentally different at some time in the past?
10. How does the Hot Big Bang Model explain the observed blackbody radiation?
11. Assuming that the Universe is filled with a completely perfect distribution of blackbody radiation, what would we expect to see looking out in space assuming that we are at rest relative to the radiation? assuming that we are moving relative to the radiation?
12. Roughly how fast is the Local Group moving relative to the blackbody radiation?

13. Why would cosmologists not expect to observe a completely perfect blackbody spectrum?
14. Aside from the dipole anisotropy associated with the Earth's motion through space, what is the evidence indicating that the blackbody spectrum is not perfectly blackbody?
15. Why physically is it that the Universe was expanding faster in the past, and that the expansion rate is decreasing?
16. Why is it that, according to the Hot Big Bang Model, a measurement of the Hubble constant provides an upper limit to the age of the Universe?
17. To what extent are the claimed values of the Hubble constant H consistent with estimates of the ages of various objects in the Universe, such as globular clusters and/or the Earth and Sun?
18. According to the Hot Big Bang Model, what two physical quantities, each in principle measureable, will determine whether the Universe expands forever or whether instead it will eventually recontract to a Big Crunch?
19. Assuming a Hubble constant of 50 km/sec/Mpc and a density parameter $\Omega = 1$, according to the Hot Big Bang Model about how old is the Universe today? Is this consistent with other estimated ages?
20. How does the observed luminous matter density compare with the critical density required to yield a Big Crunch?
21. How does the mass density infered from dynamical tests (e.g., rotation curves) compare with the critical density required to yield a Big Bang?
22. In what sense is it true that, near the Big Bang, the very notions of space and time must fail?
23. What is the highest temperature for which we have direct experimental confirmation for our theories of physics and cheistry?
24. What is believed to have been the common particles in the Universe at a temperature $T \approx 10^{11}$ K?
25. Why is it that, at early times, the Universe contained large numbers of particle-antiparticle pairs?
26. Why was it that neutrinos became decoupled from other sorts of matter at a temperature slightly below 10^{11} K?
27. Why was this decoupling of the neutrinos crucial for the development of life as we currently know it?
28. What were the principal sources of energy in the Universe at a temperature of about 10^{11} K?
29. How did the temperature of photons, neutrinos, and the like change as the Universe expanded?
30. How did the total energy of the Universe in the form of photons, neutrinos, and the like change as it expanded?
31. What is meant by primordial nucleosynthesis and when did it happen? Why was nucleosynthesis essentially nonexistent at temperatures about 1 billion K? below 400 million K?

32. What is meant by the transition from a radiation-dominated to a matter-dominated Universe, and when did that transition occur?
33. What is meant by recombination, and when did it occur?

STUDY GUIDE FOR LECTURES 35 - 39

Words and names that you should recognise and understand:

peculiar velocities	Kirschner *et al*	voids
magnitude-limited survey	celestial sphere	Center for Astrophysics Survey
right scension	declination	The Great Wall
Tully-Fisher relation	Faber-Jackson relation	Seven Samurai
The Great Attractor	Evgenii M. Lifshitz	James H. Jeans
gravitational instability	Jeans mass	Jeans length
scale-invariant spectrum	Edward Harrison	Yakov B. Zel'dovich
massive neutrinos	missing mass problem	galactic halos
Steady State Model	Bondi, Gold, and Hoyle	Perfect Cosmological Principle
continuous creation of matter	Einstein world model	Alexander A. Friedmann
cosmological constant Λ	equilibrium	unstable equilibrium
flatness problem	closed Universe	open Universe
Anthropic Principle	horizon problem	cosmic particle horizon
causal contact	finely tuned fluctuations	Inflationary Universe
Alexei A. Starobinsky	Alan Guth	vacuum energy density
Willem de Sitter	John A. Wheeler	spacetime foam
black hole evaporation		

Recommended Questions:

1. How is it that, using the Hubble Law, cosmologists attempt to estimate the distances to far away galaxies? How does the existence of peculiar velocities introduce errors into these distance estimates?
2. What was the first concrete evidence for the existence of large scale irregularities in the Universe? (There was a misprint in the notes: 1 million cubic Mpc corresponds to about 30 *quintillion* cubic lightyears.)
3. Why is it useful to perform a magnitude-limited survey?
4. What is meant by the celestial sphere? right ascension and declination?
5. What did the Center for Astrophysics (CfA) survey undertake to do? What is the principal conclusion obtained to date?
6. How do cosmologists try to estimate the distances to galaxies without using the Hubble Law?
7. What was the project undertaken by the so-called Seven Samurai? What is the principal conclusion which they reached?
8. What is meant by the gravitational instability theory?
9. Why is that, at any given time in the history of the Universe, there is a minimum size, R_J, and a minimum mass, M_J, on which structures can begin to evolve?
10. Why is it that the magnitude of the Jeans length, M_J, dropped precipitously immediately after the time of recombination?
11. What is the theoretical (i.e., non-observational) evidence indicating that, at the time of recombination, the Universe contained irregularities with an amplitude of one part in ten thousand or a bit less?

12. Why is it that the observations provided by *COBE* are interpreted as evidence indicating that such irregularities actually existed?
13. What three types of input are required to make concrete predictions as to the types of structures that should evolve in the Universe?
14. Why do the types of structure that evolve depend on the rate at which the Universe expands and the overall mass density?
15. Why do the types of structure that evolve depend on the detailed forms of the initial irregularities that are assumed?
16. Why do many cosmologists assume that the initial irregularities were characterised at least approximately by a scale-invariant spectrum?
17. Why do the types of structure that evolve depend on the detailed nature of the material of which the Universe is made?
18. Describe what is meant by the Perfect Cosmological Principle and explain how it is implemented in the Steady State Model?
19. Why is it that the Steady State Model requires a continuous creation of matter?
21. Why is that the possibility of this continuous creation is not all that easy to exclude observationally?
21. What were the motivations, both philosophical and observational, for adopting the Steady State Model?
22. Denumerate several pieces of evidence indicating that, contrary to what is assumed in the Steady State Model, the Universe has changed systematically over the course of time.
23. Why was the discovery of the microwave background a serious blow to the Steady StateModel?
24. Why did the realisation that the Hubble constant is no bigger than about 100/km/sec/ Mpc, rather than the value of about 500 km/sec/Mpc believed before 1950, make the Big Bang Model much more believable?
25. How did Einstein modify his theory of general relativity to facilitate the possibility of a static Universe?
26. What is the role of the cosmological constant in the Einstein world model? Why is it that it facilitates the possibility of a static Universe?
27. What is meant by the statement that the Einstein world model is an unstable equilibrium?
28. What physically is meant by the flatness problem?
29. In what sense is the current value of the density parameter Ω related to the spatial curvature of the Universe?
30. In what sense does $\Omega > 1$ correspond to a *closed* Universe?
31. In what sense does $\Omega < 1$ correspond to an *open* Universe?
32. What is meant by the Anthropic or Anthropogenic Principle, and how is it invoked as an explanation for the flatness problem?
33. What physically is meant by the horizon problem?

34. What is the evidence that, at the time of nucleosynthesis, the Universe was homogeneous and isotropic on scales larger than the cosmic particle horizon (at least as reckoned according to the Standard Big Bang Model)? at the time of recombination? even today?
35. What is meant by an Inflationary Universe Model?
36. How do Inflationary Models aim to explain the flatness problem? the horizon problem? the problem of finely tuned fluctuations?
37. What is meant by spacetime foam?
38. What is meant by black hole evaporation?

ESSAYS AND OTHER OPEN-ENDED QUESTIONS FOR REVIEW

Several of these questions have been broken into sections. This was done only to provide an idea of the relative importance of the different parts of the question. You are strongly encouraged to answer each question in a single coherent narrative, which addresses the different parts of the question in an order which you find logical and satisfying. The aim of these questions is to assist you in tying together pieces of the material described in these lectures which at first glance might seem comparatively disjointed.

1. Explain in words how it is that, even though atoms emit photons with very specific frequencies, the radiation emitted from an opaque object ends up being characterised very nearly by a black body spectrum. Your discussion should include comments about the roles of collisions, random motions, and temperature.

2. a. Describe how, for a relatively nearby star, an astronomer can determine more or less directly the distance d to the star, the temperature T, the luminosity L, and the radius R.

   b. Describe how, in certain cases, an astronomer can also infer the mass m of the star. Why is this not possible quite generally?

   c. Describe how, using the results from 2a, the astronomer can also estimate the distances to significantly more distant stars.

3. a. Describe the different stages in the life of a star with a mass 5 times the mass of the Sun, indicating both (i) what is going on in the star at each time (e.g., what is being fused? or is (part of) the star getting bigger? smaller?) and (ii) why what is going on is going on, i.e., what is the physics that is making the star do its thing.

   b. Explain why nuclear fusion can only happen at very high temperatures. Explain why the central parts of the star become progressively hotter as the star evolves from main sequence to red giant to supergiant, whereas the outer portions become progressively cooler. Explain why the later stages of the life of a star last a shorter time than the earlier stages.

4. a. Describe the observational evidence which is interpreted by astronomers and cosmologists as indicating that many, if not all, galaxies contain a supermassive black hole.

   b. Describe why it is that a supermassive black hole situated in the center of a galaxy might be expected to emit copious amounts of radiation. Why do some active galaxies exhibit jets, whereas others apparently do not? Why is it that active galaxies today are usually substantially less luminous than high redshift quasars, their presumed progenitors?

5. Explain the "cosmic distance ladder," i.e., the sequence of assumptions which allow an astronomer to estimate the distances to progressively more distant objects. Your discussion should touch on parallax, spectroscopic parallax, standard candles, the Tully-Fisher and Faber-Jackson relations, and the Hubble Law. Why is a determination of the correct value of the Hubble constant so hard?

6. a. Describe what is meant by the "gravitational instability" model of structure formation in the early Universe. Be as specific as you can.
   b. Why do observations of the microwave background constitute such mportant constraints on concrete implementations of this model? What are the other uncertainties which arise when a cosmologist tries to model the origins of large scale structure?

Address the following two tasks.

1. Pose, and then answer, a question similar to one of the preceding, which demonstrates something about what you have learnt from these lectures. (You may find that posing a good question is not as easy as it might seem!)
2. Describe scientifically how what you perceive in the world would differ if the speed of light were only $10,000$ km/hr rather then about $300,000$ km/sec. Let your imagination run wild, but let it be tempered by what you know about science. In particular, assume that the currently accepted theories of relativity and quantum mechanics are still true.

A brain-teaser:

As measured by the speed with which particles move, the solar corona is extremely hot, roughly 2 million K, whereas the bright, visible surface of the Sun, the photosphere, is relatively cool, about 6000 K. Why is it that the much hotter corona is so much less luminous than the photosphere that it can only be seen from the Earth during an eclipse, when the photosphere is blocked out by the Moon?